Preprints.org

Article

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

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

Article

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

Xiong Zhang [1], Weipeng Li [1], Xu Wang [1], Luyao Wang [1], Fuzhong Zheng [1], Long Wang [1] and Haisu Zhang [1,*]

[1]   School of Information and Communication, National University of Defense Technology, Wuhan 430074, China

*   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 remote sensing image data volume. One approach that has shown promising results in cross-modal retrieval of natural images is the multimodal fusion encoding method. However, remote sensing images have unique characteristics that make the retrieval task challenging. Firstly, the semantic features of remote sensing images are fine-grained, meaning they can be divided into multiple basic units of semantic expression. Additionally, these images exhibit variations in resolution, color, and perspective. Different combinations of basic units of semantic expression can generate diverse text descriptions. These characteristics pose considerable challenges for cross-modal retrieval. To address these challenges, this paper proposes a multi-task guided fusion encoder (MTGFE) based on the multimodal fusion encoding method. The model incorporates three tasks: image-text matching (ITM), masked language modeling (MLM), and the newly introduced multi-view joint representations contrast (MVJRC) task. By jointly training the model with these tasks, we aim to 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 feature 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. This method 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. Overall, this study focuses on remote sensing image-text cross-modal retrieval and introduces the MTGFE model, which combines multimodal fusion encoding with multiple tasks to enhance the model's ability to capture fine-grained correlations. Additionally, a retrieval filtering method is proposed to improve retrieval efficiency. Experimental results demonstrate the effectiveness of the proposed method.

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

---

1. Introduction

The rapid advancement of space information technology and the exponential expansion of remote sensing image data have created a pressing need for 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.

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 [1] . 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 [2], 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. To address the fine-grained and multi-perspective attributes, it becomes crucial to capture the intricate semantic correlations between images and texts, including the associations between image regions and text descriptors. Furthermore, it is necessary to overcome the influences of substantial differences in image resolution, color, and angle on the precise alignment of image-text correlations.

Recent studies on 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 [3], ResNet [4]) and vision Transformer networks [5], are commonly employed for extracting image features. Similarly, recurrent neural networks (RNNs) (e.g., LSTM [6], GRU [7]) and Transformer models (e.g., BERT [8]) 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. [9] 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.

The upper part of b illustrates multimodal semantic alignment methods [10–20] Figure 1b. These approaches aim to align image and text data in a shared 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, similarity between image and text features is determined by measuring their distance in the shared 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 [21] and InfoNCE loss [22] are utilized either directly or through intermediate variables to impose constraints on the position and distance of image and text features within the shared embedding space. The bottom half of Figure 1b depicts the method of multimodal data fusion encoding [23]. 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.
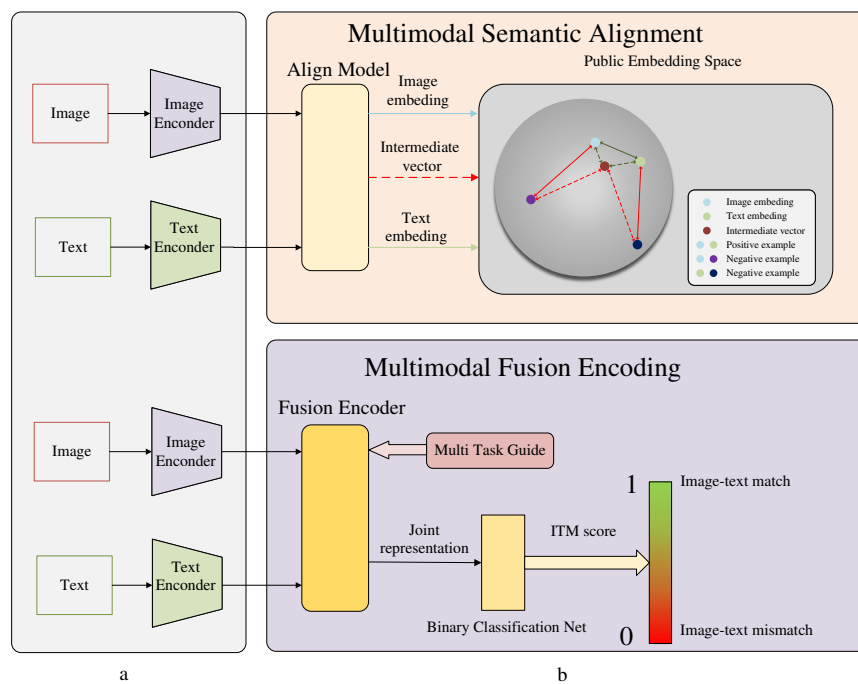
Figure 1. General framework for remote sensing image and text retrieval. (**a**) Unimodal feature extraction stage. (**b**) Multimodal interaction stage. The methods can be categorized into two groups based on the generation of a unified multimodal representation: multimodal semantic alignment and multimodal fusion encoding.

