applied sciences

MDPI

Article

Contrasting Dual Transformer Architectures for Multi-Modal Remote Sensing Image Retrieval

Mohamad M. Al Rahhal [1,*], Mohamed Abdelkader Bencherif [2], Yakoub Bazi [3], Abdullah Alharbi [4] and Mohamed Lamine Mekhalfi [5]

[1] Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia
[2] Center of Smart Robotics Research, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[3] Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4] Department of Computer Science, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[5] Digital Industry Center, Technologies of Vision Unit, Fondazione Bruno Kessler, 38123 Trento, Italy
* Correspondence: mmalrahhal@ksu.edu.sa; Tel.: +966-50804-0827

Abstract: Remote sensing technology has advanced rapidly in recent years. Because of the deployment of quantitative and qualitative sensors, as well as the evolution of powerful hardware and software platforms, it powers a wide range of civilian and military applications. This in turn leads to the availability of large data volumes suitable for a broad range of applications such as monitoring climate change. Yet, processing, retrieving, and mining large data are challenging. Usually, content-based remote sensing image (RS) retrieval approaches rely on a query image to retrieve relevant images from the dataset. To increase the flexibility of the retrieval experience, cross-modal representations based on text–image pairs are gaining popularity. Indeed, combining text and image domains is regarded as one of the next frontiers in RS image retrieval. Yet, aligning text to the content of RS images is particularly challenging due to the visual-sematic discrepancy between language and vision worlds. In this work, we propose different architectures based on vision and language transformers for text-to-image and image-to-text retrieval. Extensive experimental results on four different datasets, namely TextRS, Merced, Sydney, and RSICD datasets are reported and discussed.

Keywords: remote sensing; cross-modal retrieval; vision and language transformers; contrastive loss

Citation: Rahhal, M.M.A.; Bencherif, M.A.; Bazi, Y.; Alharbi, A.; Mekhalfi, M.L. Contrasting Dual Transformer Architectures for Multi-Modal Remote Sensing Image Retrieval. *Appl. Sci.* **2023**, *13*, 282. https://doi.org/10.3390/app13010282

Academic Editor: Dongyang Hou

Received: 13 November 2022
Revised: 18 December 2022
Accepted: 19 December 2022
Published: 26 December 2022

1. Introduction

With rapid advances in Earth observation sensors, plentiful information about the Earth's surface is now available with higher spatial, spectral, and temporal resolutions, leading to massive growth in the remote sensing (RS) image archive [1]. This vast amount of data have totally changed our perspective of monitoring the Earth's surface and has opened new horizons for a broad range of specialized applications. However, as the volume of data increases, mining a specific piece of information contained in such a large amount of data is becoming more difficult. Consequently, content-based image retrieval (CBIR), which is the task of retrieving an image that is best described by a particular query, is essential. CBIR systems plays an important role in the decision-making of various applications such as environment monitoring and disaster management.

Generally, CBIR systems includes two main steps: feature extraction and similarity matching. The first step extracts useful feature representations from a set of images in an archive, while the similarity measure aims at quantifying the similarity between a query image and the images in this archive. Thus, the performance of retrieval systems strongly relies on the quality of the extracted features as well as the similarity measure

chosen for matching. Early studies on RS image retrieval mainly focused on using hand-crafted features to represent the visual content of images. However, manually designed features might be insufficient in producing powerful representation to describe its detailed content. Yet, the recent developments based on deep-learning methods have brought crucial achievements in boosting the accuracy of CBIR systems [2] similar to other RS applications such as crop mapping from image time series [3], tree species classification [4], and cloud change detection [5] to name a few.

Currently, with the rapid generation of data across a different range of modalities such as text and audio, there is an emerging interest to go beyond the single-modal retrieval (Figure 1). Cross-modal is regarded as the next frontier in RS image retrieval, where the query can be a textual description of the content of the image or a sound describing what we want to find in the image. From the end-user perspective, the integration of different modalities in the query is more practical and increases the usability of the retrieval systems [6]. However, it poses new major challenges for processing and analysis and the problems are different from remote sensing captioning [7].

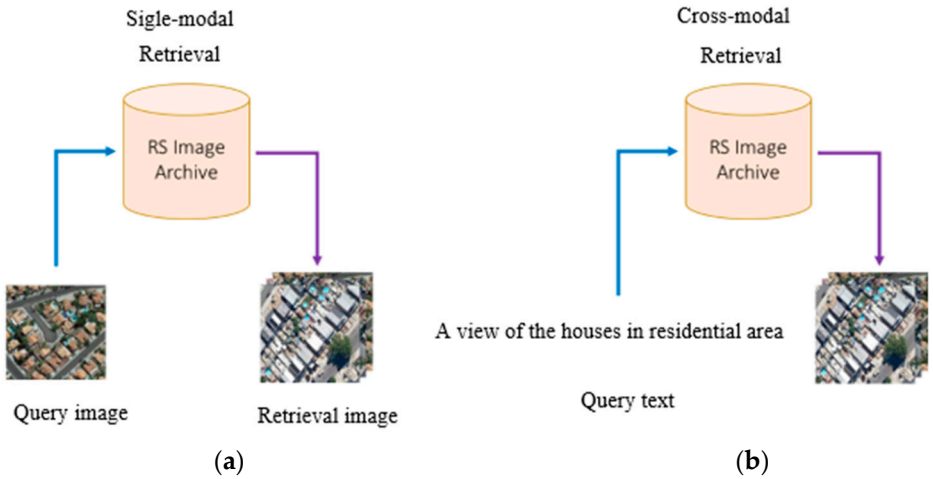


Figure 1. Content-based RS image retrieval system: (**a**) single modal, (**b**) cross-modal.

