Review

Fine-Grained Interpretation of Remote Sensing Image: A Review

Dongbo Wang , Zedong Yan , Peng Liu *

Posted Date: 25 September 2025

doi: 10.20944/preprints202509.2113.v1

Keywords: remote sensing; fine-grained; deep learning

 doi:10.20944/preprints202509.2113.v1

Article

Fine-Grained Interpretation of Remote Sensing Image: A Review

Dongbo Wang [1], Zedong Yan [2] and Peng Liu [2,*]

[1] China Nuclear Power Engineering Co., LTD, Beijing, China
[2] Aerospace Information Research Institute Chinese Academy of Sciences Beijing, China
* Correspondence: liupeng202303@aircas.ac.cn

Abstract

This article conducts a systematic review on the fine-grained interpretation of remote sensing images, delving deeply into its background, current situation, datasets, methodology, and future trends, aiming to provide a comprehensive reference framework for research in this field. In terms of fine-grained interpretation datasets, with a focus on introducing representative datasets, and analyze their key characteristics such as the number of categories, sample size, and resolution, as well as their benchmarking role in research. For methodologies, by classifying the core methods according to the interpretation level system, this paper systematically summarizes the methods, models, and architectures for implementing fine-grained remote sensing image interpretation based on deep learning at different levels such as pixel-level classification and segmentation, object-level detection, and scene-level recognition. Finally, we summarize the challenges currently faced by the research (such as the distinction of highly similar categories, cross-sensor domain migration, and high annotation costs), and look forward to future directions, emphasizing the need to enhance the generalization, support open-world recognition further, and adapt to actual complex scenarios, etc. This review aims to promote the application of fine-grained interpretation technology for remote sensing images across a broader range of fields.

Keywords: remote sensing; fine-grained; deep learning

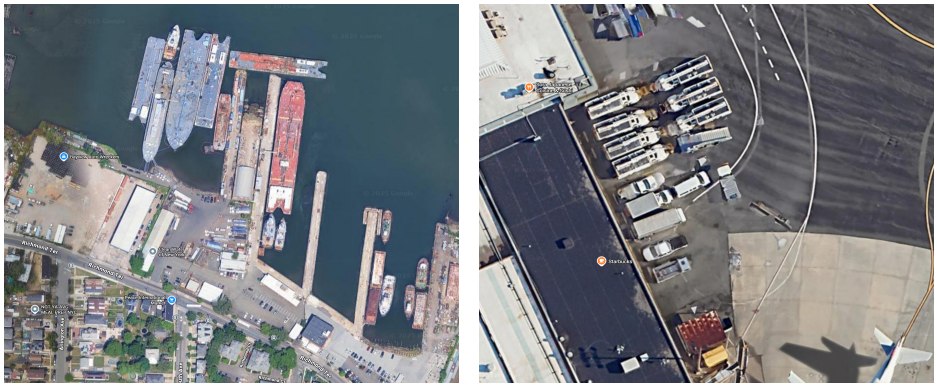
1. Introduction

Fine-grained interpretation of remote sensing images have been attracting increasing attention and are being more widely applied in many fields related to remote sensing monitoring. The core significance of fine-grained recognition and analysis of remote sensing images lies in breaking through the limitations of traditional macro interpretation and realizing accurate cognition and efficient management of the Earth's surface system through sub-class differentiation of ground objects (such as subdivision of tree species in vegetation, Different types objects as in Figure 1, and functional classification in buildings [1,2]). Its value is mainly manifested in three aspects.

Firstly, from the perspective of the data generation layer, fine-grained interpretation has greatly improved the ability to monitor and discover key features in remote sensing applications. Fine-grained recognition is a "precision sensing method" for accurately perceiving surface changes, providing high-quality data input for subsequent governance decisions. For example, in disaster scenarios, through the subdivision of building damage levels (slight cracks or structural collapse [3,4]) and land cover types in flooded areas (farmland or residential areas or roads), "risk heat maps" can be generated in real-time to support the priority allocation of rescue forces; in environmental monitoring, the differential identification of water algal species and the degree of vegetation diseases and pests (mild infection or large-area withering [5,6]) can capture "early signals" that are easily missed by macro monitoring, striving for response time for pollution control and pest prevention; even in the military field, the fine-grained distinction of target types (armored vehicles or ordinary trucks) and camouflage

 doi:10.20944/preprints202509.2113.v1

states (natural vegetation camouflage or artificial camouflage [7,8]) can improve the accuracy of battlefield situation awareness. This ability to capture "quantitative change details" upgrades dynamic monitoring from "discovering changes" to "analyzing change mechanisms", providing a data base for risk prevention and control.



(**a**) Ships (**b**) Vechiles

Figure 1. Examples of Fine-Grained Objects in Remote Sensing Images.

Secondly, fine-grained interpretation supports refined governance and decision-making optimization in the field of remote sensing. Based on fine-grained monitoring data, various governance decisions have been transformed from "extensive" to "precision". For example, in natural resource management, the subdivided data of crop types and growth stages can guide differentiated irrigation and fertilization; the distinction of tree species and health grades can optimize logging plans and ecological restoration schemes. In urban governance, the fine-grained analysis of construction land functions and building attributes provides accurate basis for floor area ratio adjustment and old city reconstruction. Methods for automatic airport detection from remote sensing images, which detect runways, terminals, etc., support precise functional classification of airport infrastructure [9]. Compared with macro classification, fine-grained data can eliminate the drawback of "homogeneous management of similar ground objects"-for example, in the governance of wetlands, if only the general category of wetland is unknown, it is difficult to formulate targeted protection measures, but after subdividing into "swamp wetlands" and "tidal flat wetlands", protection resources can be allocated according to their ecological function differences, making decisions more in line with actual needs.

Finally, fine-grained interpretation drives technological innovation and interdisciplinary integration in remote sensing-related fields. The high requirements of fine-grained analysis have forced the upgrading of the remote sensing technology system and promoted interdisciplinary collaborative innovation. To achieve sub-class differentiation of ground objects, sensor technology is constantly iterated (such as the number of bands of hyperspectral satellites increasing from dozens to hundreds), and algorithm models are continuously optimized (such as deep learning networks with attention mechanisms, which can focus on subtle features of ground objects); at the same time, it relies on interdisciplinary knowledge to build interpretation logic: plant physiology guides the interpretation of vegetation spectral features, urban planning theory constrains the judgment of building functions, ecological principles support the classification of wetland types, and so on. This cycle of "technical demand - disciplinary support - method innovation" not only helps remote sensing interpretation accuracy break through traditional bottlenecks, but also may form an integration paradigm of "remote sensing technology + domain science", further expand the application boundary of remote sensing, and provide new research tools for agriculture, ecology, urban and other fields.

