remote sensing

MDPI

Article

Remote Sensing Cross-Modal Text-Image Retrieval Based on Attention Correction and Filtering

Xiaoyu Yang [1], Chao Li [1], Zhiming Wang [1], Hao Xie [2], Junyi Mao [3] and Guangqiang Yin [1,2,4,*]

[1] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; yangxy@std.uestc.edu.cn (X.Y.); 202012090915@std.uestc.edu.cn (C.L.); zmwang@std.uestc.edu.cn (Z.W.)
[2] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China; 202222280115@std.uestc.edu.cn
[3] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia; jmao0023@student.monash.edu
[4] Kashgar Regional Electronic Information Industry Technology Research Institute, Kashi 844000, China
* Correspondence: yingq@uestc.edu.cn

Abstract: Remote sensing cross-modal text-image retrieval constitutes a pivotal component of multi-modal retrieval in remote sensing, central to which is the process of learning integrated visual and textual representations. Prior research predominantly emphasized the overarching characteristics of remote sensing images, or employed attention mechanisms for meticulous alignment. However, these investigations, to some degree, overlooked the intricacies inherent in the textual descriptions accompanying remote sensing images. In this paper, we introduce a novel cross-modal retrieval model, specifically tailored for remote sensing image-text, leveraging attention correction and filtering mechanisms. The proposed model is architected around four primary components: an image feature extraction module, a text feature extraction module, an attention correction module, and an attention filtering module. Within the image feature extraction module, the Visual Graph Neural Network (VIG) serves as the principal encoder, augmented by a multi-tiered node feature fusion mechanism. This ensures a comprehensive understanding of remote sensing images. For text feature extraction, both the Bidirectional Gated Recurrent Unit (BGRU) and the Graph Attention Network (GAT) are employed as encoders, furnishing the model with an enriched understanding of the associated text. The attention correction segment minimizes potential misalignments in image-text pairings, specifically by modulating attention weightings in cases where there's a unique correlation between visual area attributes and textual descriptors. Concurrently, the attention filtering segment diminishes the influence of extraneous visual sectors and terms in the image-text matching process, thereby enhancing the precision of cross-modal retrieval. Extensive experimentation carried out on both the RSICD and RSITMD datasets, yielded commendable results, attesting to the superior efficacy of the proposed methodology in the domain of remote sensing cross-modal text-image retrieval.

Keywords: cross-modal retrieval; remote sensing; attention weight

check for updates

Academic Editor: Haopeng Zhang

Received: 18 September 2024
Revised: 4 January 2025
Accepted: 21 January 2025
Published: 31 January 2025

Citation: Yang, X.; Li, C.; Wang, Z.; Xie, H.; Mao, J.; Yin, G. Remote Sensing Cross-Modal Text-Image Retrieval Based on Attention Correction and Filtering. *Remote Sens.* **2025**, *17*, 503. https://doi.org/10.3390/rs17030503

1. Introduction

In recent years, the rapid development and application of remote sensing technology have led to an exponential increase in the quantity of optical remote sensing images [1]. Yet, the challenge arises when faced with the task of efficiently extracting invaluable insights from such vast repositories of data. Automatic remote sensing cross-modal retrieval has become a key area of research due to the exponential growth of optical remote sensing data. The increasing role of text in human-computer interactions has further heightened interest in cross-modal text-image retrieval for remote sensing applications.

Historically, remote sensing image retrieval predominantly relied on manual annotations to label each image, utilizing query text to match these annotations [2]. Given the exponential increase in remote sensing images, manual annotation has become increasingly labor-intensive. This shift has prompted a heightened interest in automated image captioning [3,4]. For instance, to derive more nuanced descriptions, Zhao et al. [5] introduced an approach grounded in fine-grained and structured attention, aimed at harnessing the structural attributes of semantic content in remote sensing imagery. Despite the advancements in automated subtitle generation, challenges persist. Primarily, two-stage retrieval models often encounter substantial information attrition in the intermediary phase [6], thereby compromising the retrieval's precision and completeness. Furthermore, captions generated by machines may inadequately encapsulate unique semantic nuances and intricacies of remote sensing images [7]. This raises the question of whether there exist more optimal methodologies than the conventional remote sensing text-image retrieval techniques for cross-modal retrieval tasks.

Historically, when the Image-Text Retrieval challenge first gained traction, the academic community primarily aimed to map text and images into a shared subspace using an end-to-end approach. Yet, recent developments have charted new territories. For instance, SCAN [8] employs region-level and word-level feature encoding for images and text respectively, subsequently utilizing stacked cross-modal attention for affinity computations. CAMP [9] masterfully orchestrates the cross-modal messaging flow, ensuring meticulous cross-modal interactions and adeptly managing discordant pairs and non-essential data through an adaptive gating strategy. VSRN [10] devises visual representations by inferring regional relationships and global semantic connotations. This enhanced representation aptly captures pivotal objects and semantic motifs in scenes, facilitating superior alignment with associated text. SGM [11] leverages dual scene graph modalities for text and images: visual and textual scene graphs. It introduces a scene graph alignment model and harnesses two graph encoders to derive object-level and relation-level features for image-text alignment.

However, it's imperative to note that the above strategies, tailored for natural scenes, falter when applied to remote sensing imagery [12]. Yuan et al. [12] attempted to fine-tune remote sensing images using methodologies from [8,9] and related works. The results were suboptimal, leading them to formulate a multi-scale visual self-attention module to sift through extraneous image features and deploy cross-modal guidance protocols for enhanced multi-modal representations. To forge a more direct nexus between remote sensing imagery and corresponding text, Cheng et al. [13] utilized attention and gating mechanisms, optimizing data characteristics to extract more potent feature representations. Recognizing the dearth of fine-grained object perception in existing remote sensing retrieval frameworks, Yuan et al. [14] recognized the lack of fine-grained object perception in existing remote sensing retrieval frameworks. They proposed an integrated approach that combines both global and granular data through a multi-tier information fusion module. This strategy enabled a deeper understanding of objects and their interrelations, thereby improving retrieval performance.

Notwithstanding the achievements in remote sensing image-text retrieval, there remain pertinent challenges warranting further exploration. Firstly, remote sensing imagery markedly contrasts with natural scenes. Natural scenes typically feature fewer, larger, and more distinctive objects, whereas remote sensing images are characterized by numerous, smaller, and less distinct entities. As a result, extracting regional features from remote sensing images using target detection methods, and then aligning them with textual word features, becomes paramount. Although this strategy excels in natural scene datasets, the extraction of pertinent image regional features remains an outstanding challenge. Secondly, prevalent methodologies employ Bidirectional Gated Recurrent Unit (BGRU) for textual word feature extraction, which predominantly factors in immediate positional word relationships. This overlooks distant word relationships, which could offer significant insights. Lastly, while numerous scholars have championed attention mechanism strategies for discerning granular alignment between images and text, current iterations of these mechanisms warrant refinements. Specifically, the current focus with regard to merely granular alignments may be myopic, potentially obscuring cases of partial alignments in discordant sample pairs. In congruent pairs, not all attention weights bear significance. The prevailing methodologies indiscriminately treat all attention weights, inadvertently incorporating inconsequential textual prepositions.

To address the aforementioned challenges, this study introduces a cross-modal retrieval algorithm for remote sensing image text based on similarity correction and filtering (ACF). The principal contributions of this research are outlined below:

- Enhanced Image Comprehension: To bolster the model's proficiency in deciphering remote sensing images, we have adopted the visual graph neural network (VIG) [15], as the primary image feature extraction mechanism. Moreover, a multi-tier node feature fusion module has been instituted, enabling the model to understand remote sensing images both at varied granularities and in their entirety.
- Optimized Text Understanding: This study leverages the BGRU model to extract word-level vector features from the text. Subsequently, the graph attention network (GAT) is employed to compute M word-level features that represent positional relationships. The culmination of this process involves the utilization of pooling to capture the overarching features of the textual content.
- Attention Correction Unit: We introduce a novel attention correction unit. Herein, the visual area features coupled with the textual word features are processed via the cross-attention module to generate attention weights. Subsequently, global similarity metrics are employed to rectify these attention weights. A distinct attention threshold is incorporated to recalibrate the attention weight, substantially mitigating the propensity for misalignment in discordant image-text pairs, especially when such misalignments arise from specific correlations between visual area attributes and textual words.
- Attention Filtering Unit: Recognizing that not all attention-derived information holds relevance, we propose an attention filtering unit. This study aims to discern the most pertinent attention weight, resonating with the visual area features and textual word attributes, and employs a secondary attention threshold to filter out inconsequential attention. This strategic approach attenuates the influence of non-essential visual zones and words when aligning image-text pairs, thus amplifying the likelihood of accurate matches.