First, generating an informative modality-specific representation is an important step in image retrieval. Despite the influential retrieval works based on conventional deep learning models, these models generate a global semantic representation that ignores the spatial relationships between image regions. This is even more important in cross-modal text-to-image retrieval that requires modeling the global semantic concepts of the image and its corresponding text description. Second, one of the key challenges when dealing with cross-modal data is how to learn the joint representations and narrow the heterogeneity gap between the multi-modal pairs. For text-to-image retrieval, this requires an effective aligning of visual and textual data representations and modeling the relationships between each image and its corresponding text. Third, training an accurate cross-modal retrieval method strongly relies on the quality of the dataset. In RS, text-to-image retrieval methods usually reuse the existing image captioning datasets. Different from natural image datasets, these datasets are relatively small in size with more complex detailed images. Furthermore, the available datasets have a different number of captions per image, and many of these captions are redundant, and not fully descriptive.

With the above challenges in mind, we introduce a text-to-image retrieval model based on transformers. Recent developments in transformers and its variants have extended its ability to contextualize the information within and across different data modalities through the attention mechanism; therefore, obtaining more representative visual and text features that can help in achieving better text-to-image retrieval. It is worth noting that transformers have been recently introduced in image classification [8], multi-labeling [9], and change detection [10].

In this paper, we propose an efficient text–image retrieval approach based on vision and languages transformers. Our method use an embedding model composed of language and vision transformer encoders for aligning the visual representations of RS images to their related textual representations. We learn the weights of these encoders by optimizing two contrastive losses related to image-to-text and text-to-image classification. Basically, we aim at maximizing the similarity between the image and its corresponding sentence, while minimizing its similarity to unrelated sentences and vice versa. In the experiments, we use four RS benchmark datasets to validate our approach which are the Text-RS, Merced, RSICD, and Sydney datasets composed of images acquired with different sensing platforms. In addition, we propose different transformer-based architectures and introduce different types of transformer architectures.

The remainder of the paper is structured as follows: Section 1 discusses related works on single-modal and cross-modal image retrieval; Section 2 introduces the text–image matching methodology; Section 3 presents a detailed experimental analysis on four RS datasets; and finally, Section 4 draws conclusions and forecasts future developments.

2. Materials and Methods

Let us consider a $D^{(l)} = \{X_i, t_i\}_{i=1}^{N}$ be a set of N training image–text pairs in an archive. The goal of a text-to-image retrieval task is to find the most relevant image from X_i to the provided text query. Similarly, in image-to-text retrieval, we are interested in retrieving the sentence t_i that is most similar to the query image.

The suggested model's overall architecture, which is built mostly on language and vision transformers, is depicted in Figure 2. In order to generate visual and textual characteristics for the transformer encoders during training, we sample a mini-batch of images from the training set with their corresponding sentences. Next, we compute the similarity between each potential pair of text-images in the mini-batch to create a similarity matrix. Then, by maximizing an image-to-text and text-to-image contrastive classification loss, we learn the model weights. During the test, based on a query text we retrieve the most similar images by computing the similarity between the textual feature of the query and the visual features of the training images. In a similar manner, we can retrieve the most similar textual descriptions in the case of an image query. Detailed descriptions of the overall architecture are provided in next subsections.

2.1. Language and Vision Representation Encoders

The language transformer was first introduced for machine translation in 2017 [11]. Since then, more advanced variants have been proposed for text modeling such as Generative Pre-trained Transformer (GTP) [12,13] and Bidirectional Encoder Representations from Transformers (BERT) [14]. Typically, the transformer encoder is based on the so-called self-attention mechanism, which allows it to capture long-range dependencies. This mechanism makes it a better alternative to recurrent models in modeling long sequences.

In our work, we rely on these models in generating the textual feature representations. To this end, each text tx_i is first tokenized into words $tx_i = (w_1, w_2, \ldots w_m)$, where m is the maximum number of tokens. Then, the word embedding layer E_{tx} converts the words into sequence of features of dimension d_{tx}. A learnable positional embedding is added to the sequence before feeding it into the encoder to provide knowledge of the word sequence. To denote the beginning and end of the sequence, two additional special tokens, CLS and SEP, are added to the input tokens. The final matrix, Z_{tx0}, has the following representation:

$$Z_{tx0} = [w_{class};\ w_1E_{tx}; w_2E_{tx}; \ldots;\ w_mE_{tx}] + E_{pos} \quad (1)$$

where $E_{pos} \in \mathbb{R}^{(m+1)\times d_{tx}}$ and w_{class} is a special classification token. The final representation Z_{txL} is created at layer L by feeding the original representation Z_{tx0} through several identical layers. These encoder layers each start with a tiny multilayer perceptron (MLP) block and

a multi-head self-attention (MSA) block. Both blocks are joined together using LayerNorm and skip connections (LN).

$$Z'_{t\ell} = \mathrm{MSA}(\mathrm{LN}(Z_{t\ell-1})) + Z_{t\ell-1}, \ell = 1 \dots L \tag{2}$$

$$Z_{t\ell} = \mathrm{MLP}(\mathrm{LN}(Z'_{t\ell})) + Z'_{t\ell}, \ell = 1 \dots L \tag{3}$$

It is worth-recalling, that each input matrix $Z_{t\ell-1}$ for layer ℓ of the encoder is projected into queries, keys, and values using the weight matrices W^Q, W^K, $W^V \in \mathbb{R}^{d_t \times d_K}$, such that $Q = W^Q Z_{t\ell-1}$, $K = W^K Z_{t\ell-1}$, and $V = W^V Z_{t\ell-1}$ where d_K is the dimension of the key. The queries, keys, and values are used by the MSA block to generate a weighted sum of token features.