In earlier multimodal semantic alignment methods, cosine similarity was directly employed to align the global feature vectors of images and texts [14,15]. However, it is worth noting that the global similarity observed between images and texts often arises from the intricate aggregation of local similarities between image fragments (such as objects) and sentence fragments (such as words) [24]. Consequently, the global feature vectors alone struggle to capture the fine-grained correlations required to address the fine-grained and multi-angle characteristics of remote sensing images. To overcome this limitation, researchers have explored the use of fine-grained unimodal features as replacements for global features. For instance, region features [25] and patch features [23] have been utilized for images, while word features have been employed for texts [16,23,25]. These fine-grained correlations between images and texts are then established through cross-attention mechanisms between the modalities. To guide the unimodal features in the shared embedding space, previous studies, such as Yuan, Zhang, Fu, et al. [10], Zheng et al. [25], and Cheng, Zhou, Fu, et al. [16], have employed shallow cross-attention mechanisms. 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. [23] 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 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 and image region-level

information from the text. This approach facilitates more effective capture of fine-grained image-text correlations.

To address the challenges posed by imaging differences in remote sensing images, such as variations caused by weather conditions, photography equipment, and spacecraft position, this paper introduces the multi-view joint representations contrast (MVJRC) task. This task incorporates automatic contrast, histogram equalization, brightness adjustment, definition adjustment, flipping, rotation, and offset operations to simulate imaging differences. Additionally, a weight sharing twin network is designed to maximize the similarity between enhanced 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 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 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. (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.

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, 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 subtitle-based retrieval and direct measurement of image-text similarity. Shi and Zou [31] proposed an automatic subtitle generation framework for remote sensing images, demonstrating the technical feasibility of this approach. Qu et al. [32] and Lu et al. [2] contributed a publicly available remote sensing image-text dataset and proposed automatic remote sensing image subtitle generation and image-text cross-modal retrieval based on subtitles. 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. [14] employed the InfoNCE loss to map global feature vectors of images and texts to a shared embedding space, directly calculating the similarity between remote sensing images and texts. Abdullah et al. [15] 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 shared embedding space. Cheng, Zhou, Fu, et al. [16] introduced a shallow attention mechanism to combine fine-grained features of image regions and text words as intermediate features. This constrained the projection of images and texts in the shared embedding space, thereby improving the quality of semantic alignment between images and texts. Lv et al. [17] 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, Zhang, Fu, et al. [10] 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 shared embedding space using triplet-loss. Yuan, Zhang, Tian, et al. [18] proposed the multi-level information dynamic fusion (MIDF) to fuse 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, Zhou, Huang, et al. [19] 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, Zhang, Rong, et al. [20] utilized knowledge distillation to transfer the "dark knowledge" learned by the asymmetric multimodal feature matching network (AMFMN) model [10], 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 shared embedding space, significantly enhancing the efficiency of cross-modal retrieval. Li et al. [23] 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 [8]. 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 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. [5] introduced the ViT model as an image encoder based on the Transformer architecture. In this model, images are divided into multiple 16×16 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) model

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 shared embedding space while increasing the distance between dissimilar samples (negative samples). In unimodal contrastive learning, a twin 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 and He 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 twin 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

The unimodal extraction stage involves utilizing the ViT model for image encoding and the first six layers of BERT model for text encoding. In the multimodal interaction module, a 6-layer Transformer block from VLP is employed as the fusion encoder for generating joint representations. Further details regarding the structure and training tasks will be provided in Sections 3.2 and 3.3.

