remote sensing

MDPI

Review

A Review of Cross-Modal Image–Text Retrieval in Remote Sensing

Lingxin Xu [1,2], Luyao Wang [1,2], Jinzhi Zhang [1,2], Da Ha [1,2] and Haisu Zhang [2,3,*]

[1] Graduate School, National University of Defense Technology, Wuhan 430030, China; xulingxin20@nudt.edu.cn (L.X.); wangluyao@nudt.edu.cn (L.W.); zhangjinzhi@nudt.edu.cn (J.Z.); hd@nudt.edu.cn (D.H.)
[2] Hubei Provincial Key Laboratory of Data Intelligence, Wuhan 430019, China
[3] Information Support Force Engineering University, Wuhan 430030, China
* Correspondence: zhanghaisu@nudt.edu.cn

Highlights

What are the main findings?

- Identifies a convergence trend between real-valued representation and deep hashing methods in remote sensing (RS) cross-modal retrieval, propelled by large-scale vision-language pre-training (VLP) models to achieve finer-grained semantic alignment.
- Systematically synthesizes three key challenges—multi-scale semantic modeling, small object feature extraction, and multi-temporal feature understanding—that impede the effective application of cross-modal retrieval in remote sensing.

What is the implication of the main finding?

- Provides a clear technical framework tracing the evolution from global feature matching to fine-grained alignment mechanisms, serving as a foundational reference for guiding future research directions.
- Outlines prospective solutions and emerging trends, including self-supervised learning and neural architecture search, to overcome the unique challenges in RS cross-modal retrieval.

Abstract

With the emergence of large-scale vision-language pre-training (VLP) models, remote sensing (RS) image–text retrieval is shifting from global representation learning to fine-grained semantic alignment. This review systematically examines two mainstream representation paradigms—real-valued embedding and deep hashing—and analyzes how the evolution of RS datasets influences model capability, including multi-scale robustness, small object discriminability, and temporal semantic understanding. We further dissect three core challenges specific to RS scenarios: multi-scale semantic modeling, small object feature preservation, and multi-temporal reasoning. Representative architectures and technical solutions are reviewed in depth, followed by a critical discussion of their limitations in terms of generalization, evaluation consistency, and reproducibility. We also highlight the growing role of VLP-based models and the dependence of their performance on large-scale, high-quality image–text corpora. Finally, we outline future research directions, including RS-oriented VLP adaptation and unified multi-granularity evaluation frameworks. These insights aim to provide a coherent reference for advancing practical deployment and promoting cross-domain applications of RS image–text retrieval.

check for updates

Academic Editor: Salah Bourennane

Received: 20 October 2025
Revised: 4 December 2025
Accepted: 9 December 2025
Published: 11 December 2025

Citation: Xu, L.; Wang, L.; Zhang, J.; Ha, D.; Zhang, H. A Review of Cross-Modal Image–Text Retrieval in Remote Sensing. *Remote Sens.* **2025**, *17*, 3995. https://doi.org/10.3390/rs17243995

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Keywords: remote sensing; cross-modal retrieval; image–text matching; multi-scale modeling; small object detection; multi-temporal analysis

1. Introduction

With the rapid advancement of high-resolution Earth observation technology, remote sensing (RS) images have evolved into a critical data source for Earth observation and environmental monitoring. Confronted with the massive volume of data generated daily by satellites worldwide, the efficient and accurate extraction of key information from these images and its translation into comprehensible natural language descriptions have become a central challenge in the field of intelligent remote sensing interpretation. This study reveals the technological evolution from global feature matching to fine-grained alignment mechanisms, such as attention-based frameworks and fusion models supported by Transformers. These models can effectively handle specific challenges in RS, including multi-scale semantic modeling, small target feature extraction, and multi-temporal feature understanding. This study reveals the technological evolution from global feature matching to fine-grained alignment mechanisms, such as attention-based frameworks and fusion models supported by Transformers. These models can effectively handle specific challenges in RS, including multi-scale semantic modeling, small target feature extraction, and multi-temporal feature understanding. Traditional retrieval methods that rely on manual annotation are not only costly and inefficient but also struggle to meet the demands for fine-grained information perception—such as small objects and multi-temporal changes in complex scenarios—exhibiting considerable limitations in both timeliness and accuracy. Cross-modal retrieval of remote sensing images has emerged as an important research direction to address the challenges described above. This technology aims to establish deep semantic associations between RS images and natural language text, enabling precise retrieval of target images via textual descriptions, as well as automatic generation of semantic text from unlabeled images according to task-specific requirements. It overcomes the constraints of traditional tag-based retrieval methods by offering a more flexible and intelligent means of information interaction.

Unlike natural images, RS images exhibit inherent multi-scale characteristics, sparse distribution of small objects, and spatio-temporal dynamics. These attributes lead to fundamental differences from cross-modal retrieval in natural imagery, rendering direct transfers of existing methods often ineffective. Hence, there is an urgent need to develop specialized semantic alignment and representation mechanisms tailored to the unique challenges of RS data. This paper provides a systematic review of key technological advances and future directions in RS image cross-modal retrieval, with a focus on core technical frameworks. Beginning with feature representation methods, we examine two mainstream approaches: real-valued representation and deep hashing, summarizing their developments and representative works, along with their applicability and limitations in RS scenarios. Subsequently, we delve into three critical challenges that require further breakthroughs:

(1) Multi-scale semantic modeling. How to construct a unified semantic representation that remains robust to scale variations in ground objects.
(2) Small object feature extraction. How to effectively enhance the discriminability of small object features within complex backgrounds.
(3) Multi-temporal feature understanding. How to model spatio-temporal dynamic semantics in image sequences and accurately align them with natural language descriptions.

A natural "heterogeneity gap" exists between RS images and natural language text. While the low-level statistical properties of image pixel matrices and textual word vectors exhibit fundamental differences, their representations converge towards greater semantic alignment at higher abstraction levels. This characteristic complicates the establishment of semantic associations between modalities. Thus, the core of cross-modal retrieval lies in accurately measuring semantic similarity across modalities and constructing a unified vector representation space to achieve semantic alignment. The foundation of RS image cross-modal retrieval technology is the semantic alignment between visual elements and natural language. This capability allows users not only to retrieve target images efficiently using natural language queries but also to generate task-oriented textual descriptions directly from unlabeled images. Thereby, it overcomes the inefficiencies and limitations of traditional manual tagging and offers greater flexibility in application. Unlike conventional visual tasks, cross-modal retrieval systems require a deep understanding of fine-grained relationships between sentence-level text semantics and visual content, while also capturing information about ground objects and their spatial distribution characteristics. These are then translated into quantifiable and computable feature representations. This interdisciplinary effort spans remote sensing science, geographic science, and computer science, underscoring its significant research value.

The literature reviewed in this paper encompasses research directions such as deep learning-based feature embedding, cross-modal hashing, multi-scale visual modeling, small object detection and enhancement, temporal RS analysis, and the adaptation and optimization of vision-language pre-training models in the RS domain. Special emphasis is placed on studies proposing innovative solutions to the three key challenges mentioned above to provide a clear technical context and promote further research for future advancements in this field.

2. Feature Representation Method

Cross-modal image–text retrieval in remote sensing fundamentally depends on learning a unified feature representation space that can directly measure semantic similarity between images and text, thus bridging the semantic gap between the visual and textual modalities. While numerous approaches have been developed in natural image domains, their adaptation to remote sensing requires careful consideration of RS-specific characteristics such as multi-scale objects, complex backgrounds, and spatial relationships. Based on the nature of the embedding space, the prevailing feature representation methods can be categorized into two distinct paradigms: real-valued representation methods and deep hashing methods. The former learns continuous mappings into high-dimensional real-valued vector spaces. This paradigm prioritizes fine-grained semantic alignment by preserving the richness of continuous features, optimizing for similarity within a continuous metric space. The latter learns discrete mappings into compact Hamming spaces. This paradigm emphasizes retrieval efficiency for large-scale datasets by generating compact binary codes, trading some representational capacity for significant gains in storage and computational speed.

These two paradigms establish different technical routes, with the former focusing on precision-oriented alignment and the latter on efficiency-oriented retrieval. This section systematically examines both methods, delving into their mathematical foundations, core technical mechanisms of cross-modal alignment, evolutionary trajectories from natural image processing to professional applications of remote sensing, and specific applications in the field of remote sensing in recent years.

2.1. Real-Valued Representation in RS

Real-valued representation methods aim to project remote sensing images and natural language text into a common, high-dimensional continuous vector space, where the semantic similarity between modalities is quantified using distance metrics such as Euclidean or cosine distance. The core objective is to construct a shared embedding space in which semantically similar images and text are positioned closer to each other, thereby achieving fine-grained cross-modal semantic alignment. This paradigm prioritizes the preservation of rich, continuous feature information and is particularly suited for scenarios demanding high retrieval accuracy. A foundational framework for real-valued representation typically employs a dual-encoder architecture, as conceptually illustrated in Figure 1.

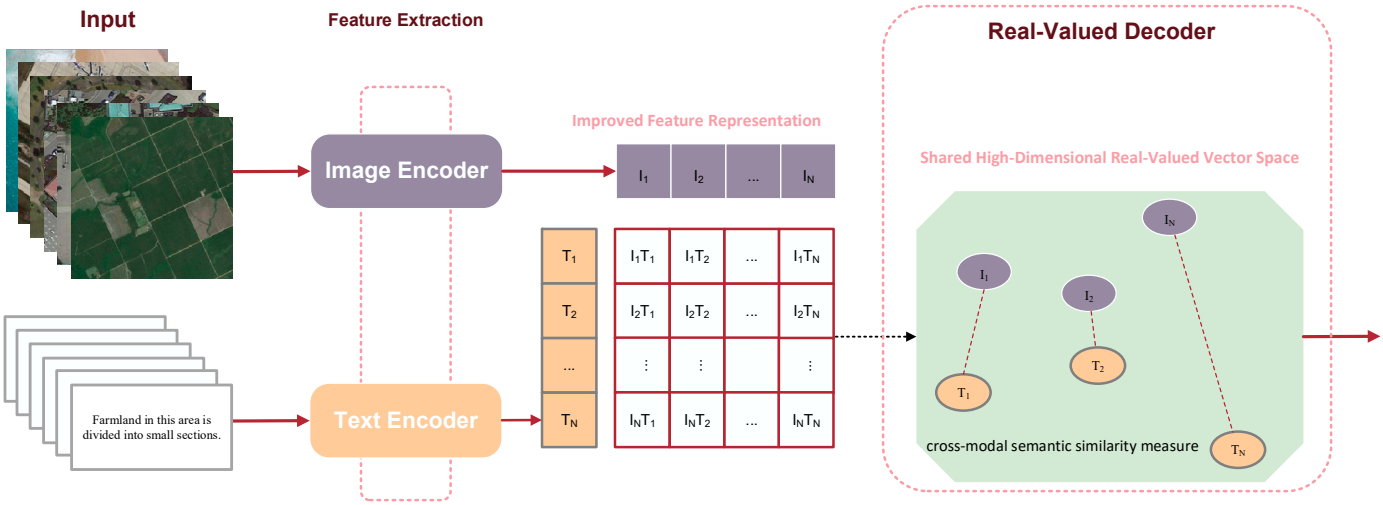

Figure 1. Framework of real-valued representation for remote sensing cross-modal retrieval.

Real-valued representation methods effectively bridge the semantic gap between visual and textual data. To realize this objective, the alignment process typically follows three key steps—feature extraction, projection to a shared space, and semantic alignment via a loss function—which are detailed as follows:

(1) Feature Extraction: An image encoder (e.g., CNN or ViT) processes an image $\boldsymbol{I}$ to extract a visual feature vector $\boldsymbol{v} \in \boldsymbol{R}^{d_v}$. Concurrently, a text encoder (e.g., LSTM or BERT) processes a sentence $\boldsymbol{T}$ to produce a textual feature vector $\boldsymbol{t} \in \boldsymbol{R}^{d_t}$.

(2) Projection to Shared Space: The visual and textual features are then projected into a common $\boldsymbol{D}$-dimensional semantic space using projection matrices. Respectively, the corresponding projection operations for visual and textual features are defined by Equations (1) and (2):

$$\boldsymbol{e_v} = \boldsymbol{W_v v} + \boldsymbol{b_v} \tag{1}$$

$$\boldsymbol{e_t} = \boldsymbol{W_t t} + \boldsymbol{b_t} \tag{2}$$

where $\boldsymbol{W_v} \in \boldsymbol{R}^{D \times d_v}$, $\boldsymbol{W_t} \in \boldsymbol{R}^{D \times d_t}$ are learnable weight matrices, and $\boldsymbol{b_v}, \boldsymbol{b_t} \in \boldsymbol{R}^{D}$ are bias terms. The outputs $\boldsymbol{e_v}, \boldsymbol{e_t} \in \boldsymbol{R}^{D}$ are the final image and text embeddings.