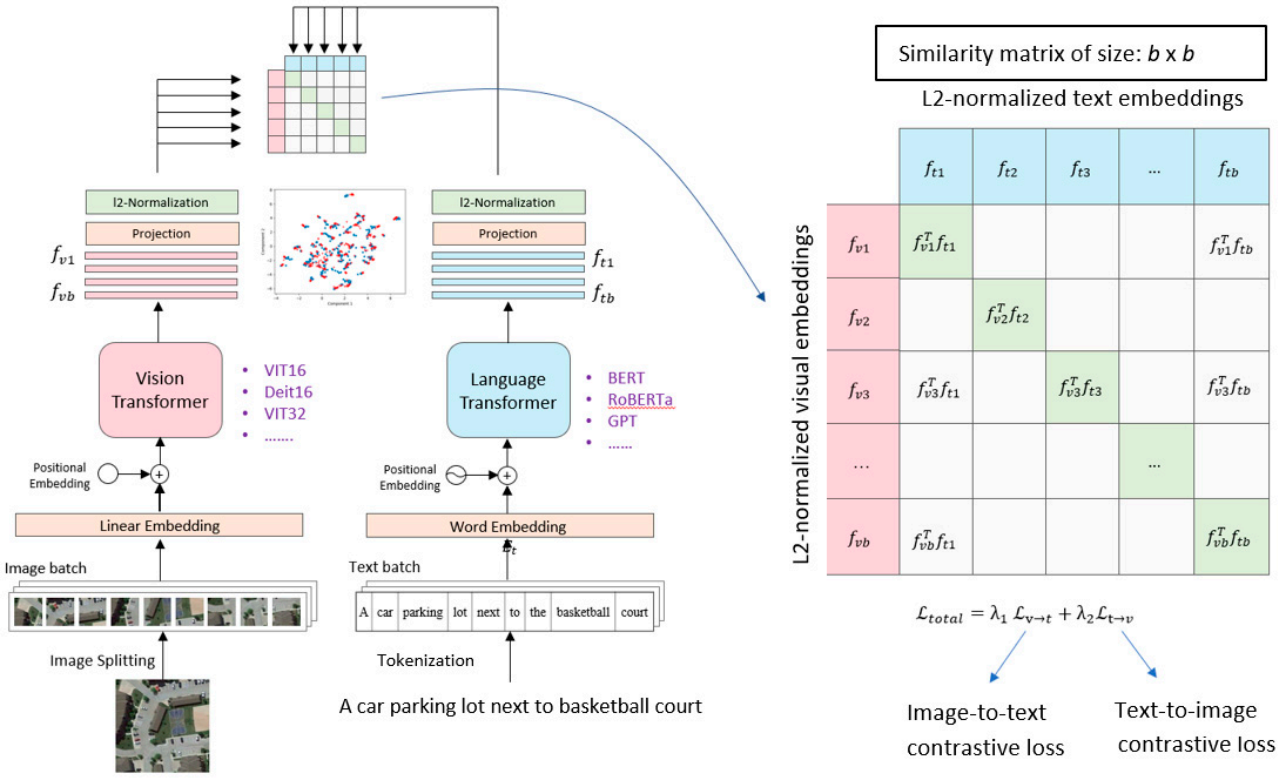


Figure 2. Overview of the proposed text–image aligning model. To produce L2-normlized visual $\{f_{vi}\}_{i=1}^{b}$ and textual $\{f_{ti}\}_{i=1}^{b}$ features, we sample a mini-batch of b images-text pairings for training the network and feed them to the vision and language transformer encoders. Then, we generate a similarity matrix of size $b \times b$ by calculating the similarity between all potential visual and textual pairs in the mini-batch.

In particular, MSA comprises h multiple independent self-attention heads operating in parallel, each head computes a different attention score using the scaled dot product similarity between the queries, keys, and values expressed by this equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right) \cdot V \tag{4}$$

All heads' outputs are combined, and a learnable weights matrix is used to project them to the appropriate dimension d_t. The MLP block copmposed of two linear layers with a GELU [15] activation in between.

In the vision encoder, the image X_{im} is reshaped into a sequence of n non-overlapping patches $X_{\mathrm{im}} = (x_1, x_2, \dots x_n)$. Each patch in the resulting sequence has the size of $c \times p \times p$ where c is the number of channels in the image and p is the patch size (e.g., $p = 16$ or 32).

These patches are then flattened and mapped to a sequence of embeddings of the model dimension d_v with a linear embedding layer $E_v \in \mathbb{R}^{(p^2c)\times d_v}$. Then, in a way similar to the language model, the positional encoding is added $E_{pos} \in \mathbb{R}^{(n+1)\times d_v}$ and a x_{class} token is appended to the patch representations. The resulting embedded sequence of patches that is fed into the encoder is:

$$Z_{v0} = [x_{class};\ x_1E; x_2E; \ldots;\ x_nE] + E_{pos} \tag{5}$$

Finally, we arrive at the final image representation Z_{vL} by using the same methods as in Equations (2) and (3). Finally, by applying the same operations as in Equations (2) and (3) we obtain the final image representation Z_{vL}. Remember that each layer of the visual transformer encoder also includes blocks from the MSA and multilayer perceptron (MLP) architecture. Prior to each block, LayerNorm (LN) is applied, along with residual.