To substantiate the superiority of the ACF approach, we orchestrated a series of comparative experiments across two remote sensing cross-modal retrieval datasets. Furthermore, a range of ablation studies were executed to dissect the efficacy of each individual module. The ensuing sections are structured as follows: Section 2 provides an overview of related work in remote sensing image text retrieval. Section 3 delves deep into the intricacies of the proposed modules. Section 4 is dedicated to a comprehensive presentation of our experimental validations, demonstrating the potency of our proposed methodology. Finally, Section 5 draws conclusions based on the research findings.

2. Related Work

2.1. Cross-Modal Text-Image Retrieval

Cross-modal retrieval is the process by which data from one modality is utilized to retrieve semantically consistent modal information from another modality [16]. Using the image and text modality as an illustration, images retrieve related texts that fall under the same category or topic. Image-text retrieval primarily follows two research trajectories: global matching and regional matching.

Global matching is designed to seamlessly embed both images and texts into a shared subspace, subsequently learning their semantic alignment through optimization using a ranking loss. Canonical correlation analysis is employed by both CCA [17] and DCCA [18] to ascertain the semantic representation of images and their corresponding texts. Given the exceptional performance of convolutional neural networks in image processing [19] and the superior performance of LSTM [20] and GRU [21] in the realm of natural language processing, R. Kiros et al. [22] innovatively introduced the CNN-LSTM architecture for the purpose of learning combined image-text embeddings. Following the advancements achieved by pre-trained models in Natural Language Processing(NLP), exemplified by BERT [23] and GPT [24], TOD-Net [25], introduced in 2021, refines text representations. This method overlays a pre-existing embedding system, altering the embedding space based on specific parameters.

Drawing inspiration from generative adversarial networks [26], analogous generative and adversarial learning strategies can be adopted for image-text matching tasks. This approach seeks to bridge the divergence between the two modalities. Tools like ACMR [27] and CM-GANs [28] incorporate a modal discriminator to discern the modal data of features, leveraging a traditional bidirectional network. When discrimination becomes unfeasible, the disparity between the two modalities is deemed to have been resolved. Additionally, GXN [29] exploits either text or visual features to produce images or captions, aiding in the diminishment of cross-modal informational gaps. Wen et al. [30] presented a cross-memory network equipped with pair recognition, designed to encapsulate shared knowledge across image and text modalities.

Furthermore, specialized mechanisms have been incorporated within global matching. DAN [31] implements an attention mechanism fortified by visual and textual elements. In the context of MTFN [32], Wang et al. conceived a reordering strategy to refine the ranking accuracy during test phases. In pursuit of holistic matching, MFM [33] employs a multifaceted representation of both images and texts, facilitating a comprehensive understanding and subsequently discerning the congruence between both modalities from various perspectives. Ji et al. [34] unveiled the Saliency-Guided Attention Network (SAN), which capitalizes on visual saliency detection, emphasizing visually significant regions or entities in images based on textual content.

The aforementioned methodologies primarily encode an entire image or text into a singular vector, and are thus categorized under global image-text matching methods. However, these techniques predominantly focus on the alignment of the overarching context of either the image or text, often neglecting the congruence between specific image regions and textual elements. This oversight is addressed by the subsequent regional image-text matching methodology.

Regional matching offers a nuanced approach to image-text pairing, associating distinct regions within an image to specific words in a text, as opposed to solely aligning overarching semantics. This method capitalizes on target detection [35] for image object identification, diverging from the conventional CNN for image feature extraction. Concurrently, the output from the text encoder transitions from a singular sentence vector to a word-centric matrix. Pioneering this approach in 2015, Karpathy et al. [36] presented a technique to identify objects in images, embedding them within a subspace. The associated similarity is determined by the cumulative similarities of each region-word pairing. This approach to determining image-text congruence has been further refined by subsequent studies [8,9,37,38], which have employed attention mechanisms to delineate the regional congruence between visual and linguistic elements. Specifically, BFAN [37] selectively omits irrelevant segments from mutual semantics, directing focus towards pertinent segments. PFAN [38] amplifies this by incorporating regional location data, introducing an integration strategy that emphasizes object location hints, augmenting the learning of visual-textual joint embeddings, and thereby achieving superior alignment. Several other studies [9,39–43] have further contributed to refining the outcomes of image-text retrieval.

A unique proposition involves constructing inherent structures and relationships among fragments within their individual modalities, and subsequently identifying distinct inter-structural semantic correlations between visual and textual elements. Specifically, relationships between visual and textual fragments are modeled by developing both a visual context-aware tree encoder (VCS-Tree) and a textual context-aware tree encoder (TCS-Tree) with mutual labels. This facilitates the concurrent learning and optimization of both visual and textual features.

Graphical structures adeptly depict object-centric relationships. However, a plethora of visual data cannot consistently be represented in grid-like formats such as visual graphs. This led to the introduction of graph neural networks (GNN) [44] as extensions of recurrent neural networks, enabling them to directly process graphs. This was further expanded upon with the introduction of the Graph Convolutional Network (GCN) [45] tailored for capturing visual relationships. Several contemporary studies [10,11,46–49] have further expanded on this foundational work, aiming to enhance either visual or textual features in the context of image-text alignment. For instance, SCG [46] develops a scene concept graph by extracting frequently co-occurring concept pairings as intrinsic scene knowledge. Subsequently, this base is expanded to incorporate additional semantic concepts, selectively merging them to enhance an image's semantic representation. CVSE [47] harnesses consensus data by calculating statistical co-occurrence correlations among semantic concepts within image datasets, employing the constructed concept correlation graph to produce consensus-aware concept representations. Meanwhile, GSMN [48] distinctly models objects, relationships, and attributes as a structured phrase. This not only identifies the correlations among objects, relationships, and attributes but also aids in recognizing the intricate congruences among structured phrases.

2.2. Remote Sensing Cross-Modal Text-Image Retrieval

RSCTIR involves using text to retrieve corresponding remote-sensing images. The heterogeneity-induced semantic gap renders the RSCTIR task particularly challenging. In terms of implementation, RSCTIR approaches can be broadly categorized into subtitle-based methods and embedding-based methods.

Subtitle-based methods are essentially two-stage retrieval approaches. In this approach, annotations are typically generated for each remote sensing (RS) image in the database through a subtitle generator. Subsequently, during the retrieval phase, the BLEU [50] metric is utilized to compute the similarity between the query text and the generated annotations. Qu et al. [51] employed multi-modal deep networks for the semantic understanding of high-resolution remote sensing images. To enhance the characterization of remote sensing images, Shi et al. [52] introduced a Deep Learning-based Remote Sensing Image Captioning (RSIC) framework, employing fully convolutional networks to semantically decompose terrestrial elements at various scales. Lu et al. [53] presented a large-scale RSIC dataset and conducted an extensive review to promote the RSIC mission. Sumbul et al. [54] introduced a summary-driven RSIC technique to address information deficits and evaluated its influence across several RS text-image datasets. Li et al. [4] formulated a truncated cross-entropy loss to mitigate the overfitting issue in RSIC. Addressing the computational demands of contemporary subtitle generators, Genc et al. [55] designed a support vector machine-based decoder that proves efficient with limited training samples. While subtitle-based RSCTIR methods [5,56–58] have matured, their two-stage nature inevitably introduces noise, potentially compromising retrieval accuracy.

Embedding-based RSCTIR techniques entail mapping RS images and text into a shared high-dimensional space. The cross-modal similarity is then ascertained using appropriate distance metrics. Abdullah et al. [59] introduced a deep bidirectional triplet network that derives joint encoding between multiple modalities and yields more robust embeddings. They implemented an average fusion approach to amalgamate features from multiple text-image pairings. Addressing multi-scale scarcity and target redundancy challenges in RSCITR, Yuan et al. [12] designed an asymmetric multi-modal feature matching network (AMFMN) and contributed a fine-grained RS image-text dataset for this task. Investigating potential associations between RS images and text, Cheng et al. [13] devised a semantic alignment module to capture more discriminative feature representations. Lv et al. [60] proposed a fusion-based correlation learning model for RS image-text retrieval, this approach bridges the heterogeneity gap by leveraging knowledge distillation. Alternatively, Yuan et al. [14] suggested a lightweight text-image retrieval model was designed to expedite RS cross-modal retrieval by employing knowledge extraction and contrastive learning, thereby improving retrieval performance.

3. Method

This section elucidates the Attention Correction and Filtering algorithm tailored for cross-modal retrieval in remote sensing image-text contexts. The comprehensive workflow is depicted in Figure 1.

