Remote Sensing Cross-Modal Text Image Retrieval

MCS

Author

Fatima Ali

Registration Number

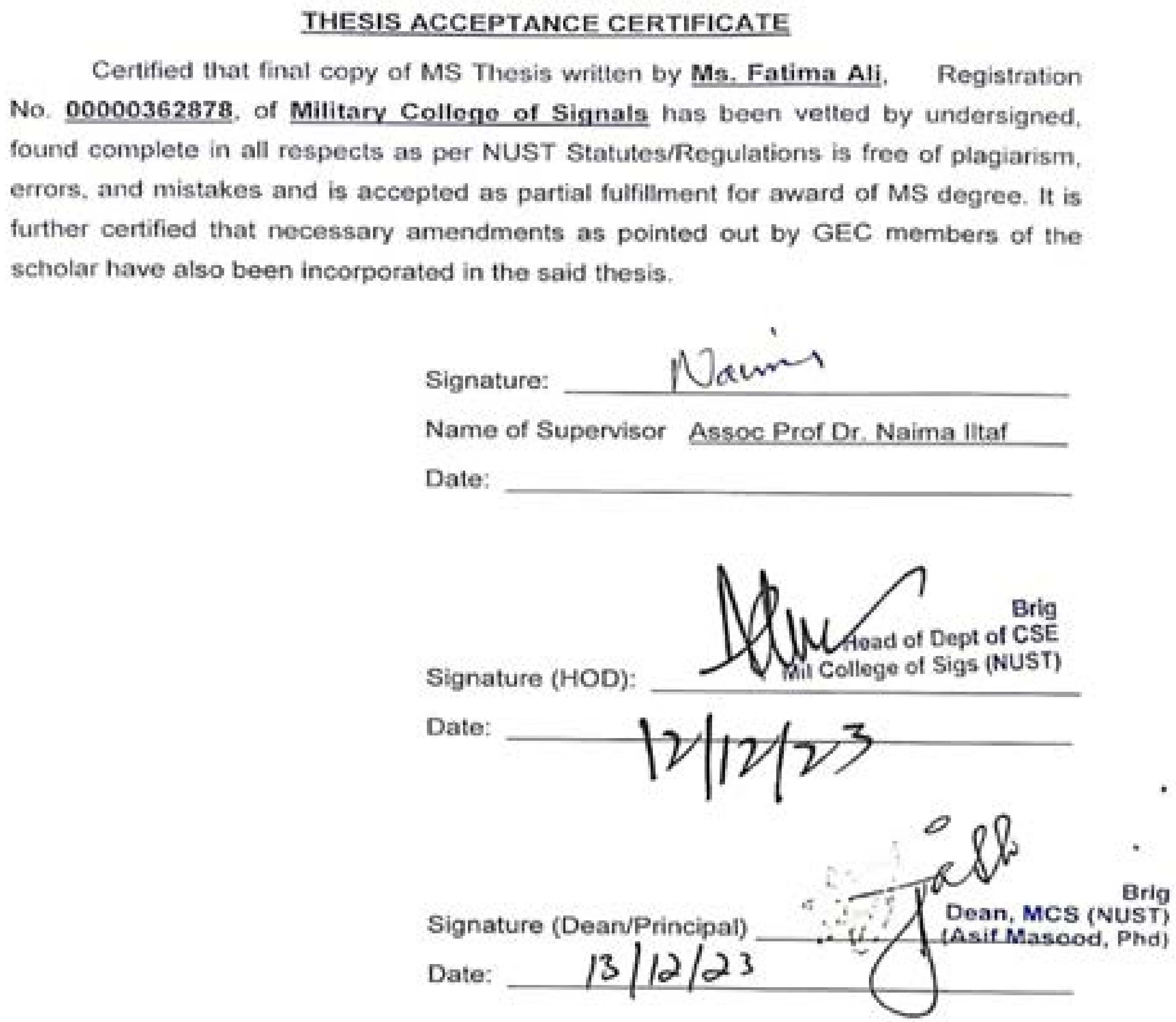
00000362878

Supervisor:

Assoc. Prof. Dr. Naima Iltaf

A thesis submitted to the faculty of Department of Computer Software Engineering, Military College of Signals (MCS), National University of Sciences and Technology (NUST), Rawalpindi in partial fulfillment of the requirements for the degree of MS in Computer Science

November 2023



THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS Thesis written by **Ms. Fatima Ali**, Registration No. **00000362878**, of **Military College of Signals** has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: ____________

Name of Supervisor Assoc Prof Dr. Naima Iltaf

Date: ____________

Brig
Head of Dept of CSE
Mil College of Sigs (NUST)

Signature (HOD): ____________

Date: 12/12/23

Brig
Dean, MCS (NUST)
(Asif Masood, Phd)

Signature (Dean/Principal) ____________

Date: 13/12/23

Declaration

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification either at this institution or elsewhere.

Name of Student: Fatima Ali

Reg. No.: 00000362878

Dedication

“In the name of Allah, the most Beneficent, the most Merciful"

I hereby dedicate this thesis to my parents who paved the way for me to become the first female MS graduate of the family and to my teachers who provided unwavering support at every stage of my academic journey.

Abstract

The advancements in technology in recent times have brought a great influx of data. The remote sensing image datasets have grown in size and numbers. With this abundance of data there comes the problem of retrieving it efficiently for various purposes. Currently, substantial endeavors are underway to articulate novel paradigms, techniques, and technologies improve this process of retrieval of remote sensing data. The heterogeneity of the data and the large semantic gap between text and image modalities makes this an inherently challenging task. Standard retrieval techniques are not effective when it comes to dealing with multi modal remote sensing data. This thesis introduces a purposefully designed framework tailored for the retrieval of targeted images with text query and vice versa. The existing techniques in the context of remote sensing text-image retrieval predominantly emphasize the utilization of high-level or macro features derived from remote sensing (RS) images, consequently resulting in the oversight of pertinent low-level or micro features that convey valuable insights into target relationships and saliency. The proposed model centers on the extraction of image features, subsequently progressing to their cohesive representation dynamic integration. It leverages macro vision features to correct micro vision features, additionally macro vision features are enhanced by micro vision features of the images. Cutting-edge deep learning methodologies are utilized to generate comprehensive representations of both image and text features. After successfully representing the image and text queries, their similarity is calculated and the results are re-ranked. This re-ranking algorithm leverages the k closest neighbors from the retrieval results to conduct a reverse search and, in the process, enhances accuracy through the integration of various bidirectional retrieval components. Predictive evaluation metric Recall is used to compare results for proposed techniques with conventional technique. The proposed solution outperformed on remote sensing datasets: RSICD dataset and RSITMD dataset for the text-image retrieval task.

Acknowledgments

All praises to Allah for His blessings and the strength for completing this thesis.

I express my sincere appreciation to Dr. Naima Iltaf, my supervisor, for her guidance and unwavering support throughout this research endeavor. Her invaluable assistance, characterized by constructive comments and insightful suggestions throughout both the experimental phase and thesis development, significantly contributed to the success of this research. I extend my gratitude to the esteemed members of my committee; Dr. Usman Zia, Dr. Hammad Afzal, and Assistant Professor Mobeena Shehzad for their valuable contributions and guidance throughout this research endeavour.

I express my gratitude to those who have consistently supported my dreams and aspirations, serving as a steadfast source of inspiration. Last, but not the least, I extend my sincere thanks to my father (Dr. Muhammad Ali Shah), and my mother (Mrs. Syeda Rafia Ali) for their unwavering care, love, and support during both challenging and exciting phases of my journey.

Table of Contents

List of Figures

List of Tables

Acronyms

Remote Sensing	RS
Remote Sensing Cross-modal Text-Image Retrieval	RSCTIR
BiLingual Evaluation Understudy	BLEU
Asymmetric Multimodal Feature Matching Network	AMFMN
Graph Convolutional Network	GCN
Remote Sensing Image Retrieval	RSIR
Computer Vision	CV
Natural Language Processing	NLP
Long Short-Term Memory	LSTM
Graph Neural Network	GNN
Siamese Bidirectional Encoder Representations from Transformers	SBERT
Multi-Grade Feature Dynamic Integration	MFDI
Remote Sensing Image Captioning Dataset	RSICD
Remote Sensing Image-Text Match Dataset	RSITMD

Chapter 1

INTRODUCTION

Remote sensing image datasets have increased exponentially with the advancement in satellite or aerial imagery. These images often capture a vast amount of data, and manually analyzing them can be time-consuming and error-prone. By using text-image retrieval techniques, analysts can quickly identify features of interest, such as land cover types, infrastructure, and natural resources. This technique can be useful in a diverse range of applications, including environmental monitoring, urban planning, military surveys, and disaster response.

Remote Sensing Cross-Modal Text-Image Retrieval (RSCTIR) is a type of cross-modal retrieval that enables rapid and adaptable extraction of information from remote sensing (RS) images. RSCTIR is an urgent research hotspot to search for relevant images based on query text or RS images provided. With the advancement in technology and the abundance of RS data the need has been realized to develop algorithms and systems that allow users to search, retrieve, and analyze remote sensing data more effectively, combining the strengths of both textual information and image content for various tasks and applications.

