remote sensing

MDPI

Article

A Fusion Encoder with Multi-Task Guidance for Cross-Modal Text–Image Retrieval in Remote Sensing

Xiong Zhang, Weipeng Li, Xu Wang, Luyao Wang, Fuzhong Zheng, Long Wang and Haisu Zhang *

School of Information and Communication, National University of Defense Technology, Wuhan 430074, China; zhangxiong17@nudt.edu.cn (X.Z.); liweipeng_prf@nudt.edu.cn (W.L.); wangxu91@nudt.edu.cn (X.W.); wangluyao@nudt.edu.cn (L.W.); zhengfuzhong21@nudt.edu.cn (F.Z.); wanglong17c@nudt.edu.cn (L.W.)
* Correspondence: zhanghaisu@nudt.edu.cn

Abstract: In recent years, there has been a growing interest in remote sensing image–text cross-modal retrieval due to the rapid development of space information technology and the significant increase in the volume of remote sensing image data. Remote sensing images have unique characteristics that make the cross-modal retrieval task challenging. Firstly, the semantics of remote sensing images are fine-grained, meaning they can be divided into multiple basic units of semantic expression. Different combinations of basic units of semantic expression can generate diverse text descriptions. Additionally, these images exhibit variations in resolution, color, and perspective. To address these challenges, this paper proposes a multi-task guided fusion encoder (MTGFE) based on the multimodal fusion encoding method, the progressiveness of which has been proved in the cross-modal retrieval of natural images. By jointly training the model with three tasks: image–text matching (ITM), masked language modeling (MLM), and the newly introduced multi-view joint representations contrast (MVJRC), we enhance its capability to capture fine-grained correlations between remote sensing images and texts. Specifically, the MVJRC task is designed to improve the model's consistency in joint representation expression and fine-grained correlation, particularly for remote sensing images with significant differences in resolution, color, and angle. Furthermore, to address the computational complexity associated with large-scale fusion models and improve retrieval efficiency, this paper proposes a retrieval filtering method, which achieves higher retrieval efficiency while minimizing accuracy loss. Extensive experiments were conducted on four public datasets to evaluate the proposed method, and the results validate its effectiveness.

Keywords: cross-modal retrieval; remote sensing images; fusion encoding method; joint representation; contrastive learning

Citation: Zhang, X.; Li, W.; Wang, X.; Wang, L.; Zheng, F.; Wang, L.; Zhang, H. A Fusion Encoder with Multi-Task Guidance for Cross-Modal Text–Image Retrieval in Remote Sensing. *Remote Sens.* **2023**, *15*, 4637. https://doi.org/10.3390/rs15184637

Academic Editor: Geert Verhoeven

Received: 24 June 2023
Revised: 6 September 2023
Accepted: 18 September 2023
Published: 21 September 2023

1. Introduction

The rapid advancement of space information technology and the exponential expansion of remote sensing image data have created a pressing need for the efficient and convenient extraction of valuable information from vast amounts of remote sensing images. In response to this demand, cross-modal retrieval between remote sensing images and text descriptions has emerged as a valuable approach. This retrieval process involves finding text descriptions that match given remote sensing images or identifying remote sensing images that contain relevant content based on text descriptions. The growing attention towards this field highlights its potential in addressing the aforementioned demand.

Recent studies on the cross-modal retrieval of remote sensing images and texts have predominantly followed a two-step approach, involving unimodal feature extraction (Figure 1a) and multimodal interaction (Figure 1b). During the unimodal feature extraction stage, remote sensing images and text data are transformed into numerical representations that capture their semantic content for further statistical modeling. Deep learning techniques, such as convolutional neural networks (CNNs) (e.g., VGGNet [1], ResNet [2]) and

vision Transformer networks [3], are commonly employed for extracting image features. Similarly, recurrent neural networks (RNNs) (e.g., LSTM [4], GRU [5]) and Transformer models (e.g., BERT [6]) are utilized for extracting textual features. In the subsequent multimodal interaction stage, the semantic consistencies between image and text features are leveraged to generate comprehensive feature representations that effectively summarize the multimodal data. Baltrusaitis et al. [7] classified multimodal feature representations into joint representations and coordinated representations. Joint representations merge multiple unimodal signals and map them into a unified representation, while coordinated representations process information independently for each modality while incorporating similarity constraints between different modalities. Following this framework, recent methods for multimodal interaction between remote sensing images and texts can be categorized into two groups: multimodal semantic alignment and multimodal fusion encoding.

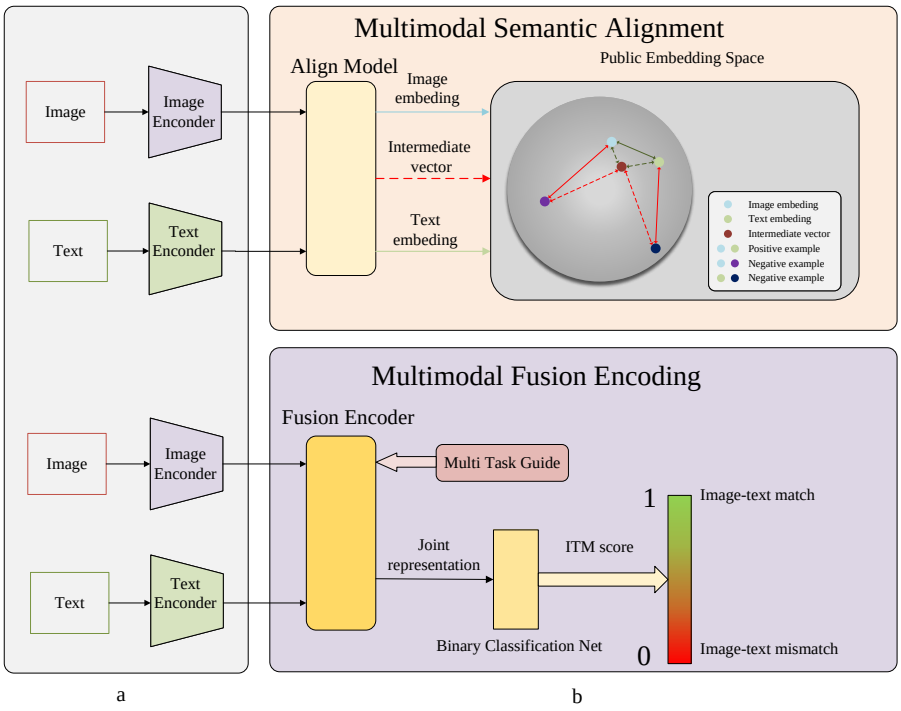

Figure 1. General framework for remote sensing image and text retrieval. (**a**) Unimodal feature extraction stage. (**b**) Multimodal interaction stage. The methods can be categorized into two groups based on the generation of a unified multimodal representation: multimodal semantic alignment and multimodal fusion encoding.

The upper part of Figure 1b illustrates multimodal semantic alignment methods [8–18]. These approaches aim to align image and text data in a public embedding space based on their semantic information. By doing so, images and texts with similar semantics are positioned closer to each other in this space. During cross-modal retrieval, the similarity between image and text features is determined by measuring their distance in the public embedding space, followed by sorting. In the context of multimodal interaction, the simple dot product or shallow attention mechanisms are commonly employed to calculate the similarity between images and texts. Triplet loss [19] and InfoNCE loss [20] are utilized either directly or through intermediate variables to impose constraints on the position and distance of image and text features within the public embedding space. The bottom half of Figure 1b depicts the method of multimodal data fusion encoding [21]. This approach involves feeding remote sensing images and text features into a unified fusion encoder to obtain joint representations of the image–text pairs. Subsequently, a binary classification task known as the image–text matching (ITM) task is performed to determine the degree of

compatibility between the image and text. During retrieval, the ITM score is employed as a measure of similarity between the image and text.

Significant advancements have been achieved in the cross-modal retrieval of natural images and texts, resulting in impressive average R@1 accuracies of 75.8% and 95.3% on the MS COCO and Flickr30k datasets, respectively [22]. However, when compared to natural images, remote sensing images possess three distinct characteristics. Firstly, they serve as objective representations of ground objects, leading to intricate and diverse semantic details within the images. This implies that remote sensing images can be dissected into multiple basic units for semantic expression. Secondly, unlike natural images, remote sensing images lack specific themes and focal points [23], which contributes to their pronounced multi-perspective nature. Consequently, the same remote sensing image can generate various descriptions from different perspectives, encompassing different combinations and permutations of the underlying fine-grained semantic units. Thirdly, remote sensing images of the same geographical area may exhibit variations in colors, brightness, resolution, and shooting angles due to factors such as weather conditions, photography equipment, and aircraft positions. These inherent characteristics pose substantial challenges in achieving effective cross-modal retrieval for remote sensing images.