(3) Semantic Alignment via Loss Function: The key to modal alignment is the training objective. The triplet loss is a widely used mechanism for this purpose. Hence, it will be employed as an example to illustrate the alignment mechanism. For a given "anchor" sample—an image $\boldsymbol{I_a}$, with a "positive" sample and its matching text $\boldsymbol{T_p}$, and a "negative" sample and its a non-matching text $\boldsymbol{T_n}$—the loss function encourages the distance between the anchor and positive to be smaller than the distance between

the anchor and negative by a margin α. To formalize this objective, the triplet loss is defined as follows in Equation (3):

$$\mathcal{L}_{triplet} = \sum_{(I,T_p,T_n)} \left[\alpha + d\left(e_v^I, e_t^{T_p}\right) - d\left(e_v^I, e_t^{T_n}\right)\right]_+ \tag{3}$$

where $d(\cdot,\cdot)$ is a distance metric, such as cosine distance, and $\alpha > 0$ is a margin hyperparameter. The hinge loss function $[x]_+$ is defined separately in Equation (4) for clarity:

$$[x]_+ = max(0, x) \tag{4}$$

This loss function directly minimizes the distance between matching pairs while pushing non-matching pairs apart, effectively achieving the alignment of image and text features in the shared space.

The evolution of these methods in natural image processing has progressed from VSE++ [1], which focused on global feature matching, to SCAN [2] with its introduction of attention mechanisms for fine-grained alignment, and further to CAMP [3], which leveraged self-supervised signals to improve alignment. Subsequent improvements in training strategies and data utilization eventually led to the adoption of pre-trained Transformer-based paradigms such as ALBEF [4], enabling end-to-end deep learning. In the remote sensing domain, research on image–text retrieval and semantic segmentation has progressively adopted embedding learning methods from natural image processing, with numerous adaptations to address the specific characteristics of RS data. As illustrated in the accompanying Figure 2, the overall trend has shifted from coarse-grained global matching toward fine-grained local alignment, with particular emphasis on improving the handling of multi-scale objects, complex backgrounds, and redundant information. In a representative study, Abdullah first introduced embedding techniques from natural image cross-modal retrieval into the RS field, proposing a Deep Bidirectional Triplet Network (DBTN) [5] to achieve global embedding alignment between images and text, thereby laying the groundwork for subsequent studies. Cheng further incorporated attention and gating mechanisms into a Deep Semantic Alignment Network (DSAN) [6], enhancing the correspondence between image regions and textual words while filtering complex background clutter in RS imagery. Liu proposed a Transformer-based Multi-modal Fusion Network (TMFN) [7], which employs Transformer encoders for global context modeling to achieve deep multi-modal feature fusion, significantly improving alignment performance for multi-scale objects and complex scenes.

Although real-valued representation methods have achieved high retrieval accuracy in RS applications, they incur substantial computational overhead when comparing high-dimensional features, especially when processing large-scale RS datasets, thus imposing high demands on hardware resources. Based on the reviewed studies, several future research directions can be suggested. The DBTN approach exhibits relatively coarse modeling of multi-object and multi-scale feature distributions in complex RS scenes; incorporating local feature interaction via attention mechanisms could improve fine-grained alignment and hard sample discrimination. DSAN shows limited capability in addressing common RS challenges such as multi-scale object coexistence, spatial relationship dependencies, and contextual ambiguity of polysemous words; integrating pre-trained language models could enhance textual context understanding. TMFN demonstrates limited adaptability in small-sample scenarios; introducing multi-scale feature reorganization modules or lightweight spatial attention in the decoder could strengthen perception of small objects and edge details. Future research on real-valued representation should therefore focus on fine-grained semantic alignment, complex contextual modeling, and improving generalization and

adaptability to advance efficient, accurate, and practical real-valued cross-modal retrieval in RS.

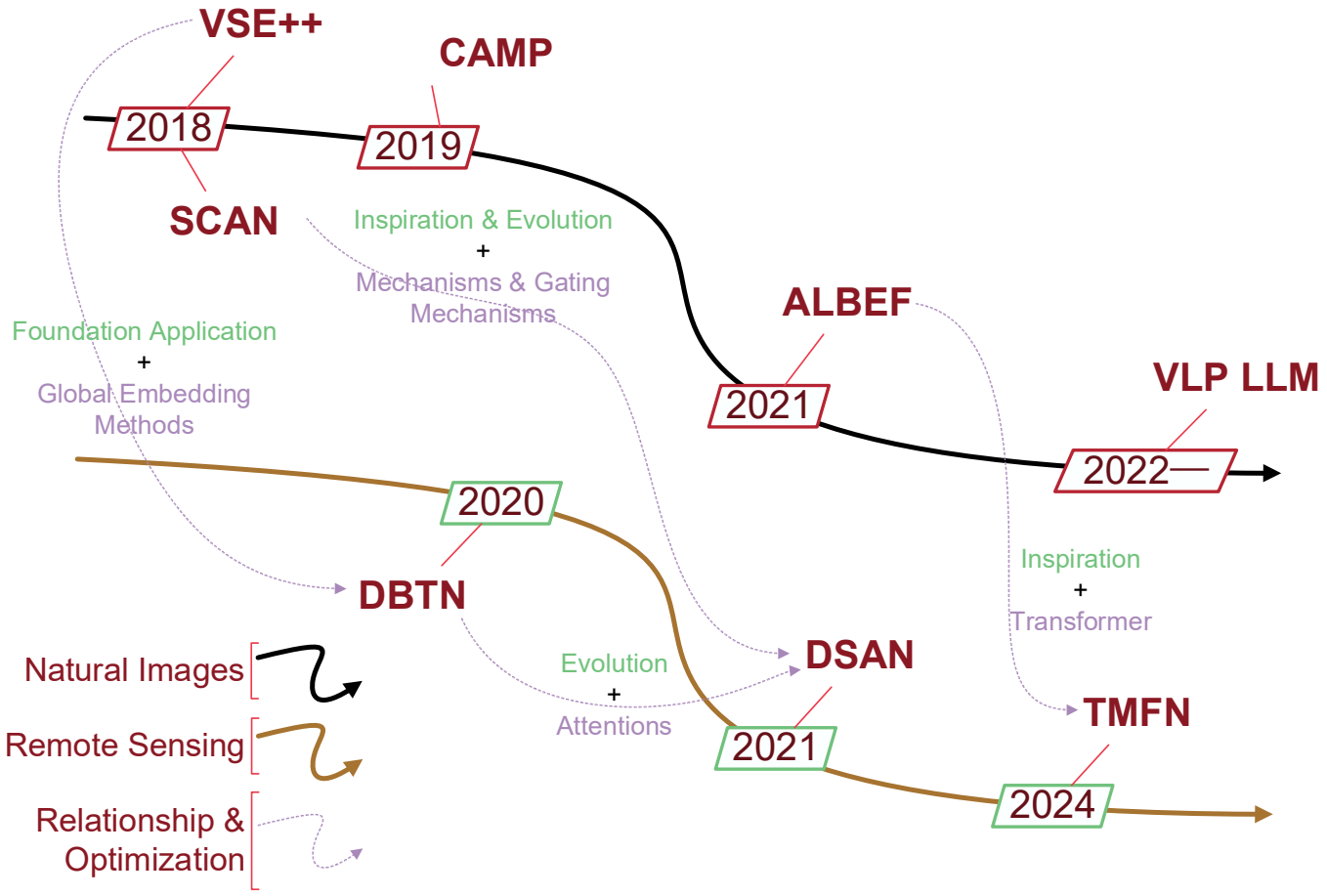

Figure 2. Evolution of real-valued representation methods: from natural image processing to remote sensing cross-modal retrieval.

2.2. Deep Hashing in RS

Deep hashing methods serve as a pivotal technical pathway for cross-modal retrieval, primarily designed to enhance semantic alignment across modalities and reduce the semantic gap between heterogeneous data sources. In contrast to the high-dimensional continuous spaces utilized by real-valued representation, these methods prioritize retrieval efficiency for large-scale datasets by learning end-to-end nonlinear mappings that encode images and text into a unified, discrete binary space [8]. This process ensures that cross-modal samples with similar semantics are represented by closely aligned binary hash codes of a fixed bit length. Similarly to other cross-modal representation learning frameworks discussed previously, these methods typically employ a dual-encoder architecture for feature extraction and transformation from images and text. The key distinction, however, lies in the introduction of a subsequent binarization step that compresses the continuous representations into compact hash codes, thereby significantly improving storage and retrieval efficiency in large-scale scenarios while maintaining considerable semantic expressiveness. A foundational framework for deep hashing, as conceptually illustrated in Figure 3, exemplifies this dual-encoder structure, with its core innovation being the end-to-end binarization process that achieves unified embedding for cross-modal representations.

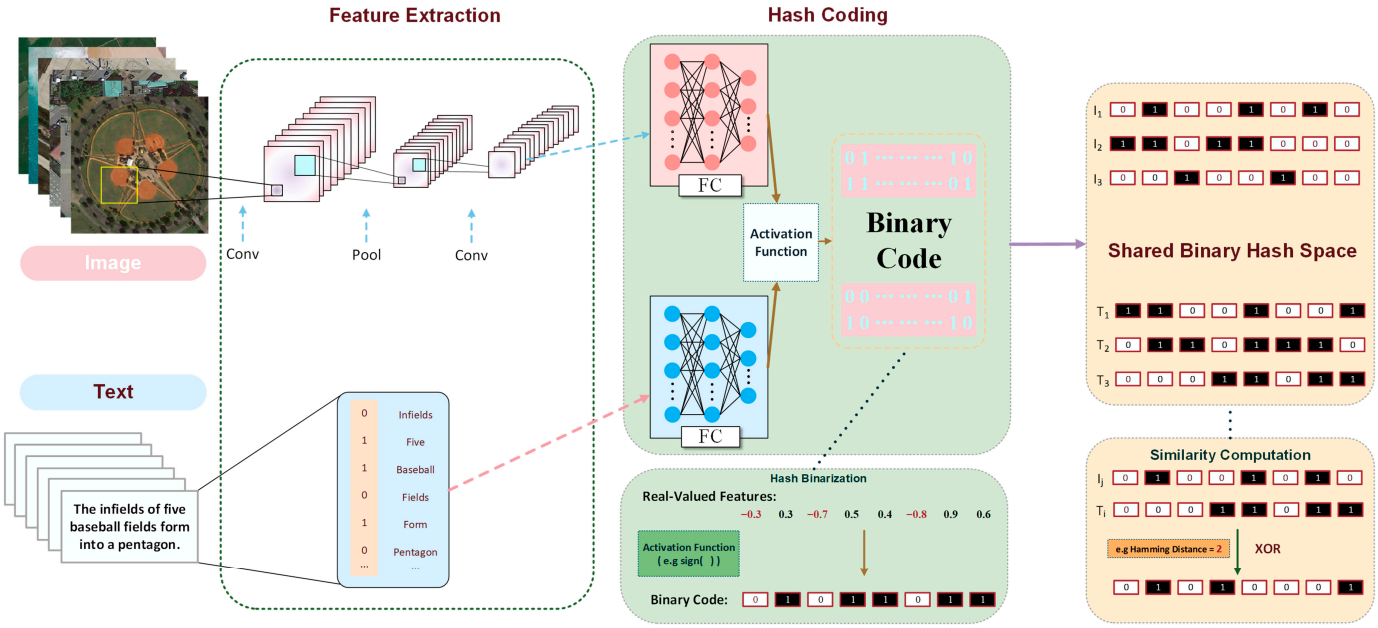

Figure 3. Framework of deep hashing for remote sensing cross-modal retrieval.

The core challenge of deep hashing methods lies in balancing the significant gains in retrieval efficiency with the minimization of inevitable information loss during the binarization process, thereby preserving critical cross-modal semantic relationships. To realize this objective, the alignment process typically follows three key steps—feature extraction and hashing code generation, semantic alignment via a similarity-preserving loss, and quantization constraint—which are detailed as follows to clarify how deep hashing methods constrain the binary codes of images and text during learning, ensuring semantically similar cross-modal samples are positioned closely in the Hamming space for effective alignment and highly efficient retrieval.

(1) Feature Extraction and Hashing Code Generation: Similarly to real-valued methods, image and text features are first extracted using dedicated encoders. Subsequently, a hashing layer, typically implemented as a fully connected layer with a K-dimensional output, which is followed by a sgn function $sgn(x)$ used to generate the binary code. The corresponding binarization operations for visual and textual features are defined by Equations (5) and (6), respectively:

$$b^v = \mathbf{sgn}(f_v(I; \theta_v)) \tag{5}$$

$$b^t = \mathbf{sgn}(f_t(T; \theta_t)) \tag{6}$$

where $sgn(x)$ outputs $+1$ if $x \geq 0$, else -1. However, the sgn function has zero gradients almost everywhere, making direct backpropagation impossible. To circumvent this, a common practice during training is to use a continuous relaxation, such as the $tanh$ function, to generate continuous outputs $u^v, u^t \in [-1, +1]^K$, which are treated as approximate hash codes for the purpose of gradient computation.