To initiate image retrieval, users are required to input text query to the RSCTIR, depending on which related remote sensing image(s) would be sourced from the database. According to the kind of such query, the RSCTIR method is systematically categorized into distinct classes. User queries may manifest either as images, text, or necessitate concurrent processing of both modalities (text and image). This prompts us to categorize RSCTIR into distinct classes.

The first kind of RSCTIR can be based on image query. In this context, the RSCTIR system requires users to provide an input image and the aim of the RSCTIR is to retrieve most relevant caption(s) from the database. The extraction of image features within this category relies exclusively on visual characteristics, encompassing color, texture, contours, and shapes. Subsequently, similarity measures are employed to quantify the likeness between the query image and the referenced captions. Similarly, the other kind of RSCTIR is based on text query, wherein the user would explain the image in words. The retrieval model must comprehend the text query and retrieve the most similar images matching the user input. In another type, retrieval is possible both ways i.e., input query text and output is image or input query image and output is text.

There are various approaches to text-image retrieval in remote sensing that exist in the literature. One of them is the caption-based method where a caption generator generates RS captions and retrieval results are obtained by calculating Bilingual Evaluation Understudy (BLEU) [1] score using query text and resultant RS captions. One example of this method is remote sensing image captioning model [2] using deep learning and a fully-connected convolutional network. Overfitting was a problem in remote sensing image captioning which was overcome by the truncated cross-entropy loss [3].

In the embedded-based method, feature distance is calculated by projecting the RS image and query text into the same high-dimensional space. The issue of multi-scale scarcity and redundancy of targets were resolved in the approach Asymmetric Multimodal Feature Matching Network (AMFMN) [4]. To resolve the heterogeneity gap by knowledge distillation, a fusion-based correlation learning model [5] was proposed. A lightweight and faster text-image retrieval model [6] was proposed that used knowledge distillation and contrast learning. Many researchers have proposed retrieval frameworks but improvement is needed to achieve better retrieval accuracy.

1.1 Motivation and Problem Statement

In Remote Sensing, one of the major issues is the difficulty in identifying and retrieving important objects in low-resolution remote sensing images. Target redundancy is an issue that needs to be dealt with to reduce computational complexity. In addition to this, large targets are described as more probable than small targets. This research minimizes these issues by leveraging high-level information to correct low-level information and vice versa. It improves the retrieval results by re-ranking the results of bidirectional retrieval.

1.2 Objectives

The major objectives of thesis are: -

- To propose a text-image retrieval system that extracts relevant remote sensing images efficiently using deep learning.

- To compare proposed algorithm with baseline and recently developed state of the art techniques to ensure that our approach outperforms the existing techniques.

1.3 Thesis Contribution

To the best of our understanding, the mechanism introduced in this paper has not been previously employed or explored in existing literature for remote sensing text-image retrieval tasks.

This work makes several noteworthy contributions, outlined as follows:

- We propose a cross-modal text-image retrieval method that includes a sentence transformer network to generate text embeddings, Res2Net for macro feature extraction of images and GCN for learning relationships between image micro features extracted by an object detection model.

- Next, the image high-level features are dynamically fused with low-level features to make full use of low-level detailed information of the image.

- Unlike existing work, we employ a post processing step of re-ranking the retrieval results in the Similarity Score Matrix.

1.4 Thesis Organization

The organizational structure of the thesis is outlined as follows:

- Chapter 1: First chapter is made up of introduction and objectives. It encompasses the contributions introduced within the context of this thesis..

- Chapter 2: In this section, explanation of the background with brief description and review of literature along with existing technique and quantitative measures to evaluate the proposed technique are discussed in this chapter.

- Chapter 3: In this chapter our proposed remote sensing cross-modal text-image retrieval technique is presented.

- Chapter 4: In this chapter, experimentation and analysis of outcomes are given to evaluate the standing of our technique with conventionally developed methods.

- Chapter 5: This chapter gives a precise conclusion and impact of our thesis for cross-modal text-image retrieval techniques in remote sensing. The future work areas are revealed in this chapter.

Chapter 2

Preliminaries

The rise of multi-modal remote sensing data has spurred research in vision-language tasks for remote sensing applications [7]. Advancements in creating more robust image features, developing joint embedding spaces, and addressing the semantic gap in multi-modal data have the potential to revolutionize tasks like remote sensing cross-modal retrieval, visual question answering [8,9] and image captioning [10,11], with wide-ranging applications across various domains.

2.1 Text-Image Retrieval Methods

Text-image retrieval is a difficult task attributable to the heterogeneity of the data and the large semantic gap among text and image modalities. There have been multiple attempts at proposing better retrieval models for natural images which are now benchmark retrieval methods [12,13,14].

In the field of remote sensing, most of the research has focused on matching similar types of data within images, termed as remote sensing image retrieval (RSIR). This is because generating textual descriptions for remote sensing images is more intricate in comparison to natural images. Additionally, remote sensing images often hold complex and ambiguous meanings from a broader perspective [15].

Text image retrieval on remote sensing data is an even complex task due to the low resolution remote sensing images that cover large areas and contain millions of pixels. Following the advent of deep learning, retrieval in the remote sensing domain has also seen progress [16]. The different methods of bidirectional text-image retrieval as shown in figure 2.1 are namely caption-based retrieval, embedding-based retrieval, attention-based retrieval, multi-modal fusion based retrieval, graph-based retrieval, hashing-based retrieval with multiple other techniques employed like contrastive learning etc.

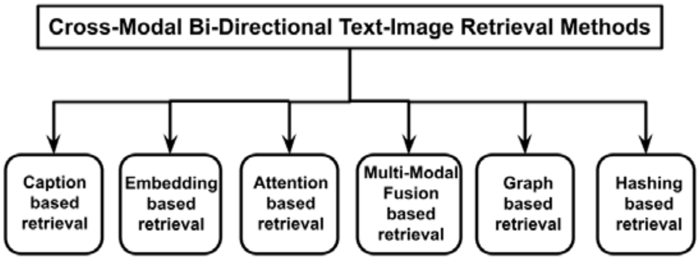

Cross-Modal Bi-Directional Text-Image Retrieval Methods
Caption based retrieval
Embedding based retrieval
Attention based retrieval
Multi-Modal Fusion based retrieval
Graph based retrieval
Hashing based retrieval

Figure 2.1: Types of cross-modal bi-directional text-image retrieval methods

2.2 Caption-Based Text-Image Retrieval

Image retrieval based on text descriptions is challenging due to the fundamental difference between visual and textual data. In recent computer vision research, significant work has been dedicated to addressing this challenge and narrowing the gap between these two modalities. Two step caption based retrieval methods are the most widely accepted and first used retrieval methods for remote sensing datasets. The captions are first generated by caption generators and then query text and generated captions are compared. Image captioning collaborates Computer Vision (CV) and Natural Language Processing (NLP) to describe images. Attention-based Encoder-Decoder model, which utilizes an attention mechanism and LSTM for word generation, is used for image captioning [17]. A mask-guided transformer network featuring a topic token to improve the accuracy and diversity of remote sensing image captioning introduces multi-head attention for feature extraction and to capture inter-object relationships, and adds a topic token in the transformer encoder to highlight high-level semantic information [18].

2.3 Embedding-Based Text-Image Retrieval

Embedding based methods for retrieval are better than caption based methods since image and textual descriptions are mapped into a common embedding space, where their similarity measurement can be done based on their Euclidean distance or cosine similarity. Unlike caption based methods, Fusion-based Correlation Learning Model (FCLM) improves the performance of image-text matching tasks by learning a common feature space and minimizing the disparity between distinct and amalgamated feature representations [19].

2.4 Attention-Based Text-Image Retrieval

The attention based methods of retrieval selectively focus on parts of an image or textual description that are most relevant to a given query. Cross-Attention Based Image Retrieval (CABIR) utilizes the cross-attention mechanism that facilitates cross-modal information interaction, directing the network through textual semantics to assign weights and selectively filter out extraneous features in image regions. This process mitigates the impact of irrelevant scene semantics on retrieval, enhancing the precision of the system [20]. FAAMI model [21] presents an approach for achieving detailed semantic alignment by effectively combining information at multiple scales. It utilizes a cross-attention network with a limited depth to capture the nuanced semantic connections between image regions of various scales and the corresponding textual descriptions.

2.5 Multi-Modal Fusion Based Text-Image Retrieval

Both textual and visual features are encoded into a joint representation space and combined to improve the similarity measure used to retrieve relevant images or textual descriptions. A single-stage solution for image retrieval by amalgamating both local and global information to create concise representations of images, the Deep Orthogonal Local and Global (DOLG) architecture, which does local representation extraction using multi-atrous convolutions and self-attention. Orthogonal components are subsequently derived from the local data and concatenated with the global representation to generate the final image representation [22]. MTGFE model [23] consists of a unimodal encoder comprising of ViT [24] and BERT [25] transformer models for images and texts respectively and a multimodal fusion encoder composed of six Transformer layers, leveraging detailed features like image segments and text tokens.

2.6 Graph-Based Text-Image Retrieval

Knowledge-aware Cross-modal Retrieval (KCR) method for remote sensing cross-modal text-image retrieval, addresses the challenge of information asymmetry between texts and images by extracting pertinent information from an external knowledge graph [26]. Graph neural networks are also being applied in the field of remote sensing text image retrieval [27]. Yao et al. introduced a macro hypergraph network and a micro hypergraph network to characterize the inter-object relationships in remote sensing images at varying scales [28].