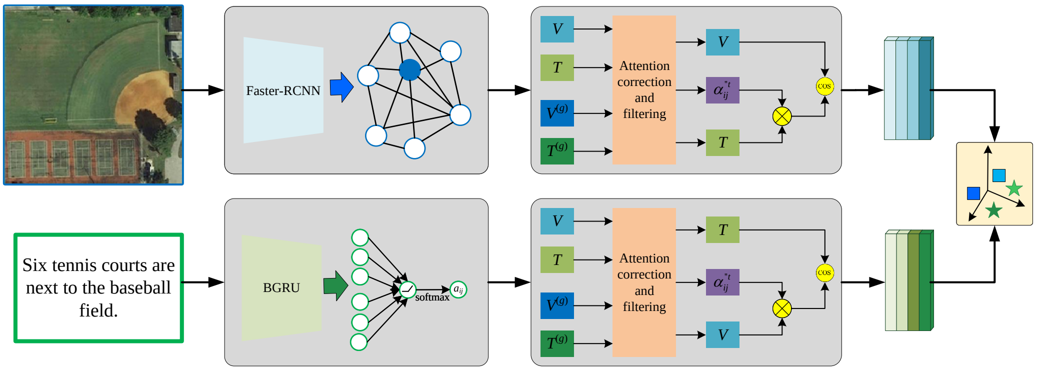


Figure 1. Overview of the proposed ACF.

Initially, in the context of image processing, the VIG is employed to extract image features. Building on this foundation, the fusion of low-level and mid-level node information produces N node details, which are subsequently treated as visual area features via the multi-scale fusion module.

For handling the textual aspect of remote sensing, a BGRU is employed to distill word-level vectors from the text. Following this, the Graph Attention Module is harnessed to derive M word-level attributes.

Subsequently, leveraging the features derived from the preceding two modules, a weight matrix is procured through the cross-attention module. This weight matrix is then scaled using global similarity metrics. To address information redundancy, only relevant weight details are retained. As such, the attention filtering module is employed to sieve out attention weights that are inconsequential to either the image or the text. To culminate, the model's training process synergistic ally integrates the Triplet loss function with global similarity measures.

3.1. Feature Extraction

3.1.1. Image Feature Extraction

Due to the complexity and particularity of remote sensing images, the object features obtained through the target detection algorithm need to be more representative. In order to extract valuable features from these object features, complex redundant operations are required, which undoubtedly increases the time required for retrieval. Process complexity. As a more flexible backbone network, the VIG [9] divides images into many blocks and treats them as nodes. Building graphs based on these nodes can better represent irregular and complex objects in the wild.

The intrinsic advantages of VIG address the challenges of meticulously extracting object features from remote sensing images. Therefore, this study adopts VIG as the backbone network for extracting remote sensing image features. Initially, each image is divided into 7 × 7 patches. Based on this, a multi-level node information fusion module is designed. The fusion of low-level and mid-level node features from the VIG results in 49 node features, which are used as visual region features. This approach enables the model to achieve a multi-level and comprehensive understanding of remote sensing images, as illustrated in Figure 2.

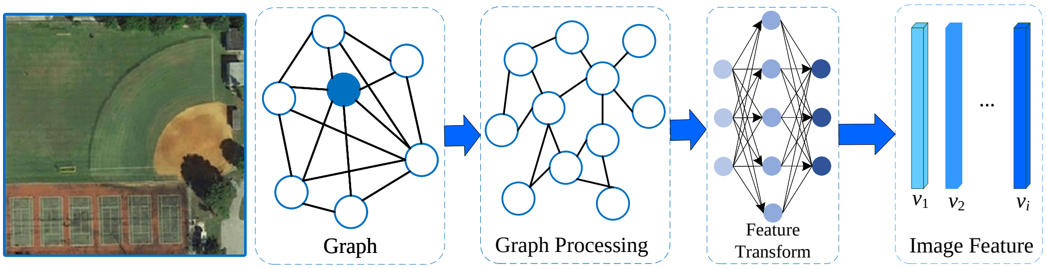


Figure 2. Image feature extraction module.

Within this module, node information from disparate levels is perceived as objects of varying dimensions. This data is then amalgamated, treating these object details as local insights. This approach compensates for the inadequacies of purely global perspectives, leading to the derivation of richer visual features, as represented in Equation (1). Subsequent to this, a global pooling mechanism is deployed to yield the visual global feature $V^{(g)}$. This feature is then synergized with the terminal layer attributes of the VIG, as depicted in Equation (2).

$$v_i = VIG(x_l, x_m), i \in [1, k] \tag{1}$$

$$v^{(g)} = VIG(x_l, x_m) \times x_t \tag{2}$$

3.1.2. Text Feature Extraction

To achieve a word-level representation of text, this study employs both the BGRU and the GAT as encoders, as depicted in Figure 3. For a specific text S , comprising m words, we represent these words using word vectors e_j. Acknowledging the significance of positional data within the sentence structure, these word vectors are channeled into the Bidirectional GRU network. This yields word feature representations, which are subsequently introduced into the GAT. This network is responsible for discerning and learning inter-word correlations, culminating in the final word features, denoted as h_j. This computational procedure is detailed in Equation (2). Subsequent to this phase, an average pooling strategy is employed to extract the global text feature, $T^{(g)}$ as elaborated upon in Equation (4).

$$h_j = GAT(\frac{\overrightarrow{GRU}(e_j, \overrightarrow{h}_{j-1}) + \overrightarrow{GRU}(e_j, \overrightarrow{h}_{j+1})}{2}), j \in [1, m] \tag{3}$$

$$T^{(g)} = \frac{1}{m} \sum_{j=1}^{m} h_j \tag{4}$$

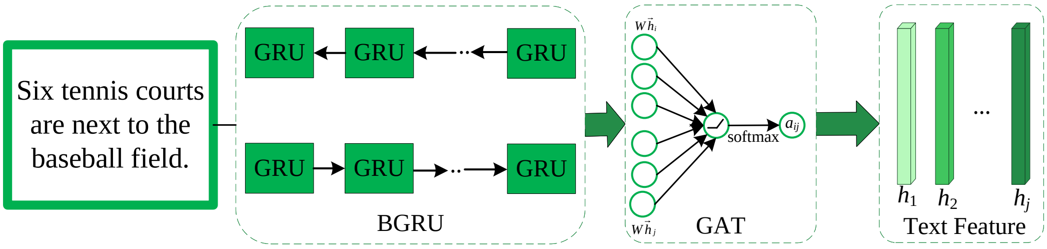


Figure 3. Text feature extraction module.

3.2. Attention Correction and Filtering

The focus of the attention mechanism should be appropriately directed towards the pertinent regions in images or relevant segments in text. The Attention Correction and Filtering module is structured in three sequential stages:

1. Cross-Attention Generation: In this phase, an attention weight matrix is derived using the cross-attention mechanism. This matrix characterizes the relationships between elements in the image and text modalities.
2. Attention Correction via Global Similarity: The initial attention weights are refined using a measure of global similarity.
   This refinement ensures that the attention mechanism is focused on semantically consistent areas of the image and corresponding segments of the text.
3. Attention Filtering for Relevance Determination: This stage identifies and retains only the most relevant attention weights. By concentrating on highly relevant areas, it eliminates non-essential regions or segments, thereby reducing noise in the attention mechanism.

The complete procedure is visually depicted in Figure 4. It's crucial to note that the aforementioned stages are bidirectional in nature. This means that the mechanism can operate in two modes: from images to text and vice versa (text-to-image). To facilitate a clearer understanding, the subsequent sections will elaborate on the image-to-text procedure, detailing each stage comprehensively.

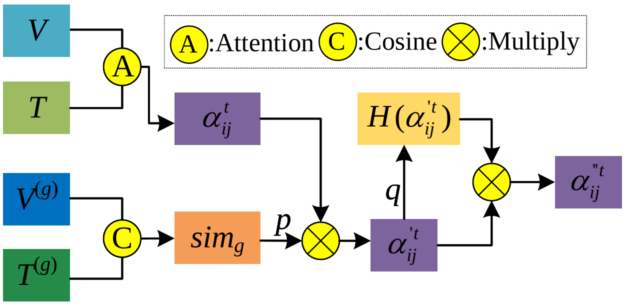


Figure 4. Attention correction and filtering flow chart.

3.2.1. Base Attention

In order to obtain the image-to-text attention weight s_{ij}, first calculate the similarity matrix between image-text pairs, which is obtained by calculating the cosine similarity between the image region feature v_i and the text word feature w_j, as shown in Equation (5).

$$s_{ij} = \frac{v_i^T w_j}{\|v_i\| \|w_j\|}, i \in [1, k], j \in [1, m] \tag{5}$$

Then regularize it to get $\overline{s_{ij}} = \frac{[s_{ij}]_+}{\sqrt{\sum_{i=1}^{k} [s_{ij}]_+^2}}$, where $[x]_+ = Max(x, 0)$. The similarity matrix α_{ij}^t is then used to calculate the attention score of each region, as shown in Equation (6).

$$\alpha_{ij}^t = \frac{\exp(\lambda \overline{s_{ij}})}{\sum_{j=1}^{m} \exp(\lambda \overline{s_{ij}})} \tag{6}$$

where λ is the inverse temperature of similarity.

