mathematics

MDPI

Article

An End-to-End Framework Based on Vision-Language Fusion for Remote Sensing Cross-Modal Text-Image Retrieval

Liu He, Shuyan Liu, Ran An, Yudong Zhuo and Jian Tao *

Department of Big Data Research and Application Technology, China Aero-Polytechnology Establishment, Beijing 100028, China; hel054@avic.com (L.H.); liusy096@avic.com (S.L.); anr051@avic.com (R.A.); zhuoyd@avic.com (Y.Z.)

* Correspondence: taoj004@avic.com

Abstract: Remote sensing cross-modal text-image retrieval (RSCTIR) has recently attracted extensive attention due to its advantages of fast extraction of remote sensing image information and flexible human–computer interaction. Traditional RSCTIR methods mainly focus on improving the performance of uni-modal feature extraction separately, and most rely on pre-trained object detectors to obtain better local feature representation, which not only lack multi-modal interaction information, but also cause the training gap between the pre-trained object detector and the retrieval task. In this paper, we propose an end-to-end RSCTIR framework based on vision-language fusion (EnVLF) consisting of two uni-modal (vision and language) encoders and a muti-modal encoder which can be optimized by multitask training. Specifically, to achieve an end-to-end training process, we introduce a vision transformer module for image local features instead of a pre-trained object detector. By semantic alignment of visual and text features, the vision transformer module achieves the same performance as pre-trained object detectors for image local features. In addition, the trained multi-modal encoder can improve the top-one and top-five ranking performances after retrieval processing. Experiments on common RSICD and RSITMD datasets demonstrate that our EnVLF can obtain state-of-the-art retrieval performance.

Keywords: remote sensing cross-modal text-image retrieval; vision-language fusion; multi-modal learning; multitask optimization

MSC: 68T07

Citation: He, L.; Liu, S.; An, R.; Zhuo, Y.; Tao, J. An End-to-End Framework Based on Vision-Language Fusion for Remote Sensing Cross-Modal Text-Image Retrieval. *Mathematics* **2023**, *11*, 2279. https://doi.org/ 10.3390/math11102279

Academic Editor: Junlin Hu

Received: 7 April 2023
Revised: 6 May 2023
Accepted: 11 May 2023
Published: 13 May 2023

1. Introduction

In recent years, remote sensing (RS) technology, which plays an important role in the satellite and unmanned surveillance aircraft industry, is developing rapidly. Under this trend, RS images have shown explosive growth [1,2], which bring challenges to multiple tasks, such as large-scale RS image recognition, detection, classification, and retrieval. Among these, remote sensing cross-modal text-image retrieval (RSCTIR) [3–7] aims to find the same or similar images in a large-scale RS image dataset according to the given natural language descriptions, and vice versa, as Figure 1. RSCTIR enables ordinary users, not limited to professionals or researchers, to achieve retrieval tasks only by natural language or visual input. It presents a better application value of human–computer interaction and information filtering, which leads to a wide range of application prospects in military intelligence generation, natural disaster monitoring, agricultural production, search and rescue activities, urban planning, and other scenarios [8,9].

The RSCTIR task, belonging to multi-modal machine learning, is becoming an emerging research area at the intersection of natural language processing (NLP) and computer vision (CV), which enables computers to understand image-text pairs in semantic terms [10]. In the RSCTIR task, there is often a huge semantic gap in feature expression between the

query and target data modality. Therefore, major research focuses on establishing connections between samples of different modalities which have the same semantics. Specifically, current RSCTIR methods are mainly divided into image caption-based methods and vision-language-embedding-based methods. The former is mainly to semantically represent each RS image to generate keyword tags that can describe the image, thus transforming RSCTIR into keyword-based text retrieval. The core of such methods is to use generative models to improve the performance of image caption tasks. For example, to make the generated tags interpretable, Bao et al. [7] designed an interpretable word-sentence framework, decomposing the task into two subtasks: word classification and ranking. In order to enable tag generators to have a more comprehensive semantic understanding of complex RS images, Devlin et al. [11] proposed a recurrent attention and semantic gating framework to generate better context features. Different from image caption-based methods, the vision-language-embedding-based methods utilize a trained image and text encoder to map the same or similar semantics sample into feature vectors with closer distances. The main point of these works is how to choose the best loss optimization strategy to minimize the distance between similar images and texts [12]. Lv et al. [13] solves the heterogeneous gap problem through knowledge distillation. On the basis of knowledge distillation, Yuan et al. [14] expects the model to be as lightweight as possible to achieve fast retrieval. Yuan et al. [5] proposed an asymmetric multi-modal feature matching network (AMFMN) and contributed a fine-grained RS image-text dataset to this task. Due to the well-studied vision-language representing the model on pre-training tasks, the embedding-based methods have become the preferred retrieval model in recent years.

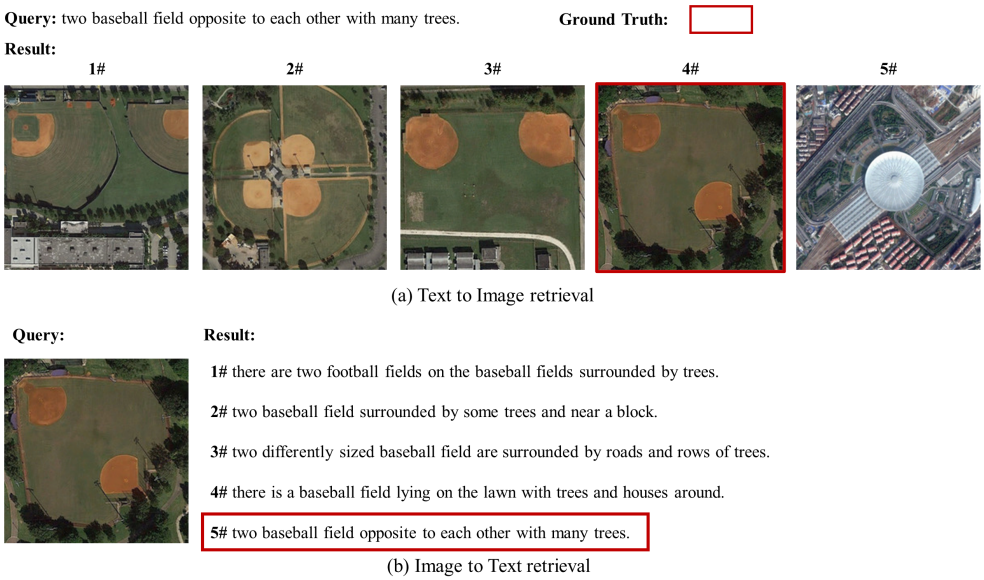

Figure 1. Example visualization results of top-5 candidates for image-text retrieval and text-image retrieval tasks.