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.
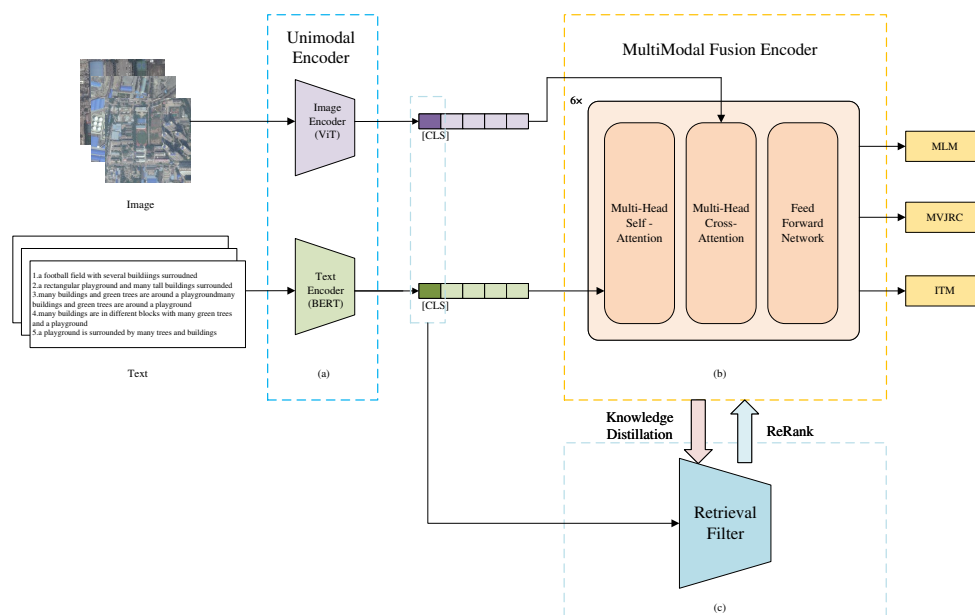

Figure 2. Overview of the MTGFE model. It comprises two components: (**a**) a unimodal encoder that utilizes the ViT and BERT (first 6 layers) models to extract features from images and texts, and (**b**) a multimodal fusion encoder that generates joint image-text representations through ITM, MLM, and 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 exploration of fine-grained correlations within each modality, resulting in robust expression capabilities for the unimodal features.

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 Ref. [5], 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 -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 [8] 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; 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, resulting in improved image encoder parameters and enhanced visual representations.

The fusion encoder is initialized with the weights of the last 6 layers of $BERT_{base}$ [8], 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, W^V, for each input token embedding. The calculation approach for each attention head is provided in Equation (1).

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

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).

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

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. Multimodal fusion encoder training task

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. MLM

The MLM task(shown in Figure 3), derived from the BERT [8] model, involves randomly masking 15% of the word elements 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 text elements 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.
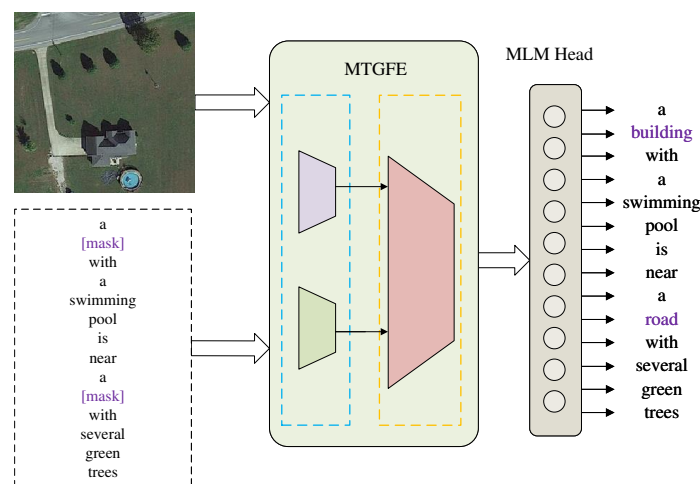

Figure 3. The illustrates the diagram of the MLM task, where the masked tokens are predicted by leveraging the contextual information from both the input image and text.