This paper systematically reviews the background, current situation and progress of fine-grained interpretation in the field of remote sensing, with a focus on the basic principles, key methods and specific application targets of fine-grained. This review focuses on summarizing different types of fine-grained remote sensing interpretation methods and conducts in-depth analyses of specific application

 doi:10.20944/preprints202509.2113.v1

scenarios. In addition, this paper also makes a profound summary of the specific challenges brought by the Fine-Grained interpretation method in the field of remote sensing applications, and explores the potential directions for future research at the same time.

2. Main Development Trends

2.1. Journal Distribution

From a journal distribution perspective, top-tier remote sensing publications have become the primary platforms for papers focused fine-grained interpretation of remote sensing images. As shown in Figure 2, between 2015 and 2025, research on fine-grained remote sensing image interpretation has been published widely across a number of high-impact journals in the geoscience and remote sensing domain. The distribution of articles shows several clear patterns:

Leading journals by output: with 455 published papers, IEEE Transactions on Geoscience and Remote Sensing serves as the primary platform for theoretical and technological innovation in this field. Its research covers the core content of the entire fine-grained interpretation chain. It also acts as a key venue for publishing achievements related to theoretical breakthroughs and technological optimization. Remote Sensing (MDPI) has published 246 papers, focusing on multi-directional application exploration and methodological research in fine-grained interpretation. Additionally, it incorporates extensive dataset validation work, providing abundant practical case support for the field. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing has 235 published papers, with a core focus on the application and implementation of fine-grained interpretation technologies. Its research emphasizes the adaptive practice of technologies in specific scenarios and pays attention to fine-grained processing methods for satellite data.

IEEE Geoscience and Remote Sensing Letters has published 113 papers, featuring short-format research. Its content focuses on single-point innovation and preliminary verification in fine-grained interpretation. It also rapidly disseminates cutting-edge innovative ideas in the field. Journal of Applied Remote Sensing has 103 published papers, emphasizing the practical verification of fine-grained interpretation methods, providing references for the engineering application of methods. With 65 published papers, International Journal of Applied Earth Observation and Geoinformation conducts research from a geospatial information perspective. It focuses on publishing achievements related to the integration of remote sensing and geospatial relationships. ISPRS Journal of Photogrammetry and Remote Sensing has 49 published papers, focusing on the integration of photogrammetry and remote sensing technologies, reflecting the in-depth linkage between fine-grained interpretation and traditional surveying and mapping technologies. Remote Sensing of Environment has 30 published papers, focusing on the high-value application of fine-grained interpretation in environmental monitoring.

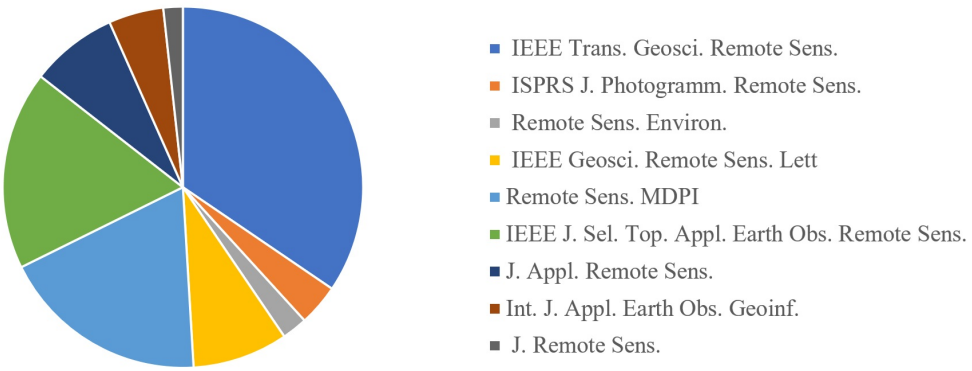


Figure 2. Distribution of Published Articles by Several Core Journals. From 2015 to 2025: IEEE Trans. Geosci. Remote Sens.(455), Remote Sens. MDPI(246), IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.(235), IEEE Geosci. Remote Sens. Lett.(113), J. Appl. Remote Sens.(103), ISPRS J. Photogramm. Remote Sens.(49), Int. J. Appl. Earth Obs. Geoinf.(65), Remote Sens. Environ.(30), J. Remote Sens.(23).

 doi:10.20944/preprints202509.2113.v1

Trends observed: while IEEE and ISPRS outlets remain dominant, the substantial number of publications in IJAEOG and Remote Sensing (MDPI) reflects a move toward more application-oriented and open-access journals, increasing global visibility. High outputs in Remote Sensing (MDPI) and IEEE JSTARS suggest that fine-grained interpretation is increasingly intersecting with computer vision, data science, and earth observation applications. TGRS and RSE continue to publish core theoretical and algorithmic advances, while IJAEOG and MDPI's Remote Sensing emphasize applied and case-driven studies.

2.2. Annually Published Articles

In Figure 3, this graph adopts a combined form of "bar chart + line chart" to intuitively present the changes in the number of published papers in the field of fine-grained remote sensing image interpretation from 2015 to 2025. The horizontal axis represents the time dimension by year, and the vertical axis denotes the annual number of published papers. Among them, the bar chart corresponds to the actual number of papers published each year, while the red line chart is used to fit the overall trend. From the perspective of data distribution, during the period 2015-2023, the annual number of papers showed a steady growth trend with a relatively moderate growth rate, reflecting the gradual development of research in this field at the technical and theoretical levels during this stage. The year 2023 marked a key turning point, after which the number of papers entered a phase of rapid growth and reached a periodic peak in 2024, demonstrating a significant surge in the research enthusiasm for this field. Although the data in the 2025 bar chart is lower than that in 2024, the actual number of published papers in 2025 has not declined because the year 2025 has not yet ended.