We thus focus on the embedding based methods, which mainly generate visual and text features from two uni-modal encoders, respectively, and optimize the encoders by minimizing vector distances with similar semantics in a high-dimensional space. In view of the high intra-class similarity and large inter-class differences of RS images, the current RSCTIR framework is mainly based on convolutional encoders to extract the global features of images [15,16] and object-detector-based encoders for the local information [17]. Finally, the two uni-modal features will be compared for loss optimization. In other words, the core of the whole framework is to build a language or visual uni-modal encoder with better feature expression ability. Although these methods have achieved good retrieval performance, they do not integrate the features of text and image in the training process by vision-language fusion. Secondly, most of them use pre-trained object detectors for image

local feature extraction. Although it can effectively improve the ability of the model to express the detailed features of RS images, the training processes for the object detectors and the cross-modal retrieval model are fragmented.

Inspired by the current mainstream vision-language pre-trained model design, we believe that an end-to-end training framework will eliminate the training gap brought by the pre-trained object detectors, and improve the effect of the model by adapting more training data. At the same time, under the whole framework, a vision-language fusion model is required for image-text semantic alignment to optimize the feature expression capabilities of uni-modal and multi-modal encoders simultaneously. By solving the above problem, the end-to-end framework based on vision-language-fusion will ultimately bring better results for RSCTIR tasks.

Driven by the above motivation, we introduced a multi-modal encoder to the current most popular uni-modal encoder architecture, which is used to learn the semantic association between two modalities. Meanwhile, we discard the pre-trained object detector, extract the local features of the image by introducing a vision transformer model, and use multi-modal fusion and text erasure strategies to enhance feature interaction between different modalities, so that the model can have the ability to represent local features based on main objects. This end-to-end multi-modal framework demonstrates excellent training convenience and shows more attractive performance. Compared with the most popular method, our experiments finally achieved about a two point improvement measured by the mR criterion, which represents the average of all calculated recalls on the RSICD and RSITMD dataset. Furthermore, the trained multi-modal encoder can further improve the top-one and top-five ranking performances after retrieval processing.

Our main contributions are as follows:

1. We design a framework with two uni-modal encoders and a multi-modal encoder for RSCTIR named EnVLF. In this framework, the uni-modal encoders are used to extract visual and text features, respectively, and the multi-modal encoder is used for modality fusion. A multi-task optimization in the training process is used to improve the feature representation for each encoder.
2. Inside the vision encoder, a shallow vision transformer model is chosen to extract the local features of images instead of a pre-trained object detector, which transforms the pipe-lined "object detection + retrieval" process into an end-to-end training process. The gap between the object detector and retrieval model training process is bridged by this end-to-end framework.
3. In the inference process, the multi-modal encoder, which learns the fusion features of image-text pairs, is used in the post-processing for reranking. It can improve the top-one and top-five ranking performances based on the results on the retrieval task.

Our method is proved to be effective. The results on two commonly used RS text-image datasets can compare with state-of-the-art (SOTA) methods. Subsequently, we first introduce related works of cross-modal retrieval and vision-language pre-training in Section 2. Then, we introduce our proposed EnVLF method and each submodule in detail in Section 3. Section 4 presents the extensive experiments we conduct to verify the effectiveness of EnVLF. Finally, our work is concluded.

2. Related Works

2.1. Cross-Modal Retrieval

Cross-modal image-text retrieval (ITR) is to retrieve relevant samples from one modality given its expression in another modality. Its key challenge is to bridge the heterogeneity gap and learn a transformation function to project multi-modal data to a common representation space, thus reducing the cross-modal retrieval task to an embedding space. The retrieval process includes the extraction of uni-modal features and the alignment of multi-modal features. Extracting features is the first and most critical process in an ITR system, which includes methods such as vision-language embedding, cross-attention, etc. Visual semantic embedding (VSE) is the most direct way to independently encode

uni-modal features. The development of VSE mainly includes two aspects of data and loss function improvement [18–21]. Cross-attention methods are mainly based on the transformer method, which improve the retrieval performance by learning contextual knowledge between modalities. For example, Cui et al. [22] proposed a method capable of simultaneously encoding cross- and intra-modality knowledge in a unified scene to enhance subsequent feature alignment tasks.

Feature alignment is also an important step, which needs to be used to calculate pairwise similarity and achieve retrieval. It can be divided into two methods: global alignment and local alignment. The global alignment method mainly uses the global feature learning model [23,24], but this will ignore the fine-grained information in images and texts, thus affecting the retrieval effect. Local alignment usually refers to corresponding the patch-level information of the image with the word information in the text. Adopting the ordinary attention mechanism [3,25] is a simple way to explore semantic region/patch word correspondences. At the same time, it is also a strongly emerging direction to combine global and local alignment. For example, Ji et al. [26] propose a step-wise hierarchical alignment network, which decomposes image-text matching into a multi-step cross-modal reasoning process, first achieving local-to-local alignment at the fragment level, and then in turn performing global-to-local and global-to-global alignment at the context level.

2.2. Vision-Language Pre-Training

Most of the early cross-modal retrieval works used pre-trained networks in the fields of natural language processing and computer vision followed by fine-tuning. However, in recent years, as pre-trained models have become a hot topic, there has been a surge of interest in developing general cross-modal pre-trained models and extending them to downstream ITR tasks [27–29]. Most current pre-training ITR methods adopt the transformer architecture as a building block. CLIP [30] and ALBEF [31] are two typical cases. CLIP uses a contrastive loss, which is one of the most effective losses in representation learning, and is pre-trained on 400 million noisy multi-modal web data points, resulting in highly general image-text features. Ultimately, CLIP achieves impressive performance on many downstream vision and language related tasks. ALBEF also uses a contrastive loss to align image and text representations before they merge. Unlike most existing methods, ALBEF requires neither labeled image data nor high resolution images. In addition, in order to better learn from a large amount of noisy data, momentum distillation is proposed, which is able to learn from the pseudo-objects of the momentum model. ALBEF also achieves SOTA performance on many downstream vision-language tasks.

Through the analysis of all these existing studies, we believe that the essence of RSCTIR is to bridge the gap in feature expression between images and texts. During the training process, the two types of features need to be fused to better learn the interaction. However, different from the pre-training process, RS retrieval has strong supervision information. Therefore, in addition to improving uni-modal feature encoding, it is a focus of our paper to use this supervision information to improve the vision-language fusion features of the model and thus enhance the ability to express cross-modal features.

3. Methods

In this section, we first outline the overall architecture of our model and the design purpose of each encoder in Section 3.1. Then, Section 3.2 introduces the optimization objective in detail. Finally, we describe the training process for the entire architecture in Section 3.3.

3.1. Model Architecture

Most traditional RSCTIR tasks are mainly to compute the vector similarity between the features of the input query (RS image or text caption) and target features stored in the database. A typical case is GaLR [4], which has been reported to achieve SOTA performance. The core of GaLR is a well-designed vision and language encoder. Specifically, the image

I and text T are encoded by vision encoder $F_{enc-V}(I)$ and language encoder $F_{enc-L}(T)$ separately as two independent vectors which are mapped into the same space. Moreover, for the visual module, GaLR specifically introduces an object detection branch followed by a graph convolution network (GCN) based on the global encoder, which may cause laborious model assembly. In the training optimization procedure, triplet loss is used for optimization.