3.2.2. Attention Correction Unit

This study employs the similarity between the global features of images and text as a constraint for attention weights. This approach aims to diminish attention to irrelevant visual areas or text segments and instead prioritize semantically relevant regions in both images and text. The mathematical representation for this global similarity sim_g is provided in Equation (7).

$$sim_g = sim(v^{(g)}, T^{(g)}) = \frac{v^{(g)T} T^{(g)}}{\|v^{(g)}\| \|T^{(g)}\|} \tag{7}$$

This study introduces an attention weight threshold, denoted as p, to evaluate the magnitude of the global similarity. By multiplying this threshold with the attention matrix, we derive a new normalized attention weight matrix $\alpha_{ij}'^t$. This process is mathematically represented in Equations (8) and (9).

$$\overline{\alpha}_{ij}^t = (sim_g - p) \times \alpha_{ij}^t \tag{8}$$

$$\alpha_{ij}'^t = \frac{\overline{\alpha}_{ij}^t}{\sum_{j=1}^{m} \overline{\alpha}_{ij}^t} \tag{9}$$

Should the global similarity prove substantial, the local similarity will be proportionally amplified following the attention correction module. Conversely, if the global similarity is minimal, the local similarity will be proportionally diminished post-attention correction. In essence, the attention correction module mitigates the potential for mismatched image-text pairings that might arise from alignments between specific image areas and distinct text words.

3.2.3. Attention Filtering Unit

Given the redundancy in attention weights, it's evident that not all attention-weight data holds significance. This study seeks to identify the attention most pertinent to the text word features, namely, the visual area features with the highest attention weight. As a foundational step, we introduce an attention weight ratio threshold, denoted as q. This threshold evaluates the relationship between a given attention weight value and the maximal attention weight value. Any weight values below this threshold q are nullified. Following this, we derive a refreshed attention weight matrix post-normalization $\alpha_{ij}''^t$, as illustrated in Equations (10) and (11).

$$H(\alpha'^{t}_{ij}) = \begin{cases} \alpha'^{t}_{ij}, \left|\alpha'^{t}_{ij} - Max(\alpha'^{t}_{i})\right| < q \\ 0, other \end{cases} \tag{10}$$

$$\alpha''^{t}_{ij} = \frac{\alpha'^{t}_{ij} H(\alpha'^{t}_{ij})}{\sum_{j=1}^{m} \alpha'^{t}_{ij} H(\alpha'^{t}_{ij})} \tag{11}$$

In corresponding text-image pairs, the attention filtering module minimizes the impact of unrelated areas and words, ensuring that the attention weight predominantly concentrates on the matching visual areas and relevant words.

3.3. Loss Function

In the Attention Correction and Filtering section, attention is directed towards relevant words or visual areas. Subsequently, the final text vector α^{t}_{i} and image vector α^{v}_{j} are determined as illustrated in Equations (12) and (13).

$$\alpha^{t}_{i} = \sum_{j=1}^{m} \alpha''^{t}_{ij} w_j \tag{12}$$

$$\alpha^{v}_{j} = \sum_{i=1}^{k} \alpha''^{t}_{ij} v_i \tag{13}$$

The matching scores of text and image can then be derived from the two-way matching, as shown in Equation (14).

$$R(I,T) = \frac{1}{k} \sum_{i=1}^{k} R(v_i, v_i) + \frac{1}{m} \sum_{j=1}^{m} R(\alpha^{v}_{j}, w_j) \tag{14}$$

In remote sensing cross-modal retrieval, the triplet ranking loss function is frequently employed. In this study, we continue to use this loss function to align images and text. Furthermore, a global similarity metric is incorporated to jointly compute the loss value. The specific calculation is provided in Equation (15), where δ represents the minimum boundary value, $\widehat{I}$ denotes the remote sensing image that does not match the text T, $\widehat{T}$ represents the text that does not match the remote sensing image I, and L_g is the loss calculated based on global similarity.

$$L = [\delta - R(\widehat{I}, T) - R(I,T)]_{+} + [\delta - R(I, \widehat{T}) - R(I,T)]_{+} + L_g \tag{15}$$

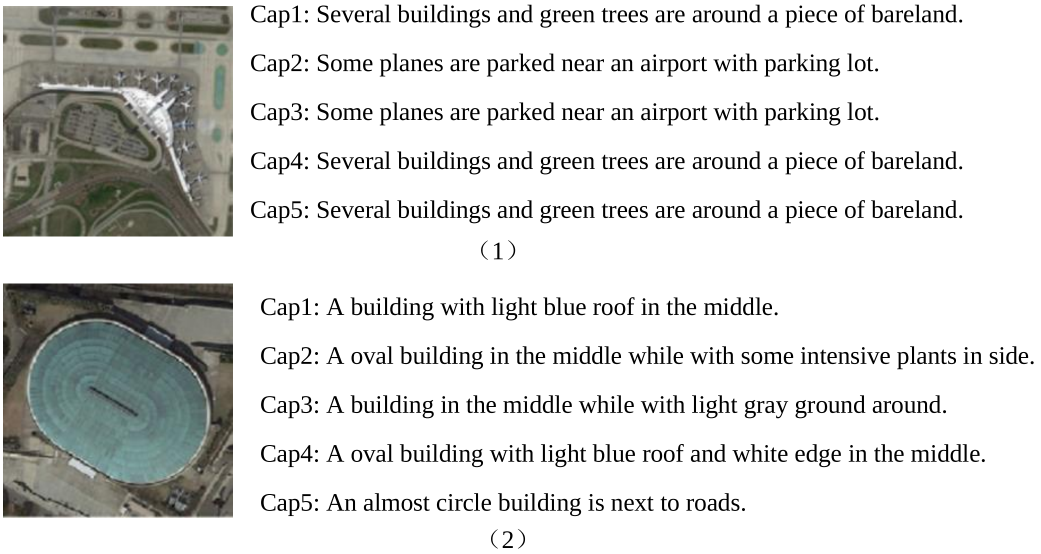
4. Experiment

This section provides an overview of the datasets utilized, the evaluation metrics, and the specifics of the experiments conducted. We will compare and analyze two widely recognized remote sensing text-image datasets and validate the efficacy of the ACF model we've developed. Furthermore, we will conduct a series of ablation studies to delve into the underlying factors contributing to the superior performance of the ACF model.

4.1. Datasets and Evaluation Metrics

In our study, two prominent remote sensing text-image datasets, RSICD [12] and RSITMD [53], were employed. The RSICD dataset comprises 10,921 samples, with each sample containing a remote sensing image accompanied by five pertinent sentence descriptions; these images have a resolution of 224 × 224. The RSITMD dataset, on the other hand, consists of 4743 samples, and akin to the RSICD, each sample features a remote sensing image coupled with five sentence descriptions, albeit with an image resolution of 256 × 256. Notably, the RSITMD dataset offers a more intricate textual representation in comparison to the RSICD dataset, as illustrated in Figure 5. For the purposes of our experiment, the datasets were partitioned into a training set (comprising 80% of the data), a validation set (10%), and a test set (10%), in alignment with the methodology proposed by Yuan et al. [12].

To evaluate the model's efficacy, this study employs the R@K and mR metrics. R@K denotes the percentage of accurate matches within the top k retrieved results. For a comprehensive assessment, the experiment utilized R@1, R@5, and R@10 as metrics. Furthermore, mR, representing the mean value across multiple R@K values (specifically for K = 1, 5, 10), was employed to provide a holistic perspective on the model's performance.



Cap1: Several buildings and green trees are around a piece of bareland.

Cap2: Some planes are parked near an airport with parking lot.

Cap3: Some planes are parked near an airport with parking lot.

Cap4: Several buildings and green trees are around a piece of bareland.

Cap5: Several buildings and green trees are around a piece of bareland.

（1）

Cap1: A building with light blue roof in the middle.

Cap2: A oval building in the middle while with some intensive plants in side.

Cap3: A building in the middle while with light gray ground around.

Cap4: A oval building with light blue roof and white edge in the middle.

Cap5: An almost circle building is next to roads.

（2）

Figure 5. (1) is a sample in RSICD, and (2) is a sample in RSITMD. Each sample in both datasets contains five descriptions, but RSITMD's descriptions are more diverse.

4.2. Implementation Details

All experiments presented in this study were executed on an NVIDIA RTX8000 GPU. Despite the varying image sizes of the two primary datasets employed, for consistency, we resized all images to a dimension of 224 × 224 before feeding them into the model. To bolster the model's resilience to variations, a series of data augmentation techniques were applied to the training set's image data, including operations such as cropping and rotation.