2.2. Model Optimization

We use global average pooling to the representation matrices Z_{tL} and Z_{vL} (while ignoring the W_{class} and x_{class} tokens) acquired from the text and image encoders, respectively, followed by L2-normalization to get the features f_t and f_v. In addition, if the resulting feature representations have different dimensions one can use an additional linear projection layer to bring them to the same dimension. In Algorithm 1, we provide the PyTorch-style pseudo code of the proposed cross-modal retrieval approach with its default parameters.

Algorithm 1: Cross-modal text–image matching PyTorch-style pseudocode

```
# Mini-batch size: default b=120,
# Regularization parameters lmabda1=lambda2=0.5
# Initial value Temperature parameter: Taux=0.07
# model.f_NLP and model.f_ViT: NLP and Vision transformer encoders
# Optimizer: SGD(lr=0.1, momentum=0.9, nesterov=True)
# Criterion=nn.CrossEntropyloss(), Number of epochs: 100
# Load mini-batch of image-text pairs of size b from the training set
for X, t in loader:
 # Feed the image-text pair into the model
  logits_image=model.f_ViT(X)
  logits_text=model.f_NLP(t)
  # L2 Normalization
  logits_image=logits_image/logits_image.norm(dim=-1,keepdim=True)
  logits_text= logits_text / logits_text.norm(dim=-1,keepdim=True)
  # Similarity matrix
  Sim_Mat= logits_image @ logits_Text.t()/Taux
  # Set class labels: 0,1,2, . . . ,b-1
  labels = torch.Tensor(np.arange(b)).long()
  # image-to-text and text-to-image classification losses (Equation (8)).
  loss=lmabda1*Criterion(Sim_Mat,labels)+lmabda2*
              Criterion(toch.transpose(Sim_Mat,0,1), labels)
   # Backward loss
   loss.backward()
   # Clip the gradients for numerical stability
   torch.nn.utils.clip_grad_norm_(model.parameters(),0.1)
   # Optimization step
   optimizer.step()
   optimizer.zero_grad()
```

2.3. Encoders Architecture

Given the limitations imposed by the small-scale nature of the available RS cross-modal datasets, we expect that learning the model from scratch may not be a good solution. Instead, we propose to transfer knowledge from backbones pre-trained on large-scale

datasets such as ImageNet (1.2 M images) and English Wikipedia (2500 M words) for vision and language tasks, respectively. In addition, as our model involves a joint learning from both images and text pairs, different types of architectures pre-trained in different manners can be investigated (Table 1). Our aim is to identify the most suitable pre-trained configuration for transferring knowledge to our task.

Table 1. Pre-trained configurations investigated for transferring knowledge to the cross-modal RS image–text retrieval task.

	Config. 1	Config. 2	Config. 3
Image encoder	ViT base (layers = 12, hidden = 768, parameters = 86M image size: 224 × 224 pixels, patch size: 32 × 32 pixels).	ViT base (layers = 12, hidden = 768, parameters = 86M, input image size: 224 × 224 pixels, patch size=16 × 16 pixels).	Deit base distilled (layers = 12, hidden = 768, parameters = 87M, input image size: 224 × 224 pixels, patch size=16 × 16 pixels).
Text encoder	BERT base (layers = 12, hidden = 512, parameters = 63M).	RoBERTa base (layers = 12, hidden = 768, parameters = 110M).	BERT base (layers = 12, hidden = 768, parameters = 110M).
Pre-training mode	CLIP model for image text matching task.	Models were learned independently in a self-supervised mode on image and language tasks (DINO for image and SimCSE for text).	Models were learned independently in a standard supervised mode on image and language tasks.

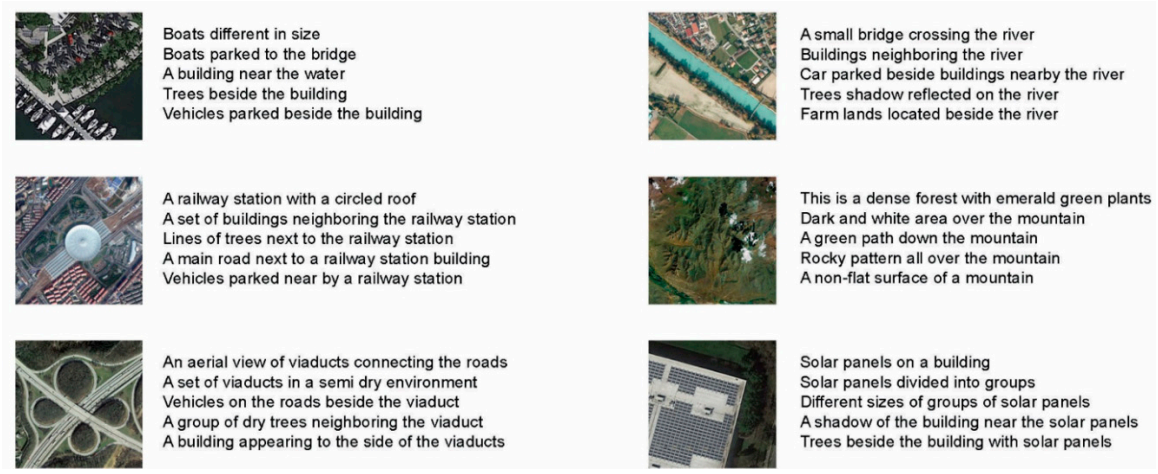
The first configuration is based on dual transformers pre-trained on matching text to computer vision images known as Contrastive Language Image Pre-training (CLIP) [16]. The related image–text encoders based on ViT32 and BERT were learned for matching 400M image–text pairs using a contrastive loss. In the second configuration, we consider transformers pre-trained independently in a self-supervised mode. For images, we use a ViT16 encoder pre-trained on ImageNet using the Distillation with no-labels (DINO) self-supervised learning approach [17]. For text, we use RoBERTa as encoder [18] pre-trained by the Simple Contrastive Learning of Sentence Embedding (SimCSE) approach [19]. In the third configuration, we use the Data efficient Transformer (DeiT16) pre-trained in a supervised mode on ImagNet [20]. While for text, we adopt the original BERT model [14].