The global similarity of image and text commonly arises from a complex aggregation of local similarities between image–sentence instances [24]. Due to the fine-grained semantic composition and multi-perspective nature of remote sensing images, it is essential to capture the intricate correlation clues between the image and text at a granular level. This includes establishing connections between specific image regions and corresponding textual words. Therefore, in order to accomplish this, researchers have explored the use of fine-grained unimodal features. For instance, region features [25] and patch features [21] have been utilized for images, while word features have been employed for texts [14,21,25]. These fine-grained correlations between images and texts are then established through cross-attention mechanisms between the modalities. However, despite utilizing high-performance unimodal encoders, simplistic interaction calculations between the features may still fall short when dealing with complex visual-and-language tasks [26]. To address this limitation, Li et al. [21] introduced a large-scale Transformer network as a multimodal fusion encoder. By leveraging multiple multi-head cross-attention modules, this approach enabled complex interaction calculations to be performed on the fine-grained features across modalities, thereby further exploring potential fine-grained correlations between the modalities.

However, existing multimodal fusion encoding models for remote sensing image–text primarily rely on the ITM task as the sole training objective, lacking precise supervision signals for capturing fine-grained correlations between images and texts. This limitation makes it challenging to provide efficient supervision for the correlation between specific words in the text and corresponding regions in the image. To address this issue, we have incorporated the masked language modeling (MLM) task from the recent vision-language pre-training (VLP) model [27–29]. In the MLM task, certain words in the text are masked, and the model is trained to predict these masked words using context information from the masked text and patch-level information from the image. This approach facilitates a more effective capture of fine-grained image–text correlations.

In addition, the variations in remote sensing image acquisition, including weather conditions, sensor configurations, and viewing angles, present challenges for models to establish fine-grained correlations between remote sensing images and textual data, as well as accurately determine their similarity. To overcome these challenges, we propose the multi-view joint representations contrast (MVJRC) task, which incorporates automatic contrast, histogram equalization, brightness adjustment, definition adjustment, flipping, rotation, and offset operations to simulate imaging differences. Additionally, a weight-sharing Siamese network is designed to maximize the similarity between augmented views of the same remote sensing image and the joint representations of the corresponding text during training. By leveraging the update gradient alternation, the model effectively utilizes

the mutual information contained in the joint representations of the same remote sensing image under different views as supervision signals. The MVJRC task successfully filters out the noise interference caused by imaging differences in remote sensing images. It achieves strong consistency in the joint representations of different views for texts and remote sensing images, facilitating the easier discrimination of paired samples. Furthermore, MVJRC enhances the complex cross-attention module between modalities by providing additional complementary signals, thereby enabling consistent fine-grained correlations.

The increasing computational complexity associated with large-scale networks can lead to reduced efficiency in measuring the similarity of multimodal data during cross-modal retrieval. While identifying negative samples with low similarity (easy negatives) is straightforward, identifying negative samples with high similarity (hard negatives) often requires a more intricate model. To address this challenge, we propose the retrieval filtering (RF) method. This method employs a small-scale network as a filter and utilizes knowledge distillation [30] to transfer the "knowledge" of similarity measurements from the complex fusion network to the filter. During retrieval, the small-scale filter is initially used to screen out easy negatives, and the top k samples with high similarity are then fed into the complex fusion encoder for similarity calculation and re-ranking. By adopting the RF method, retrieval efficiency can be significantly improved while ensuring minimal accuracy loss, even with a large sample size.

In this research, we introduced a multi-task guided fusion encoder (MTGFE) for cross-modal retrieval of remote sensing images and texts. The key contributions of this paper can be summarized as follows:

(1) The model was trained using a combination of the ITM, MLM, and MVJRC tasks, enhancing its ability to capture fine-grained correlations between remote sensing images and texts.
(2) The introduction of the MVJRC task improved the consistency of feature expression and fine-grained correlation, particularly when dealing with variations in colors, resolutions, and shooting angles of remote sensing images.
(3) To address the computational complexity and retrieval efficiency limitations of large-scale fusion coding networks, we proposed the RF method. This method filters out easy negative samples, ensuring both high retrieval accuracy and efficient retrieval performance.

The remaining part of this paper is organized as follows. In Section 2, related work on the remote sensing image–text cross-modal retrieval, text and image encoders based on Transformer, vision-language pre-training (VLP) models, and contrastive learning is summarized and analyzed. In Section 3, the system architecture of our model is described in detail, with a focus on the design of the training task. In Section 4, comparative and ablation experiments are conducted to demonstrate the superiority and effectiveness of our method. Meanwhile, the reason for the underperformance of the method is analyzed. In Section 5, the discussions and conclusions are presented.

2. Related Work

This section provides an overview of the relevant literature on remote sensing image–text cross-modal retrieval, focusing on the following topics: text and image encoders built upon the Transformer architecture, Vision-language pre-training (VLP) models, and contrastive learning methods.

2.1. Remote Sensing Image–Text Cross-Modal Retrieval

Remote sensing image–text cross-modal retrieval can be divided into two stages: image caption-based retrieval and direct measurement of image–text similarity. Shi et al. [31] proposed an automatic caption generation framework for remote sensing images, demonstrating the technical feasibility of this approach. Qu et al. [32] and Lu et al. [23] contributed a publicly available remote sensing image–text dataset and proposed automatic remote sensing image caption generation and image–text cross-modal retrieval based on captions.

However, these two-stage methods often suffer from information loss at each stage, leading to reduced retrieval accuracy. To address this issue, Rahhal et al. [12] employed the InfoNCE loss to map the global feature vectors of images and texts to a public embedding space, directly calculating the similarity between remote sensing images and texts. Abdullah et al. [13] utilized the average fused representation of five text sentences corresponding to each remote sensing image as the text feature. This approach effectively aligned the text and image features and enhanced the semantic richness of the images in the public embedding space. Cheng et al. [14] introduced a shallow attention mechanism to combine the fine-grained features of image regions and text words as intermediate features. This constrained the projection of images and texts in the public embedding space, thereby improving the quality of semantic alignment between images and texts. Lv et al. [15] divided the image-text information into complementary information and consistency information. They employed the Fully connected (FC) network to fuse the image and text information, obtaining joint features. These joint features were then used as intermediate features to independently align the image and text features with them. Yuan et al. [8] enhanced the fine-grained semantic expression ability of image features by fusing multi-scale information. The image features were used to guide the generation of text features during their interaction, followed by alignment in the public embedding space using triplet-loss. Yuan et al. [16] proposed the multi-level information dynamic fusion (MIDF) to fuse the local and global features of remote sensing images, enhancing the semantic expression capability of the images. Additionally, they introduced the multivariate re-rank (MR) algorithm to improve retrieval accuracy. Cheng et al. [17] employed a combination of channel attention, spatial attention, and position attention mechanisms to fuse multi-scale information from remote sensing images. The interaction between modalities was calculated through fine-grained alignment between image regions and text words to express their similarity. Yuan et al. [18] utilized knowledge distillation to transfer the "dark knowledge" learned by the asymmetric multimodal feature matching network (AMFMN) model [8], resulting in improved cross-modal retrieval efficiency. Mikriukov et al. [33,34] focused on using hash feature vectors instead of real value feature vectors in the public embedding space, significantly enhancing the efficiency of cross-modal retrieval. Li et al. [21] designed a remote sensing image–text cross-modal retrieval model that initially performed alignment and then fusion. They utilized vision Transformer and BERT to extract fine-grained unimodal features of image regions and text words, respectively. Through contrastive learning [35], the unimodal features were made semantically consistent. A multi-layer Transformer encoder was employed to model the correlation of more complex fine-grained features between images and texts and extract their joint features. The similarity between images and texts was modeled using the ITM task, yielding competitive results on multiple datasets.

The comparison of the studies mentioned above highlights the significance of fine-grained semantic expression in remote sensing images (e.g., through fused multi-scale features and fine-grained regional features) and the importance of modeling fine-grained interactions between modalities (such as generating intermediate features using attention mechanisms, utilizing visual features to guide text feature generation, and employing large-scale cross-attention fusion encoders) to enhance the accuracy of remote sensing image–text cross-modal retrieval. Therefore, in our approach, we specifically focused on capturing the fine-grained semantic features of the unimodal representations and selected a large-scale Transformer as the fusion encoding module between modalities.

2.2. Text and Image Encoders Based on Transformer

The Transformer architecture, originally proposed by Vaswani et al. [36], has emerged as a prominent framework in natural language processing (NLP) for tasks like machine translation. Unlike traditional RNN text encoders, Transformer utilizes bidirectional global attention and mask attention mechanisms, which are advantageous for modeling long-term dependencies in text and enabling efficient parallel computation. Building upon this architecture, Devlin et al. introduced the BERT model [6]. BERT employs MLM and next

sentence prediction (NSP) tasks for self-supervised training on large-scale text datasets, enhancing its ability to represent bidirectional text information. When utilizing BERT for text encoding, the text sentence is first decomposed into tokens using the WordPiece [37] method. The output consists of feature vectors corresponding to the tokens, along with classification labels denoted as [cls]. The token features represent the fine-grained features of individual text words, while the classification label features are often employed as features for the entire text sentence.

Dosovitskiy et al. [3] introduced the ViT model as an image encoder based on the Transformer architecture. In this model, images are divided into multiple 16 $\times$ 16 pixel patches, which are then sequentially input into the Transformer. Through self-attention calculations among these image patches, the ViT model encodes the image into fine-grained patch features, along with classification label [cls] features that can serve as global features.

The Transformer-based ViT and BERT models exhibit strong capabilities in expressing fine-grained features within each modality, and their feature structures are similar. These characteristics make them suitable choices for conducting interactive calculations. As a result, we utilized these encoders as unimodal data encoders in this study.