Through the detailed exploration of the above architecture, we found that most traditional RSCTIR methods mainly focus on improving the performance of uni-modal features extraction. A multi-modal encoder which fuses vision-language features by completing multitask training optimization objectives can further improve the performance of the entire retrieval task. Illustrated in Figure 2, we propose our EnVLF model, which introduces a multi-modal encoder based on the traditional RSCTIR framework for cross-modal semantic feature learning. It is worth noting that we do not use the object detection branch, thus, the whole training process is end-to-end. We also optimized the training process to enhance the feature representation performance of uni-modal and multi-modal encoders by simultaneously optimizing three kinds of losses: image-text-triplet (ITT), masked-language-model (MLM) and image-text-match (ITM). This process can be represented as follows:

$$\begin{aligned} f_{IL},\ f_{IGL} &= F_{enc-V}(I) \\ f_T &= F_{enc-L}(T) \\ S_{distance},\ S_{pairwise} &= F_{enc-Mul}(f_{IL},\ f_{IGL},\ f_T) \end{aligned} \tag{1}$$

f_{IL}, f_{IGL}, and f_T denote the local visual features, fused visual features, and text features extracted by uni-modal encoders F_{enc-V} and F_{enc-L} separately. Cross-modal similarity $S_{distance}$ and pair-wised image-text similarity $S_{pairwise}$ are calculated by multi-modal encoder $F_{enc-Mul}$.

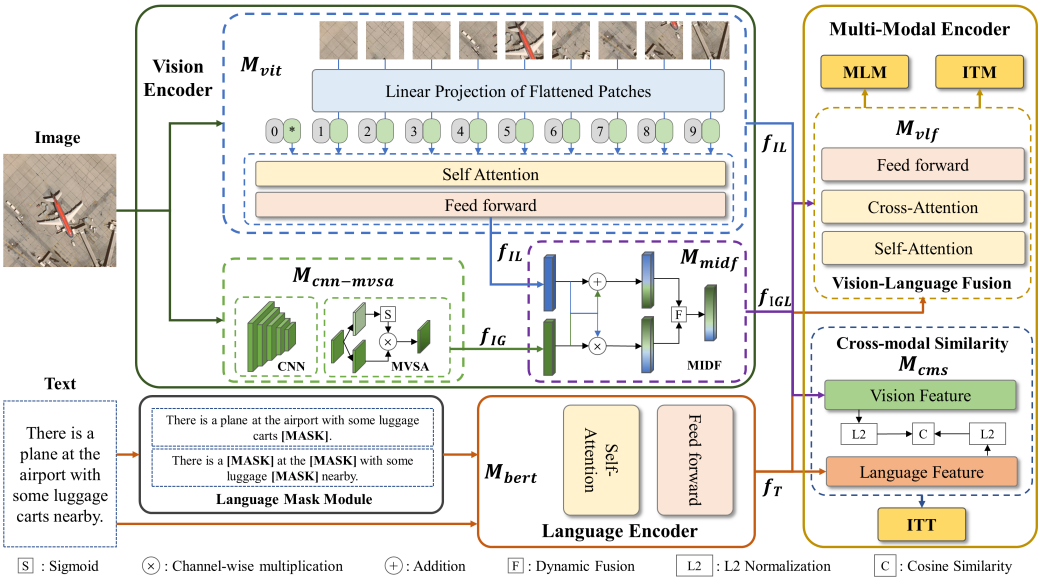

Figure 2. Proposed RSCTIR framework based on EnVLF which consists of a vision encoder, a language encoder and a multi-modal encoder. M_{vit} module is introduced into the vision encoder to achieve local object representation enhancement. The language encoder is constructed based on the BERT model with a specially designed language mask module. The multi-modal encoder aims to promote better alignment of image-text features by utilizing multitask optimization, and can also serve as a reranking module.

Next, we will describe our framework with regard to three aspects: vision encoder, language encoder and multi-modal encoder.

3.1.1. Vision Encoder

It has been found in the literature [5,7,32] that in vision-language pre-training tasks, the information carried by the vision side is much more than the language side, so a more complex vision encoder is needed to express comprehensive features. Inspired by this, we believe that a vision encoder can be divided into a global feature extraction module $M_{cnn-mvsa}$, an object-based local feature extraction module M_{vit}, and a vision fusion module M_{midf}, and in this way, the vision encoder can fully extract the multi-dimensional features of the image. This can be formalized as follows:

$$\begin{aligned} F_{enc-V} &= \left\{ M_{cnn-mvsa},\ M_{vit},\ M_{midf} \right\} \\ f_{IG} &= M_{cnn-mvsa}(I) \\ f_{IL} &= M_{vit}(I) \\ f_{IGL} &= M_{midf}(f_{IG},\ f_{IL}) \end{aligned} \tag{2}$$

where f_{IG} and f_{IL} are global and local features of the same RS image presented by $M_{cnn-mvsa}$ and M_{vit} separately. The f_{IGL} is the features fused by M_{midf}.

CNN-based methods directly map the global features of an image into the high-dimensional space, which have shown advantages in image embedding tasks. In more detail, referring to the method of GaLR, we use ResNet-18 and a multi-level visual self-attention (MVSA) module as the global features encoder $M_{cnn-mvsa}$ for RS images.

An object-detection-based model pre-trained on other RS-related datasets can identify the main objects in RS images, such as airplanes, cars, buildings, etc. Capturing local features of RS images by characterizing the relationships between these main objects can improve the RSCTIR performance [4]. However, training an effective object detector relies on well annotated RS object detection datasets, such as DOTA [33], NWPU-RESISC45 [34], and UCAS-AOD [35]. Meanwhile, it is worth noting that the difference of data annotation forms between object detection and cross-modal retrieval tasks brings a data gap which results in a pipeline mode of the training process. For this imperfection, we adopt M_{vit}, a transformer-based vision model [36], and exploit its image patch-based feature representing characteristics to replace the pre-trained object detector. By optimizing with MLM and ITC loss in training, the visual features from M_{vit} and masked text features from M_{bert} are aligned with the key semantics, which means M_{vit} has the same ability as a pre-trained object detector for image local feature extraction.

The vision transformer module M_{vit} is the key for the local feature encoder, which consists of stacked blocks that include a multi-head self-attention (MSA) layer and a multi-layer perceptron (MLP) layer.

$$\begin{aligned} v_0 &= [v_{class};\ v_1 V;\ \ldots\ v_N V] + V^{pos},\ V \in \mathbb{R}^{(P^2 \cdot C) \times H},\ V^{pos} \in \mathbb{R}^{(N+1) \times H} \\ \hat{v}^d &= MSA(LN(v^{d-1})) + v^{d-1},\ d = 1\ \ldots\ D \\ v^d &= MLP(LN(\hat{v}^d)) + \hat{v}^d,\ d = 1\ \ldots\ D \\ p &= tanh(v_0^D W_{pool}) \end{aligned} \tag{3}$$

The input image $I \in \mathbb{R}^{H \times W \times C}$ is slid into a flattened two-dimensional patch of size $(N \times (P^2 \cdot C))$, where (P, P) is the patch size and $N = HW/P^2$. N is the number of blocks that affect the length of the input sequence. Followed by the linear projection $V \in \mathbb{R}^{(P^2 \cdot C) \times H}$ and position embedding $V^{pos} \in \mathbb{R}^{(N+1) \times H}$, v is embedded into $f_{iL} \in \mathbb{R}^{N \times H}$.