For the extraction of textual features, the word vector's dimension was fixed at 300, while the features' dimension, used to compute the similarity between images and texts, was set at 512. Optimal thresholds for attention correction (p) and attention filtering (q) were identified through controlled parameter experimentation, settling at values of 0.3 and 0.1, respectively. As this study employs the triplet loss function, setting an appropriate margin is essential. Again, through parameter tuning, the most effective margin was determined to be 0.2.

Regarding the optimization of the model, the Adam optimizer was utilized. We adopted a batch size of 150 and set the initial learning rate at 0.001. A decay factor of 0.5 was applied every 20 epochs, and the model was trained for a total of 60 epochs. We leverage k-fold cross-validation to obtain an average result, and k is set to 5.

4.3. Parameter Experiment

This section delves into control experiments focused on three parameters: the attention correction threshold p, the attention filtering threshold q, and the triplet loss function margin δ. The baseline values for these parameters were initialized as follows: p was set at 0.1, q at 0.1, and δ at 0.2. Detailed experimental outcomes are visually represented in Figure 6, and the specifics of each experimental run will be elucidated in the subsequent discussions.

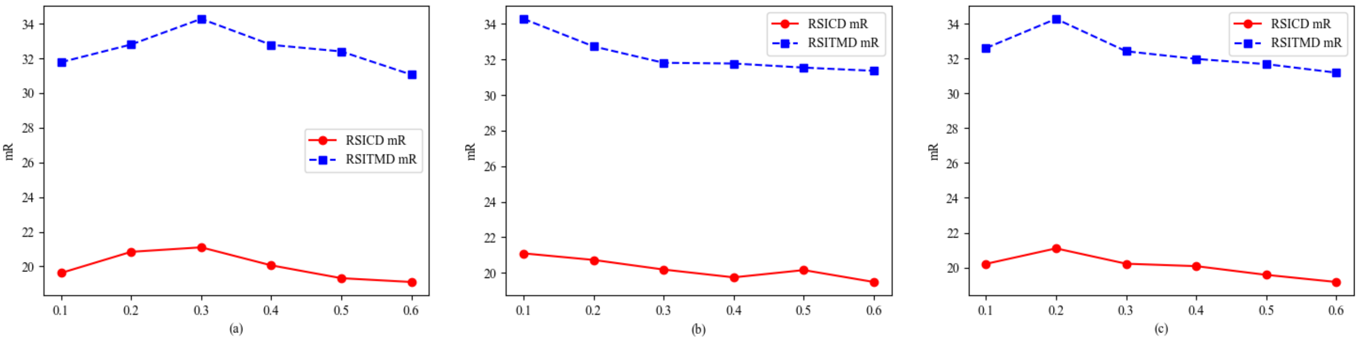


Figure 6. Parameter experiment results. (**a**) Attention correction threshold experiment results. (**b**) Attention filtering threshold experiment results. (**c**) Margin experiment results.

4.3.1. Attention Correction Threshold p

The attention correction threshold p, governs the modulation of attention weights. Its primary role is to mitigate the chances of erroneous matching in image-text pairs that are not inherently related, even if there seems to be a specific match between certain visual area features and text words. Setting p too low may render the scaling effect insignificant. Conversely, an excessively high value for p can negatively impact the matching likelihood of genuine positive samples.

To comprehensively assess the influence of p, experiments were conducted on two distinct datasets. In these experiments, the value of p was incrementally adjusted, ranging from 0.1 to 0.6. The outcomes of these experiments are tabulated in Table 1.

Table 1. Attention correction threshold experiment on RSICD dataset and RSITMD dataset.

	RSICD Dataset							**RSITMD Dataset**						
Thredhold	**Sentence Retrieval**			**Image Retrieval**			**mR**	**Sentence Retrieval**			**Image Retrieval**			**mR**
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**		**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
$p = 0.1$	6.92	19.76	30.34	5.94	20.82	33.94	19.62	13.72	30.31	44.47	9.87	37.12	55.66	31.78
$p = 0.2$	7.81	**20.71**	**31.88**	6.09	22.26	36.32	20.84	14.82	33.63	43.81	9.56	37.52	**57.52**	32.80
$p = 0.3$	**8.23**	20.31	30.47	**7.39**	**23.28**	**36.91**	**21.10**	**15.94**	**34.96**	**49.12**	**12.83**	37.74	55.53	**34.28**
$p = 0.4$	7.39	19.69	30.74	5.74	21.83	34.97	20.06	13.94	33.41	46.24	11.24	**38.36**	53.50	32.78
$p = 0.5$	5.95	18.21	31.47	5.58	21.06	33.65	19.32	13.32	33.54	46.28	10.90	38.18	56.69	32.40
$p = 0.6$	6.13	19.12	29.00	5.45	20.79	34.11	19.10	12.21	29.25	42.65	10.14	36.88	55.22	31.06

From Table 1, it is evident that while maintaining the initial values for the attention filtering threshold q and the triplet loss function margin δ, adjusting the attention correction threshold p to 0.3 enhances the model's performance across both datasets.

For the RSICD dataset, our model showcased superior performance in the image retrieval task across all three metrics. In the text retrieval task, we recorded the best results in R@1 and R@5. However, for R@10, there was a slight decrease of 0.4% and 1.41% respectively when compared to the results with $p = 0.2$.

On the RSITMD dataset, the model exhibited optimal performance in the text retrieval task for all three metrics. For the image retrieval task, while the model achieved the best results for R@1, there was a slight decrease of 0.62% for R@5 when compared to $p = 0.4$, and a 1.99% decrease for R@10 compared to $p = 0.2$. Despite these reductions, the model still demonstrates competitive and commendable results.

4.3.2. Attention Filtering Threshold q

When assessing the impact of the attention filtering threshold q, its primary role is to modulate the influence of unmatched or irrelevant visual areas and words in image-text pairs. By varying q within the range of 0.1 to 0.6 across the two datasets, we aim to understand its optimal value for maximum performance. The outcomes of these experiments can be found in Table 2.

Table 2. Attention filtering threshold experiment on rsicd dataset and rsitmd dataset.

	RSICD Dataset							RSITMD Dataset						
Thredhold	**Sentence Retrieval**			**Image Retrieval**			**mR**	**Sentence Retrieval**			**Image Retrieval**			**mR**
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**		**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
$q = 0.1$	**8.23**	**20.31**	30.47	**7.39**	**23.28**	**36.91**	**21.10**	**15.49**	**34.96**	49.12	**12.83**	37.74	55.53	**34.28**
$q = 0.2$	7.41	**20.04**	31.20	6.04	22.89	36.83	20.73	12.74	32.26	46.15	11.12	**38.21**	55.93	32.73
$q = 0.3$	**7.14**	20.40	**32.48**	**6.39**	**21.33**	33.39	20.19	12.65	31.06	44.78	11.30	36.78	54.31	31.81
$q = 0.4$	7.04	19.76	30.74	5.23	21.81	33.92	19.75	12.83	29.87	43.94	10.58	37.69	55.74	31.77
$q = 0.5$	7.50	20.21	31.38	6.37	21.06	34.47	20.16	12.26	30.58	43.41	10.25	36.70	**56.08**	31.54
$q = 0.6$	7.59	20.59	30.28	5.67	21.04	31.78	19.49	11.68	30.22	42.96	10.55	37.37	55.38	31.36

From Table 2, it's evident that without modifying the values of the attention correction threshold p and the triplet loss function margin δ, adjusting the attention filtering threshold q to 0.1 allows the model to excel in both retrieval directions on the RSICD dataset. In the context of image retrieval, optimal results were obtained. For text retrieval, peak results were noted in R@1 and R@5, although R@10 performance lagged slightly behind the outcomes achieved when q was set to 0.3.

For the RSITMD dataset, text retrieval metrics were all at their peak. In the image retrieval dimension, the model delivered the best results for R@1 and showed competitive performance for R@5—a mere 0.47% decline compared to when q was set to 0.2. For R@10, the performance drop was 0.55% in comparison to a q value of 0.5.

4.3.3. Margin δ

The triplet loss function optimizes the model by minimizing the distance between positive samples and maximizing the distance between negative samples. When the margin δ is small, the loss approaches 0, making it challenging to distinguish between positive and negative samples. Conversely, a larger margin δ suggests a greater expected distance between positive samples and a more substantial separation from negative samples. However, this can make network convergence more challenging.

Experiments were carried out on two datasets, adjusting the value of δ incrementally from 0.1 to 0.6. The findings of these experiments are presented in Table 3.

Table 3. Margin experiment on RSICD dataset and RSITMD dataset.