2.3. Vision-Language Pre-Training (VLP) Models

Vision-language pre-training (VLP) focuses on acquiring multimodal representations from large-scale image–text pairs, aiming to enhance performance in various visual and language tasks, such as image–text cross-modal retrieval, natural language for visual reasoning (NLVR), and visual question answering (VQA) [27]. In recent studies, the fusion and encoding of visual and language data have been primarily accomplished using multi-layer Transformers [38]. The training tasks include ITM [27,39,40] and MLM [27–29]. The ITM task is a binary task that determines whether an image–text pair is a match based on joint representations. On the other hand, the MLM task involves masking certain words in the text and predicting them using context and image information, which facilitates the fine-grained fusion of words and image patches. Existing methods for remote sensing image–text retrieval based on fusion encoding often solely rely on the ITM task, which may not be sufficient for capturing fine-grained correlations between modalities. To address this limitation, we introduce the MLM task from VLP-related models in this study to enable joint model training and enhance the exploration of fine-grained correlations between remote sensing images and texts.

2.4. Contrastive Learning

Contrastive learning [35] is an advanced technique for representation learning that aims to bring similar samples (positive samples) closer together in the public embedding space while increasing the distance between dissimilar samples (negative samples). In unimodal contrastive learning, a Siamese network is employed to extract features from data samples that have undergone different data augmentations, such as modifying image color and shape or introducing noise to the text. The learning objective is achieved by comparing these features with a large number of negative samples [41–43]. Chen et al. [44] proposed SimSiam, a contrastive learning method that does not require negative examples. SimSiam incorporates two modules, namely the project head and predict head, into the Siamese network with shared weights, and representation learning is performed through alternating gradient updates. For multimodal contrastive learning, methods like CLIP [45] and ALIGN [46] use matched image-text pairs as positive samples and unmatched image–text pairs as negative samples. These approaches undergo pre-training on large-scale image–text datasets and achieve competitive results in downstream tasks such as cross-modal retrieval through fine-tuning.

In contrast to previous task-oriented learning approaches, contrastive learning focuses on maximizing the mutual information [41] between pairs of instances to enhance feature consistency and expression. In the context of remote sensing images, which often exhibit significant differences in resolution, color, and angle, maintaining feature consistency and

fine-grained correlations between modalities can be challenging. To address this issue, we adopted the MVJRC method inspired by SimSiam and constructed a fusion encoding model with shared weights. The presented approach aimed to maximize the similarity of joint representations across different views and ensure consistency in fine-grained correlations between modalities.

3. Method

To achieve a fine-grained association between remote sensing images and texts, we first utilize ViT and BERT (using the first 6 layers and parameters) models to extract patches and tokens features from images and texts, respectively. Afterwards, to represent the complex interaction of fine-grained semantic units for images and texts, we employ a large-scale Transformer (initialized with the last 6 layers and parameters of the BERT) as the fusion encoder to model the fine-grained association between images and texts. To better utilize the image–text association information in the annotated data, we utilize MLM task to mine the ground truth label (real words in the manually annotated dataset) of randomly masked tokens as the supervision signal, guiding the model to learn the fine-grained association between images and texts. Meanwhile, a MVJRC task is employed to mine the joint representation of text and different imaging remote sensing images as the supervision signal, ensuring consistency between the joint representation and fine-grained association. Additionally, we use the ITM task to align the remote sensing images and texts by using the supervision signal of whether the image and text match, facilitating cross-modal retrieval between remote sensing images and texts.

Figure 2 illustrates the overall structure of the model. Initially, the visual and language features of the image–text pair are generated separately by their respective unimodal encoders. These features are then paired and passed into the fusion encoder. The model is trained jointly through the ITM, MLM, and MVJRC tasks. During cross-modal image–text retrieval, the results are ranked based on the ITM score and provided to the user. After training the fusion encoder, a small-scale multilayer perceptron (MLP) network is trained using knowledge distillation. This MLP network functions as a retrieval filter to filter out easily identifiable negative samples. Subsequently, the results are re-ranked by the fusion encoder.

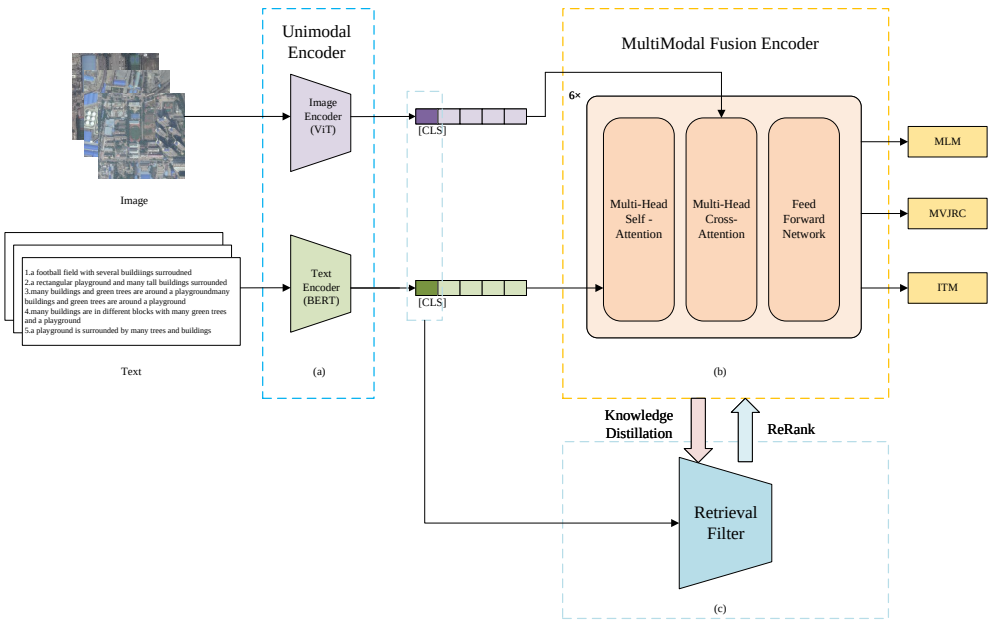

Figure 2. Overview of the MTGFE model. It comprises two components: (**a**) a unimodal encoder that utilizes the ViT and BERT (first 6 layers) models to extract features from images and texts, and (**b**) a multimodal fusion encoder (initialized using the parameters of the last 6 layers of the BERT)

that generates joint image–text representations through ITM, MLM, and MVJRC tasks. Additionally, (**c**) a retrieval filter is trained via knowledge distillation. During retrieval, the filter eliminates easy negatives, and the teacher network performs re-ranking.

3.1. Unimodal Encoder

We select the ViT and BERT models, which leverage self-attention mechanisms, as the unimodal encoders for remote sensing images and texts. These models facilitate the fine-grained semantic representation of unimodal data.

3.1.1. Image Encoder

The image encoder, which is denoted by $f_{img}(\cdot)$, adopts the ViT-B/16 model structure and is initialized using the pre-training weights on ImageNet-1k [47]. According to reference [3], a given image is segmented into multiple 16 × 16 pixel patches. After linear projection, the learnable classification labels are embedded into the special token [cls]. The encoding output is $S = f(I) = \{v_{cls}, v_1, v_2, ..., v_m\}$, where v_{cls} is the classification label feature, v_i denotes the feature of the i-th patch, and m is the number of patches.

3.1.2. Text Encoder

The first 6 layers and weights of the pre-trained $BERT_{base}$ model [6] are used as the text encoder, denoted as $f_{txt}(\cdot)$. It has 6 Transformer blocks. Given a text description T, WordPiece [37] is used at first to obtain the embedded representation of the tokens in the sentence, and a classification label token ([cls]) is added at the start, as denoted by $T = \{t_{cls}, t_1, t_2, \cdots, t_n\}$. During the execution of the MLM task (as outlined in Section 3.3), approximately 15% of the tokens are randomly masked and substituted with the special token [mask], as indicated by $T_{mask} = \{t_{cls}, t_1, t_{mask}, \cdots, t_n\}$; here, n is the length of text tokens and t_{cls} indicates the embedding representing the classification label [cls]. The encoded text features are represented as $W = f_{txt}(T) = \{w_{cls}, w_1, w_2, \cdots, w_n\}$ and $W_{mask} = f_{txt}(T_{mask}) = \{w_{cls}, w_1, w_{mask}, \cdots, w_n\}$, respectively; here, w_{cls} represents the text classification label feature, and it is often used as a global vector for the text in some downstream tasks; w_i is the feature vector of the i-th token; and w_{mask} represents the feature vector of the special token [mask].

3.2. Multimodal Fusion Encoder

The multimodal fusion encoder comprises six layers of Transformer blocks, utilizing fine-grained features such as image patches and text tokens. To enable greater gradient flow for the image encoder, the image features are independently fed into each multi-head cross-attention layer, where they serve as the key and value for attention calculations. Conversely, the text tokens are treated as the query and fed into the multi-head cross-attention layer after the computation of the multi-head self-attention layer. The multiple stacked self-attention and cross-attention layers facilitate the calculation of fine-grained correlations between text tokens and image patches, and improve image encoder parameters and enhance visual representations.