The MIDF module M_{midf} proposed by GaLR is used to achieve the dynamic fusion of global and local features in vision encoder F_{enc-V}. After M_{midf} fusing the local and global information, the comprehensive features of RS images are represented by the F_{enc-V}, which are subsequently used to interact and fuse with the text features generated by the language encoder F_{enc-L}.

3.1.2. Language Encoder

As analyzed in Section 3.1.1, the information contained in language is less than that in vision, so a lightweight language encoder that can represent text features is needed. To align with sequenced vision features generated by M_{vit}, the same language encoder as BERT [11], where a [CLS] token is appended to the beginning of the text input to summarize the sentence, is the best choice. Specifically, aiming to simplify the computational complexity, we initialize the first six layers of the BERT-based model M_{bert} as the language encoder F_{enc-L}. The language encoder transforms the input text T into a sequence of embeddings $\boldsymbol{w}_{cls}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_N$, which is fed to the multi-modal encoder for cross modal representation learning and language-vision fusion.

$$f_T = M_{bert}(T) \tag{4}$$

In order to use the semantic correlation relationship between modalities to improve the extraction of features for each modal encoder, we perform a targeted masked strategy on the target category information contained in RS images as well as the traditional random mask processing in the language mask module. The masked text is passed through the language encoder to generate the feature f_T. The local feature f_{IL} generated by vision encoder and the masked text feature f_T are fused in the multi-modal encoder. By optimizing with the MLM loss in training, the M_{bert} in the language encoder and the M_{vit} in the vision local encoder can reach a better performance in specific target recognition.

3.1.3. Multi-Modal Encoder

Considering that the semantic features of image I and text T extracted by the uni-modal encoder with vector distance constraint cannot be sufficiently aligned in traditional methods. Besides the cross-modal similarity module, a visual-language fusion module is introduced in the multi-modal encoder, which is initialized with the last six layers of the BERT-based model and uses an additional cross-attention layer to model the visual-language interaction. The multi-modal encoder can be formalized as

$$F_{enc-Mul} = \left\{ M_{cms},\ M_{vlf} \right\} \tag{5}$$

where the M_{cms} calculates the similarity between image features f_{IGL} and text features f_T by using cosine distance and optimizes the parameters in uni-modal encoders by using ITT loss. The M_{vlf} uses MLM loss to optimize the expression of visual local feature encoder M_{vit} and text feature encoder M_{bert} by fusing the features f_{IL} and f_T. Furthermore, M_{vlf} also utilizes ITM loss to simultaneously optimize the feature representation ability of the uni-modal encoders and multi-modal encoder.

Training with vision-language fusion can improve the performance of each uni-modal encoder and ultimately affects the RSCTIR results. Since the vision-language fusion model M_{vlf} requires image-text pairs as input in the inference process, it cannot be used for a large-scale data recall process in RSCTIR. Considering that the trained M_{vlf} has a better fine-grained discrimination ability for image-text pairs, so we further use it to rerank the results to improve the top-one and top-five retrieval performances.

3.2. Multitask Optimization

During the training process, the end-to-end RSCTIR framework should improve the performance of each encoder by optimizing multiple targeted tasks. Thus we designed three training objectives for the framework: ITT on the uni-modal encoders with cosine similarity, MLM loss for capturing main objects based features, and ITM on language-vision fusion for image-text pairwise similarity learning.

3.2.1. Image-Text-Triplet Loss for Uni-Modal Encoder

Triplet loss is a loss commonly used in the field of image-text alignment. It calculates the loss function value by comparing the distance between three samples (anchor, positive, and negative). The main idea is to learn a feature representation space, so that anchor samples of the same category are closer to positive samples in this space, and anchor samples of different categories are farther away from negative samples. This can be formalized as follows:

$$\mathcal{L}_{itt} = \sum_{\hat{T}} [\epsilon - cos(I,\ T) + cos(I,\ \hat{T})]_+ + \sum_{\hat{I}} [\epsilon - cos(I,\ T) + cos(\hat{I},\ T)]_+ \tag{6}$$

where ϵ represents the minimum margin designed to widen the gap between anchor and positive/negative sample pairs, $[x]_+ \equiv max(x,\ 0)$. $(I,\ T)$ is a paired image-text sample. $\hat{T}$ is the text that is not paired with the image I, and $\hat{I}$ is the RS image not paired with the text T.

The advantage of ITT loss lies in the distinction of details, that is, when the two inputs are similar, triplet loss can better model the details. Therefore, we choose the triplet loss in our RSCTIR task to constrain feature representation for uni-modal encoders. The strategy for choosing positive and negative examples is: given the N image-text pairs in one batch are positive samples, and the other $N^2 - N$ unpaired image-text pairs are negative samples. The purpose of this strategy is to make the uni-modal encoder learn a complex representation of both inter-class and intra-class differences.

3.2.2. Image-Text-Match Loss for Pairwise Similarity

In image-text matching, the model predicts whether a pair of input image and text is matched or not. Inspired by most VLP models treating image-text matching as a binary classification problem, we use this training target for the multi-modal encoder. Specifically, a special token, such as [CLS], is inserted at the beginning of the input sentence, which tries to learn a cross-modal representation. Different from the multi-modal encoder of VLP for binary classification, we add a fully connected layer to calculate a two-class probability p^{itm} for each image-text pair which can be used as a reranking model in the inference procedure. ITM loss can be formalized as follows:

$$\mathcal{L}_{itm} = \mathbb{E}_{(I,\ T)\sim D} H(\boldsymbol{y}^{itm},\ \boldsymbol{p}^{itm}(I,\ T)) \tag{7}$$

3.2.3. Masked Language Model for Object-Based Image Representation

The masking language modeling (MLM) loss strategy uses images and contextual text to predict masked words, which was originally used in language training tasks, providing better feature expression for pre-trained models. In RSCTIR, it has also been proved to have the same importance as the ITM loss strategy. In our framework, in order to allow the vision encoder to express the region features without using the object detection module, we use MLM with the patch embedding ability of the vision transformer to express the features of the main target of RS images. At the same time, we optimize the masked strategy by using two methods: targeted masked strategy and random masked strategy. The targeted masked strategy aims to enhance the generalization ability of the model, and the targeted mask mainly focuses on important information in the retrieval process, such as the type and number of objects. The text is input into the encoder through the above two masking methods, and then the model is trained to reconstruct the original mark to achieve the ability to represent the information of the main targets. The optimization objective of MLM is minimizing a cross-entropy loss:

$$\mathcal{L}_{mlm} = \mathbb{E}_{(I,\ \hat{T})\sim D} H(\boldsymbol{y}^{msk},\ \boldsymbol{p}^{msk}(I,\ \hat{T})) \tag{8}$$

where the masked text is denoted by $\hat{T}$, and the predicted probability of the model for a masked token is denoted by $\boldsymbol{p}^{msk}(I,\ \hat{T})$.