2.7 Hashing-Based Text-Image Retrieval

A Deep Unsupervised cross-modal Contrastive Hashing model comprises of a feature extraction block and a hashing block, incorporating contrastive objectives, adversarial objectives, and binarization objectives to produce cross-modal binary hash codes [29]. An advanced convolutional neural network framework for deep hashing, rooted in supervised contrastive learning, developed to enhance the discrimination of image features. The framework encompasses a module dedicated to fusing global and local features, a spatial attention mechanism, and a supervised contrastive learning approach [30].

2.8 General Text-Image Retrieval Methods

Scale-Semantic Joint Decoupling Network (SSJDN) for text-image retrieval in remote sensing, combines the ideas of "scale decoupling" and "semantic decoupling" to boosts representation capability [31]. End-to-End Framework Based on Vision-Language Fusion (EnVLF) consists of two separate uni-modal encoders for image and text, and a multi-modal encoder for fusion. The framework utilizes a vision transformer module is employed for capturing image local features, in lieu of utilizing a pre-trained object detection model, bridging the disparity in training performance between the object detector and retrieval model. The multi-modal encoder improves the ranking performance after retrieval processing [32]. MCRN achieves competitive outcomes through implementation of multitask learning for semantic alignment [33]. Parameter-Efficient Transfer Learning (PETL) model [34], transfers knowledge in the domain of visual and language understanding transitioning from the natural domain to the remote sensing (RS) domain, particularly for the task of text-image retrieval. RemoteCLIP [35] undertakes a process of continuous pretraining, where the CLIP model is tailored to be domain-specific, focusing on the remote sensing domain. Hypersphere-Based Visual Semantic Alignment framework [36], integrated with Curriculum Learning focuses on aligning remote sensing image-text pairs in a progressive manner, advancing to more challenging cases. Feature uniformity strategy minimizes instances of feature mismatches. Key-Entity Attention mechanism mitigates information imbalances among different modalities.

2.9 Text Feature Extraction

Sequential modeling networks like RNN and LSTM brought a revolution in the domain of Natural Language Processing. Word embedding generation made a significant impact on text feature extraction accuracy. BERT [25] is build upon transformer encoders able to understand the text semantics. Sentence transformer models further improved the text embedding generation. The IEFT leverages the transformer-based models and self-attention mechanism to learn images features and textual representations simultaneously, ensuring semantic consistency [37]. A composite embedding model, comprising language and vision transformer encoders, is designed to align the visual representations of remote sensing (RS) images with their corresponding textual descriptions [38]. TBFDR addresses this issue by segregating features into modal-invariant and modal-heterogeneous components, and then reconstructing the representations to preserve the information [39].

2.10 Vision Feature Extraction

The Deep Learning domain revolutionized with the arrival of the AlexNet algorithm [40] in 2012. Since then many deep convolutional network networks have improved the image feature extraction like VGGNet [41] and ResNet [42]. The introduction of vision transformers has made a significant impact in computer vision and image processing [24]. These transformer networks originally made for natural language processing tasks have made their way into vision applications too.

Chapter 3

RS Cross-Modal Text-Image Retrieval

3.1 Proposed Methodology

This chapter describes the proposed algorithm to achieve text-image retrieval with accurate predictions and high performance. In literature high-level vision feature based techniques are used but the drawback of macro vision feature based is that it does not acknowledge the object level relationships of targets in the visual data. Contrary to present literature our approach utilizes high-level information to supplement low-level information and low-level information to supplement high-level information. We apply state-of-the-art text and image feature extraction networks to our research.

The proposed model integrates denoised micro vision features extracted by graphical convolution and macro vision features extracted via residual convolution framework in Multi-Grade Dynamic Feature Integration (MDFI) block rectifying macro vision features with micro vision features and vice versa. Siamese and triplet network architectures yield sentence embeddings that capture semantically meaningful information. Sentence and image similarity are calculated and the retrieval results are re-ranked as a final step.

Looking at the literature review of RSCTIR above, RSCTIR is a research hotspot in the field of retrieval systems. The improvement of bi-directional RSCTIR systems necessitates continued efforts on the behalf of the researchers. In bi-directional cross-modal text-image retrieval systems it all comes down to creating meaningful relationships between text descriptions and image features, as shown in Figure 3.1.

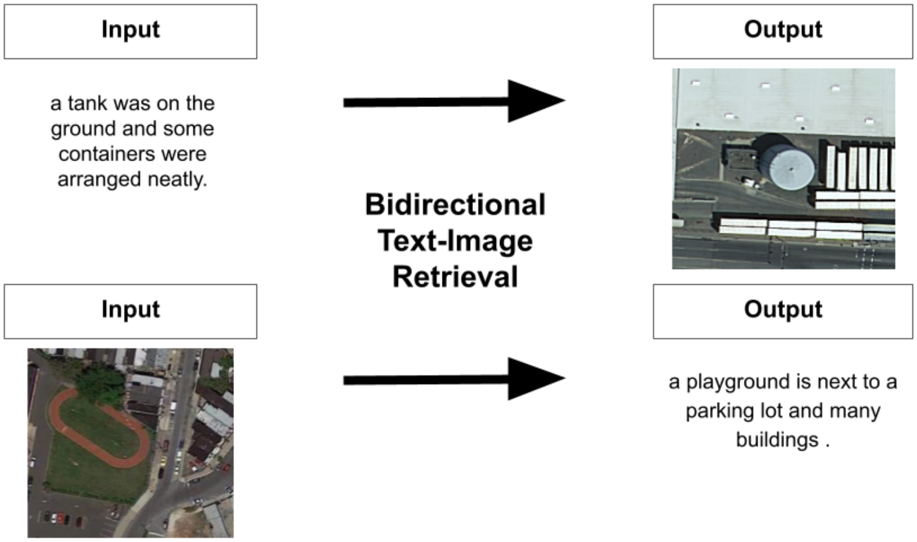

Input
a tank was on the ground and some containers were arranged neatly.
Output
Bidirectional Text-Image Retrieval
Input
Output
a playground is next to a parking lot and many buildings .

Figure 3.1: Typical bidirectional text-image retrieval system

This section elucidates the conceptual alignment of the proposed framework, wherein Res2Net serves as the tool for extracting image macro features, Graph Convolution Network (GCN) for extracting image micro features, while cutting-edge natural language processing techniques, specifically SBERT, are employed for the formulation of text embeddings.

Figure 3.2 offers a thorough illustration, providing both a comprehensive overview and a justified rationale for the incorporation of deep learning techniques within the proposed framework, aligning with the objectives of achieving dual modality goals. The main objective of the proposed framework is to generate a dynamically integrated visual feature containing both the micro and macro level information of the image and semantically significant text descriptions then do similarity calculation between image query and text description in the dataset and if query is text then with image in database.

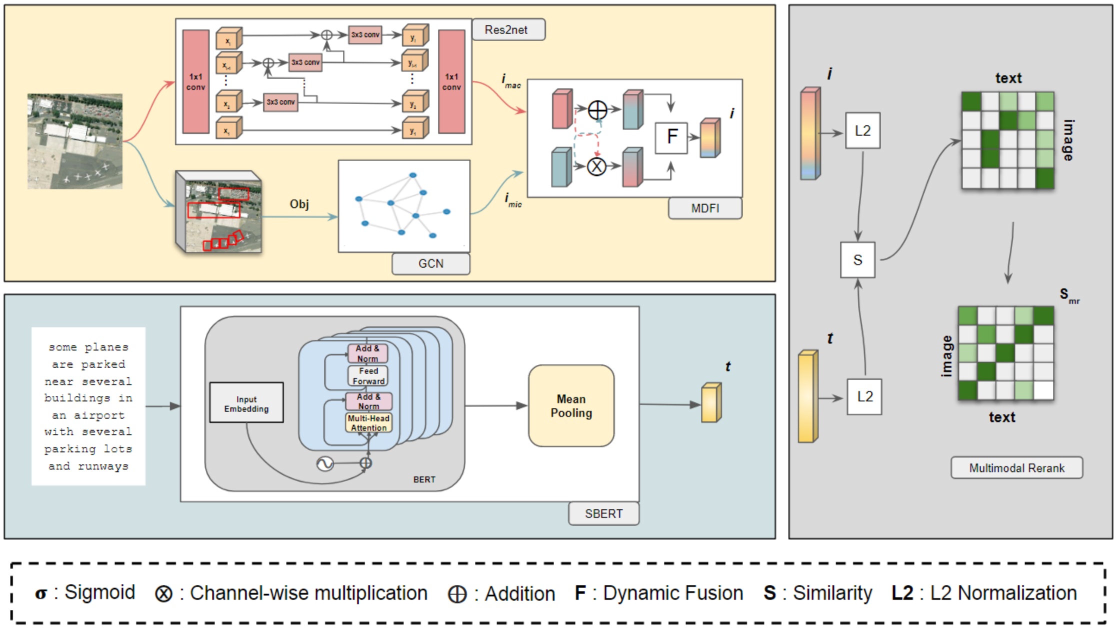


Figure 3.2: Overview of suggested framework

It comprises of three major sub-sections: (1) image features extraction (2) generation of text embedding, and (3) similarity of text and image features. In this research, we employ Res2Net to enhance the extraction of image macro features, graph convolution for learning relationship of image micro features. Subsequently, we utilize transformer-based representations to generate embeddings for text queries. Finally, we introduce a dynamic integration technique to fuse the image features and obtain a unified representation before calculating similarity with the text features. Table 3.1 records all the notations mentioned in the thesis.