The fusion encoder is initialized with the weights of the last 6 layers of $BERT_{base}$ [6], denoted as $f_{fusion}(\cdot)$. Each block in the architecture consists of three sub-layers: a multi-head self-attention layer, a multi-head cross-attention layer, and a feed-forward network (FFN) layer. Within each attention sub-layer, a residual connection is employed, where the input and output are added together prior to layer normalization [48]. Here, the input of the self-attention layer is the embedded feature of the text W; when executing the MLM task, the input is W_{mask}. The self-attention layer maintains three learnable parameter

matrices,W^Q, W^K, and W^V, for each input token embedding. The calculation approach for each attention head is provided in Equation (1).

$$Attention(Q, K, V) = \text{softmax}(\frac{(QW^Q) \cdot (KW^K)^T}{\sqrt{d_K}})VW^V \quad (1)$$

where d_K is the dim of the input key. In the fusion of multi-head attention, we need to concatenate the output of each attention head $head_i$ on the dimension of $dim = 1$ and multiply it with a learnable parameter matrix W^O , as shown in Equation (2).

$$MultiHead(Q, K, V) = \text{Concat}(head_1, head_2, \cdots, head_h)W^O \quad (2)$$

Here, h is the number of attention heads.

The calculation of the multi-head cross-attention layer is similar to that of the multi-head self-attention layer, except for that the output of text embedding W through the self-attention layer is used as Q, whereas the visual embedding S is used as K and V.

The FFN sub-layer is a FC network that utilizes the Gelu [49] activation function. This activation function applies a nonlinear transformation to the output of the cross-attention network. The hidden vector of the last layer is taken as the feature output of the fusion encoder, represented by $U = f(S, W) = \{u_{cts}, u_1, u_2, \cdots, u_n\}$, where u_{cls} is the classification label feature of the image–text joint feature, u_i is the joint feature corresponding to the i-th text token, and n is the number of input text elements.

3.3. Training Task of the Multimodal Fusion Encoder

During the training process, we incorporate three tasks, namely MLM, MVJRC, and ITM, to collectively guide the training of the multimodal fusion encoder.

3.3.1. Masked Language Modeling (MLM)

The MLM task (shown in Figure 3), derived from the BERT [6] model, involves randomly masking 15% of the tokens in a given text sentence. By incorporating the MLM task into the fusion module, the training process is transformed into a self-supervised denoising procedure. This requires the masked tokens to utilize both the unmasked contextual information (through the self-attention mechanism) and additional image information (through the cross-attention mechanism) for reconstruction. This approach strengthens the fine-grained correlations between text tokens and image patches, enhancing their alignment and coherence.

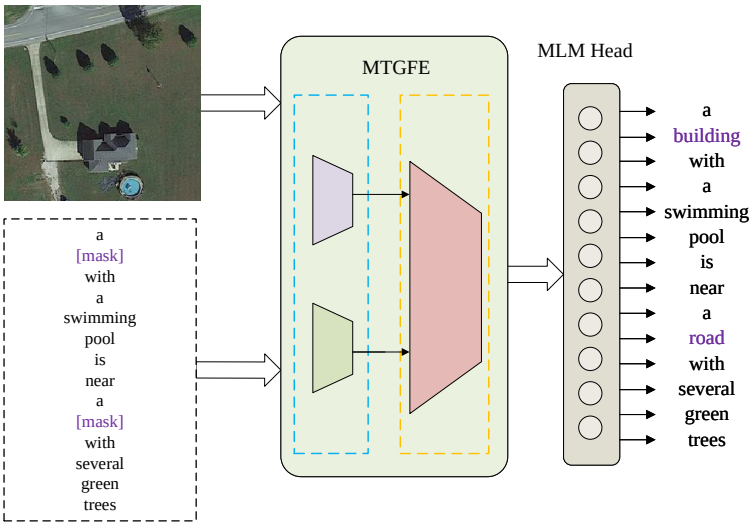

Figure 3. The diagram of the MLM task, where [mask] represents the masked token, and the purple text on the right side represents the actual values for the [mask] tokens. The goal of the task is to correctly predict these masked tokens.

A fully connected network MLM head is added after the output of the fusion encoder, and its input is the image–text joint representation, given by U. The network output employs the SoftMax function for multiple classification tasks and is then mapped to a sequence of $dim = \text{len}(vocabulary)$. Here, vocabulary is the word dictionary introduced from $BERT_{base}$, with a length of 30,522. The MLM task involves minimizing the cross-entropy loss between the predicted value and the ground truth label, which is provided in Equation (3).

$$L_{mlm} = H(y^{mask}, p^{mask}(I, T_{mask})) \tag{3}$$

where y^{mask} refers to the ground truth label of the predicted vocabulary, (I, T_{mask}) refers to the image–text pair after the masking operation, $p^{mask}(I, T_{mask})$ refers to the model prediction for the masked vocabulary, and $H(\cdot)$ refers to the cross-entropy loss function.

3.3.2. Multi-View Joint Representations Contrast (MVJRC)

To enhance the coherence of joint features and capture fine-grained correlations between a specific target and its corresponding text under varying imaging conditions, such as resolution, color, and shooting angle, we propose a weight-sharing MTGFE Siamese network (Figure 4). Various image augment operations are employed to simulate the imaging discrepancies in remote sensing images. The joint representation undergoes self-supervised training, where the objective is to maximize the similarity of the joint representations between remote sensing images captured from different perspectives and their corresponding paired text. Specifically, a project head and prediction head are added after MTGFE, which are, respectively, expressed as f_{proj} and f_{pred}. A project head (f_{proj}) has three FC layers, and each FC layer has a batch normalization (BN) layer [50]. Apart from the output layer, the activation function utilized in each BN layer is the rectified linear unit (ReLU) [51]. A prediction head (f_{pred}) is a two-layer FC layer connected by the BN layer and the ReLU activation function.

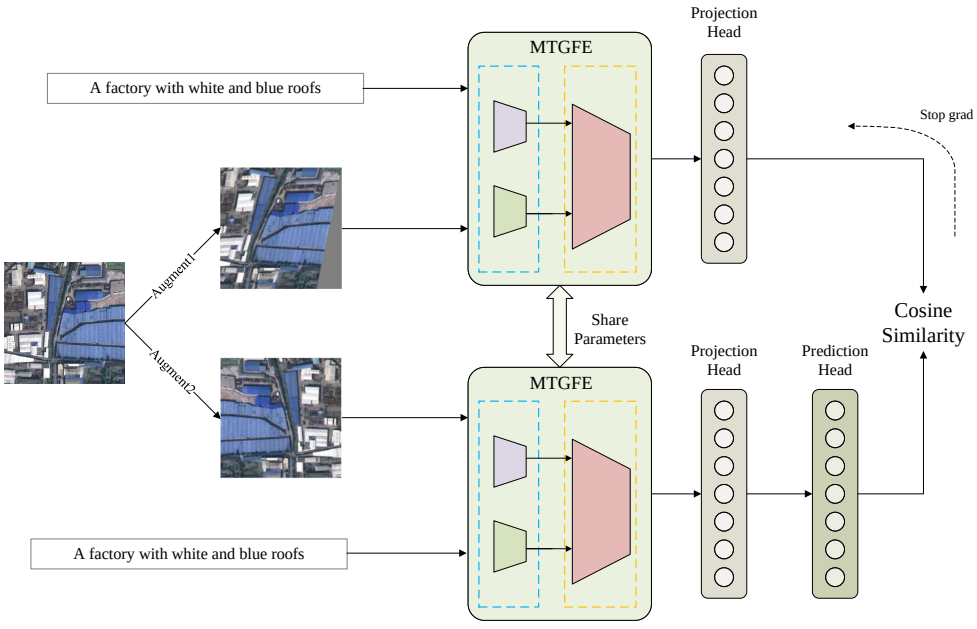

Figure 4. The MVJRC task involves setting up a Siamese network with shared parameters from MTGFE. The cosine similarity of the joint representations u_{cls} is calculated after the projection head and prediction head, and the gradient is updated alternately.

For a given image–text pair (I, T), Randaugment [52] is used for random image augment to obtain (I_1, T) and (I_2, T), whose fusion representations are denoted as U_1 and U_2, respectively, and their classification label features u_1 and u_2 are used in subsequent oper-

ations. Let $z_1 = f_{proj}(u_1)$, $z_2 = f_{proj}(u_2)$, and $p_1 = f_{pred}(z_1)$, $p_2 = f_{pred}(z_2)$, and define $S(\cdot)$ as the cosine similarity of two vectors, then

$$S(p_1, z_2) = \frac{p_1}{||p_1||_2} \cdot \frac{z_2}{||z_2||_2} \tag{4}$$

here, $||\cdot||$ represents $l_2 - norm$. The objective of the MVJRC task is to optimize the similarity between joint representations of various augments image-text pairs. The loss function for this task can be defined as follows:

$$L_{mvjrc} = -\frac{1}{2}(S(p_1, z_2) + S(p_2, z_1)) \tag{5}$$

The loss for each individual sample is computed, and then the mean loss is calculated within each minibatch. According to reference [44], to prevent the model from collapsing, a stop gradient operation (stopgrad) is introduced when updating the gradient. That is, when calculating the gradient from $S(p, z)$, only accept the gradient from p each time. The mathematical expression for the MVJRC loss is as follows:

$$L_{mvjrc} = -\frac{1}{2}(S(p_1, \text{stopgrad}(z_2)) + S(p_2, \text{stopgrad}(z_1))) \tag{6}$$

When updating the encoder parameters of the image–text pair (I, T) , the first item does not receive the gradient from z_2 , and the second item only accepts the gradient from p_2 . See Algorithm 1 for the pseudocode of MVJRC.