	RSICD Dataset							RSITMD Dataset						
Margin	**Sentence Retrieval**			**Image Retrieval**			**mR**	**Sentence Retrieval**			**Image Retrieval**			**mR**
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**		**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
$\delta = 0.1$	6.68	19.12	29.83	**6.92**	23.18	35.55	20.21	12.61	30.97	45.80	12.30	38.27	**55.66**	32.60
$\delta = 0.2$	**8.23**	**20.31**	30.47	**7.39**	**23.28**	**36.91**	**21.10**	**15.49**	**34.96**	**49.12**	**12.83**	**37.74**	55.53	**34.28**
$\delta = 0.3$	7.87	19.76	**31.56**	**6.68**	**21.48**	33.98	20.22	14.82	32.30	44.47	11.73	**38.41**	52.79	32.42
$\delta = 0.4$	6.68	17.38	30.56	6.62	22.78	36.45	20.08	12.70	32.03	44.16	10.31	37.31	55.38	31.98
$\delta = 0.5$	6.95	19.21	31.11	6.17	21.10	32.96	19.58	12.17	30.31	46.02	10.88	36.90	53.81	31.68
$\delta = 0.6$	5.76	17.84	28.82	5.56	21.35	35.66	19.17	11.50	31.86	45.58	9.16	36.81	52.26	31.19

From Table 3, we observe that by holding the initial values of the attention correction threshold p and the attention filtering threshold q constant, setting the triplet loss function margin δ to 0.2 enhances the model's performance in both retrieval directions. Specifically:

For the RSICD dataset: In the image retrieval tasks, the model achieves the best results. In text retrieval tasks, the model excels in R@1 and R@5 metrics. However, there is a decline of 1.09% in the R@10 metric when compared to $\delta = 0.3$.

For the RSITMD dataset: In text retrieval, the model outperforms the outcomes observed for other values of δ. Regarding the image retrieval task, the model's best performance is noted in the R@1 metric. However, there's a decrease of 0.67% in the R@5 metric relative to $\delta = 0.3$, and a 0.13% reduction in the R@10 metric compared to $\delta = 0.1$.

4.4. Comparison with the Other Methods

This study compares the performance of the ACF model with contemporary methods on the RSICD and RSITMD datasets. The results of this comparison are detailed in Table 4. The primary models under consideration include VSE++ [61], SCAN [8], CAMP [9], MTFN [32], CMFN [62], AMFMN [12] GaLR [63] and SWAN [64].

- VSE++ [61]: This model extracts image features using CNNs and text features using GRU. It employs the triplet loss function directly for model optimization.
- SCAN [8]: SCAN extracts image regional features via target detection and text word features using a bidirectional GRU. It subsequently aligns them finely using a cross-attention mechanism.
- CMAP [9]: CAMP utilizes a passing mechanism to adaptively control cross-modal information flow, producing the final result through cosine similarity.
- MTFN [23]: This model capitalizes on the fusion of various features to compute cross-modal similarity in an end-to-end manner.
- CMFN [62]: CMFN enhances retrieval performance by individually learning the feature interaction between query text and RS images and modeling the feature association between both modes, thus preventing information misalignment.
- LW-MCR [63]: This lightweight multi-scale cross-modal retrieval method leverages techniques such as knowledge distillation and contrast learning.
- AMFMN [12]: AMFMN employs a multi-scale self-attention module to derive image features. These features then guide text representation, and a dynamically variable triplet loss function optimizes the model.
- GaLR [64]: GaLR amalgamates image features from different levels using a multi-level information dynamic fusion module, eliminating redundancy in the process.
- SWAN [64]: SWAN uses a multi-scale fusion module to extract regional image features and then employs significant feature correlation to formulate a comprehensive image representation.

As presented in Table 4, the proposed ACF model achieves superior performance on the RSICD and RSITMD datasets compared to other methods. The experimental results on the RSICD dataset reveal that, in the text retrieval task, the ACF algorithm outperforms other methods in terms of R@1 and R@5 metrics, while the R@10 metric is slightly lower than that of the GaLR with MR method. In the image retrieval task, the ACF algorithm similarly surpasses other methods in R@1 and R@5 metrics, with the R@10 metric being slightly lower than that of the latest SWAN method. Notably, the ACF algorithm achieves an mR value of 21.10, demonstrating an improvement over other algorithms.

On the RSITMD dataset, the experimental results indicate that, in the text retrieval task, the ACF algorithm surpasses the latest SWAN algorithm across all metrics, achieving R@1, R@5, and R@10 values of 15.49%, 34.96%, and 49.12%, respectively. For the image retrieval task, the R@1 metric of the ACF algorithm is 1.59% higher than that of the SWAN algorithm, while the R@5 and R@10 metrics are slightly lower than the corresponding metrics of the SWAN algorithm. Finally, the ACF algorithm achieves an mR value of 34.28, representing a significant improvement over other algorithms. These results strongly validate the superiority of the ACF model and confirm the effectiveness of the attention weight correction and filtering method for cross-modal retrieval of remote sensing images and texts.

Table 4. Comparisons of retrieval performance on RSICD dataset and RSITMD dataset.

	RSICD Dataset							**RSITMD Dataset**						
Thredhold	**Sentence Retrieval**			**Image Retrieval**			**mR**	**Sentence Retrieval**			**Image Retrieval**			**mR**
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**		**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
VSE++	3.38	9.51	17.46	2.82	11.32	18.10	10.43	10.38	27.65	39.60	7.79	24.87	38.67	24.83
SCAN t2i	4.39	10.90	17.64	3.91	16.20	26.49	13.25	10.18	28.53	38.49	10.10	28.98	43.53	26.64
SCAN i2t	5.85	12.89	19.84	3.71	16.40	26.73	14.23	11.06	25.88	39.38	9.82	29.38	42.12	26.28
CAMP-triplet	5.12	12.89	21.12	4.15	15.23	27.81	14.39	11.73	26.99	38.05	8.27	27.79	44.34	26.20
CAMP-bce	4.20	10.24	15.45	2.72	12.76	22.89	11.38	9.07	23.01	33.19	5.22	23.32	38.36	22.03
MTFN	5.02	12.52	19.74	4.90	17.17	29.49	14.81	10.40	27.65	36.28	9.96	31.37	45.84	26.92
CMFM	5.40	18.66	28.55	5.31	18.57	30.03	17.75	10.84	28.76	40.04	10.00	32.83	47.21	28.28
LW-MCR(b)	4.57	13.71	20.11	4.02	16.47	28.23	14.52	9.07	22.79	38.05	6.11	27.74	49.56	25.55
LW-MCR(d)	3.29	12.52	19.93	4.66	17.51	30.02	14.66	10.18	28.98	39.82	7.79	30.18	49.78	27.79
AMFMN-soft	5.05	14.53	21.57	5.05	19.74	31.04	16.02	11.06	25.88	39.82	9.82	33.94	51.90	28.74
AMFMN-fusion	5.39	15.08	23.40	4.90	18.28	31.44	16.42	11.06	29.20	38.72	9.96	34.03	52.96	29.32
AMFMN-sim	5.21	14.72	21.57	4.08	17.00	30.60	15.53	10.63	24.78	41.81	11.51	34.69	54.87	29.72
GaLR w/o MR	6.50	18.91	29.70	5.11	19.57	31.92	18.62	13.05	30.09	42.70	10.47	36.34	53.35	31.00
GaLR with MR	6.59	19.85	**31.04**	4.69	19.48	32.13	18.96	14.82	31.64	42.48	11.15	36.68	51.68	31.41
SWAN	7.41	20.13	30.86	5.56	22.26	**37.41**	20.61	13.35	32.15	46.90	11.24	**40.40**	**60.60**	34.11
ACF (ours)	**8.23**	**20.31**	30.47	**7.39**	**23.28**	36.91	**21.10**	**15.49**	**34.96**	**49.12**	**12.83**	37.74	55.53	**34.28**

4.5. Ablation Study

In this section, we delve into ablation studies to assess the significance of each module within the proposed method. To ensure consistency, the hyperparameters were meticulously chosen based on prior parameter experiments. The series consists of five distinct experiments. The results for the RSICD dataset are documented in Table 5, whereas those for the RSITMD dataset can be found in Table 6.

Table 5. Ablation experiment on RSICD dataset.

M1	**M2**	**M3**	**M4**	**M5**	**Sentence Retrieval**			**Image Retrieval**			**mR**
					R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
					7.32	19.12	30.83	5.76	20.00	33.32	19.39
√					7.12	20.02	30.98	5.75	20.91	33.83	19.77
√	√				7.50	19.76	31.75	6.39	20.42	34.82	20.11
√	√	√			6.04	19.30	30.92	6.57	23.29	36.63	20.46
√	√	√	√		8.34	21.04	32.48	6.11	21.57	36.19	20.95
√	√	√	√	√	8.23	20.31	30.47	7.39	23.28	36.91	21.10

Table 6. Ablation experiment on RSITMD dataset.