Table 3.1: Notations Table

Symbols	Description
D_S	Extent of similarity
i_{mac}	Macro visual features
i_{mic}	Micro visual features
A	CNN model for macro visual feature extraction
B	GCN for micro vision feature extraction
O	Set of objects or micro features detected in images
$d_A(coord_1, coord_2)$	Distance between image co-ordinates
Σ	Sigmoid Activation Function
A	Adjacency Matrix

D	Degree Matrix of A
T	Text Features
Γ	SBERT model for text embedding generation
i_{mac}^{SA}	Self-attended macro vision features
i_{mic}^{SA}	Self-attended micro vision features
SA	Self attention block
GA	Guided attention block
i_{mac}^{SGA}	Self-guided vision macro feature
i_{mic}^{SGA}	Self-guided vision micro feature
i_{mac}^{att}	Attended macro vision feature
i_{mic}^{att}	Attended micro vision feature
i_{comb}	Combination of Attended vision feature
ω_1, ω_2	Learnable Dynamic Weights
I	Vision Features after MFDI block
L(T,I)	Triplet Loss

3.2 Vision Features Generation

The proposed method extracts the query image macro features via Res2Net which is used as backbone in the framework. Res2Net [43] is a version of convolutional neural network model presented in 2021 for visual features extraction. Unlike the usual layered layout of convolutional neural networks, it has revolutionized deep learning by establishing a hierarchical network of residual-like connections encapsulated within a singular residual block. The residual-like connections within residual blocks increase receptive fields of the neural network layers. Res2Net shines on segmentation types of tasks that capture different levels of scale within the image at a more fine-grained level. The Res2Net introduces a novel dimension, referred to as "scale," which proves to be a crucial and more impactful factor, complementing the established dimensions of depth, width, and cardinality. Because of its enhanced multi-scale capability, Res2Net generates activation maps that exhibit a tendency to encompass the entirety of objects, in contrast to ResNet, where activation maps typically focus on specific parts or regions of objects. Validation losses plummet when data is fed to Res2Net after substantial data augmentation. The relationship between micro targets in images are modeled using Graph Convolutional Network (GCN) extracted using an object detector.

3.3 Text Features Generation

The proposed method operates on dual modalities, comprising both images and textual queries. Consequently, alongside the extraction of image features, the model must also process the input text query. The textual query encapsulates the conceptual elucidation of the user input image. Simultaneously with image feature extraction, we utilized transformation-based techniques to extract text features for inputting image queries. SBERT, as described in [44], is implemented as SBERT client on a host machine transforming encoding of sentences of varying lengths into a fixed-length feature vector. SBERT is founded upon the transformer's encoder architecture, enabling it to effectively capture the linguistic semantics and contextual meanings. SBERT is trained on SNLI [45] and Multi-Genre NLI [46] corpus for around 570K and 430K sentence-pairs with entailment, contradiction and neutral labels. SBERT incorporates a pooling operation into the output of BERT [25] or RoBERTa [47] in order to generate a sentence embedding of a consistent, fixed size. Siamese and triplet networks [48], introduced in 2015, are employed in SBERT to refine BERT or RoBERTa output. SBERT is well known for being applied in scenarios characterized by computationally intensive tasks to be computed using BERT.

3.4 Dynamic Fusion of Multi-level Features

Once enriched features at each high and low level of the individual image have been obtained, the next step involves combining these features to obtain a meaningful combined learned vision feature representation. Various techniques exist for creating a joint representation of multi-level vision features, such as concatenation, element-wise addition, or multiplication, and attention mechanisms. Our approach makes use of all these approaches for dynamic integration of the vision features. The proposed Multi-Grade Dynamic Feature Integration (MDFI) block can be broken down into two key stages: feature reconstruction and dynamic integration. In the first stage, MDFI employs SA (Self Attention) and GA (Guided Attention) modules to transfigure the micro and macro vision features. The SA mechanism assesses internal similarity of the macro and micro vision features. Subsequently, in the second stage, it leverages high-level information to enhance low-level information and utilizes the latter to refine and improve the former by learnable weights obtained by linear transformation of dynamic addition of both level features.

3.5 Multimodal Rerank of Bi-directional Retrieval Results

In the context of retrieval tasks, a post-processing step has been introduced to enhance retrieval results. This method optimizes the results via amalgamating information regarding multiple variables through an additional sorting step. In a typical retrieval task, the calculation of similarity among M textual queries and N visual queries results in a similarity score matrix. The top-k outcomes or leading k images retrieved for a given text query within the complete dataset is the goal. This retrieval approach remains consistent when employing reverse search for text retrieval using images. Nonetheless, while the accuracy of this retrieval model is indisputable, it overlooks the inherent connection among bidirectional retrieval that plays a pivotal role in enhancing retrieval accuracy. Specifically, the bidirectional retrieval process should ensure that when a text and an image correspond, they are mutually retrievable. Cross-modal rerank algorithm employs the top-k positions to conduct a reverse search, the ultimate outcome of the retrieval process is determined based on the ranking placement within the reverse retrieval outcomes. While the algorithm for Reranking in Cross-Modal Retrieval offers some improvement in retrieval performance, it does not completely exploit the knowledge contained within the similarity score matrix, leaving room for further enhancement which is proposed here to employ candidates for reverse search and enhance retrieval outcomes by taking into account various ranking factors.

3.6 Mathematical Illustration

Following the detailed description of feature extraction for each modality, this section will provide a mathematical exposition of the suggested framework. In the cross-modal text image retrieval process, similarity is calculated between query text $\boldsymbol{t_q}$ and database images $\boldsymbol{I= [\ i_1,\ i_2,\ i_3,...,\ i_n\ ].}$ Consequently, the most similar image from database $\boldsymbol{i_s}$ is retrieved. As is the case with most retrieval problems, our objective is to have the model retrieve desired images that are relevant to the text queries given as input and vice versa. This process can be depicted mathematically as:

$$D_S = \cos(i, t) \qquad (1)$$

Here Degree of Similarity ($\boldsymbol{D_S}$) is represented as cosine similarity of image ***i*** and text ***t***.

In the context of macro vision feature generation, our proposed model relies on Res2Net as its foundational backbone. The input query image, initially sized at *256x256*, undergoes several pre-processing transformations as an integral part of our data augmentation technique. Res2Net is a residual network design consisting of hierarchical residual connections within residual blocks. In Res2Net the bottleneck block of the common ResNet architecture standard *3x3* filter is replaced with *3x3* convolution of ***n*** number of feature subsets split after *1x1* convolution. The equation below expresses how macro visual features for the ***ith*** image are generated:

$$i_{mac} = \alpha(i_i) \qquad (2)$$

Here $\boldsymbol{i_{mac}}$ shows the macro vision features of the input image query and Res2Net block is shown using $\boldsymbol{\alpha}$. To enhance reader comprehension, we have included an architectural diagram of the backbone used for macro vision feature extraction in our proposed model; this is shown in figure 3.3

.

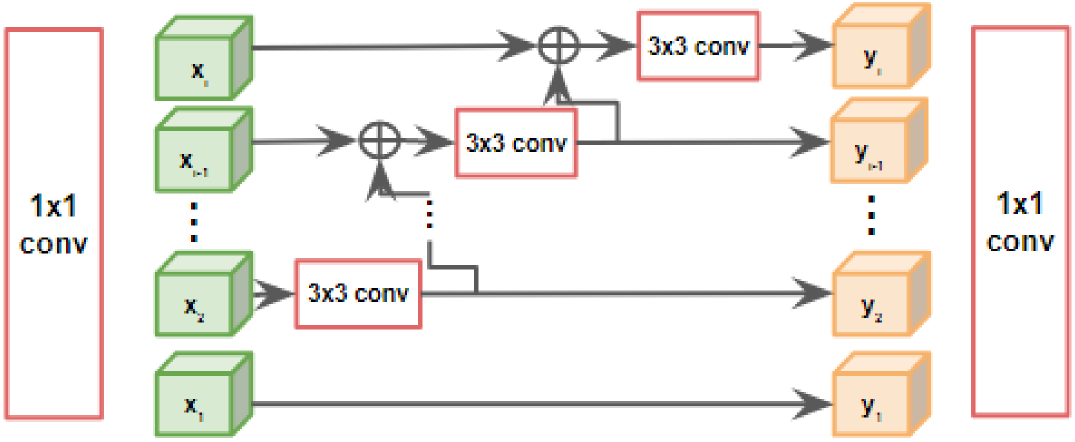


Figure 3.3: Res2Net Architecture Diagram

The relationship between the micro vision features in the image is learnt via graph convolution of the targets extracted using an object detection model trained using the

DOTA [49] dataset. The objects are represented as vertices of the graph and the relationship among them is represented by their corresponding edges. The equation below shows how micro vision features ***I_{mic}*** for the ***ith*** image are generated:

$$O = \mathrm{Detect}(i_i) \tag{3}$$

Here ***O = { Obj_1, Obj_2,...,Obj_n}*** is denoted as the set of **n** number of objects detected by the object detector in the ***ith*** image.

$$i_{mic} = \beta(O) \tag{4}$$

For the resultant ***X*** at the ***(l+1)th*** layer of the Graph Convolutional Network (GCN) denoted by ***β***, we establish the following definition:

$$X^{(l+1)} = \sigma(D^{-\frac{1}{2}}(A+I)D^{-\frac{1}{2}}X^{(l)}W^{(l)} \tag{5}$$

We define ***$W^{(l)}$*** as the weight matrix that can be learned for the ***lth*** layer, and ***σ*** represents the Sigmoid activation function. The input features ***$X^{(l)}$*** for the GCN are derived from various attributes of the target within the remote sensing image, including its position, category, probability, and area size. The normalization of matrix ***A*** is achieved through the utilization of its Degree Matrix, denoted as ***D***. The construction of the adjacency matrix ***A*** relies on the inter-target distances observed in the remote sensing image, while ***I*** denotes the identity matrix.