3. Experimental Results and Discussion

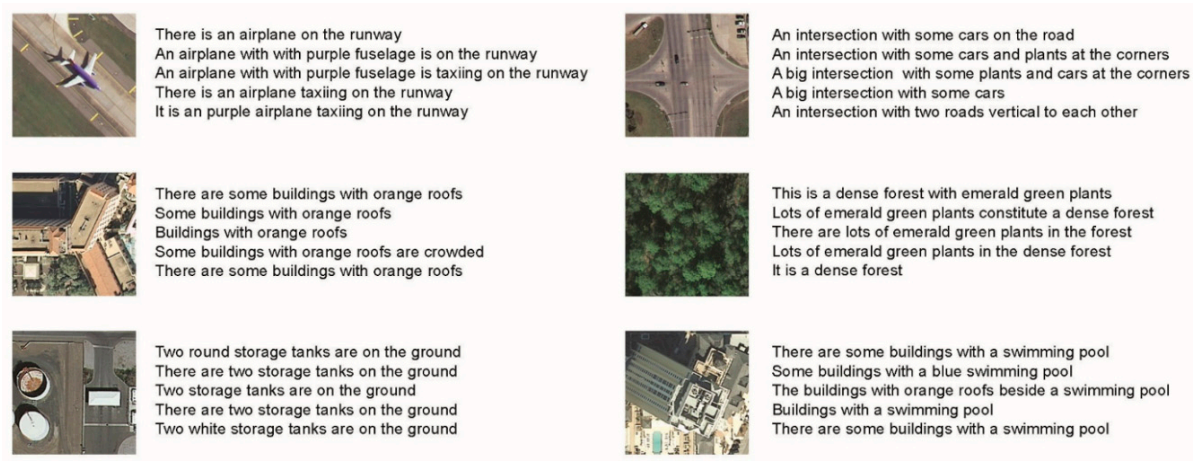
3.1. Dataset Description

In the experiments, we evaluated the proposed retrieval method on four benchmark datasets, TextRS [21], Merced [22], Sydney [22], and RSICD [23]. In the following, we provide more details about these RS text–image datasets.

TextRS caption dataset: This dataset was introduced recently by collecting images randomly from four different scene classification datasets with different image sizes and spatial resolutions. Namely, AID [24] which is composed of 30 classes, PatternNet [25] which contains 38 classes, UC-Merced [26] which has 21 classes, and finally NWPU [27] dataset which has 45 different classes. From each dataset, 16 random images were extracted from every class. This creates a new dataset with a total number of 2144 of images, each image is annotated with five different experts to ensure diversity. Figure 3a shows some images with their textual annotations from TextRS dataset.



(**a**)



(**b**)

Figure 3. Example of images with their corresponding captions (**a**) TextRS, (**b**) Merced.

Merced caption dataset: This dataset comprises 2100 images, each size of 256 × 256 pixels with a spatial resolution of 30 cm. It was obtained from the Merced Land-use dataset [26], in which the images were manually labeled to one of 21 land-use classes. Five captions are used to describe each image in this collection, totaling 10,500 descriptions. Yet many sentences are highly correlated. Some examples of this dataset are illustrated in Figure 3b.

Sydney caption dataset: This dataset was built based on the Sydney scene classification dataset, which includes RS images belonging to seven land-use classes including residential, airport, grassland, river, ocean, industrial area, and runway. It has 613 images, with a spatial resolution equal to 0.5 m.

RSICD caption dataset: This dataset is the biggest of the earlier datasets. It was gathered from a number of sources, including Tianditu, Baidu Map, MapABC, and Google Earth. The dataset contains 224 × 224 pixels of images with various spatial resolutions. There are 10,921 total images in the dataset, and each image has a unique text that describes it. RSICD has a total of 24,333 captions, which suggests that not every image has five phrases. For the sake of consistency, captions for photographs with fewer than five sentences have been duplicated. Table 2 summarizes image retrieval RS datasets

Table 2. Characteristics of the text–image retrieval RS datasets.

Dataset	# of Images	Spatial Resolution (cm)	Image Size
TextRS	2144	[0.62, 30]	256 × 256 pixels, and 600 × 600 pixels
Merced	2100	30	256 × 256 pixels
Sydney	613	50	500 × 500 pixels
RSICD	10,921	-	224 × 224 pixels

3.2. Experimental Protocol and Evaluation Metrics

To quantitatively evaluate the performance of our cross-modal retrieval model, we carried out several experiments mainly based on the first configuration, which uses language and vision encoders pre-trained on matching general image–text pairs. In particular, we analyzed the batch size effect, the generalization ability in cross-dataset settings, followed by a comparison against the second and third configurations. Then, we contrasted our results to the recent solutions proposed in the context of RS imagery.