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 task 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 value,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 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. 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 twin network Figure 4. Various image enhancement 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 between remote sensing images captured from different perspectives and their corresponding paired text. Specifically, project head and prediction head are added after MTGFE, which are respectively expressed as f_{proj} and f_{pred}. 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]. Prediction head (f_{pred}) is a two-layer FC layer connected by the BN layer and the ReLU activation function.
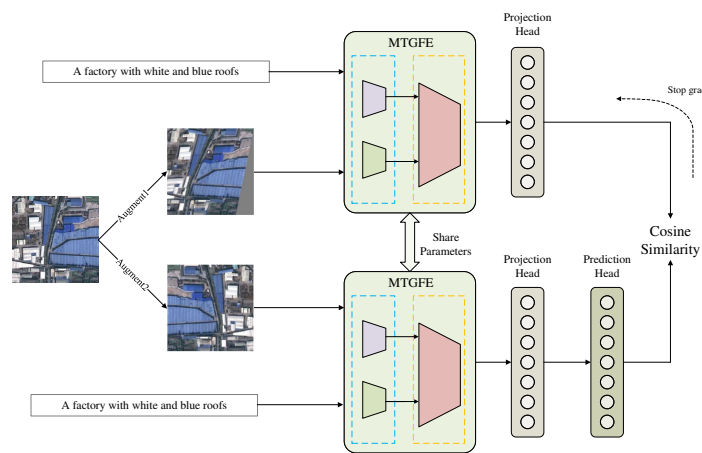

Figure 4. The MVJRC task involves setting up a twin network with shared parameters from MTGFE. The cosine similarity of the joint feature u_{cls} is calculated after the projection head and prediction head, and the gradient is updated alternately.

For a given image-text pair (I, T) , Randaugment [52] is used for random image enhancement 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 operations. Let $z_1 = f_{proj}(u_1)$, $z_2 = f_{proj}(u_2)$, $p_1 = f_{pred}(z_1)$, $p_2 = f_{pred}(z_2)$, 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 enhanced 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 Ref. [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 z 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
1: **for** (I, T) in dataloader: **do**
2:      $I_1, I_2 = \text{aug}(I), \text{aug}(I)$ \#   Image augmentation
3:      $u_1, u_2 = f(I_1, T)_{[cls]}, f(I_2, T)_{[cls]}$
4:      $z_1, z_2 = f_{proj}(u_1), f_{proj}(u_2)$
5:      $p_1, p_2 = f_{pred}(z_1), fpred(z_2)$
6:      $p_1, p_2 = f_{pred}(z_1), f_{pred}(z_2)$
7:      L = -0.5 * (S(p1,z2) + S(p2,z1))   \#   loss
8:      l.backward()   \#   Gradient return
9:      update($f, fproj, fpred$)   \#   Parameters update
10: **end for**
11: **function** $S(p, z)(\ )$ \#   Calculation of cosine similarity
12:      $z = z$.detach()   \#   Stop gradient
13:      $p = \text{normalize}(p, dim = 1)$ \# l_2-normalize
14:      $z = \text{normalize}(z, dim = 1)$ \# l_2-normalize
15:      $s = (p * z)$.sum(dim=1).mean()
16:      **return** s
17: **end function**

3.3.3. 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 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; is set to 0.

The overall loss of the MTGFE training is:

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

3.4. Retrieval filtering(RF)

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). Its input, $cat((v_{cls}, w_{cls}), dim = 1)$, is the catenation 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.
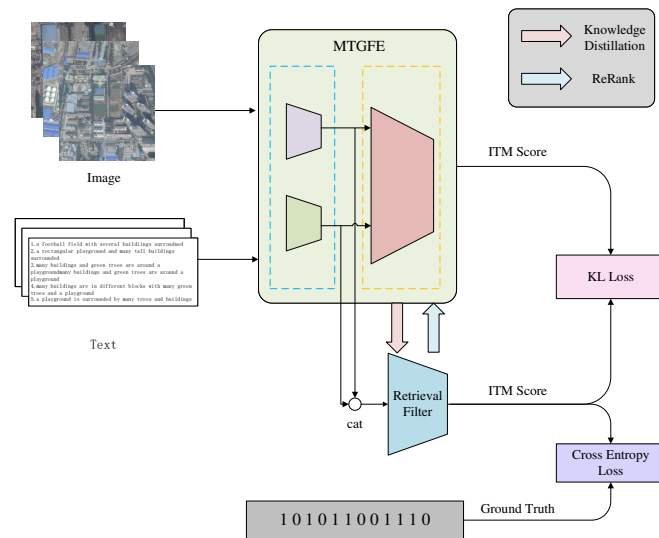

Figure 5. Retrieval filtering architecture. Knowledge distillation is utilized to transfer the knowledge from MTGFE to the filter. During the retrieval process, the filter is employed to exclude easily distinguishable negatives, while samples with higher similarity are forwarded to MTGFE for recalibration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