3.3. Training Procedure of EnVLF

The training and inference procedures of our proposed EnVLF are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Training Procedure of the Proposed EnVLF

Input:

Image-Text pairs of RS dataset $D_{IT} = \{\{I_1,\ T_1\} \cdots \{I_n,\ T_n\},\ n\ is\ the\ number\ of\ pairs\}$

Through:

Global visual feature $f_{IG} = M_{cnn-mvsa}(I;\ \theta_1)$

Local visual feature $f_{IL} = M_{vit}(I;\ \theta_2)$

Vision fusion feature $f_{IGL} = M_{midf}(I;\ \theta_3)$

Text feature $f_T = M_{bert}(T;\ \theta_4)$

Multi-modal fusion feature $f_{MM} = M_{vlf}(I,\ T;\ \theta_5)$

Repeat until convergence:

for each batch $I,\ T \in D_{IT}$ **do**

Dual visual feature extraction

$f_{IG} = M_{cnn-mvsa}(I;\ \theta_1)$

$f_{IL} = M_{vit}(I;\ \theta_2)$

$f_{IGL} = M_{midf}(I;\ \theta_3)$

Text feature extraction

$f_T = M_{bert}(T;\ \theta_4)$

Multi-model extraction

$f_{MM} = M_{vlf}(f_{IL},\ f_{IGL},\ f_T;\ \theta_5)$

Calculate the ITT loss

$l_{itt} = L(f_{IGL},\ f_T)$

Calculate the ITM loss

$l_{itm} = L(f_{MM})$

Calculate the MLM loss

$l_{mlm} = L(f_{IL},\ f_T)$

Update θ_1, θ_2, θ_3, θ_4, θ_5 **by** $l_{itt},\ l_{itm},\ l_{mlm}$

end for

return $M_{cnn-mvsa},\ M_{vit},\ M_{midf},\ M_{bert},\ M_{vlf}$

Algorithm 2 Inference Procedure of the Proposed EnVLF

Input:

Image-Text pairs of RS dataset $D_{IT} = \{\{I_1,\ T_1\} \cdots \{I_n,\ T_n\}, n\ is\ the\ number\ of\ pairs\}$

Through:

EnVLF model $f_{IGL},\ f_T = EnVLF(I,\ T;\ \theta_1)$

Initialize the similarity matrix S

Vision-Language Fusion module rerank $S_{pairwise} = M_{vlf}(f_{IGL},\ f_T;\ \theta_2)$

Calculate image-text similarity matrix:

for each batch $I,\ T \in D_{IT}$ **do**

Visual and text feature extraction

$f_{IGL},\ f_T = EnVLF(I,\ T;\ \theta_1)$

$S_{distance} = M_{cms}(f_{IGL},\ f_T)$

Append $S_{distance}$ **to** S

end for

return S

Multi-model module top-10 rerank: (optional)

$S_{pairwise} = M_{vlf}(f_{IGL},\ f_T;\ \theta_2)$

return $S_{pairwise}$

In the training process, we mainly optimize the image encoder F_{enc-V}, the text encoder F_{enc-L}, and the multi-modal encoder $F_{enc-Mul}$. First, we input the RS image-text pair batch

D_{it} for training, and then pass it through the uni-model image encoder F_{enc-V} and text encoder F_{enc-L}. F_{enc-V} contains dual modules $M_{cnn-mvsa}$ and M_{vit} to extract global features f_{IG} and local features f_{IL}, respectively, and finally visual features are formed through the vision fusion module M_{midf}. The whole process uses three loss strategies to guide the training process, and the l_{itt} loss is used to optimize the generation of f_{IGL} and f_T. l_{itm} and l_{mlm} are jointly optimized by text encoder F_{enc-L} and module M_{vlf} in multi-modal encoder $F_{enc-Mul}$, and finally after multiple epoch iterative training, the optimal EnVLF model is constructed.

For the inference process, the RS image-text test pair batch is required, which is sent to the trained EnVLF model to obtain the uni-model image and text feature f_{IGL}, f_T. Afterwards, the cosine distance is used to calculate the similarity $S_{distance}$ between query and targets, and all test datasets are traversed to generate a large matrix S, thereby completing the final recall task. At the same time, in order to utilize the pairwise similarity $S_{pairwise}$ calculated by multi-modal encoder, we choose to use it to complete the reranking of top-10 results.

4. Experiments

4.1. Datasets and Evaluation Metrics

In our experiments, we validate our EnVLF on two datasets: RSICD [9] and RSITMD [5]. Each dataset is given in the form of a large number of image-text pairs, as shown in Figure 3, where RSICD has 10,921 pairs, of which the image size is 224×224, making it the most numerous retrieval dataset currently available. RSITMD has 4743 pairs with the image size of 256×256. Compared with RSICD, RSITMD has a finer-grained text caption. In our experiments, we follow the data partitioning approach of Yuan et al. [5] and use 80%, 10%, and 10% of the dataset as the training set, validation set, and test set, respectively. For the evaluation criteria, we use R@k and mR [37] to evaluate the recall performance of EnVLF. R@k indicates the proportion of ground truth contained in the top k results recalled. Consistent with GaLR, we choose k to be 1, 5, or 10 to evaluate the results more fairly. mR represents the average value of each R@k, which can be used as the final evaluation criterion of the overall performance of the model.

(1) this is a parking apron, roads, grassland, buildings and many planes .
(2) it is gray parking apron, gray roads, green and yellow grassland, gray buildings and gray planes .
(3) several plants are parked next to terminals and near runways .
(4) several plants are parked next to terminals and near runways .
(5) this is a parking apron, roads, grassland, buildings and many planes .

(a) RSICD datasets

(1) Five baseball fields are surrounded by green trees.
(2) Five baseball fields surround the center of the building, forming a flower shape decorated with luxuriant plants.
(3) Five baseball fields surround the center of the building, forming the shape of flowers decorated with luxuriant plants.
(4) Five baseball fields are surrounded by green trees.
(5) Five baseball fields are surrounded by green trees.
Labels: "baseball fields", "five baseball", "green trees", "lush plants", "flower shape"

(b) RSITMD datasets

Figure 3. Example image-text pairs from the RSICD and RSITMD remote sensing image-text datasets, where each pair consisting of an RS image and five corresponding sentences.