(2) Semantic Alignment via Similarity-Preserving Loss: The alignment objective is to ensure that the Hamming distance between hash codes reflects their semantic similarity. A supervised hashing loss often used is based on the inner product $\langle b^v, b^t \rangle$, which is linearly related to the Hamming distance. For a pair of image I_i and text T_j, let S_{ij} be their ground-truth similarity (e.g., $S_{ij} = 1$ if relevant, $S_{ij} = -1$ otherwise). The loss function encourages the inner product of the approximate codes of a relevant

pair to be close to $\boldsymbol{K}$ and that of an irrelevant pair to be close to $-\boldsymbol{K}$. To formalize this objective, the similarity-preserving loss is defined as follows in Equation (7):

$$\mathcal{L}_{hash} = \sum_{i,j} \left(\left\langle u_i^v, u_j^t \right\rangle - KS_{ij} \right)^2 \tag{7}$$

(3) Quantization Loss: To mitigate the discrepancy between the continuous outputs $\boldsymbol{u}$ and the discrete target binary codes $\boldsymbol{b}$, a quantization loss is added to force the continuous outputs to approach the discrete values. The quantization loss is defined separately in Equation (8):

$$\mathcal{L}_{quant} = \sum_{i} \| u_i^v - b_i^v \|_2^2 + \sum_{j} \left\| u_j^t - b_j^t \right\|_2^2 \tag{8}$$

The overall training objective is, therefore, formulated as a weighted sum of the two losses, balancing semantic alignment and binarization quality, as defined in Equation (9):

$$\mathcal{L} = \mathcal{L}_{hash} + \mathcal{L}_{quant} \tag{9}$$

where λ is a hyperparameter that controls the trade-off. This composite loss function simultaneously ensures the semantic discriminability of the hash codes and reduces the error caused by the binarization approximation, effectively achieving cross-modal alignment in the Hamming space.

Building upon the foundation of effective representation learning, both self-supervised learning and deep hashing have emerged as pivotal approaches for efficient large-scale RS image–text retrieval, though serving different purposes in the retrieval pipeline. Self-supervised learning represents a transformative training paradigm that is particularly valuable in remote sensing where annotated data is scarce. Unlike supervised methods relying on manual labels, self-supervised approaches learn meaningful representations by exploiting the inherent correspondence between images and their associated text to construct learning objectives. For instance, contrastive learning frameworks such as CLIP [9] have been adapted to the remote sensing domain, where models learn to maximize agreement between paired image–text representations while minimizing similarity with negative samples, enabling semantic alignment without explicit supervision. Building on these training paradigms, deep hashing has gained prominence as a feature encoding approach that compresses learned representations into compact binary codes for efficient storage and retrieval.

The developmental trajectory of deep hashing in natural image processing began with unsupervised methods such as CMDH [10], an early cross-modal deep hashing technique. This was followed by supervised end-to-end approaches like DCMH [11], which incorporated semantic labels to guide the hash code learning process. Subsequent research introduced various training strategies to enhance hash code quality: SSAH [12] employed adversarial learning as a training paradigm to improve inter-modal alignment and bridge heterogeneity gaps, while methods like CL-CLUH [13] explored binary stream architectures, significantly influencing the cross-modal hashing field. It is important to distinguish that these methods—whether using adversarial training, contrastive objectives, or other learning strategies—represent different training paradigms applied to learn compact binary hash codes as the feature representation type. In remote sensing, researchers have recognized that mapping high-dimensional real-valued features into compact binary hash codes via deep neural networks can substantially improve retrieval efficiency for large-scale RS data, as illustrated in Figure 4, which shows the technical evolution and cross-domain influences between natural image and remote sensing fields. For instance, Che designed a triplet selection function (DTBH) [14] that constrains cross-modal semantic alignment

using hard negative examples, thereby enhancing hashing model performance through an improved training strategy. Mikriukov proposed an Unsupervised Contrastive Hashing Network (UCHN) [15], which trains the hashing module using a multi-objective loss function—combining contrastive learning as a training paradigm with hash code learning as the representation type—to effectively preserve semantic alignment between RS images and text. Huang introduced the DMFH [16] technique, which integrates multi-scale image feature extraction, redundant feature optimization, and fine-grained image–text alignment within an RS cross-modal retrieval framework and combines it with contrastive learning (as a training approach), to enable hashing-based retrieval for fine-grained, multi-scale RS data. These advances demonstrate how diverse training paradigms can be effectively integrated with binary hash code representation to address the unique challenges of remote sensing retrieval.

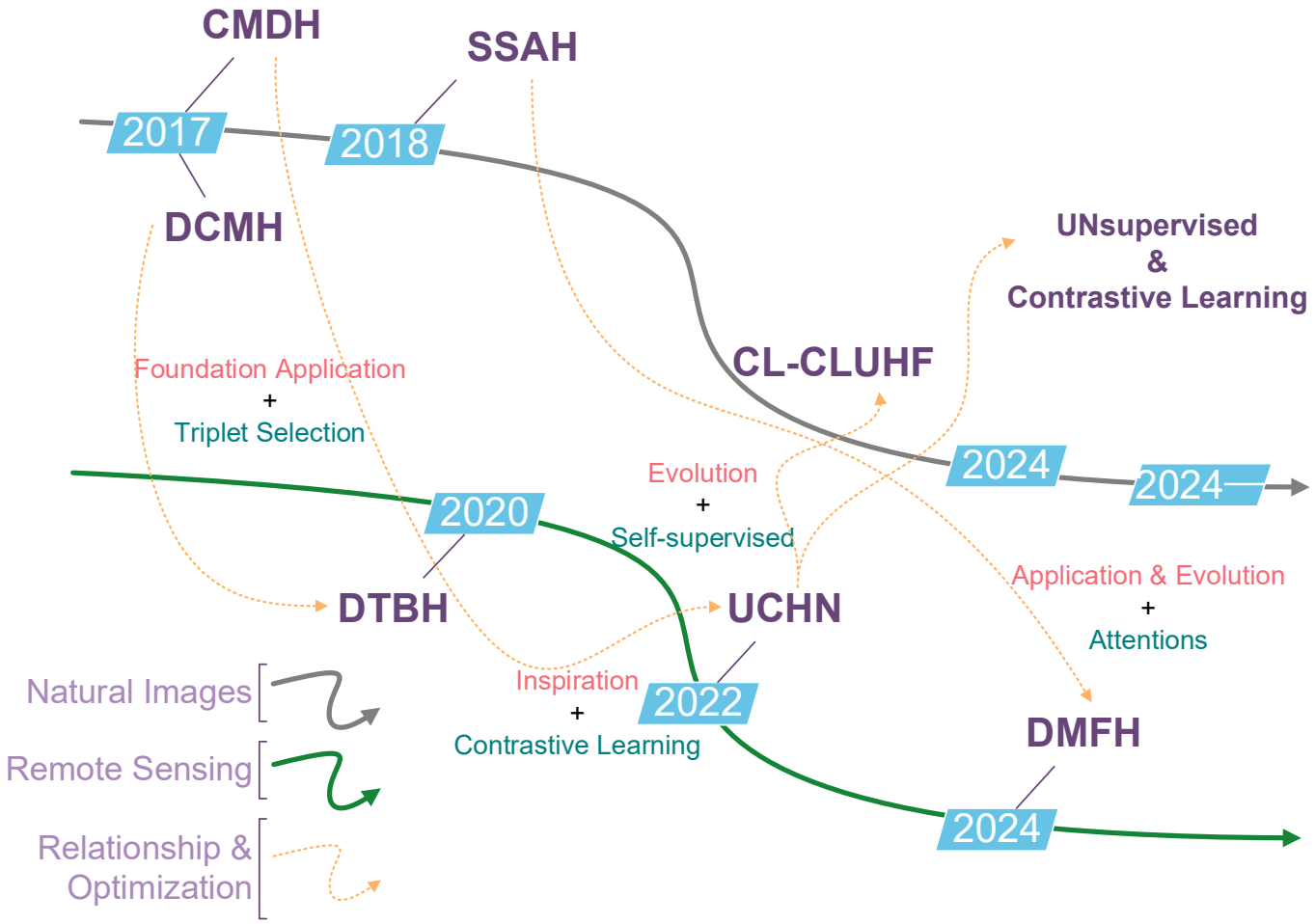

Figure 4. Evolution of deep hashing methods: from natural image processing to remote sensing cross-modal retrieval.

Despite these advances in cross-modal alignment, most existing methods focus heavily on inter-modal alignment while underemphasizing intra-modal structure preservation and fine-grained semantic discrimination. Moreover, the information loss inherent in the binarization process remains an unresolved issue, particularly given the rich detail and texture present in RS imagery, where the representational capacity of binary codes is inherently limited. Building on Chen's work, future studies could incorporate stronger quantization constraints or smoother binarization approximations to mitigate information loss. For the DMFH model, exploring domain-adaptive similarity metrics could improve robustness against noisy relationship construction in unsupervised or

weakly supervised scenarios. Therefore, future research should explore more robust semantic alignment mechanisms—such as integrating self-supervised learning or knowledge distillation—and seek to optimize binarization strategies to better balance retrieval efficiency and feature discriminability.

2.3. Dominant Paradigm and Performance Comparison

With the rise of large-scale models, deep hashing methods increasingly incorporate contrastive and self-supervised learning, which are also core components of large-scale vision-language pre-training (VLP) models [17]. Since the RS field demands finer-grained semantic associations, the two mainstream representation approaches—real-valued and hashing-based—show a trend of mutual integration in RS cross-modal retrieval. Current research is gradually converging into a paradigm dominated by fine-grained real-valued representation learning and large-scale VLP models. Influenced by this trend, researchers have developed various innovative techniques building on image and text encoders to improve RS cross-modal matching accuracy.

At the same time, data serves as the fuel for data-driven approaches, and the quality and scale of datasets directly determine the performance ceiling of retrieval models. Unlike natural images, RS imagery contains multi-scale objects, diverse viewing angles, and complex backgrounds, imposing more stringent requirements on dataset construction. To clarify the data foundation that supports current research, we summarize representative RS cross-modal datasets in Table 1, covering scale, resolution, and captioning modes. Early datasets such as UCM-Caption [18,19] and Sydney-Caption [18] were relatively small and mainly used for initial experimentation. The release of RSICD [20] significantly expanded the available training data, providing over 10,000 images with rich scene categories and becoming a standard benchmark for retrieval evaluation. Later, RSITMD [21] introduced fine-grained keyword annotations and higher-quality dense captions, enabling more precise assessment of fine-grained semantic alignment. With the advent of the large-model era, the field has further witnessed the emergence of million-scale datasets. Datasets such as RS5M [22], SkyScript [23], and GeoLangBind-2M [24] employ automated generation and filtering pipelines to dramatically scale data volume, supporting the pre-training of VLP models and promoting the shift from supervised learning on small datasets to self-supervised learning on large-scale foundational corpora. This evolution—from manually annotated small-scale datasets to massive automatically generated corpora—reveals a clear trajectory toward stronger generalization and zero-shot capabilities.

Table 1. Remote sensing datasets for cross-modal image–text retrieval.

Dataset	Number of Images	Image Size	Captioning Mode
Sydney-Caption [18]	613	500 × 500	5 sentences per image
UCM-Caption [18,19]	2100	256 × 256	5 sentences per image
RSICD [20]	10,921	224 × 224	1–5 sentences per image
RSITMD [21]	4743	256 × 256	5 sentences per image + Fine-grained keywords
NWPU-Caption [25,26]	31,500	256 × 256	5 sentences per image
RSICap [27]	2585	512 × 512	1 high-quality human-annotated caption per image
RS5M [22]	5 million	All Resolutions	Keyword filtering + BLIP-2 generation
SkyScript [23]	5.2 million+	All Resolutions	Automated generation + CLIP filtering
MMRS-1M$_{subset}$ [28]	1 million+	All Resolutions	Multi-task instruction following
GeoLangBind-2M$_{subset}$ [24]	2 million+	All Resolutions	Dataset integration + Automated generation

Therefore, a systematic overview of representative remote sensing datasets, as summarized in Table 1, is essential for understanding the evolution of data foundations, clarifying differences in annotation modes, and interpreting performance gaps among retrieval models. Early datasets typically contain simpler scenes and shorter captions, whereas subsequent ones, such as RSICD and RSITMD, introduce more complex scenes and detailed descriptions. More recently, large-scale datasets further leverage automated or semi-automated caption generation techniques, significantly enriching semantic diversity and supporting more robust model training.