M1	**M2**	**M3**	**M4**	**M5**	**Sentence Retrieval**			**Image Retrieval**			**mR**
					R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	
					11.95	28.23	41.11	10.95	34.94	51.35	29.75
√					13.76	31.02	42.57	11.05	36.08	51.81	31.05
√	√				16.59	32.08	44.49	11.55	37.70	53.05	32.61
√	√	√			12.83	33.19	48.89	11.50	37.88	54.91	33.20
√	√	√	√		14.16	34.51	48.23	12.92	38.67	54.87	33.89
√	√	√	√	√	15.49	34.96	49.12	12.83	37.74	55.53	34.28

This detailed breakdown aims to elucidate the contribution of each module to the overall effectiveness of our approach.

- M1: Incorporates the GAT for text feature extraction.
- M2: Pertains to image feature extraction supplemented with a multi-scale fusion module.
- M3: Involves the attention correction unit.
- M4: Introduces the attention filtering unit.
- M5: Adds a global similarity component.

On the RSICD dataset:

- With the inclusion of the M1 module, there was a rise in the mR score of the model by 0.38.
- Upon the integration of the M2 module, the mR score experienced an increment of 0.72. This marked a 0.34 rise compared to the addition of M1 alone.
- Introducing the M3 module further augmented the mR score by 1.07. This denotes an enhancement of 0.35 when stacked against the combined addition of M1 and M2. Notably, at this juncture, the model topped the R@5 metric in the image retrieval task.
- The addition of the M4 module propelled the mR score by 1.56, showcasing an improvement of 0.49 over the previous configuration. This configuration yielded the best performance in the realm of text retrieval.
- Finally, with all modules incorporated, the model's mR score surged by 1.71. In terms of image retrieval, the model outperformed its peers in the R@1 and R@10 metrics.

This progression underlines the cumulative efficacy of each module and their combined influence in enhancing the model's performance.

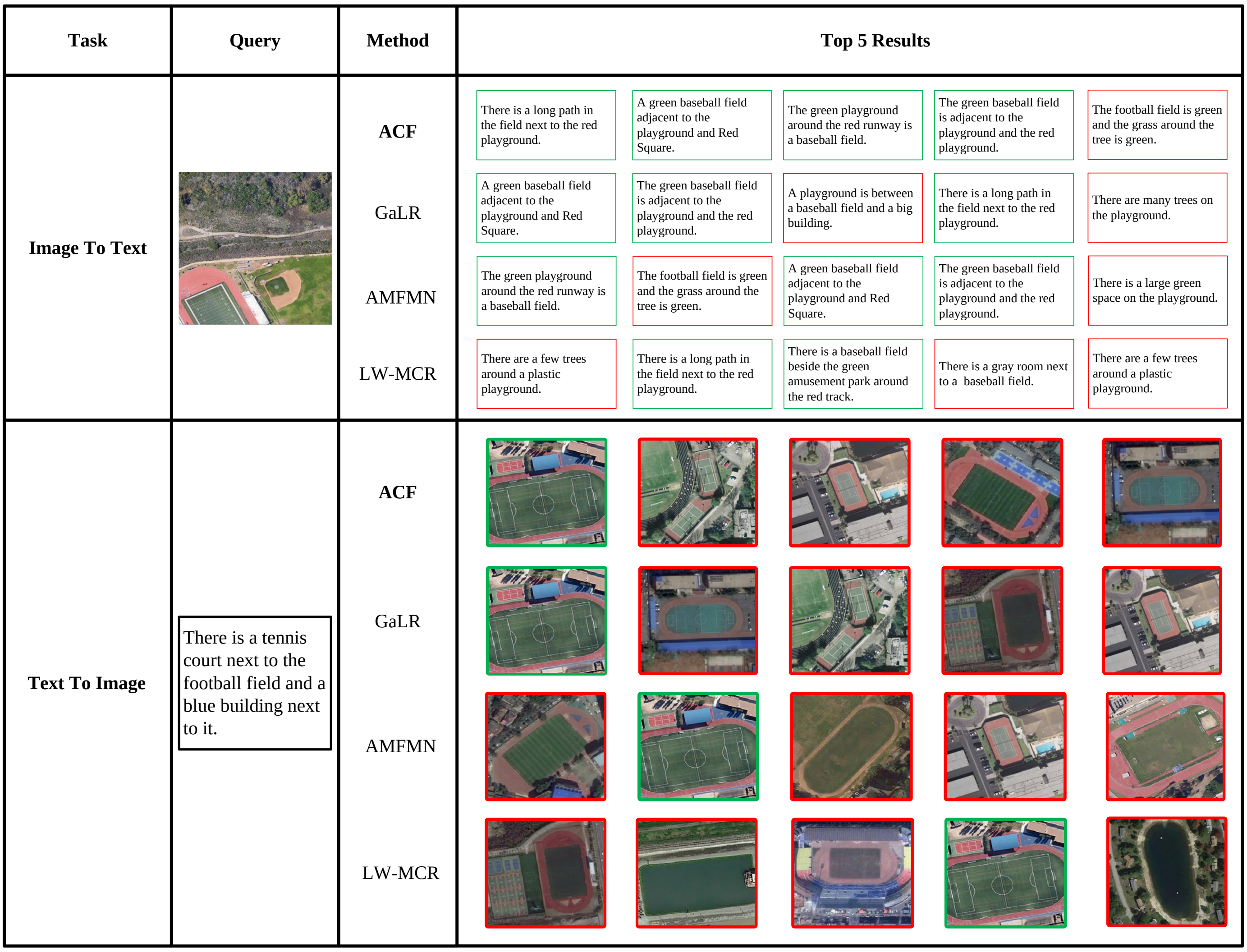
On the RSITMD dataset:

- With the integration of the M1 module, there was an increase in the mR score of the model by 1.3.
- Upon adding the M2 module, the mR score surged by 2.86, marking an enhancement of 1.56 compared to the sole addition of M1. Remarkably, during this phase, the model achieved pinnacle performance in the R@1 metric of text retrieval.
- Introducing the M3 module further augmented the mR score to 3.45. This denotes a rise of 0.59 when juxtaposed against the cumulative addition of M1 and M2.
- The inclusion of the M4 module propelled the mR score to 4.14, showcasing an improvement of 0.69 over the prior configuration. At this stage, the model clinched the top spot in the R@1 and R@5 metrics for image retrieval.
- Ultimately, when all modules were incorporated, the model's mR score reached 4.53. It stood out in the R@10 metric for image retrieval and achieved premier results in both R@5 and R@10 metrics for text retrieval.

This trajectory highlights the combined impact of each module in driving the model's performance on the RSITMD dataset to new heights.

4.6. Visual Analysis of Retrieval Results

In the following subsection, we provide a visual analysis to offer an intuitive comparison of performance disparities across several retrieval models. We have chosen the GaLR, AMFMN, and LW-MCR models to compare against our proprietary model. These experiments were performed on the RSITMD dataset, and the comparative visuals are depicted in Figure 7. Within these visuals, a green box signifies a correct match, whereas a red box indicates an incorrect match. This distinction aids in an immediate and clear understanding of each model's efficacy in retrieval tasks.



Task	Query	Method	Top 5 Results				
Image To Text		**ACF**	There is a long path in the field next to the red playground.	A green baseball field adjacent to the playground and Red Square.	The green playground around the red runway is a baseball field.	The green baseball field is adjacent to the playground and the red playground.	The football field is green and the grass around the tree is green.
		GaLR	A green baseball field adjacent to the playground and Red Square.	The green baseball field is adjacent to the playground and the red playground.	A playground is between a baseball field and a big building.	There is a long path in the field next to the red playground.	There are many trees on the playground.
		AMFMN	The green playground around the red runway is a baseball field.	The football field is green and the grass around the tree is green.	A green baseball field adjacent to the playground and Red Square.	The green baseball field is adjacent to the playground and the red playground.	There is a large green space on the playground.
		LW-MCR	There are a few trees around a plastic playground.	There is a long path in the field next to the red playground.	There is a baseball field beside the green amusement park around the red track.	There is a gray room next to a baseball field.	There are a few trees around a plastic playground.
Text To Image	There is a tennis court next to the football field and a blue building next to it.	**ACF**					
		GaLR					
		AMFMN					
		LW-MCR					

Figure 7. Visualization of retrieval results.

Based on the visualized results, it is evident that our method is proficient in retrieving accurate results even within intricate RS scenes. Our model's overall visualization underscores its ability to discern detailed and exhaustive correlations between images and textual sentences, thanks to the attention correction and filtering mechanisms. When juxtaposed with the GaLR, AMFMN, and LW-MCR models, our approach demonstrates superior retrieval outcomes.