The MTGFE's ITM output and the manually annotated ground truth label are utilized as the soft target and hard target supervision signals, respectively, for the student network filter, 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 network, p_{stu}^{itm} represents the predicted ITM value of the student network filter, and y^{itm} represents the ground truth. Finally, the distillation loss of the model can be obtained as:

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

here, α denotes a constant hyperparameter. The student and teacher networks use the same unimodal encoder to extract features. The difference between the two scenarios is that the student network 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[2], and RSITMD [10]. 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	2,100	10,500	5	21	256×256
Sydney-captions	613	6,035	5	7	500×500
RSICD	10,921	54,605	5	31	224×224
RSITMD	4,743	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 enhanced using Randaugment [52]. To simulate variations in remote sensing images, nine enhancement methods ("Identity", "AutoContrast", "Equalize", "Brightness", "Sharpness", "ShearY", "TranslateX", "TranslateY", and "Rotate") were selected. However, since strong image enhancement 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 enhancement methods, while "7" represents the amplitude of the image enhancement. 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 1,000 iterations. When training the student network, a distillation parameter of 0.2 was set, 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++ [11], SCAN [12], MTFN [13], AMFMN [10], SAM [16], LW-MCR [20], MAFA-Net [19], FBCLM [23], and GaLR [18]. 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 II, the performance metrics for VSE++, SCAN, and MTFN are obtained from Ref. [10], 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, while the validation set served as an evaluation set for the model. 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. Nonetheless, when the same set of models and parameters were applied, improved performance was demonstrated in the RSICD validation set. The R@1, R@5, and R@10 scores for text retrieval were 16.91, 44.24, and 57.86, respectively, while for image retrieval they were 20.20, 39.93, and 53.53, respectively. Additionally, the mR value was 38.78. These results could be attributed to imbalanced sampling in the validation and test sets of this particular dataset, as well as the high intra-class similarity.

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.

The comparison with the baseline models confirmed the effectiveness of the proposed method in remote sensing image-text cross-modal retrieval, as it achieved competitive performance metrics on all four datasets.

4.4. Ablation tests

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.

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 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.
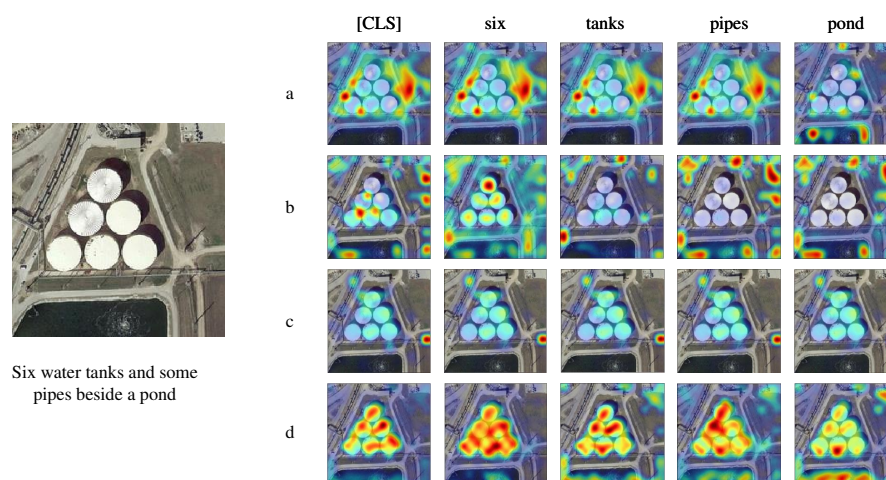

Figure 6. Attention heat maps of sentence words on the image area in the image-text fusion encoder. (a) ITM task only, (b) ITM+MLM tasks, (c) ITM+MVJRC tasks, and (d) ITM+MLM+MVJRC tasks simultaneously.

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 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.

MLM task improved the fine-grained correlation between sentence words and image regions. For example, the word "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, the words exhibited 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.