We trained the models for 100 iterations using a mini-batch size equal to 120 (larger mini-batch size given the memory constraints of our station). We employed the Stochastic Gradient Descent (SGD) with Nesterov momentum as an optimizer. We chose 0.1 for the initial learning rate and 0.9 for the default value of momentum. After 40 iterations of training, we reduced the learning rate to 0.01 and for the final 20 iterations, to 0.001. We discovered that applying gradient clipping with the gradients' maximum norm set to 0.1 is helpful for maintaining numerical stability. Each input image was reduced to 224 224 pixels, and we added the usual data augmentation techniques (such as random cropping, 50%-probability horizontal and vertical flips, and ColorJitter). We chose 80% of the image-text pairs for training and 20% for testing across all datasets. Five sentences are used to describe each image; therefore, we choose one at random to use as learning material.

We show the results in term of the Recall@K (R@K) metric with several values of (K = 1, 5, and 10), which are donated as R@1, R@5, and R@10. The following definition applies to these measures, which are frequently used to assess text-to-image and image-to-text retrieval techniques:

$$\text{R@K} = \frac{TP@K}{TP@K + FN@K} \tag{6}$$

where TP is the true positive and FN is the false negative. In addition, we provided the mean recall (mR) of all recalls (R@1, R@5, and R@10). Since each image is described by five sentences, we presented the R@1, R@5, and R@10 in terms of mean and standard deviation.

We used PyTorch to implement our models, and HP Omen Station to conduct all of the tests according to the following guidelines: NVIDIA GeForce GTX 1080 Ti graphics processing unit (GPU), 32 GB of RAM, and Intel Core (TM) i9-7920 central processing unit (CPU) at 2.9 GHz (with 11 GB GDDR5X memory).

3.3. Results Related to the First Configuration

As mentioned previously, in the first set of experiments we analyzed the performance of our model on both image-to-text retrieval and text-to-image retrieval tasks on the four different datasets using the first configuration. Table 3 shows the retrieval performance on TextRS, Merced, Sydney, and RSICD datasets along with their training time. For TextRS, in text-to-image retrieval the model achieves 24.55%, 65.66%, and 80.60% in terms of R@1, R@5, and R@10, respectively. On the other hand, for image-to-text retrieval the scores are 24.08%, 66.04%, and 80.78%, respectively. For Merced, our model achieves 19.33%, 64.00%, and 91.42%, respectively, in terms of R@1, R@5, and R@10, in text-to-image retrieval and 19.04%, 53.33%, and 77.61% in the image-to-text retrieval.

Table 3. The retrieval performance in terms of R@K on different datasets.

Dataset	Text-To-Image			Image-To-Text			Train Time
	R@1	R@5	R@10	R@1	R@5	R@10	
TextRS	24.55	65.66	80.60	24.08	66.04	80.78	5.8 h
Merced	19.33	64.00	91.42	19.04	53.33	77.61	7.9 h
Sydney	26.76	57.59	73.53	24.95	57.44	72.32	0.6 h
RSICD	9.14	28.96	44.59	10.70	29.64	41.53	102.7 h

Regarding the Sydney dataset, the text-to-image retrieval scores are 26.76%, 57.59%, and 73.53%; while image-to-text retrieval scores are 24.95%, 57.44%, and 72.32%, in terms of R@1, R@5, and R@10, respectively. The lower scores obtained for Sydney compared with the other dataset can be explained by its descriptive sentences which are longer and quite complex. Regarding the RSICD dataset, the text-to-image retrieval scores are 9.14%, 28.96%, and 44.59%; while the image-to-text retrieval scores are 10.70%, 29.64%, and 41.53%, in terms of R@1, R@5, and R@10, respectively. Similarly, we observe also that the results are lower compared with other datasets indicating that this larger dataset is more challenging, mainly due to its longer textual descriptions. It's important to observe that both the text-to-image and image-to-text retrieval tasks have retrieval performances that are extremely similar to one another, proving that the matching is successful in both directions. Additionally, the recall reveals a notable increase from R@1 to R@5 and from R@5 to R@10. This is because the probability of finding the right match among the retrieved elements increases as the number of elements is retrieved.

Table 3 compares also between the different RS datasets in terms of training time. In our experiments, TextRS, and Merced datasets take around 5.8 and 7.9 h for training, respectively. For the Sydney dataset, which has only few hundred samples, the time required for training is 0.6 h. On the contrary, for RSICD it takes 102 h as it is the largest dataset.

Figure 4 shows the attentions produced by the vision and language transformers when training on the image–text pairs. We recall that the method aligns the global feature representations of the pooled features of words and image patches. This global matching does not allow a full latent alignment using both image regions and words in a sentence as context. Instead, the matching is guided in a weakly supervised way by the pooled features. Yet, the attentions over the images and sentences produced by the transformers shows the decision is taken by emphasizing on particular regions and words.

3.4. Comparisons to the Second and Third Configurations

In this set of experiments, we investigate the two other configurations for the vision and text backbones. We recall that the two other configurations are based also on transformers, except they differ on how they are pre-trained as mentioned in Table 4. The average recalls of the three configurations as depicted in Table 4 reveal that the first configuration based on CLIP is the best on all datasets and on both retrieval tasks. This clearly provides an indication that using models pre-trained on image–text matching tasks are more suitable for knowledge transfer compared with models pre-trained independently of image and text classification tasks.