5. Conclusions

This study introduces a novel cross-modal retrieval model tailored for remote-sensing image-text associations, leveraging attention correction and filtering. The model is structured around four primary components: an image feature extraction module, a text feature extraction module, an attention correction unit, and an attention filtering unit. The image feature extraction module utilizes the VIG as its encoder, this module incorporates a multi-level node feature fusion design. This ensures the model's comprehensive understanding of remote-sensing images across multiple layers. The text feature extraction module employs both BGRU and the GAT as encoders, this module enhances the model's depth of textual comprehension. The attention correction unit addresses mismatches in image-text pairings caused by specific alignments between visual features and textual words. It accomplishes this by adjusting the attention weights. The attention filtering unit enhances the precision of cross-modal retrieval by reducing the influence of unrelated visual zones and text, thereby streamlining the matching process within image-text pairs. Experimental evaluations conducted on the RSICD and RSITMD datasets underscore the excellence of the ACF model. Furthermore, ablation studies affirm the individual effectiveness of each module.

Author Contributions: X.Y.: Conceptualization, Methodology, Software; C.L.: Software; Z.W.: Formal analysis; H.X.: Validation; J.M.: Formal analysis; G.Y.: Data curation, Supervision, Writing original draft, Writing review editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded and supported by Kashgar Science and Technology Plan Project (KS2023024).

Data Availability Statement: Please check the details through this link: https://github.com/Huey-uestc/ACF (accessed on 20 January 2025).

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

References

1. Li, Y.; Ma, J.; Zhang, Y. Image retrieval from remote sensing big data: A survey. *Inf. Fusion* **2021**, *67*, 94–115. [CrossRef]
2. Shyu, C.R.; Klaric, M.; Scott, J.G. GeoIRIS: Geospatial information retrieval and indexing system—Content mining, semantics modeling, and complex queries. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 839–852. [CrossRef]
3. Kandala, H.; Saha, S.; Banerjee, B. Exploring transformer and multilabel classification for remote sensing image captioning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
4. Li, X.; Zhang, X.; Huang, W. Truncation cross entropy loss for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5246–5257. [CrossRef]
5. Zhao, R.; Shi, Z.; Zuo, Z. High-resolution remote sensing image captioning based on structured attention. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
6. Hoxha, G.; Melgani, F.; Demir, B. Toward remote sensing image retrieval under a deep image captioning perspective. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 4462–4475. [CrossRef]
7. Hoxha, G.; Melgani, F.; Demir, B. Retrieving images with generated textual descriptions. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 5812–5815.
8. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image-text matching. In Proceedings of the Computer Vision—ECCV 2018: 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 201–216.
9. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. CAMP: Cross-modal adaptive message passing for textimage retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5764–5773.
10. Li, K.; Zhang, Y. Visual semantic reasoning for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4654–4662.
11. Wang, S.; Wang, R.; Yao, Z. Cross-modal scene graph matching for relationship-aware image-text retrieval. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp.1508–1517.
12. Yuan, Z.; Zhang, W.; Fu, K. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *arXiv* **2022**, arXiv:2204.09868. [CrossRef]
13. Cheng, Q.; Zhuo, Y.; Fu, P. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 4284–4297. [CrossRef]
14. Yuan, Z.; Zhang, W.; Tian, C. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
15. Han, K.; Wang, Y.; Guo, J. Vision GNN: An Image is Worth Graph of Nodes. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 1–16.
16. Peng, Y.; Huang, X.; Zhao, Y. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2372–2385. [CrossRef]
17. Hardoon, D.R.; Szedmak, S.; Shawe, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **2004**, *16*, 2639–2664. [CrossRef] [PubMed]
18. Andrew, G.; Arora, R.; Bilmes, J. Deep canonical correlation analysis. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1247–1255.
19. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]

20. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
21. Cho, K.; Merrienboer, B.V.; Gulcehre, C. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Comput. Sci.* **2014**, *1*, 1–15.
22. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:1411.2539.
23. Devlin, J.; Chang, M.W.; Lee, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
24. Radford, A.; Narasimhan, K.; Salimans, T. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://hayate-lab.com/wp-content/uploads/2023/05/43372bfa750340059ad87ac8e538c53b.pdf (accessed on 25 January 2025).
25. Matsubara, T. Target-oriented deformation of visual-semantic embedding space. *IEICE Trans. Inf. Syst.* **2021**, *104*, 24–33. [CrossRef]
26. Goodfellow, I.; Pouget, J.; Mirza, M. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
27. Wang, B.; Yang, Y.; Xu, X. Adversarial cross-modal retrieval. In Proceedings of the ACM Multimedia Conference, Mountain View, CA, USA, 23–27 October 2017; pp. 154–162.
28. Peng, Y.; Qi, J. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2019**, *15*, 1–24. [CrossRef]
29. Gu, J.; Ha, J.C.; Joty, S.R. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7181–7189.
30. Wen, X.; Han, Z.; Liu, Y.S. CMPD: Using cross memory network with pair discrimination for image-text retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2427–2437. [CrossRef]
31. Nam, H.; Ha, J.W.; Kim, J. Dual attention networks for multimodal reasoning and matching. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 299–307.
32. Wang, T.; Xu, X.; Yang, Y. Matching images and text with multi-modal tensor fusion and re-ranking. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 12–20.
33. Ma, L.; Jiang, W.; Jie, Z. Matching image and sentence with multi-faceted representations. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2250–2261. [CrossRef]
34. Ji, Z.; Wang, H.; Han, J. Saliency-guided attention network for image-sentence matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5754–5763.
35. Ren, S.; He, K.; Girshick, R. Faster rcnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
36. Karpathy, A.; Li, F.F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
37. Liu, C.; Mao, Z.; Liu, A. Focus your attention: A bidirectional focal attention network for image-text matching. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 3–11.
38. Wang, Y.; Yang, H.; Qian, X. Position focused attention network for image-text matching. *arXiv* **2019**, arXiv:1907.09748.
39. Zhang, Q.; Lei, Z.; Zhang, Z. Context-aware attention network for image-text retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3536–3545.
40. Chen, H.; Ding, G.; Liu, X. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12655–12663.
41. Ji, Z.; Chen, K.; Wang, H. Step-wise hierarchical alignment network for image-text matching. *arXiv* **2021**, arXiv:2106.06509.
42. Liu, Y.; Wang, H.; Meng, F. Attend, Correct And Focus: A Bidirectional Correct Attention Network For Image-Text Matching. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 2673–2677.
43. Ge, X.; Chen, F.; Jose, J.M. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 5185–5193.
44. Scarselli, F.; Gori, M.; Tsoi, A.C. The graph neural network model. *IEEE Trans. Neural Netw.* **2008**, *20*, 61–80. [CrossRef]
45. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
46. Shi, B.; Ji, L.; Lu, P. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), Macao, China, 10–16 August 2019; pp. 5182–5189.
47. Wang, H.; Zhang, Y.; Ji, Z. Consensus-aware visual-semantic embedding for image-text matching. In Proceedings of the 2020 European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 18–34.
48. Liu, C.; Mao, Z.; Zhang, T. Graph structured network for image-text matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10921–10930.

49. Nguyen, M.D.; Nguyen, B.T.; Gurrin, C. A deep local and global scene-graph matching for image-text retrieval. *arXiv* **2021**, arXiv:2106.02400.
50. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
51. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding ofhigh 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.
52. Shi, Z.; Zou, Z. Can a machine generate humanlike language descriptions for a remote sensing image? *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3623–3634. [CrossRef]
53. Lu, X.; Wang, B.; Zheng, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [CrossRef]
54. Sumbul, G.; Nayak, S.; Demir, B. SD-RSIC: Summarization-driven deep remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6922–6934. [CrossRef]
55. Hoxha, G.; Melgani, F. A novel SVM-based decoder for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
56. Wang, Q.; Huang, W.; Zhang, X. Word–sentence framework for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 10532–10543. [CrossRef]
57. Wang, B.; Zheng, X.; Qu, B. Retrieval topic recurrent memory network for remote sensing image captioning. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 256–270. [CrossRef]
58. Zhang, Z.; Zhang, W.; Yan, M. Global visual feature and linguistic state guided attention for remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
59. Abdullah, T.; Bazi, Y.; Rahhal, M.M.A.; Mekhalfi, M.L.; Rangarajan, L.; Zuair, M. TextRS: Deep bidirectional triplet network for matching text to remote sensing images. *Remote Sens.* **2021**, *12*, 405. [CrossRef]
60. Lv, Y.; Xiong, W.; Zhang, X. Fusion-based correlation learning model for cross-modal remote sensing image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
61. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv* **2017**, arXiv:1707.05612.
62. Yu, H.; Yao, F.; Lu, W. Text-Image Matching for Cross-Modal Remote Sensing Image Retrieval via Graph Neural Network. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2023**, *16*, 812–824. [CrossRef]
63. Yuan, Z. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–19. [CrossRef]
64. Pan, J.; Ma, Q.; Cong, B. Reducing Semantic Confusion: Scene-aware Aggregation Network for Remote Sensing Cross-modal Retrieval.. In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, Thessaloniki, Greece, 12–15 June 2023; pp. 398–406.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.