Having outlined the foundational datasets, we now turn our attention to the specific research efforts and methodologies. Wang proposed a fusion model based on rank decomposition (MTFN) [29] to effectively compute image–text similarity. Yuan introduced the AMFMN method, which incorporates a Multi-scale Visual Self-Attention (MVSA) [21] module to extract salient features addressing multi-scale redundancy and intra-class similarity issues in RS images, and achieves fine-grained alignment via a visually guided dynamic triplet loss to tackle positive sample ambiguity. Subsequently, Yuan introduced the Lightweight Multi-Scale Cross-modal Retrieval (LW-MCR) [30] method, which addresses the challenges of computational efficiency and multi-scale redundancy in remote sensing imagery. LW-MCR employs a concise architecture that integrates multi-scale visual features via bilinear pooling and dynamically filters redundant information through a Visual Self-Attention (VSA) mechanism, while using lightweight group convolution for text encoding. To enhance performance without increasing parameters, the model incorporates a hidden supervision optimization approach based on knowledge distillation, allowing it to learn dark knowledge from a teacher network like AMFMN [21], and further leverages unlabeled data through contrastive learning for semi-supervised boosting. Yuan also proposed MCRN [31], a unified framework for managing RS retrieval tasks across multiple sources, including text, audio, and visual. In order to solve the semantic heterogeneity caused by multiple data sources, MCRN introduces a shared mode transfer module (SPTM) based on pattern memory, which dynamically selects the appropriate modal transformation matrix through the gating mechanism to realize the semantic representation that is not constrained by a specific modality, which effectively alleviates the problem of annotation scarcity in RS scenarios. Based on cross-modal retrieval capabilities, Yuan further introduced a comprehensive framework for evaluating multi-modal semantic localization performance [32]. This work systematically establishes quantitative evaluation metrics for semantic positioning (SeLo) tasks, including significant area ratio, attention diversion distance, and discrete attention distance, which measure positioning quality at both the pixel and area levels. The authors contributed AIR-SLT, a multi-semantic, multi-scenario test set containing 22 large RS images and 59 test cases to provide a standardized evaluation benchmark for SeLo tasks. Pan proposed a Scene-aware Aggregation Network (SWAN) [33], which optimizes both visual representation and textual semantic enhancement to improve fine-grained scene perception, reduce semantic confusion, and boost cross-modal alignment accuracy. Zheng proposed FAAMI [34], which addresses multi-scale challenges in RS images by constructing multi-scale feature representations through cross-layer feature connections. It incorporates a Feature Consistency Enhancement Module (FCEM) to improve semantic consistency across layers and employs a shallow cross-attention network for fine-grained alignment between image regions and text words. Concurrently, Pan introduced PIR [35], which leverages prior knowledge from remote sensing scene recognition to guide unbiased vision and text representations. PIR utilizes Progressive Attention Encoders (PAEs), including a Spatial-PAE for external knowledge integration and the Temporal-PAE for cyclic text enhancement, along with Vision Instruction Representation (VIR) and Language Cycle Attention (LCA). A cluster-wise attribution loss reduces semantic confusion zones, leading

to significant improvements on benchmark datasets. Zhang developed HVSA [36], which employs an adaptive alignment strategy based on curriculum learning to align image–text pairs from easy to hard samples. It incorporates a feature uniformity loss for robust embedding on the unit hypersphere and a Key-Entity Attention (KEA) mechanism to handle information imbalance. Ji introduced a retrieval method based on Knowledge-assisted Momentum Contrastive Learning (KAMCL) [37], which reinforces key concept discrimination through a Knowledge-Assisted Learning (KAL) framework and uses a Hierarchical Aggregator (HA) to capture multi-level visual information, achieving breakthroughs in retrieval accuracy and inference efficiency on datasets including RSICD. Liu proposed RemoteCLIP [38], which converts heterogeneously annotated RS data into a unified image–text pair format, scaling pre-training data by a factor of 12 using annotated drone imagery and significantly improving performance in zero-shot classification, cross-modal retrieval, and object counting. Zhang developed GeoRSCLIP [22], a CLIP-based model [9] fine-tuned on the large-scale RS vision-language dataset RS5M, which considerably improves zero-shot classification, cross-modal retrieval, and semantic localization. Ji proposed the EBAKER method, adopting a "filter before align" strategy that dynamically prunes weakly relevant sample pairs during training via Eliminate Before Align (EBA) to reduce noise, combined with a Keyword Explicit Reasoning (KER) module to explicitly reason about key concept differences in RS text [39], enabling fine-grained alignment using foundation models without additional pre-training. SkyScript [23] bridges the critical gap in remote sensing by introducing a large-scale, semantically diverse vision-language dataset, constructed via geo-coordinate-based alignment of open satellite imagery from Google Earth Engine with rich semantic tags from OpenStreetMap. The dataset comprises 2.6 million image–text pairs spanning 29,000 distinct tags, enabling continual pre-training of the SkyCLIP model. This model demonstrates superior zero-shot transfer capabilities, achieving an average 6.2% accuracy gain in scene classification across seven benchmarks, along with advancements in fine-grained attribute classification and cross-modal retrieval. Xiong introduces GeoLangBind [24], an extension of DOFA [40,41], which is a neural plasticity-inspired hypernetwork that dynamically adapts to different sensor wavelengths, enabling joint training across five Earth observation modalities. By incorporating a wavelength-aware dynamic encoder, it handles variable spectral channels effectively, achieving unified representation learning. This approach demonstrates strong performance in zero-shot classification and cross-modal retrieval tasks, providing a scalable solution for heterogeneous data integration in remote sensing applications. iEBAKER [42] extends EBAKER by proposing an improved EBA strategy with two schemes (joint and split) to eliminate weakly correlated pairs more robustly, along with a Sort After Reversed Retrieval (SAR) strategy for similarity optimization. The enhanced Keyword Explicit Reasoning (KER) module explicitly models key concept distinctions, facilitating fine-grained alignment without requiring additional pre-training.

The above methods employ strategies such as basic embedding, fine-grained alignment, multimodal fusion, data augmentation, scene-aware optimization, and pre-training adaptation. Validating the trends implied by these methods, recent research is increasingly shifting toward Vision–Language Pre-training (VLP) models. These models rely on massive image–text corpora to learn generalizable and transferable representations. Representative works include RemoteCLIP, which first adapted the CLIP paradigm to remote sensing and demonstrated the effectiveness of the pre-training–fine-tuning paradigm; SkyCLIP, which gains strong zero-shot transferability from the rich semantic tags of SkyScript; and RSGPT and EarthGPT, which further extend retrieval capabilities toward comprehensive multimodal RS assistants capable of handling complex reasoning and generation. This shift

reflects a broader movement from handcrafted architectures toward large-scale pre-trained foundation models tailored for RS scenarios.

To facilitate intuitive comparison, we evaluate representative studies on two general RS cross-modal datasets: RSICD and RSITMD. RSICD contains 10,921 RS images with manual annotations covering 20 scene categories including farmland, urban areas, and water bodies. RSITMD builds upon RSICD by providing five high-quality dense captions per image, greatly expanding both the quantity and richness of image–text pairs. Performance results on these datasets are summarized in Table 2. To provide a more rigorous and interpretable comparison of existing methods, we adopt mean recall (mR) as the primary evaluation metric for both image-to-text and text-to-image retrieval tasks. In this context, mR is calculated separately for each retrieval direction by averaging Recall@K (R@K) over multiple cutoff values (e.g., K = 1, 5, 10). R@K measures the percentage of queries for which a correct match appears within the top-K retrieved results. While metrics such as R@1 and R@10 are commonly reported, they each present limitations. R@1, which reflects the accuracy of the top-ranked result, can be highly sensitive to noise and exhibit significant variability across different queries. On the other hand, R@10 is often considered too lenient, as it allows for correct matches to appear anywhere within a relatively large set of top retrievals. Crucially, this metric effectively captures a model's robustness in addressing the key challenges to remote sensing image–text retrieval, particularly those involving multi-scale scene modeling and the extraction of small-object features. In remote sensing scenarios, models often struggle to rank images containing tiny objects or complex multi-scale backgrounds at the absolute top position (R@1) due to visual ambiguity and background noise. However, robust fine-grained alignment mechanisms ensure that these "hard" samples remain within the top candidates (e.g., R@5 and R@10). Therefore, a consistent improvement in mR serves as a quantitative proxy for a model's ability to mitigate semantic dilution caused by both small objects and scale variations.

Table 2. Performance of main methods. Mean recall (%) for cross-modal retrieval.

Method	Year	Backbone	Image-to-Text Retrieval		Text-to-Image Retrieval	
		Vision Encoding/ Text Encoding	RSICD Dataset	RSITMD Dataset	RSICD Dataset	RSITMD Dataset
			mR = (R@1 + R@5 + R@10)/3		mR = (R@1 + R@5 + R@10)/3	
VSE++$_{BMVC}$ [1]	2018	ResNet/Bi-GRU	10.12	25.88	10.75	23.78
SCAN$_{ECCV}$ [2]	2018	ResNet/Bi-GRU	12.86	25.44	15.61	27.11
CAMP-triplet$_{ICCV}$ [3]	2019	ResNet/Bi-GRU	13.04	25.59	15.73	26.80
MTFN$_{ACM}$ [29]	2019	ResNet/Bi-GRU	12.43	24.78	17.19	29.06
LW-MCR-d$_{TGRS}$ [30]	2022	ResNet/Bi-GRU	11.91	26.33	17.40	29.25
AMFMN$_{TGRS}$ [21,30]	2022	ResNet/Bi-GRU	14.62	25.74	18.21	33.69
SWAN$_{ACM}$ [33]	2023	ResNet/Glove + Bi-GRU	19.47	30.80	21.74	37.41
HVSA$_{TGRS}$ [36]	2023	ResNet18/Bi-GRU	20.07	30.29	20.26	36.03
FAAMI$_{Sensors}$ [34]	2023	DetNet/BERT	21.33	33.55	25.02	38.44
PIR$_{ACM}$ [35]	2023	Swin-T + ResNet/BERT	25.43	37.39	23.48	39.09
KAMCL$_{TGRS}$ [37]	2023	ResNet/Bi-GRU	26.01	33.97	26.20	38.32
RemoteCLIP$_{TGRS}$ [38]	2024	ResNet + ViT/CLIP	35.50	48.08	35.02	50.68
GeoRSCLIP$_{TGRS}$ [22]	2024	ViT/CLIP	39.49	51.18	38.26	52.43
EBAKER$_{ACM}$ [39]	2024	ViT/CLIP	41.75	52.07	39.64	54.57
SkyCLIP$_{AAAI}$ [23]	2024	ViT/CLIP	23.70	30.75	19.97	30.58
iEBAKER$_{ESWA}$ [42]	2025	ViT/CLIP	46.72	55.46	43.41	55.65
GeoLangBind-L$_{arXiv}$ [24]	2025	ViT/CLIP	23.54	29.57	23.59	35.98

The practical relevance of mR becomes clearer when considering the inherently variable difficulty levels in RS retrieval. Multi-scale scenes may lead to inconsistent ranking behavior across R@1, R@5, and R@10; averaging recall values naturally captures this fluctuation and thus provides an integrated assessment of robustness to scale variation. Similarly, small objects typically offer only weak local cues, making them difficult to retrieve under strict thresholds but still detectable at broader ranks. By aggregating performance across these thresholds, mR directly reflects a model's capacity to recover fine-grained information that might otherwise be suppressed by background complexity. For multi-temporal scenarios—where subtle temporal changes are easily overshadowed by global similarity—mR further reduces the bias introduced by any single rank cutoff, offering a more balanced and comprehensive view of retrieval quality. By averaging performance across several ranks, mR offers a more balanced and stable assessment of ranking quality, smoothing out fluctuations inherent to individual cut-off points and providing a more reliable measure of overall semantic alignment capability.

Analysis of the results in Table 2 reveals several key trends and performance trade-offs. Early baselines such as VSE++ and SCAN, which rely on ResNet and Bi-GRU backbones, exhibit relatively low performance, with VSE++ attaining only 10.12% mR on the RSICD dataset for image-to-text retrieval. These models are trained from scratch on limited remote sensing data and lack sophisticated alignment mechanisms, which restricts their ability to bridge the semantic gap effectively. In contrast, methods that incorporate fine-grained alignment mechanisms—such as AMFMN and PIR—achieve notable improvements; PIR, for example, reaches 25.43% mR on RSICD by explicitly modeling local interactions between image regions and text words. A more substantial performance leap is observed with vision-language pre-training (VLP)-based models. RemoteCLIP and GeoRSCLIP, for instance, achieve mR scores exceeding 35% and 39%, respectively, on RSICD, demonstrating that large-scale pre-training significantly enhances the ability to handle the semantic complexity and scene diversity of remote sensing imagery. However, these gains come at the cost of increased computational complexity. Lightweight architectures such as LW-MCR strike a balance between efficiency and accuracy, achieving competitive performance—around 26% mR on RSITMD—with substantially fewer parameters, making them suitable for resource-constrained deployment environments. The most recent method, iEBAKER, achieves state-of-the-art performance with 46.72% mR on RSICD. It integrates a CLIP backbone with a noise-filtering strategy and an explicit keyword reasoning module, illustrating that coupling foundation models with domain-specific alignment strategies offers a promising path forward for remote sensing cross-modal retrieval.

3. Key Challenges

Although cross-modal retrieval research in natural images has produced numerous effective techniques, directly transferring these approaches to the remote sensing domain is nontrivial due to the distinctive characteristics of RS imagery. The performance patterns observed in Table 2 further highlight this gap: models lacking explicit multi-scale modeling—such as early ResNet–BiGRU baselines—exhibit large discrepancies between R@1 and higher-rank recalls, reflecting limited adaptability to scale variation and complex background clutter. Approaches that incorporate fine-grained alignment or multi-scale feature aggregation (e.g., AMFMN, PIR) mitigate these issues by producing more balanced recall distributions. Likewise, architectures designed to preserve small-object features—such as attention-based or Transformer-based models—achieve stronger gains at stricter cutoff thresholds, indicating improved sensitivity to subtle local cues. VLP-based models show particularly notable improvements under both small-object and multi-

temporal conditions, suggesting enhanced capability in modeling fine-grained scene semantics and temporal dynamics.

These empirical differences underline the fundamental challenges inherent in RS cross-modal retrieval. RS images exhibit substantial variation in scale, contain numerous small yet semantically critical objects, and often encode complex spatio-temporal patterns. Such characteristics amplify the semantic gap between visual and textual modalities, making alignment markedly more difficult than in natural-image scenarios. Correspondingly, the key difficulties can be summarized as follows:

(1) Modal Heterogeneity. Remote sensing images acquired by different sensors exhibit inherent differences in physical properties and scale characteristics, such as variations in ground object features and spatial resolution, which lead to diverse requirements for natural language expression.
(2) Semantic Granularity Difference. Descriptions of natural image scenes typically operate at a relatively fixed scale, whereas RS text often needs to describe ground objects across multiple granularities.
(3) Spatio-temporal Dynamics. RS images possess significant dynamic characteristics, and the extraction of change features differs considerably from that of static natural images.