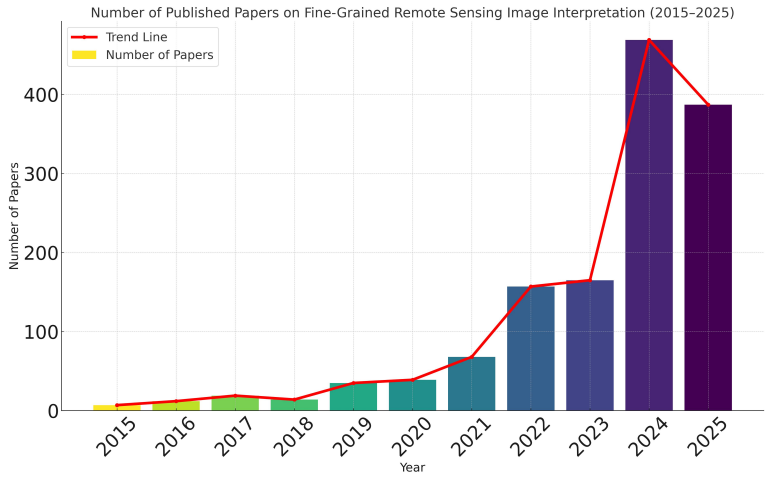


Figure 3. Number of Published Articles Annually in Selected Journals.

In view of the current development trend, it is expected that fine-grained remote sensing image interpretation will remain a research hotspot in 2025 and the coming years. With the continuous development of multi-source data fusion, artificial intelligence (especially deep learning technology), the accuracy and efficiency of interpretation are expected to be further improved. Interdisciplinary research will also deepen, promoting remote sensing technology to move from macro observation to micro fine-grained interpretation and providing core technical support for the digital and intelligent transformation of various industries.

2.3. Keyword Co-Occurrence Network

In Figure 4, this is a keyword co-occurrence network diagram in the field of fine-grained remote sensing interpretation. Nodes of different colors represent different categories: blue for methods/models, red for tasks, green for datasets/sensors, and yellow for applications. The lines between nodes indicate the associations among keywords, while the size of each node reflects the importance or frequency of occurrence of the corresponding keyword. The diagram covers a wide range of keywords,

 doi:10.20944/preprints202509.2113.v1

spanning from data acquisition (e.g., satellite sensors such as WorldView and Landsat), processing methods (e.g., self-supervised learning and graph convolutional networks), tasks (e.g., fine-grained ship recognition and change detection) to multi-domain applications (e.g., agricultural monitoring and disaster monitoring). It presents the complex interconnections among various elements in this field.

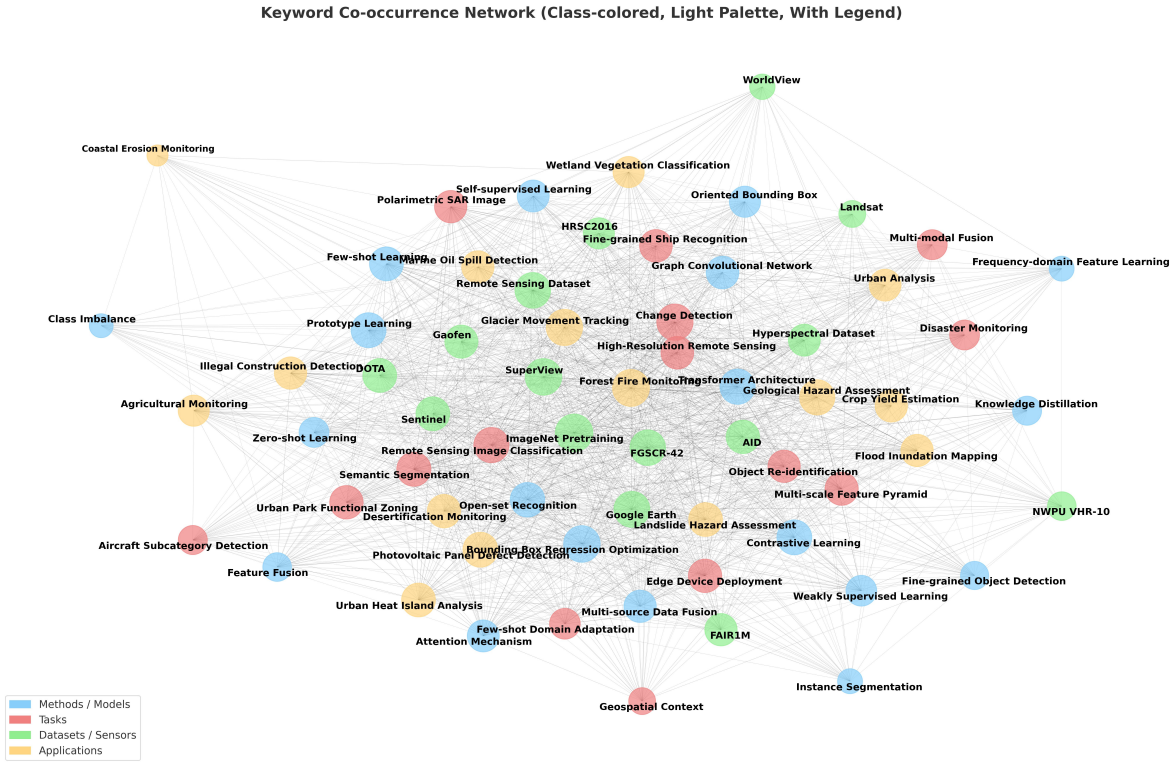


Figure 4. Keyword Co-occurrence Network

In the methods/models category (blue nodes), the largest nodes correspond to deep learning-related technologies—specifically Transformer architecture and contrastive learning. These two keywords not only have the highest frequency of occurrence but also connect to the most other nodes (e.g., linking to "fine-grained object recognition" in tasks and "hyperspectral datasets" in data), becoming the core driving forces of technical innovation in the field.

In the tasks category (red nodes), "fine-grained object recognition" and "change detection" are the largest nodes. Their prominent size and dense connecting lines (e.g., linking to multiple processing methods and application scenarios) confirm that high-precision, detail-oriented interpretation tasks have become the primary research focus.

In the datasets/sensors category (green nodes), "hyperspectral datasets" and "high-resolution satellite data (e.g., WorldView)" are the largest nodes. Their frequent co-occurrence with methods and tasks reflects that multi-source, high-precision data has become the basic support for advancing research, with increasing attention paid to data quality and diversity.

In the applications category (yellow nodes), "agricultural monitoring" and "urban planning" are the most prominent nodes. Their strong associations with core tasks indicate that practical application in key fields is the main orientation of research, and the integration between technical methods and industry needs is becoming increasingly close.