Algorithm 1 MVJRC Task Pseudocode

```
# f: MTGFE Net, our fusion encoding model
# f_proj: Projection head
# f_pred : Prediction head
    for (I, T) in dataloader: do
        I_1, I_2 = aug(I), aug(I)    # Image augmentation
        u_1, u_2 = f(I_1, T)_[cls], f(I_2, T)_[cls]
        z_1, z_2 = f_proj(u_1), f_proj(u_2)
        p_1, p_2 = f_pred(z_1), f_pred(z_2)
        L = −0.5 ∗ (S(p1, z2) + S(p2, z1))    # loss
        l.backward()    # Gradient return
        update( f, f_proj, f_pred)    # Parameters update
    end for
    function S(p, z)( )    # Calculation of cosine similarity
        z= z.detach()    # Stop gradient
        p = normalize(p, dim = 1)    # l_2-normalize
        z = normalize(z, dim = 1)    # l_2-normalize
        s = (p ∗ z).sum(dim = 1).mean()
        return s
    end function
```

3.3.3. Image–Text Matching (ITM)

In order to assess the similarity between images and text and determine their compatibility, we employ the ITM head to perform a linear mapping of the joint representation onto the [0,1] interval. A higher value approaching 1 indicates a greater image–text similarity. During the cross-modal retrieval process for remote sensing images and text, the ITM score serves as the ranking criterion and is presented to the user. The ITM head is a FC layer that outputs $dim = 2$. Linear mapping is utilized to project the jointly represented classification label feature u_{cls} into a 2D prediction p^{itm} . The ITM loss quantifies the disparity between the minimized prediction and the ground truth label (whether the images and

texts match in the manually annotated dataset) in terms of probability distribution, which can be defined as Equation (7).

$$L_{itm} = H(y^{itm}, p^{itm}) \tag{7}$$

where y^{itm} is the true matching value of the given image–text pair and $H(\cdot)$ denotes the cross-entropy loss function. In training, the probability of y^{itm} for the input image–text pair (I, T) is set to 1; the negative examples are randomly selected for each image and text in the minibatch and denoted as $(I, \hat{T})$ and $(\hat{I}, T)$, respectively, which is set to 0.

The overall loss of the MTGFE training is as follows:

$$L = L_{mlm} + L_{mvjrc} + L_{itm} \tag{8}$$

3.4. Retrieval Filtering (RF)

Knowledge distillation, a machine learning technique, trains a compact model to mimic a larger, complex one. It involves transferring knowledge from the larger "teacher" model to the smaller "student" model. In order to improve the efficiency of MTGFE cross-modal retrieval, after model training, a simple FC network is designed as a retrieval filter (Figure 5). Knowledge distillation transfers knowledge from the MTGFE (teacher model) to the retrieval filter (student model). The input of the retrieval filter is the concatenation of image and text classification label features, which includes three FC layers. Following the initial two FC layers, BN and the ReLU activation function are applied, which are consistent with the ITM head architecture. The final linear layer then transforms the output of the hidden layer into a two-dimensional vector.

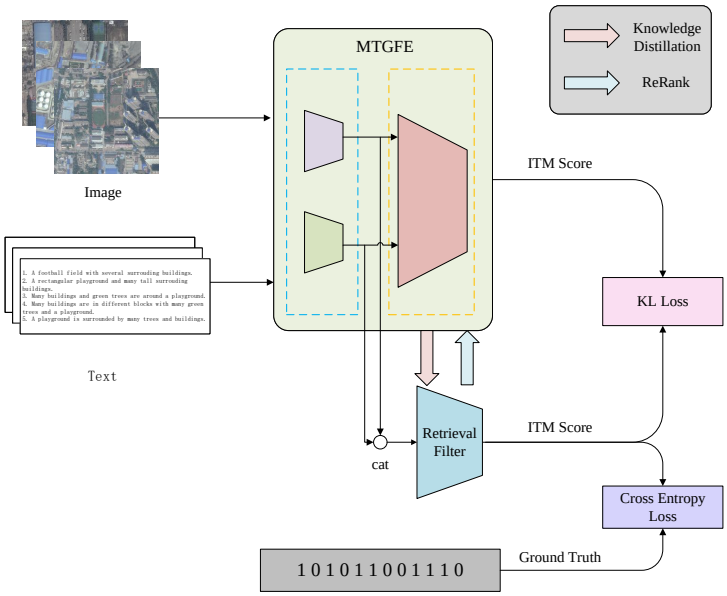

Figure 5. Retrieval Filtering architecture. Knowledge distillation is utilized to transfer the knowledge from MTGFE to the retrieval filter. During the retrieval process, the retrieval filter is employed to exclude easily distinguishable negatives, while samples with higher similarity are forwarded to MTGFE for recalibration and ranking.

The MTGFE's ITM output and the manually annotated ground truth label (whether the images and texts match in the manually annotated dataset) are utilized as the soft target and hard target supervision signals, respectively, for the student model, considering the same set of image–text samples. The distribution biases between these signals are

calculated using the Kullback–Leibler (KL) loss and cross-entropy loss. The calculation methods are as follows:

$$L_{soft} = KL(p_{tea}^{itm}, p_{stu}^{itm}) \tag{9}$$

$$L_{hard} = H(y^{itm}, p_{stu}^{itm}) \tag{10}$$

where $KL(\cdot)$ represents the KL divergence loss, p_{tea}^{itm} represents the ITM output of the teacher model, p_{stu}^{itm} represents the predicted ITM value of the student model, and y^{itm} represents the ground truth label. Finally, the distillation loss of the model can be obtained as follows:

$$L_{distll} = L_{soft} + \alpha L_{hard} \tag{11}$$

Here, α denotes a constant hyperparameter. The student and teacher models use the same unimodal encoder to extract features. The difference between the two scenarios is that the student model only inputs the image and text classification label features, v_{cls} and w_{cls}.

4. Experimental Results and Analysis

To substantiate the efficacy of the proposed method in remote sensing image–text cross-modal retrieval tasks, we performed comprehensive experiments on four publicly available datasets. Furthermore, we conducted ablation tests to provide additional validation for the presented approach. It is important to mention that, in Section 4.5, we exclusively employed the retrieval filtering method to evaluate its effectiveness, whereas the remaining experimental results were computed using MTGFE.

4.1. Datasets and Evaluation Indicators

In the experiments, we used four publicly available remote sensing image–text datasets: UCM-captions [32], Sydney-captions [32], RSICD [23], and RSITMD [8]. The basic information of each dataset is given in Table 1.

Table 1. Basic information of datasets.

Dataset	Images	Captions	Captions per Image	No. of Classes	Image Size
UCM-captions	2100	10,500	5	21	256 × 256
Sydney-captions	613	3065	5	7	500 × 500
RSICD	10,921	54,605	5	31	224 × 224
RSITMD	4743	23,715	5	32	256 × 256

In the evaluation, we employed recall at K (R@K), where K represents the rank position (1, 5, and 10), as the performance metric. R@K measures the percentage of correct samples within the top K ranked results for a given query. Additionally, we introduced the mR indicator, which represents the arithmetic mean of R@K values, to evaluate the performance of the proposed method.

4.2. Implementation Details

The experiments were performed on four NVIDIA GeForce RTX 3090 GPUs. All images were standardized to a size of 224 × 224 pixels and augmented using Randaugment [52]. To simulate variations in remote sensing images, nine augment methods ("Identity", "AutoContrast", "Equalize", "Brightness", "Sharpness", "ShearY", "TranslateX", "TranslateY", and "Rotate") were selected. However, since strong image augment transformations can disrupt the matching relationship between remote sensing images and texts, we applied relatively mild Randaugment function parameters, specifically (2,7). Here, '2' indicates that two methods were randomly chosen from the aforementioned sequence of image augment methods, while "7" represents the amplitude of the image augment. For image, text, and fused representations, the dimensions of the token and patch features

were set to 768. We utilized PyTorch's DistributedDataParallel tool for distributed training and incorporated distributed BN. During the training of the multimodal fusion encoder, a batch size of 32 was employed, and the training process spanned 60 epochs. The AdamW optimizer [53] with a weight decay of 0.02 was employed, and a cosine schedule was applied to decay the learning rate from 0.0001 during the first 1000 iterations. When training the student network, the distillation hyperparameter α was set to 0.2, the batch size was adjusted to 128, and the optimizer parameters remained unchanged.

4.3. Experimental Results and Analysis

During the experiments, we conducted a comparative analysis of the proposed method against the most up-to-date models, including VSE++ [9], SCAN [10], MTFN [11], AMFMN [8], SAM [14], LW-MCR [18], MAFA-Net [17], FBCLM [21], and GaLR [16]. Table 2 provides an overview of the performance of the proposed method as well as the baseline models on four datasets: UCM-captions, Sydney-captions, RSICD, and RSITMD. The superior results are highlighted in bold. In this context, "text retrieval" refers to the task of matching relevant textual descriptions with images based on specific criteria, while "image retrieval" denotes the task of matching relevant remote sensing images with textual descriptions using specific criteria.