4.2. Implementation Details

All our experiments are performed on a single Tesla V100 GPU. For images, we unified the size of all images to $3 \times 256 \times 256$ and sent them to the image encoder, and then applied data enhancement methods such as random rotation and random cropping. We directly chose ResNet-18 to apply to the global encoder. For the local encoder, we used a 6-layer

standard VIT model, and the final generated image and text uni-model features were both 512. For the text model, we used the first six layers of BERT to build a text encoder, and the last six layers to build a multi-model encoder. The batch size we applied in the training procedure was 64 and adjusted to 128 during validation. The learning rate was initialized to 2×10^{-4}, and decayed by a factor of 0.7 every 15 epochs of validation following the iterative process. We trained the entire model for 30 epochs and optimized the model with the Adam optimizer. The rest of the parameters except the learning rate were kept as default values and were not adjusted during the entire procedure.

4.3. Comparisons With the SOTA Methods

We compare the excellent work based on RS cross-modal retrieval task on RSICD and RSITMD datasets: including VSE++ [23], SCAN [12], CAMP [38], MTFN [39], AMFMN [5], LW-MCR [14], and GaLR [4].

VSE++ is one of the pioneers of using uni-modal encoders to extract image and text features for cross-modal retrieval. A triplet loss is used to optimize the training objective.

SCAN enhances VSE++ by using an object detector to extract local features. The object detector has been proved to achieve excellent performance in local feature extraction.

CAMP propose a message passing mechanism, which adaptively controls the information flow for message passing across modalities. The triplet loss and the BCE loss method are used as control groups.

MTFN designs a multi-modal fusion network based on the idea of rank decomposition to improve the retrieval performance by a reranking process.

AMFMN utilizes an asymmetric approach based on triplet loss that uses visual features to guide text presentation.

LW-MCR takes advantage of methods such as knowledge distillation and contrast learning for lightweight retrieval models.

GaLR optimizes the representation matrix and the adjacency matrix of local features by using GCN. The quantitative analysis on multiple RS text-image datasets demonstrates the effectiveness of the proposed method for RS retrieval.

For the methods above, we use the results reported in the literature [14]. Following the same strategy for our experiments as these works, we provide the performance of our final model for comparison.

EnVLF: We use the whole EnVLF framework in training progress, and use the uni-modal encoder for RS cross-modal retrieval.

Table 1 shows that the performance of the EnVLF is particularly impressive on both RSICD and RSITMD datasets, and we can obtain conclusions as follows:

Table 1. Performance comparison with other models on the RSICD and RSITMD datasets.

Model	RSICD Dataset							RSITMD Dataset						
	Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	3.38	9.51	17.46	2.82	11.32	18.10	10.43	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN t2i	4.39	10.90	17.64	3.91	16.20	26.49	13.25	10.18	28.53	38.49	10.10	28.98	43.53	26.64
SCAN i2t	5.85	12.89	19.84	3.71	16.40	26.73	14.23	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP-triplet	5.12	12.89	21.12	4.15	15.23	27.81	14.39	11.73	26.99	38.05	8.27	27.79	44.34	26.20
CAMP-bce	4.20	10.24	15.45	2.72	12.76	22.89	11.38	9.07	23.01	33.19	5.22	23.32	38.36	22.03
MTFN	5.02	12.52	19.74	4.90	17.17	29.49	14.81	10.40	27.65	36.28	9.96	31.37	45.84	26.92
LW-MCR(b)	4.57	13.71	20.11	4.02	16.47	28.23	14.52	9.07	22.79	38.05	6.11	27.74	49.56	25.55
LW-MCR(d)	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-soft	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	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	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
EnVLF	**7.78**	**20.52**	**31.56**	**6.09**	**22.33**	**36.65**	**20.82**	13.42	**33.11**	**47.12**	11.15	**38.63**	**57.58**	**33.50**

The performance of EnVLF on the RSICD dataset, which is a large RS image-to-text dataset, strongly demonstrates the effectiveness of our proposed method. For the EnVLF model, the mR score is even 1.86 points higher than the most outstanding models, reaching 20.82. Whether it is text-to-image retrieval or image-to-text retrieval, EnVLF model shows the best performance compared with the previous SOTA method on R@1, R@5, and R@10 results.

RSITMD has more fine-grained representation in text than RSICD, which makes the retrieval task even more challenging. However, EnVLF still performs well on the RSITMD dataset. The mR of EnVLF finally reached 33.50, which is 2.09 points ahead of GaLR. In more detail, the most difficult R@1 results for text-to-image and image-to-text retrieval of our EnVLF are slightly lower than SOTA. This may be due to the difference between the training set and the test set. Perhaps the semantic distribution of the test set is a little simpler than that of the training set. While compared with AMFMN, our model is more complex, leading to poor results, which can be a future research point. However, it is worth noting that the R@5 and R@10 results are well ahead of the SOTA.

We further compare the qualitative results of our EnVLF with GaLR in Figure 4. For text-to-image retrieval, we chose an example of the playground. The results show that EnVLF hit the ground truth of the image at the top-one position, and in the top-five results, most of the results centered on the playground. While for GaLR, the hit is not only realized in the top-three position, but most of the results are centered around the baseball field, which indicates that EnVLF can learn more matching, semantically similar features through our cross-modal design. The image-to-text retrieval results are shown below. We can find that EnVLF can hit more ground truth texts in the top-five results, and all of them rank within the top two, while GaLR only hit one low-ranked ground truth text.

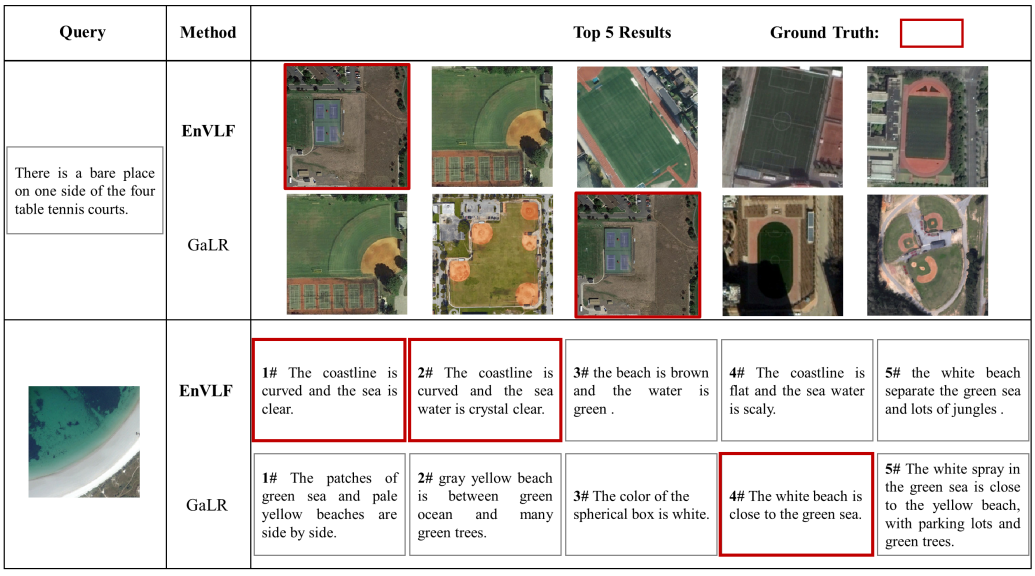

Figure 4. Visual comparison of retrieval results between EnVLF and the state-of-the-art GaLR on the RSITMD test set for two retrieval tasks. Among the top-5 candidates, the red box surrounds the ground truth.