In summary, current research in RS image cross-modal retrieval centers on three key directions: multi-scale modeling, small object feature capturing, and multi-temporal feature understanding. Multi-scale Modeling: Variations in acquisition methods and resolutions from different sensors lead to significant appearance differences in the same ground object across scales. To ensure stable target recognition in multi-scale imagery, it is essential to design model architectures with cross-scale perception capabilities. Such architectures should extract scene structures at large scales while preserving edge details at small scales, thereby achieving consistent and robust multi-scale semantic representation. Small Object Feature Capturing: High-resolution RS images often contain numerous small objects that are tiny and sparsely distributed yet semantically important. Traditional feature extraction networks tend to overlook such objects, leading to semantic information loss. Effective small object feature capturing must emphasize the highlighting of small object regions within complex backgrounds, preserve feature information of tiny ground objects, and enhance their discriminability during cross-modal alignment through mechanisms such as local enhancement and fine-grained attention. Multi-temporal Feature Understanding: The revisit cycle of satellites endows RS images with strong temporal characteristics and spatio-temporal semantics. This entails not only spatial semantics at fixed moments but also dynamic semantics describing change features over time. Multi-temporal models must be capable of analyzing spatio-temporal differential features across multi-temporal images and accurately aligning them with natural language descriptions. Although change detection technology for RS imagery is relatively mature, natural language understanding for multi-temporal RS images remains underexplored in the cross-modal domain. Existing models still struggle to accurately interpret consistency and difference information across multiple RS images.

3.1. Multi-Scale

Scale variation is a common issue in RS image analysis tasks such as land cover classification, object detection [43], and semantic segmentation. In the context of RS image cross-modal retrieval, scale variation involves challenges in multi-scale feature learning [44], cross-modal semantic alignment, and spatial resolution differences. With the increasing application of deep learning in remote sensing, constructing an effective multi-scale feature representation system has become a key technical challenge for improving RS image

understanding and retrieval performance. To address this need, researchers have proposed various multi-scale feature extraction strategies under deep network frameworks, primarily falling into three technical routes: image pyramids, parallel branch networks, and feature pyramid networks. These approaches contribute to multi-scale feature extraction in remote sensing imagery to varying degrees, offering important technical support for cross-modal retrieval, as conceptually illustrated in Figure 5, which provides a schematic overview of these three technical pathways.

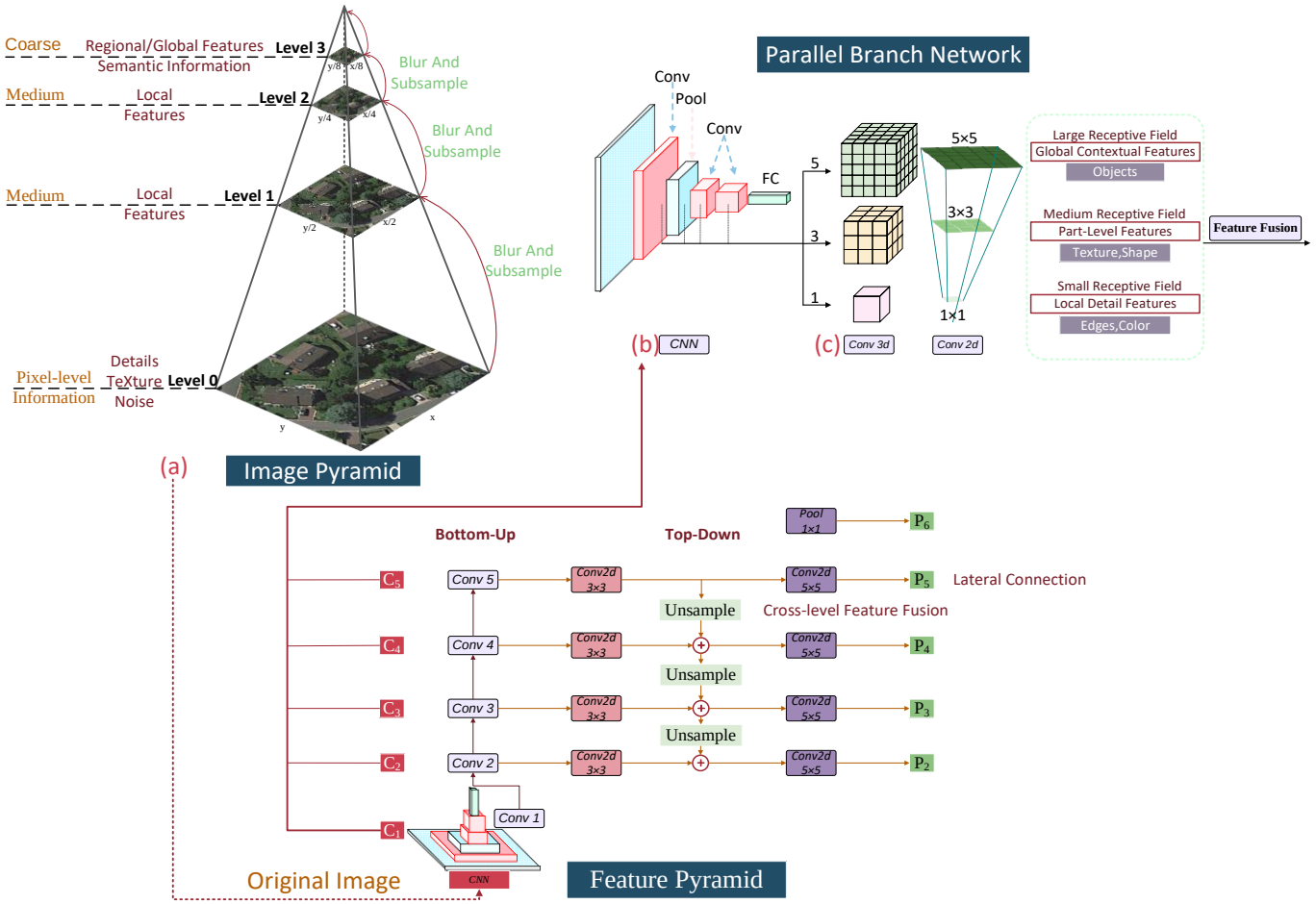

Figure 5. A schematic of the mainstream frameworks for multi-scale feature representation in remote sensing. (**a**) The hierarchical structure of image pyramids and their application in Feature Pyramid Networks. (**b**) The CNN module of parallel branch networks and its application in Feature Pyramid Networks. (**c**) The multi-dimensional convolutional module within parallel branch networks.

The image pyramid is an intuitive approach for constructing multi-scale feature representations [45]. By combining Gaussian kernel convolution with down-sampling, the original RS image is decomposed into a series of sub-images at different resolutions [46], simulating scale space variation at the input stage. Both traditional methods based on hand-crafted features and deep learning approaches [47,48] have widely used image pyramids to mitigate the impact of scale differences on object recognition and matching. However, within deep network frameworks, the need for all network layers to respond to pyramid inputs significantly increases computational and memory overhead during training and inference, reducing the practical utility of this method. Existing improvements mainly focus on enhancing algorithmic efficiency for specific tasks. For example, in RS object detection, Singh proposed the SNIPER model [49], which crops and fine-tunes only the candidate region sub-images in the pyramid, greatly reducing redundant computa-

tion. Dollar employed reinforcement learning to automatically identify relevant pyramid levels and spatial regions, progressively screening regions of interest and further lowering training and inference costs [50]. Nevertheless, traditional image pyramids remain computationally intensive and memory-consuming for real-time applications involving large-scale RS data, and their practicality within deep network frameworks requires further improvement.

Parallel network branches typically start from a feature map of a certain CNN layer, constructing multi-scale feature representations in parallel using convolution kernels of different sizes. HE first introduced the Spatial Pyramid Pooling (SPP) layer in RS object detection, forming multiple receptive fields through different tiling and pooling schemes (dividing image patches into 3×3, 2×2, and 1×1 bins for pooling) [51] to enhance adaptability to the diverse scales of ground objects in RS images. Chen further proposed the Atrous Spatial Pyramid Pooling (ASPP) method [52], replacing spatial pooling operations with parallel dilated convolutions of different dilation rates. Methods in the literature [53] adopted a similar idea, using receptive field variations among branches to construct multi-scale representations while introducing innovations in branch design and information fusion. Compared to image pyramids, parallel branches introduce only a small number of additional branches at specific network layers, significantly reducing storage and computational demands caused by multi-resolution inputs, and can obtain multi-scale representations in a single forward pass. Li embedded parameter-shared TridentBlocks into ResNet residual blocks, enabling the network to adaptively learn scale-invariant features through weight sharing and further compressing computational overhead by activating only a single path during inference [54]. Parallel network branches strike a balance between multi-scale information richness and system lightweightness in RS cross-modal retrieval, offering an efficient solution to the scale variation problem.

Feature pyramid methods represent a compromise between image pyramids and parallel branches: while the former processes multi-resolution images in parallel at the input stage and the latter extracts multi-scale features in parallel within the network structure, feature pyramids "serially" fuse features from different depth layers to approximate multi-scale responses. Sermanet first demonstrated the role of cross-layer connections for scale adaptability: shallow features offer high spatial resolution and rich details, while deep features provide strong semantic discriminability; combining both through cross-layer fusion is feasible [55]. FPN is a well-known architecture in this category, which improved feature inconsistency in cross-layer connections by gradually upsampling and linearly combining features from deep to shallow layers, enabling interactive fusion and enhancing feature consistency [56]. This design introduces minimal extra computational cost, making it highly suitable for large-scale RS image retrieval. Based on FPN, numerous studies have investigated improvements for cross-layer feature fusion. Liu proposed the PANet architecture [57], enhancing spatial information at each layer by adding a fusion path from shallow to deep layers; Kong and Pang modified the layer-by-layer signal transmission approach [58,59], both adopting the idea of first computing a unified fusion feature and then reconstructing the feature pyramid layer by layer, differing mainly in fusion feature computation and reconstruction methods. Tan employed more complex inter-layer connections and stacked the feature pyramid architecture multiple times to achieve better feature representation [60]. Ghiasi abandoned manual fusion network design in favor of reinforcement learning to adaptively select feature layers for fusion [61]. In RS cross-modal retrieval, feature pyramid methods have become an important technical route for constructing multi-scale feature representations and achieving fine-grained image–text matching due to their advantages of low computational overhead and end-to-end trainability.

The three multi-scale feature construction methods each have their merits: image pyramids are intuitive but computationally expensive; parallel network branches balance feature expression and computational efficiency; feature pyramid methods achieve an optimal trade-off between computational efficiency and feature richness through cross-layer connections. In the field of RS image cross-modal retrieval, combining the strengths of these methods to build an efficient multi-scale feature representation system remains an important research direction. Furthermore, considering the particularities of RS images, integrating multi-scale feature extraction with cross-modal semantic alignment and developing targeted multi-scale feature learning methods will be key to enhancing the performance of RS image cross-modal retrieval.

3.2. Small Objects

In RS image cross-modal retrieval, the effective extraction and representation of small objects remains a significant challenge. Although small object feature extraction is inherently related to multi-scale analysis, it represents a distinct and indispensable challenge in remote sensing (RS) image–text retrieval, warranting its treatment as a core section parallel to multi-scale modeling. This necessity arises from three fundamental reasons. First, scale variation primarily concerns differences in object appearance across resolutions, while small object detection uniquely focuses on the extremely limited pixel footprint, high background interference, and low signal-to-noise ratio of small targets, which cause severe semantic dilution in cross-modal alignment. Second, in text descriptions of RS imagery, small objects such as vehicles, ships, aircraft, or small buildings often carry disproportionately critical semantic information—yet they are easily lost even when multi-scale features are adequately modeled. This makes small object preservation an independent bottleneck in retrieval accuracy. Third, compared with multi-scale feature extraction, effective small object retrieval requires additional mechanisms such as low-level feature retention, contextual noise suppression, resolution recovery, and fine-grained attention focus, which are not naturally solved by multi-scale methods alone. For these reasons, small object modeling constitutes a complementary technical route that directly determines the upper limit of fine-grained semantic retrieval in RS imagery, justifying its independent discussion in this review.

From the perspective of image feature extraction, existing RS image–text retrieval models primarily employ Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) as backbone architectures. CNNs, with their strong local feature extraction capabilities, dominated early research [62]; however, their inherent local receptive fields limit global contextual perception. In contrast, ViT models do not rely on convolutional operations, thereby reducing computational complexity. A ViT captures global image features through self-attention mechanisms [63], yet in practice, it often over-attends to large objects, leading to the loss of small object information. Current technical challenges in optimizing small object extraction in RS images include: (a) effectively suppressing noise interference during small object feature extraction; (b) designing more efficient feature fusion mechanisms to balance local and global information representation; and (c) reducing model complexity while improving computational efficiency.