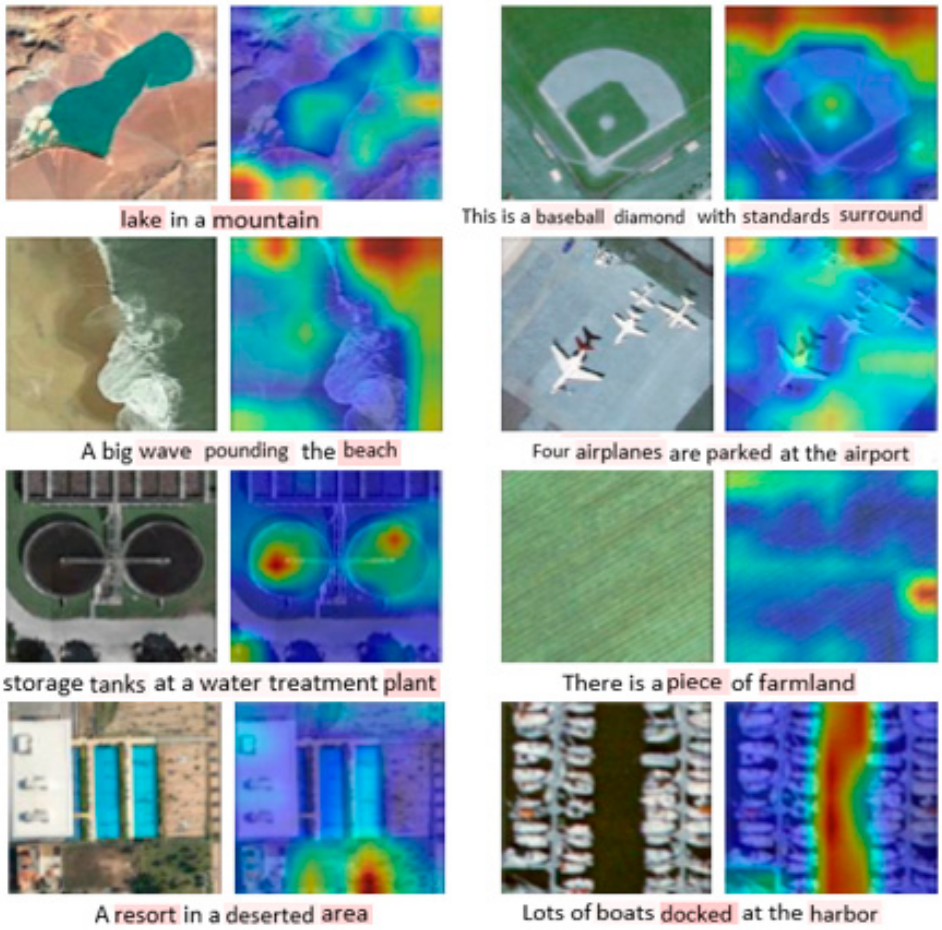


Figure 4. Examples for original image its corresponding attention (image and text) for dataset: (**a**) TextRS and (**b**) Merced dataset.

Table 4. Average mR retrieval results for TextRS, Merced, and Sydney datasets based on three different configurations: Config. 1: models pre-trained on text–image matching; Config. 2: models pre-trained in a self-supervised mode; Config. 3: models pre-trained in a supervised mode.

	Text to Image			**Image to Text**		
	Config. 1	**Config. 2**	**Config. 3**	**Config. 1**	**Config 2**	**Config. 3**
TextRS	**56.93**	52.34	51.56	**56.96**	52.74	51.18
Merced	**58.25**	49.62	50.22	**49.99**	48.91	47.26
Sydney	**52.62**	51.52	51.32	**51.57**	49.46	50.97

3.5. Comparisons to State-Of-The-Art RS Methods

In order to assess the effectiveness of our retrieval method, we compare it against different methods published recently, Tables 5–8 show the test results of the model on four datasets: TextRS, UCM, Sydny, and RSICD. Table 5 compares the proposed transformer-based model to the most recent retrieval techniques in terms of R@K for the TextRS dataset. The work proposed by [28] is based on a CNN and a Bi-LSTM for image and text encoding, respectively. The model was learned using a contrastive loss. While [21] used a bidirectional triplet network composed of an LSTM and pre-trained CNN. It is evident that the new methods significantly outperform the current techniques, while Table 6 contrasts the suggested model for the UCM dataset with recently published techniques such as VSE++ SCAN, MTFN, and AMFMN. Work [29] proposed a multi-modal feature matching model

which used a multi-scale visual self-attention module to extract visual features and a triplet loss for training. Finally, [30] proposed a model which incorporates an alignment module to address the semantic relationships between text and images with the help of attention mechanism and gate function. The obtained findings demonstrate the robustness of our model against SOTA models for valid comparison. Tables 7 and 8 states the comparison results. Table 7 shows the Sydney dataset results are reasonable compared with the most recent research. The performance of the RSCID dataset clearly demonstrates the robustness of the suggested method, which is presented in Table 8.

Table 5. Comparison between the state-of-the-art retrieval methods in the TextRS dataset in terms of R@K.

Approach	Text Retrieval			mR	Image Retrieval			mR
	R@1	R@5	R@10		R@1	R@5	R@10	
Bi-LSTM [28]	19.02	55.25	71.72	48.66	22.95	59.52	77.23	53.23
Triplet [21]	12.55	41.62	62.09	38.75	12.55	39.53	59.53	37.20
Ours	24.55	65.66	80.60	56.93	24.08	66.04	80.78	56.96

Table 6. Comparison between the state-of-the-art retrieval methods in the Merced dataset in terms of R@K.

Approach	Text Retrieval			mR	Image Retrieval			mR
	R@1	R@5	R@10		R@1	R@5	R@10	
VSE++ [31]	12.38	44.76	65.71	40.95	10.10	31.80	56.85	32.92
SCAN [32]	12.85	47.14	69.52	43.17	12.48	46.86	71.71	43.68
MTFN [33]	10.47	47.62	64.29	40.79	14.19	52.38	78.95	48.51
AMFMN-soft [29]	12.86	51.90	66.67	43.81	14.19	52.38	78.95	48.51
AMFMN-fusion [29]	16.67	45.71	68.57	43.65	12.86	53.24	79.43	48.51
AMFMN-sim	14.76	49.52	68.10	44.13	13.43	51.81	76.48	47.24
Ours	19.33	64.00	91.42	58.25	19.04	53.33	77.61	49.99

Table 7. Comparison between the state-of-the-art retrieval methods in the Sydney dataset in terms of R@K.