Table 2. Experimental results of remote sensing image–text cross-modal retrieval on UCM-captions, Sydney-captions, RSICD, RSITMD datasets, and comparison with baseline models.

Approach	UCM-Captions Dataset							Sydney-Captions Dataset						
	Text Retrieval			Image Retrieval			mR	Text 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++	12.38	44.76	65.71	10.1	31.8	56.85	36.93	24.14	53.45	67.24	6.21	33.56	51.03	39.27
SCAN	14.29	45.71	67.62	12.76	50.38	77.24	44.67	18.97	51.72	74.14	17.59	56.9	76.21	49.26
MTFN	10.47	47.62	64.29	14.19	52.38	78.95	44.65	20.69	51.72	68.97	13.79	55.51	77.59	48.05
SAM	11.9	47.1	76.2	10.5	47.6	93.8	47.85	9.6	34.6	53.8	7.7	28.8	59.6	32.35
AMFMN	16.67	45.71	68.57	12.86	53.24	79.43	46.08	29.31	58.62	67.24	13.45	60	81.72	51.72
LW-MCR	13.14	50.38	79.52	18.1	47.14	63.81	45.35	20.69	60.34	77.59	15.52	58.28	80.34	52.13
MAFA-Net	14.5	56.1	**95.7**	10.3	48.2	80.1	50.82	22.3	60.5	76.4	13.1	61.4	81.9	52.6
FBCLM	28.57	63.81	82.86	27.33	72.67	94.38	61.6	25.81	56.45	75.81	27.1	**70.32**	**89.68**	57.53
MTGFE	**47.14**	**78.1**	90.95	**40.19**	**74.95**	**94.67**	**71**	**44.83**	**68.97**	**86.21**	**38.28**	69.31	83.1	**61.52**

Approach	RSICD Dataset							RSITMD Dataset						
	Text Retrieval			Image Retrieval			mR	Text 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	18.1	10.43	10.38	27.65	39.6	7.79	24.87	38.67	24.83
SCAN	5.85	12.89	19.84	3.71	16.4	26.73	14.24	11.06	25.88	39.38	9.82	29.38	42.12	26.27
MTFN	5.02	12.52	19.74	4.9	17.17	29.49	14.81	10.4	27.65	36.28	9.96	31.37	45.84	26.92
SAM	12.8	31.6	47.3	11.5	35.7	53.4	32.05	-	-	-	-	-	-	-
AMFMN	5.39	15.08	23.4	4.9	18.28	31.44	16.42	10.63	24.78	41.81	11.51	34.69	54.87	29.72
LW-MCR	4.39	13.35	20.29	4.3	18.85	32.34	15.59	9.73	26.77	37.61	9.25	34.07	54.03	28.58
MAFA-Net	12.3	35.7	**54.41**	12.9	32.4	47.6	**32.55**	-	-	-	-	-	-	-
FBCLM	13.27	27.17	37.6	**13.54**	**38.74**	**56.94**	31.21	12.84	30.53	45.89	10.44	37.01	57.94	32.44
GaLR	6.59	19.9	31	4.69	19.5	32.1	18.96	14.82	31.64	42.48	11.15	36.68	51.68	31.41
MTGFE	**15.28**	**37.05**	51.6	8.67	27.56	43.92	30.68	**17.92**	**40.93**	**53.32**	**16.59**	**48.5**	**67.43**	**40.78**

In Table 2, the performance metrics for VSE++, SCAN, and MTFN are obtained from reference [8], while the results of the other models are cited from their respective original papers. For the UCM-captions, Sydney-captions, and RSICD datasets, we followed the partitioning of the training set, validation set, and test set as defined by the dataset contributors. During the training phase, the model parameters were adjusted solely using the training set. The performance data presented in Table 2 are exclusively derived from

the test set. However, for the RSITMD dataset, the contributors only provided a division of the data into a training set and a test set. Thus, after training our model on the provided training set, the model's performance was measured on the test set.

Results on UCM-captions: The performance of the proposed approach on the UCM-captions dataset is displayed in the upper left section of Table 2. The mR metric of the method surpassed that of the best model by 9.4%. Except for the R@10 score in text retrieval, the method outperformed the baseline models, showcasing its overall superior performance. Notably, the R@1 scores for both text and image retrieval were 18.57% and 12.86% higher than those of the other models, respectively, indicating that our method exhibited a higher likelihood of returning accurate results at the top-1 position.

Results on Sydney-captions: The performance of the proposed method on the Sydney-captions dataset is presented in the upper right section of Table 2. The results reveal that the average R@K of our method surpassed that of the best baseline model by 3.99%. Specifically, the R@1, R@5, and R@10 scores for text retrieval, as well as R@1 for image retrieval, outperformed those of the best baseline model by 15.52%, 8.47%, 8.62%, and 11.8%, respectively. These findings align with the outcomes obtained from the UCM-captions dataset, which also exhibited a substantial enhancement in terms of R@1 performance.

Results on RSICD: The performance of our model on the RSICD dataset is presented in the lower left section of Table 2. It is evident that our model performed well, exhibiting superior text retrieval R@1 performance compared to other models. However, there were still some performance gaps observed in relation to other indicators when compared to the optimal baseline model.

Results on RSITMD: The performance on the RSITMD dataset can be observed in the lower right section of Table 2. For this dataset, our proposed model achieved higher values for all R@K indicators and mR compared to the other baseline models. This suggests that our model was more effective in capturing the image–text similarity relationships in datasets with richer text semantics and lower text repeatability.

Our experimental results across four datasets showcase the competitiveness of our method against baseline models. In the retrieval task, R@1 is significantly more important than R@5 and R@10, as users prefer the model to return the desired result as the first result, rather than filtering through the results. Except for Image Retrieval on the RSICD dataset, our method outperformed all other models in terms of R@1 on all four datasets, providing strong evidence of its superior performance. However, it falls short in other RSICD dataset metrics. To analyze the reasons, we conducted experiments on the validation set of RSICD using the same model and parameters. The R@1, R@5, and R@10 scores for Text Retrieval and Image Retrieval are 16.91, 44.24, and 57.86 and 20.20, 39.93, and 53.53, respectively, with an mR of 38.78. These results significantly outshine baseline models, suggesting potential dataset imbalances as the cause.

Furthermore, we scrutinized the RSICD dataset, which is similar to the UCM-captions and Sydney-captions datasets. These datasets were specifically curated for the purpose of generating captions for remote sensing images. The objective of the image caption is to generate sentences that are similar to the annotated text. In these datasets, although each image has five textual captions, these five sentences are often repetitive. Additionally, there are instances where different remote sensing images have the same or similar textual descriptions. In cross-modal retrieval, these text–image similarities may align semantically but are frequently deemed incorrect in evaluations, failing to contribute to metric. Yuan et al. [8] also noted this limitation of the dataset and quantified the diversity of data samples by using the ratio of inconsistent sentences to the number of images. The scores for UCM-captions, Sydney-captions, and RSICD datasets stand at 0.97, 1.83, and 1.67, respectively. However, cross-modal retrieval requires discerning the similarity between different samples, and needs more diverse samples to improve the discriminative ability of the model. To explore datasets that are more suitable for cross-modal retrieval between remote sensing images and text, Yuan et al. [8] contributed a more diverse remote sensing image–text dataset called RSITMD, with an increased ratio of inconsistent sentences to the

number of images to 4.60. In this dataset, our proposed method demonstrates a significant advantage over baseline models.

We further analyzed the performance of different models in Table 2. While baseline models endeavor to address fine-grained associations between multimodal data through multimodal semantic alignment and multimodal fusion coding, issues persist. Models such as VSE++, SCAN, MTFN, AMFMN, SAM, LW-MCR, MAFA-Net, and GaLR grapple with insufficiently complex interactions between modalities, limiting their performance. The work on multimodal fusion encoding, exemplified by FBCLM, uses a large-scale fusion encoder to mine complex associations between modalities, demonstrating optimal performance across multiple datasets. However, it does not utilize different training tasks to mine more supervised signals to further promote fine-grained correlation between modalities, which limits the performance of the fusion coding model. Our approach combines three supervised tasks—MLM, MVJRC, and ITM—to extract richer supervised signals and attain superior multimodal fine-grained associations. By aggregating local similarities between images and texts through a large-scale cross-attention network, the accuracy of cross-modality retrieval is improved. We further analyze the contribution of these three tasks in Section 4.4.

Although methods based on large-scale fusion encoders exhibit superior performance in remote sensing image–text cross-modal retrieval, their computational overhead hampers the retrieval speed. On the other hand, multi-modal semantic alignment methods can extract remote sensing image and text features offline and obtain the similarity between images and texts through simple calculations, thereby possessing superior retrieval speed. To compensate for the low retrieval efficiency of large-scale fusion encoders, we attempt to transfer the knowledge learned by the fusion encoder about the association between images and text to a small-scale model to improve retrieval efficiency. The details and arguments of this approach are presented in Section 4.5.