To address these challenges, researchers have proposed various improvement schemes, focusing primarily on two directions: enhancing low-level feature retention and constructing hybrid architectures, as conceptually illustrated in Figure 6 which provides a schematic overview of the hybrid architecture for enhanced low-level feature retention. To mitigate the loss of small object information in ViT models, researchers have adopted various strategies to strengthen low-level feature retention. AGPCNet [64] innovatively adopted a cross-layer feature fusion strategy to effectively recover lost low-level details. DNANet [65]

achieved deep semantic mining through progressive interaction between high-level and low-level features. IRSTFormer [66] designed a hierarchical vision transformer structure combined with downsampling operations to aggregate multi-scale features. CourtNet [67] introduced dense blocks into ViTs, significantly improving small object representation by retaining semantic features from each Transformer block. Although these methods alleviate small object information loss to some extent, they still face challenges such as noise introduction and the lack of standardized feature fusion strategies, which limit model generalization. Hybrid structure methods effectively combine the advantages of CNNs and ViTs to complement local details with global information. IAANet [68] proposed connecting local patch outputs from the CNN with the original Transformer, significantly enhancing small object feature expression. The method in [69] adopted a strategy using a CNN for local information extraction and a ViT for global context, further strengthening small object capture capability. However, such hybrid structures often increase model complexity, and issues such as parameter redundancy and imbalanced fusion mechanisms persist; small object detection performance in complex scenes still has room for improvement. Recently, a limited number of studies have begun exploring high-resolution feature representations, enhancing model perception of tiny objects by introducing high-resolution feature maps or multi-scale self-attention modules [70]. Some scholars have introduced high-resolution feature recovery mechanisms into the Transformer architecture, preserving fine-grained feature maps from certain network layers or upsampling down-sampled feature maps to restore edge details of small objects in cross-modal representations [71]. Such high-resolution auxiliary mechanisms partially counteract the adverse effects of end-to-end deep feature extraction on small objects, improving the discriminability of retrieval models for small objects.

Building on optimized small object feature extraction, semantic alignment becomes a key challenge in enhancing retrieval accuracy for small objects in RS cross-modal retrieval. Since small objects typically occupy very few pixels in RS images [72], their semantic information is easily interfered with by background noise, and they may lack clear contextual correlation in text descriptions, leading to deviations in cross-modal feature mapping and difficulties in ensuring semantic consistency. To address this, researchers have explored three main directions: multi-modal feature interaction, attention-guided alignment, and multi-scale feature fusion coupled with semantic decoupling. Some studies promote multi-modal feature interaction by constructing a joint embedding space. RSRefSeg [73] innovatively fused CLIP and SAM foundation models, converting global/local text semantic embeddings into visually activated features through an AttnPrompter module to guide segmentation mask generation; RSRefSeg achieved b- and cIoU/gIoU scores of 76.05/63.68 and 77.24/64.67, respectively, enabling semantic-consistent expression for 20 classes of small objects on the RRSIS-D dataset. Xiao proposed a Grouped Feature Concern Unit (EFC) [74] combined with a lightweight fusion strategy that enhances inter-layer feature correlation, validated on multiple datasets with up to 1.7% mAP improvement in small object detection accuracy, effectively reconstructing and transforming strong and weak information at each pyramid layer while reducing redundant feature fusion and preserving small object information. Concurrently, several studies based on attention mechanisms have proposed dynamic weight allocation, hierarchical attention architectures, and multi-attention mechanisms to enhance semantic focus on small object regions, offering new ideas for small object semantic alignment. The PBT framework uses a hierarchical visual Transformer to separate target responses from background context during encoding and adopts progressive decoding to achieve pixel-level semantic alignment [75], achieving the highest Intersection over Union (IoU) values on three datasets: NUDT-SIRST, IRSTD-1k, and IRSTD-Air (90.20/72.65/78.45). BAFNet [76] employs a dual-stream

attention mechanism, achieving complementary learning through global semantic flow and local detail flow, effectively reducing the false detection rate via a boundary-aware supervision strategy, and attaining SOTA performance on four RS small object detection datasets: AI-TOD, VisDrone, DIOR, and LEVIR-Ship. To address scale diversity in small objects within remote sensing images, multi-scale feature fusion and semantic decoupling techniques have been widely applied. MwdpNet [77] is a novel multi-level weighted deep perception network in which a multi-level weighted fusion strategy fully utilizes shallow feature information to improve detection performance, especially for small targets, achieving an average precision (AP) of 39.3% on a self-built dataset. SEB-YOLO employs an SPD-CONV module [78] to reconstruct the downsampling process, preserving global features and reducing feature loss, combined with a bilinear interpolation upsampling strategy in the bidirectional feature pyramid (Bi-FPN) to optimize feature fusion, resulting in a 4.0% to 5.3% improvement in mean average precision (mAP) for small object detection across different datasets compared to the baseline model. The performance of the above method in improving the accuracy of small target detection is presented in Table 3.

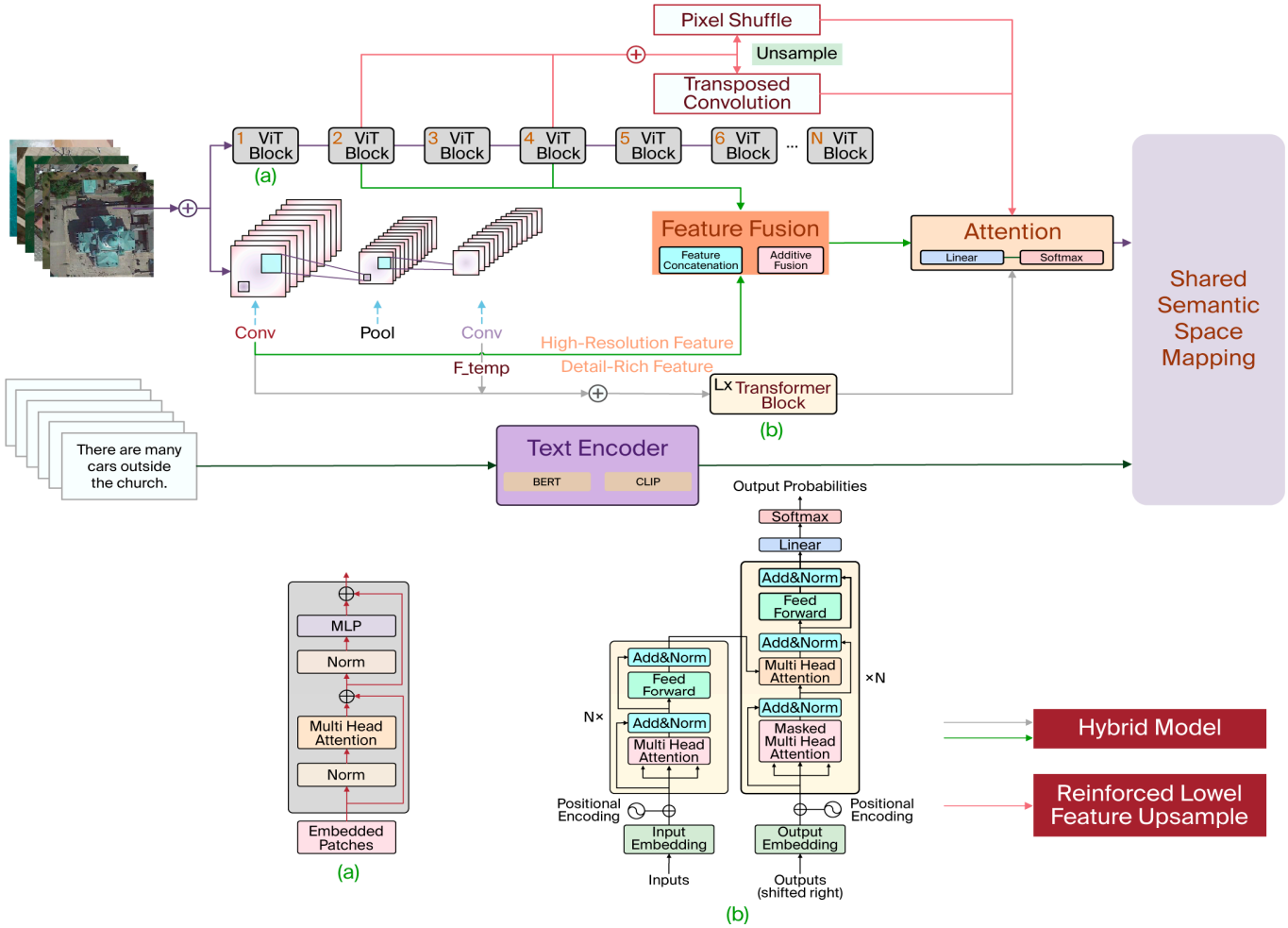

Figure 6. A schematic diagram of the hybrid architecture used to enhance the retention of low-level features. (**a**) The detailed structure of a Vision Transformer block. (**b**) The detailed structure of a Transformer block.

Table 3. Performance of methods in terms of accuracy gain for small object retrieval.

Method	Metric	Dataset	Baseline Methods	Accuracy Gain
RSRefSeg [73]	cIoU	RRSIS-D	RMISN	+0.74
			FIANet	+0.33
	gIoU		RMISN	+2.4
			FIANet	+0.66
EFC [74]	mAP	VisDrone	GFL	+0.017
		COCO	RetinaNet	+0.011
			GFL	+0.01
PBT [75]	loU	NUDT-SIRST	UIU-Net	+1.38
		IRSTD-1k		+1.76
		IRSTD-Air		+0.71
BAFNet [76]	AP	AI-TOD	CAF2ENet-S	+2.4
		VisDrone	CMDNet	+1.6
MwdpNet [77]	mAP	Dataset 1	MDSSD	+0.007
		Dataset 2	YOLOV6-M	-0.004
		Dataset 3	R-FCN	+0.012
SEB-YOLO [78]	mAP	NWPU VHR-10	Original YOLOv5	+0.04
		RSOD		+0.053

Analysis of future research trends focuses on the following key directions: (1) developing adaptive feature selection mechanisms integrated with neural architecture search to refine dynamic semantic weight allocation for adaptive cross-modal feature matching; (2) exploring lightweight hybrid structures to reduce computational overhead while maintaining performance; (3) incorporating prior knowledge to guide model learning, enhancing the accuracy and robustness of small object detection [79]; and (4) establishing comprehensive evaluation frameworks that assess both low-level detail preservation and high-level semantic alignment (via cross-modal retrieval accuracy and semantic consistency measures such as BLEU [80] and ROUGE [81]). To complement the above technical directions, recent studies have also begun exploring auxiliary strategies—such as super-resolution reconstruction, semantic consistency analysis, and attention visualization—to better assess the fidelity and robustness of small object representations in cross-modal retrieval. Through continuous iteration of multi-dimensional evaluation systems and synergistic optimization of core technologies, future work is expected to further improve the expressive completeness and retrieval accuracy of small objects in remote sensing cross-modal retrieval, laying a solid foundation for intelligent remote sensing interpretation and multi-modal human–computer interaction.

3.3. Multi-Temporal

Information extraction from multi-temporal remote sensing imagery is predominantly accomplished through change detection technology, a core methodology for monitoring alterations in geographic elements and tracking the status of natural resource utilization. As a pivotal research direction at the intersection of computer vision and remote sensing science [82], temporal change detection involves the analysis of multi-temporal observational data of dynamic scenes to achieve precise identification and quantitative assessment of changes in target states. The fundamental concept, as defined in multi-temporal analysis, relies on image sequences acquired at different times over the same geographical area to reveal the dynamics of surface objects. This technology has demonstrated significant application value in recent years across critical areas including video action recognition,

dynamic monitoring of urban land cover, and temporal analysis of multi-source remote sensing imagery

In video behavior analysis research, Wang proposed a temporal difference modeling-based framework (TDN) [83] for video action recognition. This method innovatively designs a dual-scale difference module (TDM): the short-term module (S-TDM) enhances local motion representation through RGB differences between adjacent frames, while the long-term module (L-TDM) extracts global temporal structures using cross-segment feature differences. It is the first to integrate different operators into end-to-end training [84], effectively addressing temporal modeling challenges in video action recognition. Tan proposed a new spatio-temporal prediction framework—a parallelizable spatio-temporal predictive learning framework (TAU)—which replaces traditional RNN structures with a factorized attention mechanism [85], innovatively decomposing temporal attention into intra-frame static attention (SA) and inter-frame dynamic attention (DA) to capture spatial structures and temporal evolution separately. It designs a divergence regularization loss based on pixel-wise MSE loss to strengthen inter-frame dynamic constraints. Liu proposed the AdaTAD framework [86], which achieves efficient end-to-end temporal action detection training by designing a temporal information adapter (TIA) that integrates depth-wise separable temporal convolutions to explicitly aggregate contextual information from adjacent frames, enhancing action boundary perception capability.

Urban land cover dynamic detection and multi-source remote sensing image temporal analysis share significant methodological commonality: both require spatio-temporal sequence modeling to achieve precise state evolution tracking of surface targets. Tang proposed the ClearSCD model [87], a multi-task learning framework that comprehensively leverages semantic and change relationships for semantic change detection in high spatial resolution remote sensing images. It innovatively interprets semantic features from different time phases as mutually beneficial relationships and integrates cellular automata to transform abrupt interface changes into gradual transitions while preserving clear interface assumptions, addressing the challenge of capturing subtle and complex changes in urban environments by combining semantic context with dynamic relationship modeling. Quan proposed a dual-stage multi-modal fusion framework that fuses optical and SAR data into three-band inputs via principal component analysis [88], retaining key spectral and structural features; it introduces a channel attention mechanism to adaptively weight multi-modal information, enhancing complementary feature expression and effectively solving the heterogeneity fusion challenge between optical and SAR images. Soni focused on unified multi-temporal modeling, proposing the first remote sensing dialog model supporting multi-temporal sequences [89]. It processes cross-scale images through an adaptive high-resolution module and designs a channel attention-driven data fusion architecture, effectively overcoming the limitations of traditional models with fixed input resolutions.