We then explore the reasons why EnVLF exhibits more attractive performance compared with GaLR, which we believe can be attributed to the following factors: Firstly, EnVLF adds an M_{vit} local vision encoder on the basis of the global vision encoder. However, GaLR contains a global encoder and an object detector whose effect has been proved to be closely related to the number of objects contained in the image [4]. When there are fewer objects, the detector will show weaker performance. Secondly, the language encoder of EnVLF also adopts the transformer model and carefully designs a targeted masked strategy, thus improving the expression of the main features, which are not considered in GaLR.

Finally, GaLR only uses the triplet loss to optimize the vision and language encoder, while EnVLF introduces a multi-modal encoder which can better perform semantic alignment across modalities and uses three kinds of losses during the optimization.

4.4. Reranking Process Based on Multi-Modal Encoder

In $EnVLF$, the cross-modal similarity calculation module M_{cms} can be used to efficiently find the top-N similar results for large-scale data retrieval. Furthermore, in order to take full advantage of the vision-language fusion module M_{vlf} that learns pairwise image-text feature representations, we train it to be used as a reranking process. In our experiment, the $EnVLF$ model only uses each uni-modal encoder for feature extracting, and M_{cms} in multi-modal encoder for feature similarity calculating. However, the results show that the recall rate of the model on the more difficult top-one and top-five is always lower, so based on the results of $EnVLF_{rerank}$, we use the $EnVLF_{rerank}$ model to calculate the pairwise similarity between the query and top-ten recalled candidates. The results in Table 2 shows the reranking process can further improve the mR criterion by about 0.1 points. During our experiment, we found that the performance of the reranking process is not stable, but a slightly higher result can be achieved on most of the top-one and top-five results, especially in some difficult cases, such as Figure 5; although $EnVLF$ recalls similar candidates, $EnVLF_{rerank}$ can still rerank the ground truth to top-one, and can even find more top-five correct retrieval results, which shows the potential of the reranking process. Thus, we believe that with more careful design in the future, the reranking process can play a greater role.

Table 2. Comparison results of the rerank module on the RSICD dataset.

Model	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
$EnVLF$	7.78	20.52	31.56	**6.09**	22.33	36.65	20.82
$EnVLF_{rerank}$	**7.99**	**20.98**	31.56	5.73	**22.80**	36.65	**20.95**

4.5. Ablation Studies of Structures

We conducted ablation experiments on the RSITMD dataset for the cross-modal retrieval task to validate each module separately. Specifically, four control experiments were designed to explore the influence of every proposed module on retrieval performance, and we kept the language encoder F_{enc-L} the same in all experiments. The ablation experiment modules are as follows:

$EnVLF_{glb}$: Use the vision encoder F_{enc-V} only containing the global feature extractor $M_{cnn-mvsa}$ for visual features and multi-modal encoder $F_{enc-Mul}$ only containing M_{cms} for cosine similarity calculating.

$EnVLF_{uni}$: Use the full vision encoder F_{enc-V} containing the global feature extractor $M_{cnn-mvsa}$, local feature extractor M_{vit} and vision fusion module M_{midf} for visual features, and multi-modal encoder $F_{enc-Mul}$ only containing M_{cms} for cosine similarity calculating.

$EnVLF_{RanMask}$: Use the full vision encoder F_{enc-V} and full multi-modal encoder $F_{enc-Mul}$ containing M_{cms} and vision-language fusion module M_{vlf}. We do not interfere in the selection of masks in this experiment, and a random masked strategy is used for text preprocessing.

$EnVLF_{TarMask}$: The overall model is the same as $EnVLF_{RanMask}$, but our proposed targeted masked strategy proposed is used for the text preprocessing to guide the text encoder to learn more targeted features that are conducive to the image-text retrieval process. At the same time, we do not remove the random masked strategy module.

In these models, the comparison between $EnVLF_{glb}$ and $EnVLF_{uni}$ shows the performance of the M_{vit} module without vision-language fusion. The results of $EnVLF_{RanMask}$ compared with $EnVLF_{uni}$ shows the influence of vision-language fusion module M_{vlf} in multi-modal encoder $F_{enc-Mul}$. The control group ($EnVLF_{RanMask}$, $EnVLF_{TanMask}$) veri-

fies the superiority of the targeted masked strategy with vision transformer M_{vit} in local feature extraction.

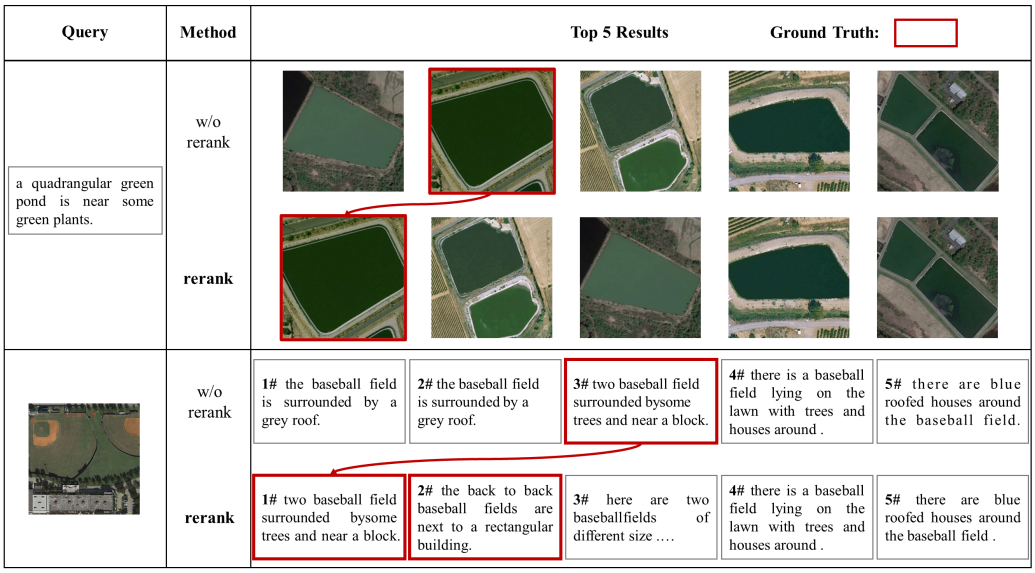

Figure 5. Visual comparison of the rerank module for two retrieval tasks on the RSITMD test set. Among the top-5 candidates, the red box surrounds the ground truth.

Table 3 shows the results of the above three groups of experiments.