The distance ***d_A*** between two objects in an image with coordinates ***$coord_1$*** and ***$coord_2$*** should be utilized to strengthen relationship between close targets and is represented by the following equation:

$$d_A(coord_1, coord_2) = \mathrm{e}^{-\| coord_1 - coord_2 \|_2^2}(1 - \| coord_1 - coord_2 \|_2^2) \tag{6}$$

In addition to extracting visual features from the query image, our proposed model simultaneously incorporates the textual input query. To achieve our ultimate objective, it is imperative to extract comprehensive features from the query text. To fulfill this role, we employ a transformer based model Siamese Bidirectional Encoder Representation from Transformer (SBERT) [44], as it excels in capturing linguistic nuances and contextual information. SBERT eliminates the need for local implementation and saves computational resources. In mathematical terms, the formulation of text query features can be represented as follows:

$$t = \gamma(t_i) \tag{7}$$

Here ***t*** represents a text query feature having dimension dependent on the SBERT model. In the case of "all-MiniLM-L6-v2" model the embedding dimension is *384*. The schematic representation of our linguistic model is depicted in the figure 3.4.

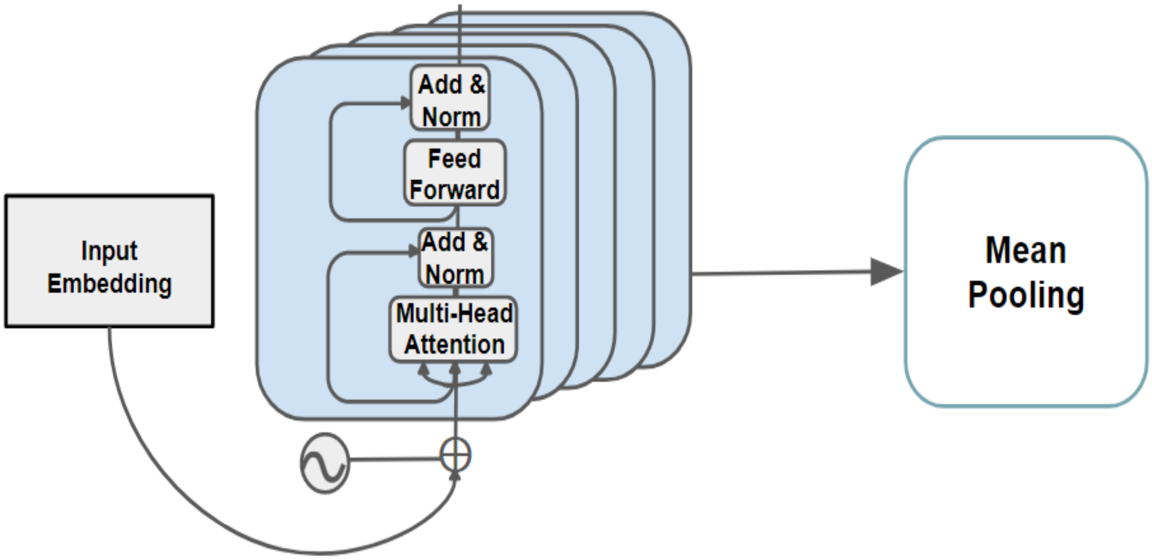


Figure 3.4: SBERT Architecture Diagram

Following the extraction of features from both individual modalities, the next step is to combine the micro and macro vision features. The micro vision features influence macro vision features, reflecting the details in the visual domain. Various techniques in the literature, such as element-wise multiplication, concatenation, and dynamic addition, are commonly used to merge vision features. However, our approach explores an attention based feature integration mechanism for this fusion. We employ a two step process involving attention mechanisms and dynamic integration of multi-level vision features. Self attention is applied on micro and macro features individually by internal similarity calculation and then guided attention is used to learn information by similarity calculation of the multi-level vision features. Then the transformed vision features are superimposed and learnable weights are obtained by linear transformation. The initial step involves passing the vision features through ***SA*** block, as defined by the equation:

$$i_{mac}^{SA} = \mathrm{SA}(i_{mac}), \qquad i_{mic}^{SA} = \mathrm{SA}(i_{mic}), \qquad (8)$$

Now the self-attended vision features are passed through ***GA*** block resulting in self-guided vision features shown as:

$$i_{mac}^{SGA} = \mathrm{GA}\left(i_{mac}^{SA}, i_{mic}^{SA}\right), \qquad i_{mic}^{SGA} = \mathrm{GA}\left(i_{mic}^{SA}, i_{mac}^{SA}\right) \tag{9}$$

Approaching the end of the first step of the dynamic feature integration block information interaction between micro and macro features is carried out as:

$$i_{mac}^{att} = i_{mac}^{SGA} \otimes i_{mic}^{SGA}, \qquad i_{mic}^{att} = i_{mic}^{SGA} \oplus i_{mac}^{SGA} \tag{10}$$

While multi-level features exhibit improved performance following interaction, there remains a requirement to obtain a joint representation of image features. To achieve this, we initially combine these two level features to create a unified vision information representation, denoted as $\boldsymbol{i_{comb}}$. Subsequently, we derive learnable dynamic weights, denoted as $\boldsymbol{\omega}$ through a linear transformation applied to $\boldsymbol{i_{comb}}$. These dynamic weights are then employed to produce the fused feature representation, denoted as $\boldsymbol{i}$.

$$i_{comb} = i_{mac}^{att} \oplus i_{mic}^{att} \tag{11}$$

$$\omega_1, \omega_2 = \mathrm{Softmax}(\sigma(i_{comb} W_x) W_y) \tag{12}$$

$$i = (\omega_1 \otimes i_{mac}^{att}) \oplus (\omega_2 \otimes i_{mic}^{att}) \tag{13}$$

In the given context, where $\boldsymbol{W_x}$ and $\boldsymbol{W_y}$ represent weight matrices and ***Softmax*** activation function is applied to activate the resultant fused vision feature as a result of the MDFI module proposed in this study vision features masked by information at micro level and macro level information is guided by micro level features of the images.

For the post-processing re-ranking method, i2t and t2i ranking information is fused in the primary similarity score matrix in addition to a crucial component to improve reranking performance. In the i2t case, $\boldsymbol{R_{i2t}(i, k)}$ is established as a query within the primary similarity score matrix $\boldsymbol{S_p}$, where $\boldsymbol{i}$ and $\boldsymbol{k}$ refer to the selection of the top $\boldsymbol{k}$ closest neighbors.

$$t_1, \ldots, t_m, \ldots, t_k = R_{i2t}(i, k) \tag{14}$$

Here, $\boldsymbol{t_m}$ represents the text that exhibits the ***mth*** level similarity to the query image. Here ranking position ***P*** is initially acquired for each candidate text, where ***P*** falls within the range of (***0, 1, ..., M)***. To effectively harness this ranking information, we introduce the i2t component, denoted as $\boldsymbol{c_{i2t}}$, to make full use of this valuable data.

$$c_{i2t} = e^{-\xi(\mathrm{P}+1)} \tag{15}$$

The parameter $\boldsymbol{\xi}$ represents the ranking coefficient. The primary objective concerning this operation is standardizing the ranking statistics in the image-to-text (i2t) direction, where the highest-ranked text candidates receive a top i2t component. Subsequently, utilize the candidates identified in the preceding part for reverse searching. When conducting a query using a specific text, denoted as **t_m**, we define **$R_{t2i}(t_q, k)$** as the querying process within the similarity score matrix **S_P** using **t_m**.

$$i_1, \dots, i_n, \dots, i_k \;=\; R_{t2i}(\mathrm{t_m}, k) \qquad (16)$$

Where **k** denotes the first **k** closest neighbors, and **i_n** is the closest neighbor image. For each retrieved image, the **k** closest neighbor text is identified, and consequently, the **k** closest neighbor image through reverse retrieval approach is determined. Simultaneously, the position **L_q** of the query image **I_q** within the **k** closest neighbor images is established. Accordingly, we formally define the t2i component as follows:

$$c_{t2i} \;= \begin{cases} e^{-\xi(\mathrm{P}+1)}, & i \; in \; R_{t2i}(\mathrm{t_m}, k) \\ \qquad 1, & other \end{cases} \qquad (17)$$