Building on the above research, change detection technology is transitioning from single-task approaches to a paradigm integrating multiple techniques, with core trends evolving from traditional change detection to a joint task system combining temporal segmentation, super-resolution reconstruction, and weakly supervised learning; methodological innovations exhibit deep integration of cross-domain technologies such as diffusion models with change prior guidance, spiking neural networks with edge computing, and end-to-end vector-image learning. A generalized structural framework for change detection within the context of cross-modal retrieval of multi-temporal remote sensing data is illustrated in Figure 7.

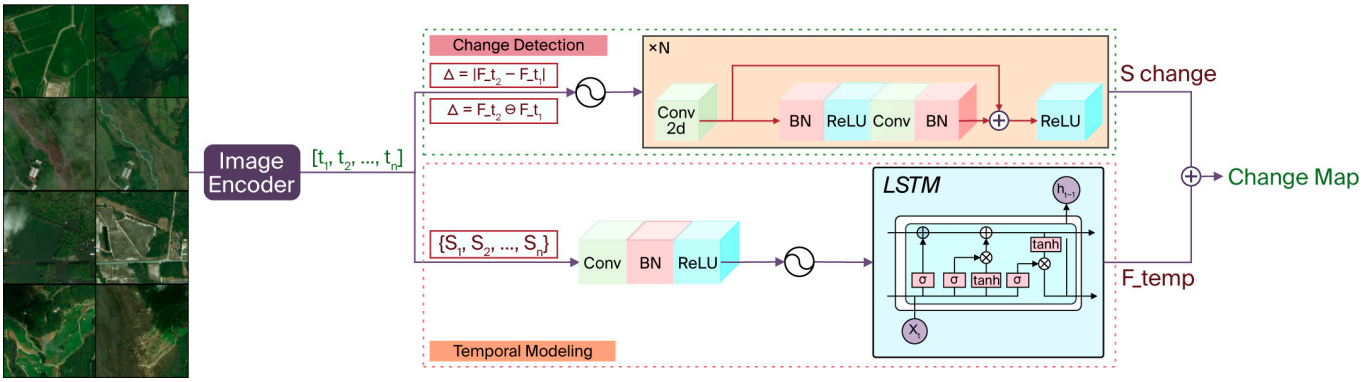

Figure 7. A general structural framework for cross-modal retrieval in the context of multi-temporal remote sensing data.

Although change detection technology is relatively mature, research on natural language understanding for multi-temporal remote sensing images is still in its infancy. Yuan proposed a new task combining change detection with visual question answering (VQA)—change detection-based visual question answering (CDVQA) [90]—aiming to help non-expert users understand surface change information in multi-temporal remote sensing images through natural language interaction. Notably, CDVQA answers questions by classifying predefined answer categories, focusing on identifying correct answer types from multi-modal inputs rather than generating new text, making it a discriminative task with a gap compared to generative natural language understanding. Similarly, Yang introduced change captioning to the remote sensing field, pioneering the requirement for models to automatically generate textual descriptions of changes across time [91]. This new paradigm provides intuitive language-level explanations for remote sensing change detection, holding potential value in applications such as environmental monitoring. The above cross-modal strategy attempts demonstrate new possibilities for multi-temporal remote sensing interpretation, but current related data resources and research remain very limited. Models face significant challenges in understanding fine-grained changes and aligning visual and linguistic representations, making it difficult to ensure complete and accurate generated descriptions and answers.

Overall, multi-temporal image–text cross-modal fusion is still in its early stages, urgently requiring in-depth research in large-scale annotated dataset construction, model architecture design, and evaluation systems to further enhance model capabilities in understanding and expressing cross-modal multi-temporal information. To address the cross-modal understanding needs of multi-temporal remote sensing images, future research directions should focus on (1) developing unified evaluation frameworks that combine generative semantic verification (using metrics such as BLEU-4 and ROUGE-L to assess the alignment between generated temporal descriptions and actual change features) with multi-granularity change detection verification (quantifying model sensitivity to subtle changes through pixel-level and object-level accuracy assessments); (2) integrating change detection technology with vision-language models to enable end-to-end learning of spatio-temporal dynamics; and (3) constructing large-scale multi-temporal image–text paired datasets with fine-grained temporal annotations to support robust model training and comprehensive performance evaluation. These advances will balance local change detail perception with global semantic relationship understanding, ultimately enabling more accurate and interpretable multi-temporal remote sensing analysis.

The challenges analyzed in this chapter reveal three fundamental bottlenecks that currently constrain cross-modal image–text retrieval in remote sensing: multi-scale semantic

modeling, small object feature preservation, and multi-temporal semantic understanding. Although these problems share conceptual connections—for instance, both multi-scale and small object issues stem from spatial variability—they correspond to different forms of semantic ambiguity and require tailored architectural solutions. Multi-scale modeling focuses on stabilizing semantic representation across varying spatial resolutions; small object detection targets the recovery of weak, noise-sensitive fine-grained features; and multi-temporal understanding emphasizes the alignment of dynamic scene changes with textual semantics.

4. Conclusions and Future Trends

4.1. Overall Progress and Technological Evolution

Recent advances demonstrate substantial improvements in remote sensing cross-modal retrieval performance. Real-valued representation methods have evolved from global matching networks to attention-enhanced models capable of capturing fine-grained visual–textual correspondences. In parallel, deep hashing methods have progressed toward more semantically consistent binary encoding with contrastive learning and adversarial training strategies. The convergence of these two representation paradigms—high-accuracy real-valued embeddings and high-efficiency hashing codes—reflects the community's effort to balance retrieval precision with scalability for massive satellite archives. Meanwhile, the emergence of RS-specific large-scale datasets (e.g., RS5M and SkyScript) and foundation models (RemoteCLIP, GeoRSCLIP, and iEBAKER) has significantly strengthened the generalization capability of cross-modal retrieval systems, marking a critical step toward universal remote sensing vision-language models.

4.2. Current Limitations and Open Problems

Despite these advances, several limitations impede practical deployment. First, the strong dependence on supervised or weakly supervised annotations remains problematic, given the high cost and inherent subjectivity of remote sensing captioning. Automatically generated large-scale datasets mitigate this issue but introduce potential semantic inconsistencies that propagate through training. Second, while multi-scale modeling, small object preservation, and multi-temporal semantic understanding have received extensive research attention, no unified architecture effectively addresses all three challenges simultaneously. Scale variation, tiny target ambiguity, and temporal dynamics represent distinct forms of semantic complexity that require complementary but specialized mechanisms. Third, model evaluation remains insufficiently comprehensive: existing benchmarks measure retrieval ranking quality but rarely assess fine-grained localization ability, semantic correctness, or temporal reasoning. Without richer and more realistic evaluation protocols, performance indicators can be misleading and hinder the field's progress toward operational systems.

4.3. Future Research Directions

To advance cross-modal retrieval toward robust, real-world applicability, future efforts should prioritize the following directions:

(1) Richer and More Adaptive Semantic Representation. Integrating domain knowledge—such as geographic priors, sensor characteristics, and hierarchical land-cover semantics—into model architectures may enhance interpretability and alleviate annotation scarcity. The incorporation of neural architecture search (NAS) and adaptive feature selection may further balance performance and efficiency.

(2) Unified Multi-Scale, Small Object, and Temporal Modeling Frameworks. Future retrieval systems should aim to jointly model large-scale spatial structures, fine-

grained object details, and temporal change dynamics. This may be achieved by combining hierarchical feature pyramids, high-resolution refinement modules, and temporal reasoning blocks into an end-to-end cross-modal learning pipeline.

(3) Next-Generation Datasets and Self-Supervised Learning. Building high-quality multi-temporal, multi-sensor, and fine-grained annotated datasets is crucial for capturing real-world complexity. In parallel, self-supervised, semi-supervised, and few-shot learning frameworks will reduce dependence on large-volume human annotations and improve cross-scene generalization.

(4) Comprehensive Evaluation Systems for Real-World Deployment. Future benchmarks should incorporate multi-dimensional evaluation metrics covering retrieval accuracy, semantic consistency, robustness to atmospheric/sensor variations, interpretability, and computational efficiency. For multi-temporal understanding, hybrid metrics combining BLEU/ROUGE with change-detection accuracy will be essential.

4.4. Outlook

Cross-modal retrieval is poised to become a foundational technology for next-generation intelligent Earth observation. As the field transitions from handcrafted architectures toward RS-tailored foundation models, future retrieval systems will increasingly achieve precise semantic grounding, real-time performance, and robust generalization across diverse environments. Ultimately, breakthroughs will likely emerge from holistic system-level designs that simultaneously optimize semantic fidelity, scalability, and operational reliability—driving remote sensing image interpretation toward greater automation, intelligence, and practical impact in environmental monitoring, urban planning, and disaster response.

Author Contributions: Conceptualization, L.X. and H.Z.; methodology, L.X. and H.Z.; validation, L.X. and J.Z.; formal analysis, L.X. and D.H.; investigation, L.X. and L.W.; resources, H.Z. and L.W.; data curation, L.X.; writing—original draft preparation, L.X.; writing—review and editing, H.Z. and L.W.; visualization, L.X. and J.Z.; supervision, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are derived from publicly available sources. No new data were created or analyzed.

Acknowledgments: The authors wish to thank all the reviewers who helped improve this manuscript.

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

References

1. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. *arXiv* **2018**, arXiv:1707.05612.
2. Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. *arXiv* **2018**, arXiv:1803.08024. [CrossRef]
3. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5763–5772. [CrossRef]
4. Li, J.; Selvaraju, R.R.; Gotmare, A.D.; Joty, S.; Xiong, C.; Hoi, S. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. *arXiv* **2021**, arXiv:2107.07651. [CrossRef]
5. Abdullah, T.; Bazi, Y.; Al Rahhal, M.M.; Mekhalfi, M.L.; Rangarajan, L.; Zuair, M. TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images. *Remote Sens.* **2020**, *12*, 405. [CrossRef]
6. Cheng, Q.; Zhou, Y.; Fu, P.; Xu, Y.; Zhang, L. A Deep Semantic Alignment Network for the Cross-Modal Image-Text Retrieval in Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4284–4297. [CrossRef]

7. Liu, Y.; Gao, K.; Wang, H.; Yang, Z.; Wang, P.; Ji, S.; Huang, Y.; Zhu, Z.; Zhao, X. A Transformer-Based Multi-Modal Fusion Network for Semantic Segmentation of High-Resolution Remote Sensing Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104083. [CrossRef]
8. Ma, X.; Zhang, T.; Xu, C. Multi-Level Correlation Adversarial Hashing for Cross-Modal Retrieval. *IEEE Trans. Multimed.* **2020**, *22*, 3101–3114. [CrossRef]
9. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020. [CrossRef]
10. Liong, V.E.; Lu, J.; Tan, Y.-P.; Zhou, J. Cross-Modal Deep Variational Hashing. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4097–4105. [CrossRef]
11. Jiang, Q.-Y.; Li, W.-J. Deep Cross-Modal Hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3270–3278. [CrossRef]
12. Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; Tao, D. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. *arXiv* **2018**, arXiv:1804.01223.
13. Xu, M.; Luo, L.; Lai, H.; Yin, J. Category-Level Contrastive Learning for Unsupervised Hashing in Cross-Modal Retrieval. *Data Sci. Eng.* **2024**, *9*, 251–263. [CrossRef]
14. Chen, Y.; Lu, X. A Deep Hashing Technique for Remote Sensing Image-Sound Retrieval. *Remote Sens.* **2019**, *12*, 84. [CrossRef]
15. Mikriukov, G.; Ravanbakhsh, M.; Demir, B. Deep Unsupervised Contrastive Hashing for Large-Scale Cross-Modal Text-Image Retrieval in Remote Sensing. *arXiv* **2022**, arXiv:2201.08125.
16. Huang, J.; Feng, Y.; Zhou, M.; Xiong, X.; Wang, Y.; Qiang, B. Deep Multiscale Fine-Grained Hashing for Remote Sensing Cross-Modal Retrieval. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6002205. [CrossRef]
17. Chen, F.; Zhang, D.; Han, M.; Chen, X.; Shi, J.; Xu, S.; Xu, B. VLP: A Survey on Vision-Language Pre-Training. *Mach. Intell. Res.* **2023**, *20*, 38–56. [CrossRef]
18. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep Semantic Understanding of High Resolution Remote Sensing Image. In Proceedings of the International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016; pp. 1–5. [CrossRef]
19. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279. [CrossRef]
20. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
21. Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; Sun, X. Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4404119. [CrossRef]
22. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and A Large Vision-Language Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642123. [CrossRef]
23. Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; Rajagopal, R. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing. In Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5805–5813. [CrossRef]
24. Xiong, Z.; Wang, Y.; Yu, W.; Stewart, A.J.; Zhao, J.; Lehmann, N.; Dujardin, T.; Yuan, Z.; Ghamisi, P.; Zhu, X.X. GeoLangBind: Unifying Earth Observation with Agglomerative Vision-Language Foundation Models. *arXiv* **2025**, arXiv:2503.06312.
25. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
26. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5629419. [CrossRef]
27. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Li, X. RSGPT: A Remote Sensing Vision Language Model and Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 272–286. [CrossRef]
28. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. EarthGPT: A Universal Multimodal Large Language Model for Multisensor Image Comprehension in Remote Sensing Domain. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5917820. [CrossRef]
29. Wang, T.; Xu, X.; Yang, Y.; Hanjalic, A.; Shen, H.T.; Song, J. Matching Images and Text with Multi-Modal Tensor Fusion and Re-Ranking. *arXiv* **2020**, arXiv:1908.04011. [CrossRef]
30. Yuan, Z.; Zhang, W.; Rong, X.; Li, X.; Chen, J.; Wang, H.; Fu, K.; Sun, X. A Lightweight Multi-Scale Crossmodal Text-Image Retrieval Method in Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5612819. [CrossRef]
31. Yuan, Z.; Zhang, W.; Tian, C.; Mao, Y.; Zhou, R.; Wang, H.; Fu, K.; Sun, X. MCRN: A Multi-Source Cross-Modal Retrieval Network for Remote Sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *115*, 103071. [CrossRef]
32. Yuan, Z.; Zhang, W.; Li, C.; Pan, Z.; Mao, Y.; Chen, J.; Li, S.; Wang, H.; Sun, X. Learning to Evaluate Performance of Multimodal Semantic Localization. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5631918. [CrossRef]