Overall, the field of fine-grained remote sensing interpretation shows a trend of multi-dimensional coordinated development, with clear core nodes standing out in the keyword network. The cross-integration among these core nodes (e.g., Transformer connecting to hyperspectral datasets and fine-grained object recognition) is driving the field toward a more refined, intelligent, and application-oriented direction.

 doi:10.20944/preprints202509.2113.v1

3. The Datasets for Fine-Grained Interpretation

3.1. Current Status of the Dataset

Remote sensing datasets play an extremely important role in the research of fine-grained remote sensing image interpretation. The current fine-grained remote sensing interpretation datasets can roughly be divided into three categories: pixel-level, target(object)-level, and scene-level.

Pixel-level datasets are mostly used for land cover classification and feature change detection, such as TREE [10], Belgium Data [11], and FUSU [12]. The data sources are mostly airborne or ground systems (such as the LiCHy hyperspectral system). Emphasize the subtle distinctions in spectral dimensions, such as the spectral differences among land covers. **Object-level datasets** mainly are constructed for individual targets such as ships, aircraft, and buildings, for example, HRSC2016 [13], FGSCR-42 [14], ShipRSImageNet [15], and MFBFS [16]. They are often high resolution (0.1–6m) remote sensing images. There are a large number of categories, emphasizing the subtle differences between similar categories (such as ship models, aircraft models). The data sources mainly include Google Earth, WorldView, GaoFen series satellites, etc. **Scene-level datasets**, they have the widest coverage and are applied in remote sensing scene classification and retrieval, such as AID [17], NWPU-RESISC45 [18], PatternNet [19], MLRSNet [20], Million-AID [21], and MEET [22], etc. They are with wide resolution range (0.06–153 m). The sample size is large (ranging from tens of thousands to millions of images). The sources mainly include Google Earth, Bing Maps, Sentinel, OpenStreetMap, etc. In Table 1, common datasets for fine-grained remote sensing image interpretation is summarized

Overall, the existing datasets have basically covered typical fine-grained objects and scenarios such as ships, aircraft, buildings, vegetation, and land use/cover, providing important support for related research.

Table 1. Summary of Datasets for Fine-Grained Remote Sensing Image Interpretation

Dataset Name	Resolution	Content	Categories	Total Images	Source
FGSCR-42 [14]	0.1-4.5	Ship	42	9320	GoogleEarth, ISPRS, GanFen etc.
FGSD [23]	0.3-2	Ship	43	4736	GoogleEarth
ShipRSImageNet [15]	0.12-6	Ship	50	3435	WorldView-3, GaoFen-2 etc.
MFBFS [16]	1-4	Building	3	11005	GaoFen-2
UBC [24]	0.5-0.8	Building	61	800	SuperView, GaoFen-2
UBC-v2 [25]	0.5-1	Building	12	11336	SuperView, GaoFen-2, GaoFen-3
DFC2023 [26]	0.5-1	Building	12	300k	SuperView, GaoFen-2, GaoFen-3
MTARSI [27]	0.3-2	Aircraft	20	9598	GoogleEarth
MAR20 [28]	0.3-2	Aircraft	20	3842	GoogleEarth
FAIR1M [29]	0.3-0.8	Air planes, Ships Vehicles, Courts Road	37	15000	Gaofen, GoogleEarth
OpenEarthSensing [30]	0.3-10	Objects and Scenes	189	157674	Different Public Datasets
TREE [10]	0.68	Tree Species	12	-	LiCHy Hyperspectral system
Belgium Data [11]	0.68	Tree Species	7	1450	LiCHy Hyperspectral system
FUSU [12]	0.2-0.5	Land Use Change	17	62752	Google Earth, Sentinel
MEET [22]	2025	Scene	80	1033778	OpenStreetMap
NWPU [18]	0.2-30	Scene	45	31500	GoogleEarth
AID [17]	0.5-8	Scene	30	10000	GoogleEarth
RSD46-WHU [31]	0.5-2	Scene	46	117000	GoogleEarth
MLRSN [20]	0.1-10	Scene	46	109161	GoogleEarth
Million-AID* [21]	0.5-153	Scene	51	10000	GoogleEarth
PatternNet [19]	0.06-4.7	Scene	38	30400	Different Public Datasets
OPTIMAL-31 [32]	-	Scene	31	1860	GoogleEarth, Bing maps
RSI-CB256 [33]	0.3-3	Scene	35	24000	GoogleEarth, Bing maps
RSI-CB128 [33]	0.3-3	Scene	45	36000	GoogleEarth, Bing maps
SR-RSKG [34]	0.2-30	Scene	70	56000	GoogleEarth etc.
Multiscene [35]	0.3-0.6	Scene	36	100000	GoogleEarth, OpenStreetMap
WH-MAVS [36]	1.2	Scene	14	47137	GoogleEarth

3.2. Existing Deficiencies of Datasets

Despite substantial progress, current fine-grained datasets face several limitations.

High intra-class similarity: many fine-grained categories are visually very close, such as different ship or aircraft models, or subtle differences in tree species. This creates severe classification challenges, often leading to model confusion and reduced generalization ability.

 doi:10.20944/preprints202509.2113.v1

Limited modality diversity: most datasets are dominated by optical imagery, whereas multi-modal data integrating SAR, LiDAR, and hyperspectral information are rare. This restricts the ability to fully exploit complementary signals and hampers progress in multi-sensor fusion research.

Geographic imbalance: the majority of existing datasets are constructed from images over China, the United States, and Europe. Large regions, particularly in Africa, South America, and parts of Southeast Asia, remain underrepresented, which reduces global applicability and introduces domain bias.

Annotation bottlenecks: fine-grained annotation requires expert knowledge and is highly time-consuming. As a result, dataset expansion is slow, and some benchmarks are at risk of being "oversaturated," where algorithmic improvements may reflect overfitting to benchmark idiosyncrasies rather than true generalization.

3.3. Future Outlook of Fine-Grained Datasets

Future dataset development for fine-grained remote sensing interpretation is expected to follow several important directions:

1. Multi-modal integration. Most existing benchmarks are dominated by optical imagery, which captures rich spectral and spatial details but is often limited by weather, lighting, and occlusion. To address these challenges, constructing datasets that integrate optical, SAR, LiDAR, and hyperspectral modalities will be critical. SAR can penetrate clouds and provide structural backscatter features, LiDAR captures accurate 3D geometry and elevation information, and hyperspectral imaging offers detailed spectral signatures for material identification. By combining these complementary data sources, future datasets will enable models to recognize fine-grained categories even under challenging conditions (e.g., distinguishing tree species in dense canopies or identifying military targets under camouflage). Multi-modal benchmarks will also foster the development of fusion-based algorithms that better reflect real-world operational requirements.