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.
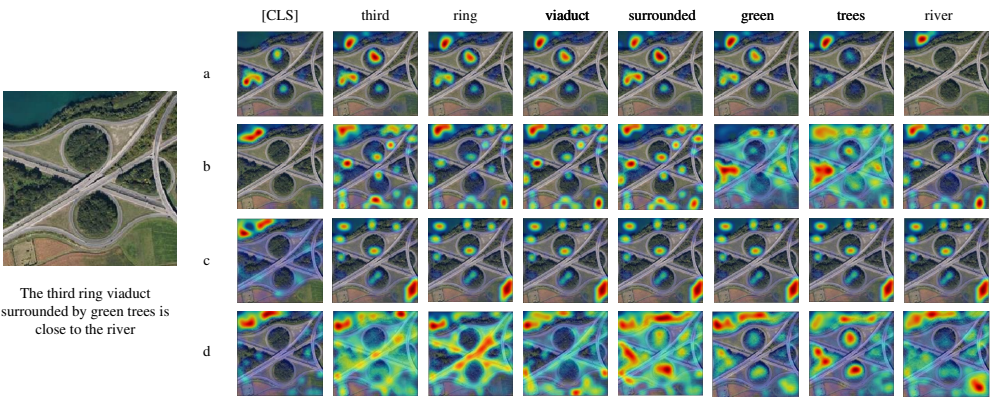

Figure 7. Evaluation of correlation quality between text words and image regions for image-text pairs with more complex semantics. (a) Results obtained using the ITM task, (b) Results obtained using ITM+MLM tasks, (c) Results obtained using ITM+MVJRC tasks, and (d) Results obtained using ITM+MLM+MVJRC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

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.

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 indicated that using only the ITM task resulted in poor retrieval accuracy, with an mR of 38.89. However, when combining the ITM and MLM tasks, all retrieval accuracy indicators showed improvement, with a notable mR enhancement of 2.25. In contrast, incorporating the MVJRC task alongside the ITM task did not lead to improved retrieval performance. When all three tasks were combined, there was a slight improvement compared to the ITM and MLM combination, with an increase of 0.93 in mR. Comparing the aforementioned experiments, we observed that when employing a large-scale fusion encoder, relying solely on the ITM task yielded comparable or even superior accuracy to existing advanced methods. This finding validates the significant impact of complex fine-grained interactions between modalities on cross-modal retrieval accuracy. The MLM task contributed to improved retrieval accuracy by establishing more precise fine-grained correlations. Additionally, by enhancing the consistency of joint representations, the MVJRC task further strengthened the model's ability to differentiate remote sensing images with imaging variations, thereby enhancing the model's generalization capability.

4.5. Retrieval filtering experiments

As described in Section 3.3, we conducted 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 1,093 images and 5,465 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

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. In particular, the MVJRC task improves the model's capability to differentiate the semantics of remote sensing images under different imaging conditions by enhancing the consistency of joint representation features. 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. Moving forward, we aim to explore downstream tasks, such as VQA and reasoning for remote sensing images, leveraging the joint representations obtained in this study.

Conflicts of Interest: The authors declare no conflicts 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. Zeng, Y.; Zhang, X.; Li, H.; Wang, J.; Zhang, J.; Zhou, W. X$\hat{2}$-VLM: All-In-One Pre-trained Model For Vision-Language Tasks, 2022, [arxiv:cs/2211.12402].
2. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *56*, 2183–2195. doi:10.1109/TGRS.2017.2776321.
3. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015, [arxiv:cs/1409.1556].
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition, 2015, [arxiv:cs/1512.03385].
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]* **2021**, [arxiv:cs/2010.11929].
6. Greff, K.; Srivastava, R.K.; Koutnik, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks & Learning Systems* **2017**, *28*, 2222–2232. doi:10.1109/TNNLS.2016.2582924.

7. 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, 2014, [arxiv:cs, stat/1406.1078].

8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* **2019**, [arxiv:cs/1810.04805].

9. Baltrusaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2019**, *41*, 423–443. doi:10.1109/TPAMI.2018.2798607.

10. 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 Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–19, [arxiv:cs/2204.09868]. doi:10.1109/TGRS.2021.3078451.

11. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives, 2018, [arxiv:cs/1707.05612].

12. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching, 2018, [arxiv:cs/1803.08024].

13. 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. Proceedings of the 27th ACM International Conference on Multimedia; ACM: Nice France, 2019; pp. 12–20. doi:10.1145/3343031.3350875.

14. Rahhal, M.M.A.; Bazi, Y.; Abdullah, T.; Mekhalfi, M.L.; Zuair, M. Deep Unsupervised Embedding for Remote Sensing Image Retrieval Using Textual Cues. *Applied Sciences* **2020**, *10*, 8931. doi:10.3390/app10248931.