The term $\mathrm{c_{t2i}}$ denotes the secondary similarity confirmation that comes into play during the process of reverse retrieval. Its primary function is to serve as a corrective factor for **c_{i2t}**. Additionally, we introduce the concept of "significance components" denoted as **c_S** to quantitatively assess the level of confirmation regarding the predicted similarity by our model. In the context of reverse retrieval, when evaluating a candidate text **t_q**, a higher degree of certainty is associated with a greater proportion of similarity between this text and the images compared to the overall similarity with all images. To operationalize this notion, we compute the relevant ratio and interpret it as a confidence-indicating component.

$$c_P \;= \frac{\cos(\mathrm{t_m}, i_n)}{\sum_{n=0}^{N} \cos(\mathrm{t_m}, i_n)} \qquad (18)$$

A confidence score **c_P** is computed to gauge the model's confidence in the primary similarity. This **c_P** is then employed as a weighting factor in the computation of the final similarity score. Ultimately, we combine and adjust the contributions of the three reranking components to derive the similarity score denoted as multi-modal rerank **S_{mr}**.

$$S_{mr} \;= \mathrm{c_{i2t}} + \omega_{c_1}\mathrm{c_{t2i}} + \omega_{c_2}\mathrm{c_P} \qquad (19)$$

S_{mr} is thus the result of the secondary re-ranking of the similarity score matrix.

3.7 Learning Methodology

The key objective of our proposed method is to effectively retrieve highly similar images from a stored image database. This retrieval process is performed in response to a query that integrates both visual and text features. To improve the model's performance, we update its weights based on a loss value. Specifically, we employ the triplet loss method as in [50] to quantify the similarity in the characteristics between the textual features of the query and the visual features of the target image. or query image and target caption within the dataset. As implied by its name, computing the triplet loss value, which quantifies similarity, necessitates three key components. These components are: the positives, the negatives and the anchor. The negatives are mismatched samples from the dataset and the positives are the most similar ones to the query in the dataset. The anchor is the query and the goal is to decrease its distance from positive samples and increase from negative samples. In summary, the triplet loss in our proposed model, computed during the training phase, can be expressed as shown:

$$L(T,I) = \sum_{\mathrm{T_n}} [\varepsilon + \mathrm{S}(\mathrm{I},\mathrm{T_n}) - \mathrm{S}(\mathrm{I},\mathrm{T})]_+ + \sum_{\mathrm{I_n}} [\varepsilon + \mathrm{S}(\mathrm{I_n},\mathrm{T}) - \mathrm{S}(\mathrm{I},\mathrm{T})]_+ \quad (20)$$

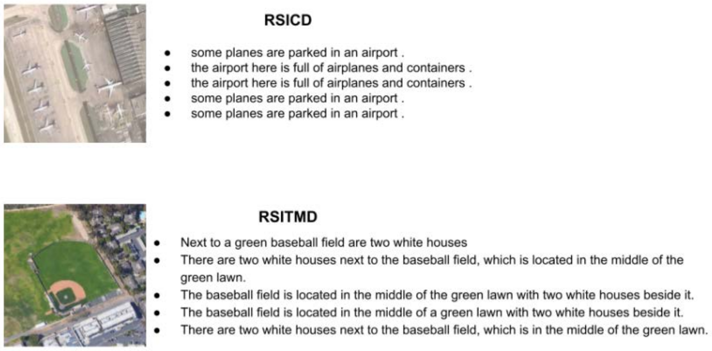
Here, ***L(T,I)*** is the triplet loss of paired sample pairs of text ***T*** and image **I**. ***ε*** is the minimum margin which is the minimum separation that is ideal between the positive and negative samples in the dataset. ***[x]***$_+$***= max(x,0).*** The negative sample pairs are $\boldsymbol{I_n}$ and $\boldsymbol{T_n}$ and the aim is to make anchor sample points farther from the negative samples and closer to the positive samples. The initial summation encompasses all negative text $\boldsymbol{T_n}$ with respect to a particular image ***I***, while the following summation accounts for all negative images $\boldsymbol{I_n}$ concerning a particular text ***T***. If the joint embedding space positions ***I*** and ***T*** in closer proximity to each other than any negative pairs by a margin of ***ε***, the hinge loss becomes zero. In practical implementation in consideration of computational efficiency, the approach involves not performing summation across all negative samples; the focus is typically placed on the challenging negative instances within a mini-batch during stochastic gradient descent.

Chapter 4

Experimental Results and Analysis

4.1 Implementation Details

Experimentations are carried out using a NVIDIA GPU. To keep the image size uniform across both datasets we uniformly scale them to *256x256* pixels. Some data augmentation steps like rotation and flip are also performed on images to improve the variation of images consequently the model robustness. The training epochs are set to *70*. The optimizer used is Adam optimizer with triplet loss. The critical threshold of similarity is *0.8* and the relationship boost factor is *1.15* for adjacency matrix optimization. The text vector representation dimension is *384* and common text and image embedding space is *512*. The training batch size is configured to *100* and validation batch size is configured to *70*. The initial learning rate is configured to *0.0002* which drops by *0.7* after every *20* epochs and margin for triplet loss is *0.2*.



RSICD

- some planes are parked in an airport .
- the airport here is full of airplanes and containers .
- the airport here is full of airplanes and containers .
- some planes are parked in an airport .
- some planes are parked in an airport .

RSITMD

- Next to a green baseball field are two white houses
- There are two white houses next to the baseball field, which is located in the middle of the green lawn.
- The baseball field is located in the middle of the green lawn with two white houses beside it.
- The baseball field is located in the middle of a green lawn with two white houses beside it.
- There are two white houses next to the baseball field, which is in the middle of the green lawn.

Figure 4.1: Samples from RSICD and RSITMD datasets and their five corresponding captions

4.2 Dataset Description

Table 4.1 illustrates the numerical analytics of the two remote sensing datasets employed for both the training and testing phases in the evaluation of the proposed framework. The Remote Sensing Image Captioning Dataset is composed of *10921* remote sensing images and 5 captions per image. The RSICD dataset comprises imagery sourced from Google Earth, Baidu Map, Tianditu and MapABC. The image dimension size is *224x224* pixels. The Remote Sensing Image-Text Match Dataset contributed by Yuan et al. is composed of *4743* remote sensing images and *5* captions per image. The RSITMD dataset is comprises imagery sourced from Google Earth and the RSICD dataset. The image size is *256x256* pixels. The RSITMD dataset contains one to five keywords for each image which can provide a finer grained dataset for retrieval. Sample images and their corresponding captions of RSITMD and RSICD datasets are shown in figure 4.1.

Table 4.1: Details of Datasets

Datasets	**RSITMD Dataset**	**RSICD Dataset**
Composition	Images, Sentences, Keywords	Images, Sentences
Image Sources	RSICD dataset, Google Earth	Google Earth, Baidu Map, MapABC, Tianditu
Total no. of Images	4743	10921
Total no. of Captions	23,715	54,605
No. of captions per image	5	5
Average Length of Texts	10.25	10.55
Maximum Length of Texts	34	34
Image pixels	256x256	224x224

4.3 Quantitative Analysis

Consistent with prior academic research endeavors in the literature, the proposed model underwent evaluation with a focus on the recall metric concerning the retrieval of top-k images. It is represented by ***R@K***. According to the text image retrieval method, recall at k denotes the percentage of images effectively retrieved in response to a dual-modality search. Recall is defined as the ratio of correctly identified images to the total number of images that were successfully retrieved during the entire retrieval process.

$$R@k = \frac{(\text{number of retrieved items among the top k that are relevant})}{(\text{total number of relevant items})} \quad (21)$$

Mean recall denoted by ***mR*** represents the mean of multiple recall at k (***R@k***) values, providing a more intuitive reflection of the overall model performance.

$$mR = (\frac{1}{k}) * \sum (R@i), i = 1 \; to \; k \quad (22)$$

Here, ***k*** represents the number of values of ***k*** at which Recall is calculated, typically ***k=1, 5,*** and ***10***. ***R@i*** represents the Recall value obtained at the ***ith*** value of ***k***.

4.4 Compared Methods

Following the selection of the evaluation metric, we performed a quantitative comparison between the proposed framework and the present baseline within the domain of RSCTIR. The operational approach of the foundational frameworks is elucidated here to the best of our comprehension:

- VSE++ [12]: VSE++ uses GRU to extract text features and CNN for image feature extraction. Triplet loss was used as an optimization method.

- CAMP [13]: CAMP calculates similarity after implementation of a cross modal message passing mechanism to monitor the text-image association.

- SCAN [50]: SCAN discovers the full latent alignments between images and texts.

- LW-MCR [51]: LW-MCR improves inference time by knowledge distillation and contrast learning.
- MTFN [14]: MTFN proposes a fusion based model to calculate the similarity for the cross modal data.

4.5 Results

The numerical results obtained from the proposed method, quantifying its performance as recall at k ***R@K*** and mean recall ***mR*** for the RSITMD dataset is recorded in table 4.2 and for the RSICD dataset it is presented in table 4.3 and compared with baseline approaches. The model rankings highlighted in red, signify the highest-ranked model retrieval results and in blue color represents the retrieval performance ranked as the second-best.