2. Global coverage and domain diversity. Current datasets are geographically imbalanced, with most samples collected from regions such as China, the United States, and parts of Europe. This geographic bias restricts the generalization ability of models to unseen domains. Expanding datasets to cover diverse climates, cultures, and ecosystems—for instance, tropical rainforests in South America, arid deserts in Africa, or island regions in Oceania—will help mitigate domain bias. In addition, datasets should incorporate varying socio-economic environments (urban, rural, coastal, industrial) to ensure broader representativeness. Such global and cross-domain coverage will make fine-grained datasets more reliable for worldwide applications such as biodiversity monitoring, agricultural assessment, and disaster response.

3. Temporal and dynamic monitoring. Most existing benchmarks are static snapshots, which limits their use for monitoring changes over time. However, many fine-grained tasks are inherently dynamic, such as crop phenology, urban expansion, forest succession, and water resource fluctuation. Incorporating time-series data will allow researchers to capture temporal evolution and model long-term trends. For example, crop species might be indistinguishable at a single time point but reveal distinct spectral or structural patterns when tracked across multiple growth stages. Similarly, urban construction stages or seasonal flooding patterns can only be fully captured in temporal datasets. Building fine-grained time-series benchmarks will thus support more realistic monitoring and predictive modeling tasks.

4. Efficient annotation strategies. The creation of fine-grained datasets is constrained by the costly and time-consuming nature of expert annotations, especially when subtle distinctions (e.g., between aircraft variants or tree species) require domain expertise. To reduce labeling costs, future work should explore weakly supervised learning (using coarse labels or incomplete annotations), self-supervised learning (leveraging large-scale unlabeled imagery), and crowdsourcing platforms that engage non-experts under expert validation. Additionally, incorporating knowledge graphs and generative augmentation can help generate pseudo-labels or synthetic samples to expand datasets efficiently. These strategies will make it feasible to construct large-scale fine-grained benchmarks in a scalable and sustainable way.

 doi:10.20944/preprints202509.2113.v1

5. Open-world and zero-shot benchmarks. In real-world applications, remote sensing systems often encounter novel classes that were not present in the training data. However, most current datasets assume closed-world settings, where the label space is fixed. Future benchmarks should explicitly support open-world recognition and zero-shot learning, where models can detect and reason about unseen categories by leveraging semantic embeddings, textual descriptions, or external knowledge bases. Initiatives such as OpenEarthSensing [30] exemplify this trend, providing benchmarks that require models to generalize to novel classes and handle uncertain environments. Such benchmarks will be vital for practical deployments in tasks like disaster monitoring, where emergent phenomena (e.g., new building types or unusual environmental events) cannot be predefined.

In summary, fine-grained datasets at the pixel,object and scene levels have substantially advanced research in remote sensing interpretation, enriching both the scale and complexity of available benchmarks. Nevertheless, limitations such as high intra-class similarity, modality constraints, geographic imbalance, and annotation costs continue to hinder broader applicability. The future of fine-grained dataset construction will rely on multi-modality, global-scale diversity, temporal dynamics, efficient labeling strategies, and open-world settings, enabling more generalizable, intelligent, and application-ready solutions for remote sensing interpretation.

4. Methodology Taxonomy

Remote sensing image interpretation refers to the comprehensive technical process of analyzing, identifying, and interpreting the spectral, spatial, textural, and temporal characteristics of objects or phenomena in remote sensing images. Essentially, it serves as a "bridge" between remote sensing data and practical Earth observation applications. According to the granularity and objectives of information extraction, it can be divided into three core levels: **pixel-level, object-level, and scene-level**. Each level is interrelated yet has a clear differentiated positioning, while fine-grained interpretation is an in-depth extension of the demand for "subclass distinction" based on these levels.

Pixel-level interpretation, as the foundation of remote sensing interpretation, focuses on the semantic attribution of individual or local pixels. Its core tasks include pixel-level classification (e.g., distinguishing basic ground objects such as farmland, water bodies, and buildings) and semantic segmentation (delineating pixel-level boundaries of ground objects). Traditional methods rely on spectral features (e.g., the low near-infrared reflectance of water bodies) or simple texture features, which are suitable for macro ground object classification in medium- and low-resolution images (e.g., large-scale land use classification). With the development of high-resolution remote sensing technology, pixel-level interpretation has gradually advanced toward "fine-grained attribute distinction." For example, it can distinguish different crop varieties in hyperspectral images and identify building roof materials in high-resolution optical images. This demand for "subclass segmentation under basic ground objects" has become the prototype of fine-grained interpretation at the pixel level.

Object-level interpretation centers on "discrete ground object targets" and requires both spatial localization of targets (e.g., bounding box annotation) and category judgment. Typical applications include ship detection, aircraft recognition, and building extraction. Traditional object-level interpretation focuses on "presence/absence" and "broad category distinction" (e.g., distinguishing "ships" from "aircraft"). However, practical scenarios often require more refined target classification: for instance, ships need to be distinguished into "frigates" and "destroyers," aircraft into "passenger planes" and "military transport planes," and buildings into "historic protected buildings" and "ordinary residential buildings." This type of "subclass identification under the same broad category" has driven object-level interpretation toward fine-grained development, which needs to overcome the technical challenge of "feature confusion between highly similar targets" (e.g., the similar outlines of different ship models).

Scene-level interpretation takes the "entire image scene" as the analysis unit. By integrating pixels, targets, and contextual information, it judges the overall semantics of the scene (e.g., "airport," "port," "urban residential area") and supports regional-scale applications (e.g., urban functional zone division, disaster scene assessment). Traditional scene-level interpretation focuses on "broad scene

 doi:10.20944/preprints202509.2113.v1

category distinction" (e.g., distinguishing "forests" from "cities"). However, refined applications require more detailed scene subclass division: for example, "urban residential areas" need to be subdivided into "high-density high-rise communities" and "low-density villa areas," "wetlands" into "swamp wetlands" and "tidal flat wetlands," and "airports" into "military-civilian joint-use airports" and "civil airports." This "functional/morphological subclass identification under broad scene categories" has become the core demand for fine-grained scene-level interpretation.