Compared with $EnVLF_{glb}$, the mR criterion of the model $EnVLF_{uni}$ increased by 0.77 points after adding the vision transformer M_{vit} for local feature extraction without multi-modal fusion.

Subsequently, we show the results of $EnVLF_{RanMask}$. Through adding the vision-language fusion module M_{vlf} to the multi-modal encoder $F_{enc-Mul}$, the mR criterion is further improved by 0.44 points compared with $EnVLF_{uni}$.

The results of control-experiment group ($EnVLF_{RanMask}$, $EnVLF_{TarMask}$) show that the object-related information masked strategy can improve the representation for local features, especially for the top-10 recall results, through $EnVLF_{TarMask}$ can reach 1.69 points of improvement. This experiment can also prove that the vision and language transformers trained by multi-modal fusion show a comparable ability with pre-trained object-detector-based methods.

Table 3. Ablation experiments for different modules on the RSITMD dataset.

Model	M_{mvsa}	M_{vit}			I2T Retrieval			T2I Retrieval			mR
		w/o mlm	Random Mask	Targeted Mask	R@1	R@5	R@10	R@1	R@5	R@10	
$EnVLF_{glb}$	√	√			12.83	29.42	44.25	11.64	38.89	54.82	31.98
$EnVLF_{uni}$	√	√			13.86	32.37	45.20	11.49	38.60	54.96	32.75
$EnVLF_{RanMask}$	√		√		**14.60**	32.52	44.69	**12.39**	**39.03**	55.90	33.19
$EnVLF_{TarMask}$	√		√	√	13.42	**33.11**	**47.12**	11.15	38.63	**57.58**	**33.50**

5. Conclusions

In this paper, we proposed an end-to-end RS cross-modal retrieval framework named EnVLF, which consists of three modules: a vision encoder and a language encoder for uni-model feature extraction, and a multi-modal encoder for vision-language fusion. Specifically, for the vision encoder, we introduce a vision transformer module trained with a multi-modal encoder to achieve the ability of object detection, which transforms the pipelined training process into an end-to-end process. By optimizing multiple target tasks

for training, EnVLF can obtain competitive retrieval performance and bridge the training gap between object detection and retrieval tasks. In addition, the trained multi-modal encoder can improve the top-one and top-five ranking performances after retrieval processing. The experiments and analysis on multiple RS text-image datasets demonstrate the effectiveness of our EnVLF method for RS retrieval.

Visual grounding and image captioning, as two other typical tasks of multi-modal machine learning, have potential benefits for text-to-image retrieval and image-to-text retrieval separately. Therefore, in future work, we will further improve by means of multi-level cross-modal feature learning and generative transformer modals, which are proved to be efficient in visual grounding and image captioning tasks [40,41]. Furthermore, obtaining accurate results from noisy data will also be part of our follow-up research to improve the practical value of RSCTIR [42].

Author Contributions: Conceptualization, L.H.; methodology, L.H.; software, S.L.; validation, S.L.; investigation, R.A.; data curation, Y.Z.; writing—original draft preparation, L.H.; writing—review and editing, J.T. and S.L.; supervision, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in the literature [5,9,33–35].

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Li, Y.; Ma, J.; Zhang, Y. Image retrieval from remote sensing big data: A survey. *Inf. Fusion* **2021**, *67*, 94–115. [CrossRef]
2. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (Cits), Kunming, China, 6–8 July 2016; pp. 1–5.
3. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 18–24 July 2021; pp. 5583–5594.
4. Yuan, Z.; Zhang, W.; Tian, C.; Rong, X.; Zhang, Z.; Wang, H.; Fu, K.; Sun, X. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
5. Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; Sun, X. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv* **2022**, arXiv:2204.09868.
6. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
7. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32897–32912.
8. Ning, H.; Zhao, B.; Yuan, Y. Semantics-consistent representation learning for remote sensing image–voice retrieval. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
9. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [CrossRef]
10. Park, D.H.; Darrell, T.; Rohrbach, A. Robust change captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 4624–4633.
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
12. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
13. Lv, Y.; Xiong, W.; Zhang, X.; Cui, Y. Fusion-based correlation learning model for cross-modal remote sensing image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
14. Yuan, Z.; Zhang, W.; Rong, X.; Li, X.; Chen, J.; Wang, H.; Fu, K.; Sun, X. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–19. [CrossRef]
15. Chaudhuri, U.; Banerjee, B.; Bhattacharya, A.; Datcu, M. CMIR-NET: A deep learning based model for cross-modal retrieval in remote sensing. *Pattern Recognit. Lett.* **2020**, *131*, 456–462. [CrossRef]
16. Chen, Y.; Lu, X. A deep hashing technique for remote sensing image-sound retrieval. *Remote Sens.* **2019**, *12*, 84. [CrossRef]

17. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
18. Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; Lin, J. Learning cross-modal retrieval with noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5403–5413.
19. Wang, L.; Li, Y.; Huang, J.; Lazebnik, S. Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 394–407. [CrossRef]
20. Chun, S.; Oh, S.J.; De Rezende, R.S.; Kalantidis, Y.; Larlus, D. Probabilistic embeddings for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8415–8424.
21. Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; Shen, Y.D. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *16*, 1–23. [CrossRef]
22. Cui, Y.; Yu, Z.; Wang, C.; Zhao, Z.; Zhang, J.; Wang, M.; Yu, J. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 797–806.
23. Faghri, F.; Fleet, D.; Kiros, J.; Fidler, S.V. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv* **2017**, arXiv:1707.05612.
24. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 18–24 July 2021; pp. 4904–4916.
25. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 104–120.
26. Ji, Z.; Chen, K.; Wang, H. Step-wise hierarchical alignment network for image-text matching. *arXiv* **2021**, arXiv:2106.06509.
27. Sun, S.; Chen, Y.C.; Li, L.; Wang, S.; Fang, Y.; Liu, J. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 6–11 June 2021; pp. 982–997.
28. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
29. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv* **2019**, arXiv:1908.08530.
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 18–24 July 2021; pp. 8748–8763.
31. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S.C.H. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9694–9705.
32. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18166–18176.
33. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
34. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
35. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Québec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Huang, Y.; Wu, Q.; Song, C.; Wang, L. Learning semantic concepts and order for image and sentence matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6163–6171.
38. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. Camp: Cross-modal adaptive message passing for text-image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5764–5773.
39. Wang, T.; Xu, X.; Yang, Y.; Hanjalic, A.; Shen, H.T.; Song, J. Matching images and text with multi-modal tensor fusion and re-ranking. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 12–20.
40. Chang, S.; Ghamisi, P. Changes to Captions: An Attentive Network for Remote Sensing Change Captioning. *arXiv* **2023**, arXiv:2304.01091.

41. Zhan, Y.; Xiong, Z.; Yuan, Y. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
42. Mikriukov, G.; Ravanbakhsh, M.; Demir, B. An Unsupervised Cross-Modal Hashing Method Robust to Noisy Training Image-Text Correspondences in Remote Sensing. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2556–2560.