Table 4.2 shows that the proposed approach works best for recall at k=5 and 10 for image-text retrieval and recall at k=1 text-image retrieval for RSITMD dataset. Table 4.3 depicts that the proposed approach works best for recall for text-image retrieval and image-text retrieval for the RSICD dataset except for recall at k=10 for image-text retrieval where SCAN performs the best and second-best in retrieval performance when k=5 in text-image retrieval.

Table 4.2: Comparison of R@K and mR of baseline approaches with proposed method on RSITMD dataset

Framework	Image-Text Retrieval			Text-Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
VSE++	7.20	20.22	33.40	5.50	16.33	**34.16**	19.46
CAMP	**9.31**	24.98	37.01	**7.99**	**26.01**	**33.66**	**23.16**
SCAN	**10.22**	**25.91**	36.73	5.26	23.43	31.24	22.13
LW-MCR	5.09	21.68	35.62	6.06	**24.75**	31.42	20.77
MTFN	8.60	23.65	**38.28**	6.79	19.57	32.85	21.62
Proposed Method	9.06	**28.74**	**42.18**	**9.82**	21.77	32.90	**24.07**

Table 4.3: Comparison of R@K and mR of baseline approaches with proposed method on RSICD dataset

Framework	Image-Text Retrieval			Text-Image Retrieval			mR
	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
CAMP	3.42	8.12	19.25	**4.33**	18.06	**28.64**	**13.63**
SCAN	3.85	6.89	19.84	3.71	16.40	26.73	12.90
LW-MCR	**5.59**	5.20	16.44	2.30	15.32	26.24	11.84
MTFN	4.02	**9.52**	**19.74**	3.90	17.17	26.49	13.47
Proposed Method	**5.90**	**10.40**	16.90	**5.90**	**17.60**	**29.00**	**14.28**

The figure 4.2 clearly shows that the proposed method works best for text-image retrieval at k=1 and best for image-text retrieval at k=5. The figure 4.3 makes it obvious that the proposed method works best for text-image retrieval at k=1 and best for image-text retrieval at k=10.

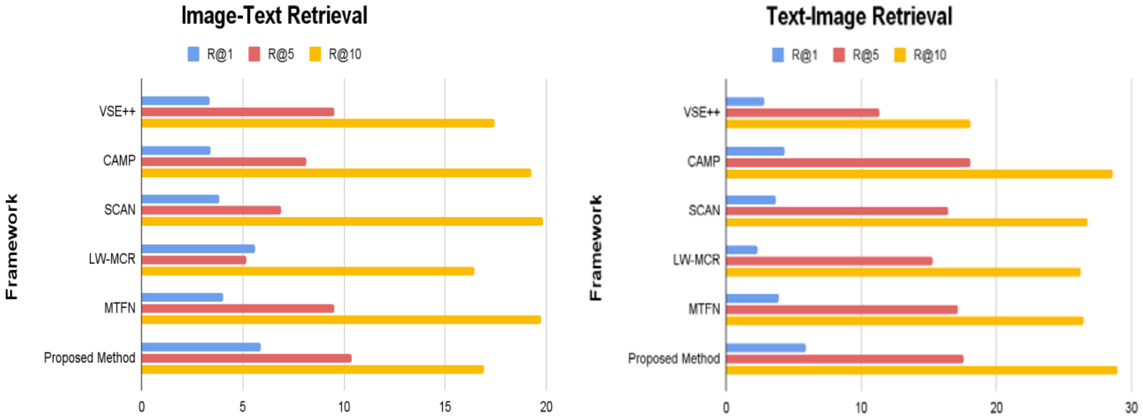


Figure 4.2: Comparison of recall at k (k=1,5 and 10) text-image retrieval and image-text retrieval results for RSICD dataset

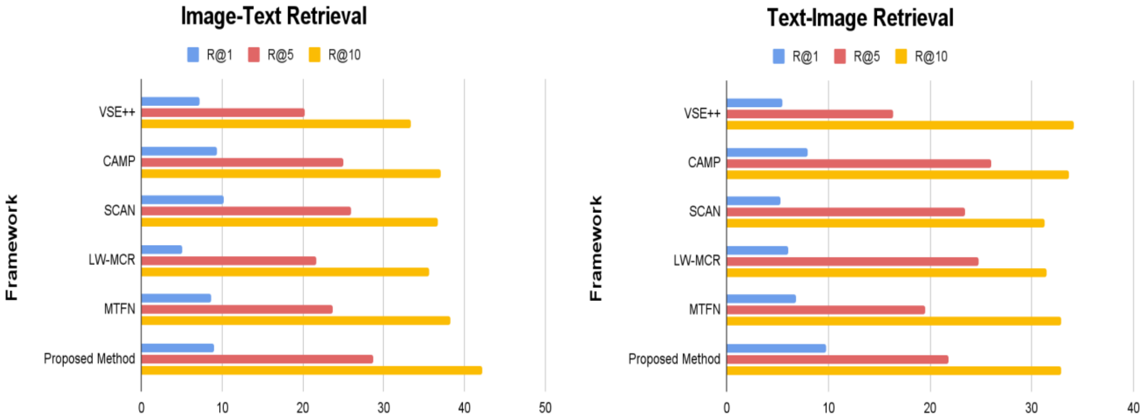


Figure 4.3: Comparison of recall at k (k=1,5 and 10) text-image retrieval and image-text retrieval results for RSITMD dataset

It can be observed in figure 4.4 the superior performance of the proposed method in comparison to previous approaches in terms of mean recall for both the datasets. Previous approaches used high-level vision features only for retrieval and the utilization of low-level vision features besides the high-level features has shown the improvement in the overall mean recall consistently for both the remote sensing datasets.

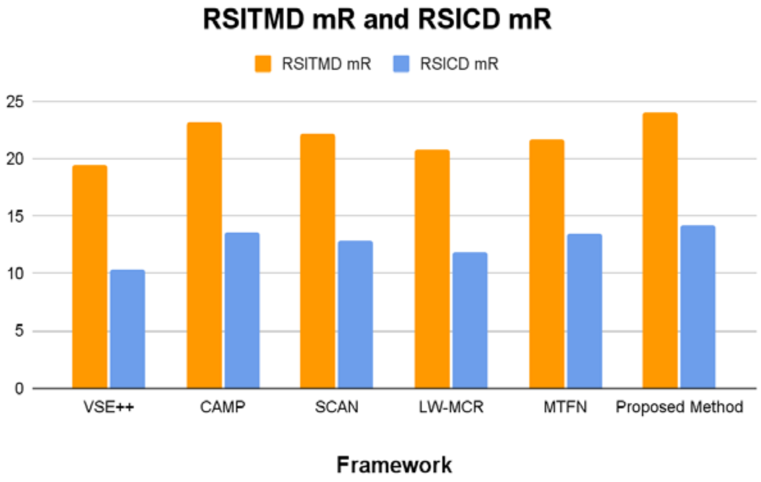


Figure 4.4: Comparison of mean recall text-image retrieval and image-text retrieval results for RSITMD dataset and RSICD dataset

4.5.1 Time Consumption Analysis

In the present subsection, we conduct a comparative analysis of the retrieval performance of our proposed model in relation to other methods. To facilitate a meaningful time-based comparison, we employ a key evaluation metric namely evaluation time (ET). In this context, ET represents the time taken to calculate the similarity between all the images and the texts within different remote sensing data test sets. It is imperative to highlight that these experiments are carried out under conditions of minimal system load, ensuring equitable and unbiased results.

Table 4.4 presents a comparison of individual methods concerning evaluation time. The proposed approach excels in terms of computational efficiency than other methods except for LW-MCR being a light weight model demonstrates notable advantages in terms of retrieval time but lags significantly behind the proposed model assessed in terms of its overall performance.

Table 4.4: Time Consumption comparison of the proposed method with previous methods

Time Consumption	VSE++	CAMP	SCAN	LW-MCR	MTFN	Proposed Method
Evaluation Time of RSICD (s)	22.81	29.35	64.15	**18.29**	45.96	20.43
Evaluation Time of RSITMD (s)	4.38	6.72	11.69	**4.12**	8.55	5.24

Chapter 5

Conclusion and Future Work

Remote sensing image-text retrieval is a task that involves complex understanding of image and text semantics. Our approach attains state-of-the-art retrieval performance in comparison to other retrieval frameworks in remote sensing. Contrary to previous work in this field our method employs micro level visual features to supplement the higher level visual features in the remote sensing datasets not solely the macro level visual features. The proposed framework employs residual network Res2Net for visual data features extraction at a macro level and graph convolution to understand the relationship between targets at the micro level in the visual data. These both level visual data are later dynamically fused to provide a vision feature consisting of quality visual information. The utilization of micro features for supplementation of macro features and similarly the usage of macro features for guiding micro features improves the model performance significantly. Moreover following the popularity of transformer architectures the proposed model utilizes sentence transformer SBERT framework to understand the text semantics. Further retrieval results are re-ranked in the similarity matrix. The proposed technique is compared against Recall scores on the RSITMD and RSICD remote sensing datasets. The effectiveness of the proposed framework has been validated on real-world datasets considering prediction quality.