4.4. Ablation Studies

For the RSITMD dataset, we performed ablation tests to analyze the contributions of the ITM, MLM, and MVJRC tasks proposed by the fusion encoder in terms of fine-grained image–text correlation and cross-modal retrieval. We examined four different task combinations: ITM, ITM + MLM, ITM + MVJRC, and ITM + MLM + MVJRC.

4.4.1. Visualization of Fine-Grained Correlations in Word–Patch

In order to assess the contributions of different tasks to fusion representation, we extracted the attention values of each input word to the corresponding image region from the fifth cross-attention layer of the multimodal fusion encoder. These values were then used to generate a visual heat map illustrating the word–patch correlation. Darker colors indicate a higher correlation between the query word and the image region. Figure 6 presents the word–patch correlation heat map for a selected image and the sentence “Six water tanks and some pipes beside a pond” under various task combinations. It should be noted that the words displayed in the map are the result of contextual self-attention processing, thus encompassing contextual information.

The MLM task improved the fine-grained correlation between sentence words and image regions. For example, the words “six” and “pond” accurately matched the six white water tanks and the nearby pond, respectively, although some noise was present in the attention. However, when combining the ITM and MVJRC tasks, the correct association between words and image regions was not achieved. Only when all three tasks (ITM, MLM, and MVJRC) were used together did the words exhibit a strong correlation with the image regions. The global classification label [cls] was linked to a region that semantically matched the entire sentence. Words like “six” (referring to 6 water storage tanks), “tanks”, and “pond” (referring to the nearby pond) were correctly associated with their respective image regions. Compared to scenario b, the correlation between words and the image was more specific and accurate, demonstrating the effectiveness of the proposed MVJRC task

in filtering out irrelevant correlations. Regarding the word "pipes", except for scenario a, none of the other task combinations correctly associated it with an image region. This could be attributed to the low resolution of the target, which made detection challenging, and the lack of relevant samples in the training data.

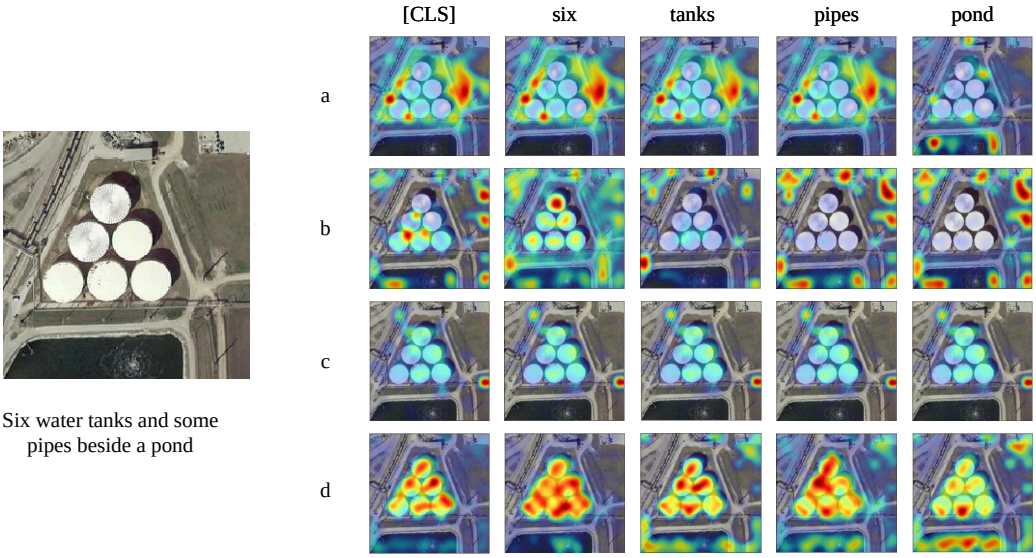

Figure 6. Attention heat maps of sentence words on the image area in the image–text fusion encoder. (**a**) ITM task only, (**b**) ITM + MLM tasks, (**c**) ITM + MVJRC tasks, and (**d**) ITM + MLM + MVJRC tasks simultaneously.

We conducted additional testing of the proposed method using image–text pairs that had more diverse and detailed text semantics. Figure 7 illustrates an example where the input text described a "viaduct" scene with multiple objects and included information about its surroundings. The results demonstrated that our method effectively improved the correlation between the text and the image. Even for non-target vocabulary such as "ring", "surrounded", and "green", our method successfully associated them with the appropriate image regions.

Based on the visual analysis of the image–text correlation discussed above, it was observed that the supervision signal provided by the ITM task for fine-grained image–text correlation was not precise enough, leading to overlapping correlation effects. On the other hand, the MLM task played a crucial role in enhancing the fine-grained correlation between images and texts by providing more refined and accurate supervision signals. When combining the ITM and MVJRC tasks, the correlation effects between images and texts intersected, resulting in improved correlation effects compared to when only the ITM and MLM tasks were combined. The addition of the MVJRC task enhanced the mutual information for fine-grained correlation between modalities and improved the consistency of joint representation. By strengthening the consistency of fine-grained correlations between remote sensing images from different perspectives and the associated text, the correlation effects between remote sensing images and texts were significantly enhanced.

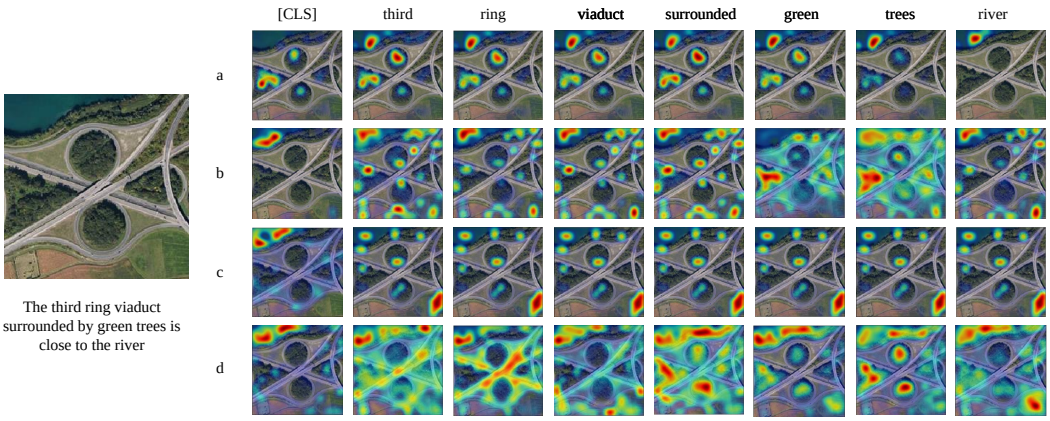

Figure 7. Evaluation of correlation quality between text words and image regions for image–text pairs with more complex semantics. (**a**) Results obtained using the ITM task, (**b**) results obtained using ITM + MLM tasks, (**c**) results obtained using ITM + MVJRC tasks, and (**d**) Rresults obtained using ITM + MLM + MVJRC tasks simultaneously.

4.4.2. Impact of Task Combinations on Retrieval Accuracy

We conducted experiments on the RSITMD dataset, evaluating the contributions of four different task combinations: ITM, ITM + MLM, ITM + MVJRC, and ITM + MLM + MVJRC. The results of these experiments are presented in Table 3.

Table 3. Retrieval accuracies of different task combinations on the RSITMD dataset.

Task	Text Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
ITM	15.71	35.62	50.44	13.41	44.78	65.66	37.6
ITM + MLM	16.37	**38.05**	52.88	16.46	47.92	**67.43**	39.85
ITM + MVJRC	12.39	33.19	49.56	10.66	40.35	61.64	34.63
ITM + MLM + MVJRC	**17.92**	31.19	**53.32**	**16.59**	**48.5**	**67.43**	**40.78**

The experimental results demonstrate that employing the ITM task alone yields a remarkable mR of 38.89, surpassing the accuracy metrics of the current state-of-the-art methods. It has validated the promoting effect of complex fine-grained interactions between modalities on the accuracy of cross-modal retrieval. When combining the ITM and MLM tasks, all retrieval accuracy metrics show significant improvement, with an increase of 2.25 in mR. This underscores the beneficial impact of complex fine-grained intermodal interactions on cross-modal retrieval accuracy. However, when combining the ITM and MVJRC tasks, the MVJRC task does not contribute to the retrieval performance, and there is a noticeable decrease in all retrieval accuracy metrics compared to using only the ITM task. When combining the ITM, MLM, and MVJRC tasks, the performance either slightly improves or remains the same compared to the combination of ITM and MLM, with a 0.93 increase in mR. The MVJRC task does not provide a significant improvement in retrieval accuracy. The impact of adding the MVJRC task to ITM and ITM + MLM on the retrieval accuracy aligns with the visual analysis results in Section 4.4.1, indicating that the MVJRC task does not provide a significant gain in image–text association on top of the ITM task and may even introduce some noise. After adding the MVJRC task to the combination of ITM + MLM, the visualization of fine-grained correlations between remote sensing image regions and text words is significantly enhanced, but the contribution to retrieval accuracy metrics is not as evident. In some subjective retrieval experiments, the combination of ITM, MLM, and MVJRC tends to return samples that match retrieval conditions but are not ground truth samples in the dataset. While this may enhance user experience, it does not necessarily improve the retrieval accuracy metrics. We attribute this to the limitations of the dataset in terms of sample diversity. The dataset exhibits high intra-class similarity,

where remote sensing images of the same scene, such as deserts, airports, and parking lots, have minimal differences, allowing many remote sensing images in the same scene to have the same text description. Additionally, the dataset contains significant category ambiguity in remote sensing images. For instance, the same remote sensing image can be classified as airport, barren land, or airplane, which further complicates the measurement of image–text matching in the dataset. Therefore, exploring datasets and metrics that are more suitable for cross-modal retrieval between remote sensing images and text is necessary for future work.