15. 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 Sensing* **2020**, *12*, 405. doi:10.3390/rs12030405.

16. 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *14*, 4284–4297. doi:10.1109/JSTARS.2021.3070872.

17. Lv, Y.; Xiong, W.; Zhang, X.; Cui, Y. Fusion-Based Correlation Learning Model for Cross-Modal Remote Sensing Image Retrieval. *IEEE Geoscience and Remote Sensing Letters* **2022**, *19*, 1–5. doi:10.1109/LGRS.2021.3131592.

18. 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 Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–16. doi:10.1109/TGRS.2022.3163706.

19. 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 Journal of Automatica Sinica* **2022**, *9*, 1532–1535. doi:10.1109/JAS.2022.105773.

20. 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 Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–19. doi:10.1109/TGRS.2021.3124252.

21. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Boston, MA, USA, 2015; pp. 815–823. doi:10.1109/CVPR.2015.7298682.

22. van den Oord, A.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding, 2019, [arxiv:cs, stat/1807.03748]. doi:10.48550/arXiv.1807.03748.

23. Li, H.; Xiong, W.; Cui, Y.; Xiong, Z. A Fusion-Based Contrastive Learning Model for Cross-Modal Remote Sensing Retrieval. *International Journal of Remote Sensing* **2022**, *43*, 3359–3386. doi:10.1080/01431161.2022.2091964.

24. Huang, Y.; Wang, W.; Wang, L. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, 2017; pp. 7254–7262. doi:10.1109/CVPR.2017.767.

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. *Applied Sciences* **2022**, *12*, 12221. doi:10.3390/app122312221.

26. Kim, W.; Son, B.; Kim, I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, 2021, [arxiv:cs, stat/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, 2021, [arxiv:cs/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, 2019, [arxiv:cs/1908.03557]. doi:10.48550/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, 2020, [arxiv:cs/2004.00849]. doi:10.48550/arXiv.2004.00849.

30. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network, 2015, [arxiv:cs, stat/1503.02531]. doi:10.48550/arXiv.1503.02531.

31. Shi, Z.; Zou, Z. Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3623–3634. doi:10.1109/TGRS.2017.2677464.

32. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep Semantic Understanding of High Resolution Remote Sensing Image. 2016 International Conference on Computer, Information and Telecommunication Systems (CITS); IEEE: Kunming, China, 2016; pp. 1–5. doi:10.1109/CITS.2016.7546397.

33. Mikriukov, G.; Ravanbakhsh, M.; Demir, B. Deep Unsupervised Contrastive Hashing for Large-Scale Cross-Modal Text-Image Retrieval in Remote Sensing, 2022, [arxiv:cs/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:2202.13117 [cs]* **2022**, [arxiv:cs/2202.13117].

35. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06); IEEE: New York, NY, USA, 2006; Vol. 2, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv:1706.03762 [cs]* **2017**, [arxiv:cs/1706.03762].

37. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, Ł.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; Dean, J. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. p. 23.

38. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners Are Image-Text Foundation Models, 2022, [arxiv:cs/2205.01917].

39. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, 2019, [arxiv:cs/1908.02265].

40. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers, 2019, [arxiv:cs/1908.07490]. doi:10.48550/arXiv.1908.07490.

41. Tian, Y.; Krishnan, D.; Isola, P. Contrastive Multiview Coding. *arXiv:1906.05849 [cs]* **2020**, [arxiv:cs/1906.05849].

42. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv:1911.05722 [cs]* **2020**, [arxiv:cs/1911.05722].

43. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]* **2020**, [arxiv:cs, stat/2002.05709].

44. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. *arXiv:2011.10566 [cs]* **2020**, [arxiv:cs/2011.10566].

45. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision **2021**. doi:10.48550/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:2102.05918 [cs]* **2021**, [arxiv:cs/2102.05918].

47. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation through Attention, 2021, [arxiv:cs/2012.12877].

48. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization, 2016, [arxiv:cs, stat/1607.06450].

49. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs), 2020, [arxiv:cs/1606.08415].

50. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]* **2015**, [arxiv:cs/1502.03167].

51. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. *Journal of Machine Learning Research* **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. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE: Seattle, WA, USA, 2020; pp. 3008–3017. doi:10.1109/CVPRW50498.2020.00359.

53. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization, 2019, [arxiv:cs, math/1711.05101].