In the context of forthcoming research endeavors, we intend to:

- Incorporate cutting-edge methodologies for image feature extraction, such as leveraging vision transformers for vision feature extraction.
- Evaluate the proposed system on other datasets including natural image datasets and remote sensing datasets.
- Improve retrieval accuracy further by employing an improved retrieval model.

References

[1] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2002, pp. 311–318.

[2] Z. Shi and Z. Zou, “Can a machine generate humanlike languagedescriptions for a remote sensing image?” IEEE Trans. Geosci. RemoteSens., vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[3] X. Li, X. Zhang, W. Huang, and Q. Wang, “Truncation crossentropy loss for remote sensing image captioning,” IEEE Trans.Geosci. Remote Sens., vol. 59, no. 6, pp. 5246–5257, Jun. 2021, doi:10.1109/TGRS.2020.3010106.

[4] Z. Yuan et al., “Exploring a fine-grained multiscale method for crossmodal remote sensing image retrieval,” IEEE Trans. Geosci. RemoteSens., vol. 60, pp. 1–19, 2022, doi: 10.1109/TGRS.2021.3078451.

[5] Y. Lv, W. Xiong, X. Zhang, and Y. Cui, “Fusion-based correlation learning model for cross-modal remote sensing image retrieval,”IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022, doi:10.1109/LGRS.2021.3131592.

[6] Z. Yuan, “A lightweight multi-scale crossmodal text-image retrieval method in remote sensing,” IEEE Trans. Geosci. Remote Sens., vol. 60,2021, Art. no. 5612819, doi: 10.1109/TGRS.2021.3124252.

[7] C. Wen, Y. Hu, X. Li, Z. Yuan, X. X. Zhu, Vision-language models in remote sensing: Current progress and future trends, arXiv preprint arXiv:2305.05726 (2023).

[8] Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, M. A. Al Zuair, F. Melgani, Bi-modal transformer-based approach for visual question answering in remote

sensing imagery, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–11.

[9] C. Chappuis, V. Mendez, E. Walt, S. Lobry, B. Le Saux, D. Tuia, Language transformers for remote sensing visual question answering, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 4855–4858.

[10] W. Nanal, M. Hajiarbabi, Captioning remote sensing images using transformer architecture, in: 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2023, pp. 413–418.

[11] Z. Ren, S. Gou, Z. Guo, S. Mao, R. Li, A mask-guided transformer network with topic token for remote sensing image captioning, Remote Sensing 14 (12) (2022) 2939.

[12] F. Faghri, D. J. Fleet, J. R. Kiros, S. Fidler, Vse++: Improving visual-semantic embeddings with hard negatives (2018). arXiv:1707.05612.

[13] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, J. Shao, Camp: Cross-modal adaptive message passing for text-image retrieval (2019). arXiv:1909.05506.

[14] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, J. Song, Matching images and text with multi-modal tensor fusion and re-ranking (2020). arXiv:1908.04011.

[15] T. Abdullah, L. Rangarajan, Towards multimodal data retrieval in remote sensing (2020).

[16] W. Zhou, H. Guan, Z. Li, Z. Shao, M. R. Delavar, Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions,

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023).

[17] W. Nanal, M. Hajiarbabi, Captioning remote sensing images using transformer architecture, in: 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2023, pp. 413–418.

[18] Z. Ren, S. Gou, Z. Guo, S. Mao, R. Li, A mask-guided transformer network with topic token for remote sensing image captioning, Remote Sensing 14 (12) (2022) 2939.

[19] Y. Lv, W. Xiong, X. Zhang, Y. Cui, Fusion-based correlation learning model for cross-modal remote sensing image retrieval, IEEE Geoscience and Remote Sensing Letters 19 (2021) 1–5.

[20] F. Zheng, W. Li, X. Wang, L. Wang, X. Zhang, H. Zhang, A cross- attention mechanism based on regional-level semantic features of images for cross-modal text-image retrieval in remote sensing, Applied Sciences 12 (23) (2022) 12221.

[21] F. Zheng, X. Wang, L. Wang, X. Zhang, H. Zhu, L. Wang, H. Zhang, A fine-grained semantic alignment method specific to aggregate multiscale information for cross-modal remote sensing image retrieval, Sensors 23 (20) (2023) 8437.

[22] M. Yang, D. He, M. Fan, B. Shi, X. Xue, F. Li, E. Ding, J. Huang, Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features, in: Proceedings of the IEEE/CVF International conference on Computer Vision, 2021, pp. 11772–11781.

[23] X. Zhang, W. Li, X. Wang, L. Wang, F. Zheng, L. Wang, H. Zhang, A fusion encoder with multi-task guidance for cross-modal text–image retrieval in remote sensing, Remote Sensing 15 (18) (2023) 4637.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale (2021). arXiv:2010.11929.

[25] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018) p. 4171–4186. arXiv:1810.04805. URL http://arxiv.org/abs/1810.04805

[26] L. Mi, S. Li, C. Chappuis, D. Tuia, Knowledge-aware cross-modal text-image retrieval for remote sensing images, in: Proceedings of the Second Workshop on Complex Data Challenges in Earth Observation (CDCEO 2022), 2022.

[27] H. Yu, F. Yao, W. Lu, N. Liu, P. Li, H. You, X. Sun, Text-image matching for cross-modal remote sensing image retrieval via graph neural network, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2022) 812–824.

[28] F. Yao, X. Sun, N. Liu, C. Tian, L. Xu, L. Hu, C. Ding, Hypergraph-enhanced textual-visual matching network for cross-modal remote sensing image retrieval via dynamic hypergraph learning, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2022) 688–701.

[29] G. Mikriukov, M. Ravanbakhsh, B. Demir, Deep unsupervised contrastive hashing for large-scale cross-modal text-image retrieval in remote sensing, arXiv preprint arXiv:2201.08125 (2022).

[30] M. Huang, L. Dong, W. Dong, G. Shi, Supervised contrastive learning based on fusion of global and local features for remote sensing image retrieval, IEEE Transactions on Geoscience and Remote Sensing (2023).

[31] C. Zheng, N. Song, R. Zhang, L. Huang, Z. Wei, J. Nie, Scale-semantic joint decoupling network for image-text retrieval in remote sensing, ACM

Transactions on Multimedia Computing, Communications and Applications 20 (1) (2023) 1–20.

[32] L. He, S. Liu, R. An, Y. Zhuo, J. Tao, An end-to-end framework based on vision-language fusion for remote sensing cross-modal text-image retrieval, Mathematics 11 (10) (2023) 2279.

[33] Z. Yuan, W. Zhang, C. Tian, Y. Mao, R. Zhou, H. Wang, K. Fu, X. Sun, Mcrn: A multi-source cross-modal retrieval network for remote sensing, International Journal of Applied Earth Observation and Geoinformation 115 (2022) 103071.

[34] Y. Yuan, Y. Zhan, Z. Xiong, Parameter-efficient transfer learning for remote sensing image-text retrieval, IEEE Transactions on Geoscience and Remote Sensing (2023).

[35] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, J. Zhou, Remoteclip: A vision language foundation model for remote sensing, arXiv preprint arXiv:2306.11029 (2023).

[36] W. Zhang, J. Li, S. Li, J. Chen, W. Zhang, X. Gao, X. Sun, Hypersphere-based remote sensing cross-modal text-image retrieval via curriculum learning, IEEE Transactions on Geoscience and Remote Sensing (2023).

[37] X. Tang, Y. Wang, J. Ma, X. Zhang, F. Liu, L. Jiao, Interacting enhancing feature transformer for cross-modal remote sensing image and text retrieval, IEEE Transactions on Geoscience and Remote Sensing (2023).

[38] M. M. A. Rahhal, M. A. Bencherif, Y. Bazi, A. Alharbi, M. L. Mekhalfi, Contrasting dual transformer architectures for multi-modal remote sensing image retrieval, Applied Sciences 13 (1) (2022) 282.

[39] H. Zhang, Y. Sun, Y. Liao, S. Xu, R. Yang, S. Wang, B. Hou, L. Jiao, A transformer-based cross-modal image-text retrieval method using feature decoupling and reconstruction, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 1796–1799.

[40] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, Vol. 25, Curran Associates, Inc., 2012, pp. 1106–1114.

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015) p.770 – 778. arXiv:1512.03385. URL http://arxiv.org/abs/1512.03385

[43] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE transactions on pattern analysis and machine intelligence 43 (2) (2019) 652–662.

[44] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). arXiv:1908.10084. URL http://arxiv.org/abs/1908.10084

[45] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 632–642. doi:10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075

[46] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans,31 Louisiana, 2018, pp. 1112–1122. doi:10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101

[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). arXiv:1907.11692. URL http://arxiv.org/abs/1907.11692

[48] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, CoRR abs/1503.03832 (2015). arXiv:1503.03832. URL http://arxiv.org/abs/1503.03832

[49] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, Dota: A large-scale dataset for object detection in aerial images, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3974–3983. doi:10.1109/CVPR.2018.00418.

[50] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching (2018). arXiv:1803.08024.

[51] Z. Yuan, W. Zhang, X. Rong, X. Li, J. Chen, H. Wang, K. Fu, X. Sun, A lightweight multi-scale crossmodal text-image retrieval method in remote sensing, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–19.

—

—

—

—

38

38

—

—

—

—