4.5. Retrieval Filtering Experiments

In order to alleviate the problem of low retrieval efficiency for a large-scale fusion encoder, as described in Section 3.3, we conducted a validation of our proposed retrieval filtering method on the RSICD dataset. To accomplish this, the study utilized the MTGFE model trained on the RSICD dataset as the teacher network. We then performed joint training to train the student network filter by leveraging the ITM output of the teacher network along with the ground truth labels. A total of 30 epochs were trained with a parameter of 128. During the testing phase, the study implemented a process where the first 128 samples of the filter's evaluation results were forwarded to the teacher network. The teacher network then recalculated the similarity ranking based on these samples and returned the updated ranking. The combined retrieval indicators are shown in Table 4. The RSICD test set comprised 1093 images and 5465 texts. The average search time for text retrieval from images was reduced from 472.10 ms to 24.70 ms, while the average search time for image retrieval from texts was reduced from 94.41 ms to 14.27 ms. Remarkably, the average retrieval accuracy mR decreased by only 0.88, demonstrating that the retrieval filtering method substantially enhanced the model's retrieval speed while maintaining a minimal loss in accuracy.

Table 4. Performance of model migration on the RSICD dataset.

Method	Text Retrieval				Image Retrieval				mR
	R@1	R@5	R@10	Time (ms)	R@1	R@5	R@10	Time (ms)	
MTGFE	15.28	37.05	51.6	472.1	8.67	27.56	43.92	94.41	30.68
MTGFE + Filter	13.82	36.32	50.41	24.7	8.27	27.17	42.8	14.27	29.8

The retrieval filtering experiments in this study exclusively comprised simple knowledge distillation experiments. Further investigations, including hyperparameter optimization, parameter distillation, and the exploration of combination strategies between teacher and student networks, have the potential to significantly enhance the performance of retrieval filtering.

5. Conclusions

To address the challenges posed by the fine-grained and multi-perspective features, as well as the significant imaging variations in remote sensing images, this study incorporates the MLM task into existing multimodal fusion coding models and introduces the novel MVJRC task. By combining the ITM, MLM, and MVJRC tasks, the model's ability to capture fine-grained correlations between remote sensing images and texts is enhanced. Furthermore, this paper proposes the retrieval filtering method to tackle the issue of low retrieval efficiency in large-scale fusion encoders. Experimental evaluations on four public datasets confirm the effectiveness of the proposed method in improving the accuracy and speed of cross-modal retrieval, leading to overall enhanced performance.

The limitation of this study is that the current remote sensing image–text datasets may not be suitable for high-performance cross-modal retrieval. The complex relationship between remote sensing images and texts also requires better evaluation metrics to judge the performance of cross-modal retrieval. This makes it difficult to effectively validate some of the methods we proposed, such as the MVJRC task, in experimental metrics. Additionally,

conducting additional knowledge distillation experiments may enhance the efficiency of cross-modal retrieval between remote sensing images and texts. Finally, exploring the concept of good joint representation has yielded various downstream tasks in VLP model studies, thereby opening up possibilities for the joint learning of remote sensing images and texts in applications such as visual question answering, multi-temporal remote sensing image comprehension, and remote sensing image object segmentation.

In future endeavors, we will focus on annotating more diverse remote-sensing image-text datasets and specifying cross-modal retrieval evaluation metrics. Furthermore, our research will extend to exploring joint learning techniques and cross-modal retrieval tasks, leveraging high-performance fusion encoders for analyzing multi-temporal remote-sensing images alongside textual data.

Author Contributions: Conceptualization, X.Z., H.Z. and W.L.; methodology, X.Z., W.L. and H.Z.; software, X.Z. and W.L.; validation, X.Z., X.W. and L.W. (Luyao Wang); formal analysis, L.W. (Long Wang); investigation, X.Z. and F.Z.; resources, H.Z. and L.W. (Luyao Wang); data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, H.Z., X.W. and L.W. (Luyao Wang); visualization, L.W. (Long Wang) and X.Z.; supervision, L.W. (Long Wang); project administration, X.Z.; funding acquisition, H.Z. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 62102423).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MTGFE	multi-task guided fusion encoder
ITM	image–text matching
MLM	masked language modeling
MVJRC	multi-view joint representations contrast
VLP	vision-language pre-training
RF	retrieval filtering
FC	fully connected
MLP	multilayer perceptron
FFN	feed-forward network
BN	batch normalization
ReLU	rectified linear unit

References

1. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
3. 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 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
4. Greff, K.; Srivastava, R.K.; Koutnik, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Networks Learn. Syst.* **2017**, *28*, 2222–2232. [CrossRef]
5. Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078.
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
7. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef]
8. 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. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]
9. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *arXiv* **2018**, arXiv:1707.05612.

10. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. *arXiv* **2018**, arXiv:1803.08024.
11. 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. [CrossRef]
12. Rahhal, M.M.A.; Bazi, Y.; Abdullah, T.; Mekhalfi, M.L.; Zuair, M. Deep Unsupervised Embedding for Remote Sensing Image Retrieval Using Textual Cues. *Appl. Sci.* **2020**, *10*, 8931. [CrossRef]
13. Abdullah, T.; Bazi, Y.; Al Rahhal, M.M.; Mekhalfi, M.L.; Rangarajan, L.; Zuair, M. TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images. *Remote Sens.* **2020**, *12*, 405. [CrossRef]
14. Cheng, Q.; Zhou, Y.; Fu, P.; Xu, Y.; Zhang, L. A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4284–4297. [CrossRef]
15. 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.* **2022**, *19*, 1–5. [CrossRef]
16. 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]
17. Cheng, Q.; Zhou, Y.; Huang, H.; Wang, Z. Multi-Attention Fusion and Fine-Grained Alignment for Bidirectional Image-Sentence Retrieval in Remote Sensing. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1532–1535. [CrossRef]
18. 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.* **2022**, *60*, 1–19. [CrossRef]
19. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
20. Van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2019**, arXiv:1807.03748.
21. Li, H.; Xiong, W.; Cui, Y.; Xiong, Z. A Fusion-Based Contrastive Learning Model for Cross-Modal Remote Sensing Retrieval. *Int. J. Remote Sens.* **2022**, *43*, 3359–3386. [CrossRef]
22. Zeng, Y.; Zhang, X.; Li, H.; Wang, J.; Zhang, J.; Zhou, W. X^2-VLM: All-In-One Pre-trained Model For Vision-Language Tasks. *arXiv* **2022**, arXiv:2211.12402.
23. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
24. Huang, Y.; Wang, W.; Wang, L. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7254–7262. [CrossRef]
25. Zheng, F.; Li, W.; Wang, X.; Wang, L.; Zhang, X.; Zhang, H. A Cross-Attention Mechanism Based on Regional-Level Semantic Features of Images for Cross-Modal Text-Image Retrieval in Remote Sensing. *Appl. Sci.* **2022**, *12*, 12221. [CrossRef]
26. Kim, W.; Son, B.; Kim, I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *arXiv* **2021**, arXiv:2102.03334.
27. Li, J.; Selvaraju, R.R.; Gotmare, A.D.; Joty, S.; Xiong, C.; Hoi, S. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arXiv* **2021**, arXiv:2107.07651.
28. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:1908.03557.
29. Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; Fu, J. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *arXiv* **2020**, arXiv:2004.00849.
30. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
31. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
32. 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. [CrossRef]
33. Mikriukov, G.; Ravanbakhsh, M.; Demir, B. Deep Unsupervised Contrastive Hashing for Large-Scale Cross-Modal Text-Image Retrieval in Remote Sensing. *arXiv* **2022**, arXiv:2201.08125.
34. Mikriukov, G.; Ravanbakhsh, M.; Demir, B. An Unsupervised Cross-Modal Hashing Method Robust to Noisy Training Image-Text Correspondences in Remote Sensing. *arXiv* **2022**, arXiv:2202.13117.
35. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
37. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.

38. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners Are Image-Text Foundation Models. *arXiv* **2022**, arXiv:2205.01917.
39. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv* **2019**, arXiv:1908.02265.
40. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv* **2019**, arXiv:1908.07490.
41. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Multiview Coding. *arXiv* **2020**, arXiv:1906.05849.
42. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv* **2020**, arXiv:1911.05722.
43. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2002.05709.
44. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. *arXiv* **2020**, arXiv:2011.10566.
45. 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. *arXiv* **2021**, arXiv:2103.00020.
46. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv* **2021**, arXiv:2102.05918.
47. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation through Attention. *arXiv* **2021**, arXiv:2012.12877.
48. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
49. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2020**, arXiv:1606.08415.
50. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
51. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. *J. Mach. Learn. Res.* **2011**, *15*, 315–323.
52. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 3008–3017. [CrossRef]
53. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**, arXiv:1711.05101.