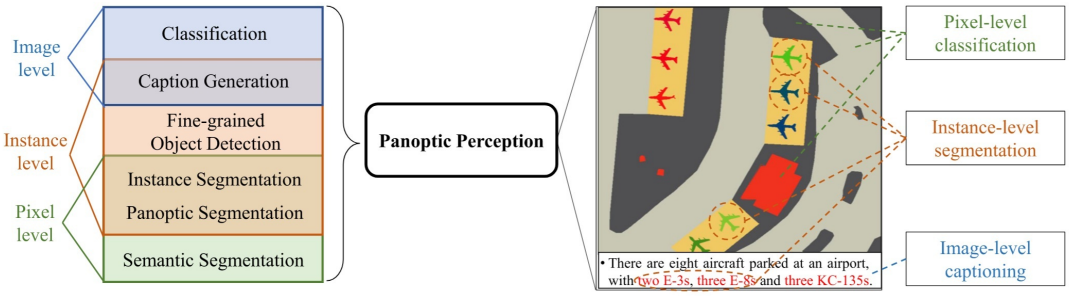


Figure 5. Examples of Fine-Grained Interpretation at Different Levels [37]. It proposed a new "panoptic perception" task, which showed the similar conception of image interpretation at different level. The instance-level is similar to the object-level in this paper.

Traditional interpretation at the three aforementioned levels all have the limitation that "macro categorization cannot meet refined application needs": pixel-level interpretation struggles to distinguish highly similar ground object subclasses (e.g., spectral confusion between different tree species), object-level interpretation fails to identify subdivided types under the same broad category (e.g., different equipment models), and scene-level interpretation cannot divide functionally differentiated scenes (e.g., farmland with different utilization types). Against this backdrop, fine-grained remote sensing image interpretation has emerged.

Fine-grained interpretation is not a new paradigm independent of the three levels, but a technical deepening centered on the goal of "subclass distinction" based on each level. Its core value lies in breaking the bottleneck of semantic ambiguity of ground objects in the same broad category. In the following sections, a comprehensive and in-depth review of the methodologies for these three levels of fine-grained interpretation will be presented.

4.1. Fine-Grained Pixel-Level Classification or Segmentation

The fine-grained remote sensing interpretation at the pixel level mainly includes the classification at the pixel level and the semantic segmentation of the remote sensing images. Generally, there are more studies on pixel-level classification for hyperspectral images and more on semantic segmentation for high-resolution multispectral images. In recent years, research based on spatial-spectral joint classification has become popular. Many classification applications also take advantage of the correlation between pixels and feature consistency, and the boundary between segmentation and classification has gradually blurred. Whether it is classification or segmentation, the challenges faced by pixel-level fine-grained interpretation mainly come from the similarity of the spectral characteristics of subclass pixels within a large category. Some subclasses are even almost indistinguishable in the spectral dimension and can only be distinguished by features such as spatial texture or consistency in the temporal dimension.

 doi:10.20944/preprints202509.2113.v1

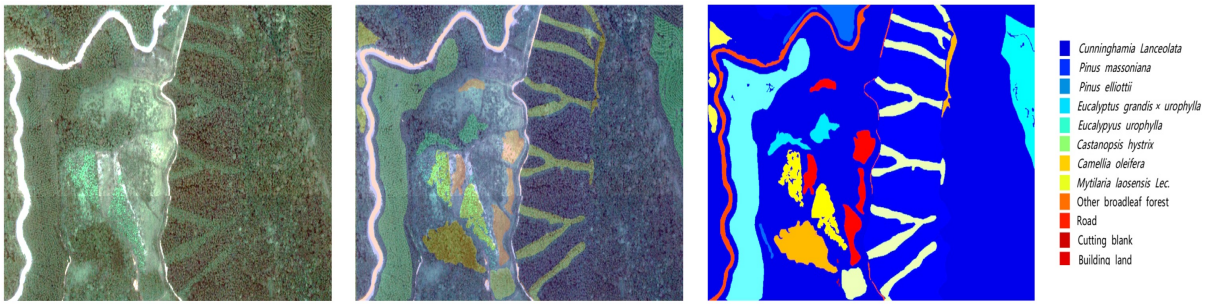


Figure 6. Examples of Fine-Grained Pixel-Level Classification of Remote Sensing Images [10,38].

For example, as in Figure 6 from GSFF dataset [10], which originally has 12 different land-cover classes, containing 9 forest vegetation categories. However, we find that these nine types of vegetation are very difficult to distinguish because their spectra are all very similar. Distinguishing these nine types of vegetation is a typical fine-grained classification problem. In addition, in natural image research like ImageNet, this fine-grained pixel-level classification based on similar spectra is not common. This is one of the significant differences between fine-grained research in the field of remote sensing and traditional computer vision.

Common methods for fine-grained classifying or segmenting pixel-level remote sensing images include **novel data representation methods, coarse-fine category relationship modeling method, multi-source data fusion method, advanced data annotation optimization methods, etc.**

4.1.1. Novel Data Representation

This category of methods focus on breaking through the limitations of traditional and single features from spectrum through innovative feature extraction mechanisms, enabling more accurate capture of key information required for fine-grained classification (such as morphological structures, subtle texture differences, and edge textures etc.)

Spatial-Spectral Joint Representation. Spatial-spectral joint representation is one of the most popular methods in fine-grained classification at pixel-level. For examples, CASST [39] establishes long-range mappings between spectral sequences (inter-band dependencies) and spatial features (neighboring pixel correlations) through a dual-branch Transformer and cross-attention; GRetNet [40] introduces Gaussian multi-head attention to dynamically calibrate the saliency of spectral-spatial features, enhancing the discriminability of fine-grained differences (e.g., spectral peak shifts in closely related tree species); in [41], CenterFormer focuses on the spatial-spectral features of target pixels through a central pixel enhancement mechanism to reduce background interference; E2TNet [42] designs an efficient multi-granularity fusion module to balance global correlations of coarse/fine-grained spatial-spectral features; FGSCNN [43] fuses high-level semantic features with fine-grained spatial details (e.g., edge textures) through an encoder-decoder architecture. Some studies do not simply jointly extract features in the spatial-spectral dimension, but introduce new spatial features such as gradients. For example, in [44], it proposes G2C-Conv3D, which weighted combines traditional convolution with gradient centralized convolution to simultaneously capture pixel intensity semantic information and gradient changes, so that it supplements intensity features with gradient information to improve the model's sensitivity to subtle structures such as edges and textures. Spatial-Spectral Joint Representation break the independent modeling of spectral and spatial features, capturing their intrinsic correlations via mechanisms like attention and Transformer to improve classification accuracy in complex scenes with fine-grained classes.