Approach	Text Retrieval			mR	Image Retrieval			mR
	R@1	R@5	R@10		R@1	R@5	R@10	
SAM t-i [30]	9.60	34.60	55.80	33.53	5.80	32.70	48.10	30.57
VSE++ [31]	24.4	53.45	67.24	48.36	6.21	33.56	51.03	30.27
MTFN [33]	20.69	51.72	68.97	47.13	13.79	55.51	77.59	48.05
SCAN [32]	20.69	55.17	67.24	47.7	15.52	57.59	76.21	49.77
AMFMN-soft [29]	20.69	51.72	74.14	48.85	15.17	58.62	80.00	51.26
AMFMN-fusion [29]	24.14	51.72	75.86	50.57	14.83	56.55	77.89	49.76
AMFMN-sim	29.31	58.62	67.24	51.72	13.45	60.00	81.72	51.72
Ours	26.76	57.59	73.53	52.62	24.95	57.44	72.32	51.57

Table 8. Comparison between the state-of-the-art retrieval methods in the RSICD dataset in terms of R@K.

Approach	Text Retrieval			mR	Image Retrieval			mR
	R@1	R@5	R@10		R@1	R@5	R@10	
AMFMN-fusion [29]	4.90	18.28	31.44	18.21	5.39	15.08	23.40	14.62
VSE++ [31]	3.38	9.51	17.46	10.12	2.82	11.32	18.10	10.75
MTFN [33]	5.02	12.52	19.74	12.43	4.90	17.17	29.49	17.19
AMFMN [29]	5.39	15.08	23.40	14.62	4.90	18.28	31.44	18.21
SCAN [32]	5.85	12.89	19.84	12.86	3.71	16.40	26.73	15.61
SAM t-i [28]	6.59	19.85	31.04	19.16	4.69	19.48	32.13	18.77
CABIR [34]	8.59	16.27	24.13	16.33	5.42	20.77	33.85	20.01
Ours	9.14	28.96	44.59	27.56	10.70	29.64	41.53	27.29

4. Conclusions

In this work, we proposed a language based and vision-based framework for RS text-to-image retrieval. To generate visual and textual representations, we used vision and language transformer encoders. By maximizing a bidirectional contrastive loss associated with text-to-image and image-to-text classification, we were able to align these representations. The experimental results on four different RS text–image datasets confirm the promising ability of the proposed model compared with recent RS text–image retrieval methods. For future developments, we propose to improve the retrieval accuracy by performing a full latent alignment using image regions and words in a sentence as a context instead of a global alignment of the image and sentence representations.

Author Contributions: Methodology, M.M.A.R., Y.B. and A.A.; software, M.M.A.R., Y.B. and M.L.M.; formal analysis, A.A. and M.L.M.; writing—original draft preparation, M.A.B.; funding acquisition, A.A. and M.L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Research Supporting Project number (RSP2022R444), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors extend their appreciation to the Research Supporting Project number (RSP2022R444), King Saud University, Riyadh, Saudi Arabia.

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

References

1. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
2. Hoxha, G.; Melgani, F.; Demir, B. Toward Remote Sensing Image Retrieval Under a Deep Image Captioning Perspective. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4462–4475. [CrossRef]
3. Gella, G.W.; Bijker, W.; Belgiu, M. Mapping Crop Types in Complex Farming Areas Using SAR Imagery with Dynamic Time Warping. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 171–183. [CrossRef]
4. Hu, B.; Li, Q.; Hall, G.B. A Decision-Level Fusion Approach to Tree Species Classification from Multi-Source Remotely Sensed Data. *ISPRS Open J. Photogramm. Remote Sens.* **2021**, *1*, 100002. [CrossRef]
5. Winiwarter, L.; Anders, K.; Höfle, B. M3C2-EP: Pushing the Limits of 3D Topographic Point Cloud Change Detection by Error Propagation. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 240–258. [CrossRef]
6. 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]
7. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5629419. [CrossRef]
8. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]

9. Bashmal, L.; Bazi, Y.; Al Rahhal, M.M.; Alhichri, H.; Al Ajlan, N. UAV Image Multi-Labeling with Data-Efficient Transformers. *Appl. Sci.* **2021**, *11*, 3974. [CrossRef]
10. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection With Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607514. [CrossRef]
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
12. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems, Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6–12 December 2020*; Curran Associates Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 1877–1901.
13. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. *Language Models Are Unsupervised Multitask Learners. Technical Report*; OpenAI: San Francisco, CA, USA, 2019.
14. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
15. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2020**, arXiv:1606.08415.
16. 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 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021.
17. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. *arXiv* **2021**, arXiv:2104.14294.
18. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
19. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv* **2021**, arXiv:2104.08821.
20. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation through Attention. *arXiv* **2020**, arXiv:2012.12877.
21. 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]
22. 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.
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. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
25. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
26. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; ACM: New York, NY, USA, 2010; pp. 270–279.
27. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
28. 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]
29. 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]
30. 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]
31. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *arXiv* **2017**, arXiv:1707.05612.
32. Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. *arXiv* **2018**, arXiv:1803.08024.

33. 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; Association for Computing Machinery: New York, NY, USA, 15 October 2019; pp. 12–20.
34. 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]