33. Pan, J.; Ma, Q.; Bai, C. Reducing Semantic Confusion: Scene-Aware Aggregation Network for Remote Sensing Cross-Modal Retrieval. In Proceedings of the ACM International Conference on Multimedia Retrieval, Thessaloniki Greece, 12–15 June 2023; pp. 398–406. [CrossRef]
34. Zheng, F.; Wang, X.; Wang, L.; Zhang, X.; Zhu, H.; Wang, L.; Zhang, H. A Fine-Grained Semantic Alignment Method Specific to Aggregate Multi-Scale Information for Cross-Modal Remote Sensing Image Retrieval. *Sensors* **2023**, *23*, 8437. [CrossRef] [PubMed]
35. Pan, J.; Ma, Q.; Bai, C. A Prior Instruction Representation Framework for Remote Sensing Image-Text Retrieval. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 611–620. [CrossRef]
36. Zhang, W.; Li, J.; Li, S.; Chen, J.; Zhang, W.; Gao, X.; Sun, X. Hypersphere-Based Remote Sensing Cross-Modal Text–Image Retrieval via Curriculum Learning. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5621815. [CrossRef]
37. Ji, Z.; Meng, C.; Zhang, Y.; Pang, Y.; Li, X. Knowledge-Aided Momentum Contrastive Learning for Remote-Sensing Image Text Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625213. [CrossRef]
38. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. [CrossRef]
39. Ji, Z.; Meng, C.; Zhang, Y.; Wang, H.; Pang, Y.; Han, J. Eliminate Before Align: A Remote Sensing Image-Text Retrieval Framework with Keyword Explicit Reasoning. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 1662–1671. [CrossRef]
40. Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A.J.; Hanna, J.; Borth, D.; Papoutsis, I.; Saux, B.L.; Camps-Valls, G.; Zhu, X.X. Neural Plasticity-Inspired Multimodal Foundation Model for Earth Observation. *arXiv* **2025**, arXiv:2403.15356.
41. Xiong, Z.; Wang, Y.; Yu, W.; Stewart, A.J.; Zhao, J.; Lehmann, N.; Dujardin, T.; Yuan, Z.; Ghamisi, P.; Zhu, X.X. DOFA-CLIP: Multimodal Vision-Language Foundation Models for Earth Observation. *arXiv* **2025**, arXiv:2503.06312.
42. Zhang, Y.; Ji, Z.; Meng, C.; Pang, Y.; Han, J. iEBAKER: Improved Remote Sensing Image-Text Retrieval Framework via Eliminate Before Align and Keyword Explicit Reasoning. *arXiv* **2025**, arXiv:2504.05644. [CrossRef]
43. Yan, S.; Song, X.; Liu, G. Deeper and Mixed Supervision for Salient Object Detection in Automated Surface Inspection. *Math. Probl. Eng.* **2020**, *2020*, 3751053. [CrossRef]
44. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
45. Zha, H.; Chen, X.; Wang, L.; Miao, Q. (Eds.) *Computer Vision: CCF Chinese Conference, CCCV 2015, Xi'an, China, September 18–20, 2015, Proceedings, Part II; Communications in Computer and Information Science*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 547, ISBN 978-3-662-48569-9.
46. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
47. Hao, Z.; Liu, Y.; Qin, H.; Yan, J.; Li, X.; Hu, X. Scale-Aware Face Detection. *arXiv* **2017**, arXiv:1706.09876. [CrossRef]
48. Tomè, D.; Monti, F.; Baroffio, L.; Bondi, L.; Tagliasacchi, M.; Tubaro, S. Deep Convolutional Neural Networks for Pedestrian Detection. *Signal Process. Image Commun.* **2016**, *47*, 482–489. [CrossRef]
49. Singh, B.; Davis, L.S. An Analysis of Scale Invariance in Object Detection—SNIP. *arXiv* **2018**, arXiv:1711.08189.
50. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 304–311. [CrossRef]
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *arXiv* **2014**, arXiv:1406.4729.
52. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
53. Kim, S.-W.; Kook, H.-K.; Sun, J.-Y.; Kang, M.-C.; Ko, S.-J. Parallel Feature Pyramid Network for Object Detection. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11209, pp. 239–256, ISBN 978-3-030-01227-4.
54. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z.-X. Scale-Aware Trident Networks for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6053–6062. [CrossRef]
55. Sermanet, P.; LeCun, Y. Traffic Sign Recognition with Multi-Scale Convolutional Networks. In Proceedings of the International Joint Conference on Neural Networks, San Jose, CA, USA, 31 July–5 August 2011; pp. 2809–2813. [CrossRef]
56. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]

57. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 8759–8768. [CrossRef]
58. Kong, T.; Sun, F.; Huang, W.; Liu, H. Deep Feature Pyramid Reconfiguration for Object Detection. *arXiv* **2018**, arXiv:1808.07993. [CrossRef]
59. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 821–830. [CrossRef]
60. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]
61. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 13–19 June 2019; pp. 7029–7038. [CrossRef]
62. Rahhal, M.M.A.; Bazi, Y.; Abdullah, T.; Mekhalfi, M.L.; Zuair, M. Deep Unsupervised Embedding for Remote Sensing Image Retrieval Using Textual Cues. *Appl. Sci.* **2020**, *10*, 8931. [CrossRef]
63. Zhang, X.; Li, W.; Wang, X.; Wang, L.; Zheng, F.; Wang, L.; Zhang, H. A Fusion Encoder with Multi-Task Guidance for Cross-Modal Text–Image Retrieval in Remote Sensing. *Remote Sens.* **2023**, *15*, 4637. [CrossRef]
64. Zhang, T.; Li, L.; Cao, S.; Pu, T.; Peng, Z. Attention-Guided Pyramid Context Networks for Detecting Infrared Small Target Under Complex Background. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [CrossRef]
65. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense Nested Attention Network for Infrared Small Target Detection. *IEEE Trans. Image Process.* **2023**, *32*, 1745–1758. [CrossRef] [PubMed]
66. Chen, G.; Wang, W.; Tan, S. IRSTFormer: A Hierarchical Vision Transformer for Infrared Small Target Detection. *Remote Sens.* **2022**, *14*, 3258. [CrossRef]
67. Peng, J.; Zhao, H.; Zhao, K.; Wang, Z.; Yao, L. CourtNet: Dynamically Balance the Precision and Recall Rates in Infrared Small Target Detection. *Expert Syst. Appl.* **2023**, *233*, 120996. [CrossRef]
68. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior Attention-Aware Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013. [CrossRef]
69. Li, C.; Huang, Z.; Xie, X.; Li, W. IST-TransNet: Infrared Small Target Detection Based on Transformer Network. *Infrared Phys. Technol.* **2023**, *132*, 104723. [CrossRef]
70. He, L.; Liu, S.; An, R.; Zhuo, Y.; Tao, J. An End-to-End Framework Based on Vision-Language Fusion for Remote Sensing Cross-Modal Text-Image Retrieval. *Mathematics* **2023**, *11*, 2279. [CrossRef]
71. Zhang, X.; Liu, Q.; Chang, H.; Sun, H. High-Resolution Network with Transformer Embedding Parallel Detection for Small Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2023**, *15*, 4497. [CrossRef]
72. Nikouei, M.; Baroutian, B.; Nabavi, S.; Taraghi, F.; Aghaei, A.; Sajedi, A.; Moghaddam, M.E. Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications. *Intell. Syst. Appl.* **2025**, *27*, 200561. [CrossRef]
73. Chen, K.; Zhang, J.; Liu, C.; Zou, Z.; Shi, Z. RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models. *arXiv* **2025**, arXiv:2501.06809. [CrossRef]
74. Xiao, Y.; Xu, T.; Yu, X.; Fang, Y.; Li, J. A Lightweight Fusion Strategy with Enhanced Interlayer Feature Correlation for Small Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4708011. [CrossRef]
75. Yang, H.; Mu, T.; Dong, Z.; Zhang, Z.; Wang, B.; Ke, W.; Yang, Q.; He, Z. PBT: Progressive Background-Aware Transformer for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5004513. [CrossRef]
76. Song, J.; Zhou, M.; Luo, J.; Pu, H.; Feng, Y.; Wei, X.; Jia, W. Boundary-Aware Feature Fusion with Dual-Stream Attention for Remote Sensing Small Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5600213. [CrossRef]
77. Ma, D.; Liu, B.; Huang, Q.; Zhang, Q. MwdpNet: Towards Improving the Recognition Accuracy of Tiny Targets in High-Resolution Remote Sensing Image. *Sci. Rep.* **2023**, *13*, 13890. [CrossRef]
78. Hui, Y.; You, S.; Hu, X.; Yang, P.; Zhao, J. SEB-YOLO: An Improved YOLOv5 Model for Remote Sensing Small Target Detection. *Sensors* **2024**, *24*, 2193. [CrossRef]
79. Chen, L.; Su, L.; Chen, W.; Chen, Y.; Chen, H.; Li, T. YOLO-DHGC: Small Object Detection Using Two-Stream Structure with Dense Connections. *Sensors* **2024**, *24*, 6902. [CrossRef] [PubMed]
80. 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—ACL '02, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: St, Stroudsburg, PA, USA, 2001; p. 311. [CrossRef]
81. Lin, C.-Y. *ROUGE: A Package for Automatic Evaluation of Summaries*; Association for Computational Linguistics: Barcelona, Spain, 2004.

82. Lin, Z.; Cheng, M.-M.; He, R.; Ubul, K.; Silamu, W.; Zha, H.; Zhou, J.; Liu, C.-L. (Eds.) *Pattern Recognition and Computer Vision: 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18–20, 2024, Proceedings, Part V; Lecture Notes in Computer Science*; Springer Nature: Singapore, 2025; Volume 15035, ISBN 978-981-97-8619-0.
83. Wang, L.; Tong, Z.; Ji, B.; Wu, G. TDN: Temporal Difference Networks for Efficient Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904. [CrossRef]
84. Geng, X.; Kang, B.-H. (Eds.) *PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28–31, 2018, Proceedings, Part I; Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2018; Volume 11012, ISBN 978-3-319-97303-6.
85. Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; Li, S.Z. Temporal Attention Unit: Towards Efficient Spatiotemporal Predictive Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18770–18782. [CrossRef]
86. Liu, S.; Zhang, C.-L.; Zhao, C.; Ghanem, B. End-to-End Temporal Action Detection with 1B Parameters Across 1000 Frames. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 18591–18601. [CrossRef]
87. Tang, K.; Xu, F.; Chen, X.; Dong, Q.; Yuan, Y.; Chen, J. The ClearSCD Model: Comprehensively Leveraging Semantics and Change Relationships for Semantic Change Detection in High Spatial Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2024**, *211*, 299–317. [CrossRef]
88. Quan, Y.; Zhang, R.; Li, J.; Ji, S.; Guo, H.; Yu, A. Learning SAR-Optical Cross Modal Features for Land Cover Classification. *Remote Sens.* **2024**, *16*, 431. [CrossRef]
89. Soni, S.; Dudhane, A.; Debary, H.; Fiaz, M.; Munir, M.A.; Danish, M.S.; Fraccaro, P.; Watson, C.D.; Klein, L.J.; Khan, F.S.; et al. EarthDial: Turning Multi-Sensory Earth Observations to Interactive Dialogues. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025. [CrossRef]
90. Yuan, Z.; Mou, L.; Xiong, Z.; Zhu, X.X. Change Detection Meets Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5630613. [CrossRef]
91. Yang, Y.; Liu, T.; Pu, Y.; Liu, L.; Zhao, Q.; Wan, Q. Remote Sensing Image Change Captioning Using Multi-Attentive Network with Diffusion Model. *Remote Sens.* **2024**, *16*, 4083. [CrossRef]

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.