Morphological Representation. Different from spatial-spectral joint representation, there are also studies exploring new feature extraction methods that are completely different from convolution operations or transform operations to deal with fine-grained classification problems. In [45], the authors propose SLA-NET, which combines morphological operators (erosion and dilation) with trainable structuring elements to extract fine morphological features (e.g., contours and compactness) of tree

 doi:10.20944/preprints202509.2113.v1

crowns; in [46], it designs a dual-concentrated network (DNMF) that separates spectral and spatial information before fusing morphological features to enhance the robustness of tree species classification; morphFormer [47] models the interaction between the structure and shape of trees/minerals through spectral-spatial morphological convolution and attention mechanisms. This kind of method focuses on the geometric morphology of objects (e.g., crown shape and texture distribution), compensating for the inability of traditional convolution to capture non-Euclidean features.

Edge and Area Representation. These methods Focus on edge continuity and regional integrity, addressing the fragmentation of classification results in traditional methods. In [48], PatchOut adopts a Transformer-CNN hybrid architecture and a feature reconstruction module to retain large-scale regional features while restoring edge details, enabling patch-free fine land-cover classification; SSUN-CRF [49] combines a spectral-spatial unified network with a fully connected conditional random field to smooth the edges of classification results and enhance regional consistency; Edge feature enhancement framework (EDFEM+ESM) [50] improves the segmentation accuracy of mineral edges through multi-level feature fusion and edge supervision.

Advantages of Novel Data Representation mainly lies in: 1) Strong fine-grained feature capture: Innovative representation mechanisms accurately capture key information such as morphology, spatial-spectral correlations, gradient changes, and edge textures, significantly improving the discriminability of closely related categories (e.g., tree species and minerals). 2) Flexible model adaptability: Modular designs (e.g., morphological modules and attention modules) can be embedded into mainstream architectures like CNN and Transformer, compatible with diverse scene requirements. Their Limitations: High model complexity: Modules like multi-scale fusion and morphological transformation increase parameter scales and computational loads, imposing strict requirements on training data volume and hardware computing power.

4.1.2. Modeling Relationships Between Coarse and Fine Classes

This category of methods reduces the reliance of fine-grained tasks on annotated data by modeling the hierarchical relationship between coarse-grained categories (e.g., “vegetation”) and fine-grained categories (e.g., “oak” and “poplar”), using prior knowledge of coarse categories to guide fine category classification.

Typical methods are: in [51], it uses GAN and DenseNet, where the generator learns coarse category distributions and the discriminator distinguishes fine category differences to achieve semi-supervised fine-grained classification; coarse-to-fine joint distribution alignment framework [52] matches cross-domain coarse category distributions and then calibrates fine category feature differences through coupled VAE and adversarial learning; CSSD [53] maps patch-level coarse-grained information to pixel-level fine category classification through central spectral self-distillation, solving the “granularity mismatch” problem; CPDIC [54] framework aligns cross-domain coarse-fine category distributions using calibrated prototype loss to enhance domain adaptability; fine-grained multi-scale network [55] combines superpixel post-processing to iteratively optimize fine category boundaries from coarse classification results; CFSSL [56] performs coarse classification with a small number of labels, then uses high-confidence pseudo-labels to guide fine-grained classification of small categories.

Advantages of Modeling Relationships Between Coarse and Fine Classes: 1) High data efficiency: By reusing coarse category knowledge (e.g., spectral commonalities of “vegetation”), the demand for annotated samples for fine-grained categories (e.g., specific tree species) is reduced, making it particularly suitable for few-shot scenarios. 2) Strong generalization ability: Hierarchical modeling mitigates the interference of intra-fine-category variations (e.g., different growth stages of the same tree species) on classification, improving the model’s adaptability to scene changes. The limitations are: 1) Risk of hierarchical bias: Unreasonable definition of hierarchical relationships between coarse and fine categories (e.g., incorrectly classifying “shrubs” as a subclass of “arbor”) can lead to systematic bias in fine-grained classification. 2) Limited cross-domain adaptability: In scenes with severe spectral variation (e.g., vegetation in different seasons), differences in feature distribution between coarse and fine categories may disrupt hierarchical relationships, reducing classification accuracy.

 doi:10.20944/preprints202509.2113.v1

4.1.3. Multi-Source Data Integration

The core of this category of methods is to break through the information dimensional limitations of single-source data by fusing complementary data sources (e.g., hyperspectral and LiDAR, remote sensing and crowdsourced data), thereby improving the robustness and accuracy of fine-grained classification.

The fusion of hyperspectral data with LiDAR data or hyperspectral data with geographic information data is one of the two most common methods for fine-grained classification based on data fusion. In [57] proposes a coarse-to-fine high-order network that fuses spectral features of hyperspectral data and 3D structural information of LiDAR to capture multi-dimensional attributes of land cover through hierarchical modeling; in [58], it designs a multi-scale and multi-directional feature extraction network that integrates spectral-spatial-height features of hyperspectral and LiDAR data to enhance category discriminability in complex scenes; In [59], Sentinel-1 radar images (capturing microwave scattering characteristics of flooded areas) are combined with OpenStreetMap crowdsourced data (providing semantic labels of urban functional zones) to improve the accuracy of fine-grained urban flood detection.

Advantages of Multi-source Data Integration Methods: 1) Information complementarity: Multi-source data provide multi-dimensional information such as spectral, spatial, structural, and semantic, compensating for the lack of discriminability of single-source data in complex scenes (e.g., vegetation coverage and urban heterogeneous areas). It is applicable to diverse scenarios such as forests, cities, and hydrology, especially outstanding in distinguishing fine-grained subcategories (e.g., different tree species and flood-submerged buildings/roads). Their Limitations: Challenges of data heterogeneity: Differences in spatial resolution (e.g., 10m for hyperspectral vs. 1m for LiDAR), coordinate systems, and noise levels among different data sources require complex registration and preprocessing steps, increasing the difficulty of method implementation.

4.1.4. Advanced Data Annotation Strategies.

This category of methods focuses on reducing the reliance of fine-grained classification on large-scale accurately annotated data, optimizing annotation efficiency through strategies such as few-shot learning and semi-supervised annotation, and addressing the practical pain points of "high annotation cost and scarce samples".

The most common approach is to introduce active learning or incremental learning methods into fine-grained remote sensing image classification. LPILC [60] algorithm, based on linear programming, enables incremental learning with only a small number of new category samples without requiring original category data, adapting to dynamically updated classification needs; CSSD [53] uses central spectral self-distillation, taking the model's own predictions as pseudo-labels to reduce dependence on manual annotation; CFSSL [56] screens high-confidence pseudo-labels through "breaking-tie" sampling (BT criterion) to reduce the impact of noisy annotations on the model.

The Advantages of Advanced Data Annotation Strategies are: 1) Significantly reduced annotation cost: these strategies can reduce manual annotation, making them particularly suitable for scenarios requiring professional knowledge for annotation, such as hyperspectral data. Their Limitations are: The performance of few-shot/incremental learning highly depends on the robustness of pre-trained models. If the initial model has biases (e.g., a tendency to misclassify certain categories), it will continuously affect the classification of new categories.

4.2. Fine-Grained Object-Level Detection

In the context of remote sensing, fine-grained object detection refers to the task of not only identifying major target categories such as vehicles, airplanes, and ships, but also distinguishing their more detailed subcategories. As illustrated in the Figure 7, conventional object detection merely recognizes broad categories like Vehicle, Airplane, or Ship. In contrast, fine-grained object detection is able to further differentiate vehicles into Van, Small Car, and Other Vehicle; airplanes into A330, A321,

 doi:10.20944/preprints202509.2113.v1

A220, and Boeing 737; and ships into Tugboat and Dry Cargo Ship, among others. This enables a more precise and detailed recognition and classification of targets in remote sensing imagery.

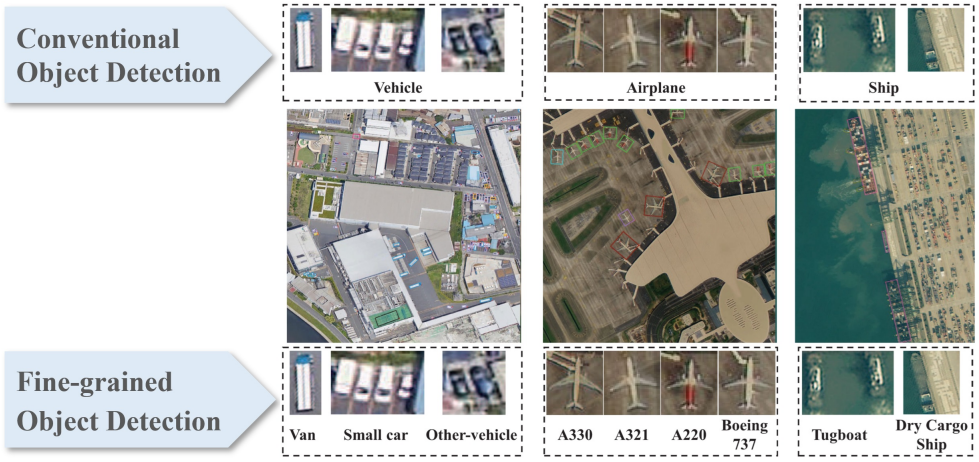


Figure 7. Comparison Between Ordinary Remote Sensing Target Detection and Fine-grained Remote Sensing Target Detection [61].

Object detection, a core task in computer vision, aims to localize and classify objects in images. It has mainly evolved into two dominant paradigms: two-stage detectors and one-stage detectors, each with distinct architectural designs and trade-offs between accuracy and speed. Most of the target detection methods in the field of remote sensing are derived from these two types of methods in the field of computer vision. The following subsections will respectively review and summarize the improvements of the two types of methods (two-stage and one-stage) for fine-grained object detection tasks.

4.2.1. Two-Stage Detectors

Two-stage methods separate object detection into two sequential steps: (1) generating region proposals (potential object locations) and (2) classifying these proposals and refining their bounding boxes. This modular design typically achieves higher accuracy but at the cost of computational complexity.

R-CNN (Region-based Convolutional Neural Networks) [62] introduced the first two-stage framework. It uses selective search to generate region proposals, extracts features via CNNs, and applies SVMs for classification. Despite its pioneering nature, redundant computations make it inefficient. Fast R-CNN [63] addressed R-CNN's inefficiencies by sharing convolutional features across proposals, using a RoI (Region of Interest) pooling layer to unify feature sizes, and integrating classification and regression into a single network. Faster R-CNN [64] revolutionized the field by replacing selective search with a Region Proposal Network (RPN), a fully convolutional network that predicts proposals directly from feature maps. This made two-stage detection end-to-end trainable and significantly faster. Mask R-CNN [65] extended Faster R-CNN by adding a branch for instance segmentation, demonstrating the flexibility of two-stage architectures in handling complex tasks beyond detection. Cascade R-CNN [66] improved bounding box regression by iteratively refining proposals with increasing IoU thresholds, addressing the mismatch between training and inference in standard two-stage methods.

R-CNN-based methods rely on a two-stage framework to address the core challenge of distinguishing highly similar targets (e.g., aircraft subtypes, ship models) in remote sensing images. In the research of fine-grained object detection, the two-stage structure is more popular than the one-stage structure. Two-stage object detection architectures can be further decomposed into a feature extraction backbone (Backbone) with a feature pyramid network (FPN), a region proposal network (RPN) for candidate regions, a region of interest alignment module (RoIAlign) for precise feature mapping, and task-specific heads for object classification (Cls), bounding box regression (Reg), and optional mask

 doi:10.20944/preprints202509.2113.v1

prediction (Mask Branch). Table 2 summarizes the improvements of fine-grained object detection on these components. Below is a detailed analysis of their improvements with different methods.

Table 2. Summary of the Improvements of Fine-grained Object Detection on Different Components.

Method	Backbone (+FPN)	RPN	RoIAlign	Bbox Cls/Reg	Mask Branch	Purpose	Reference
EIRNet	✓	✓	×	✓	×	Bidirectional feature fusion via DFF-Net; Optimize proposals with Mask-RPN (reuse attention mask); Mine interclass relations for ship fine-grained Cls	[67]
MRAN	✓	✓	✓	✓	✓	Fuse RGB/multispectral/LiDAR features; Proposal generation via attention scores; Optimize RoI sampling for small trees; Multisource feature-driven Cls/Reg; Refine tree canopy segmentation	[68]
PCLDet	✓	✓	×	✓	✓	Prototype learning for fine-grained features; Class-balanced sampler (CBS) for long-tail data; ProtoCL loss for Bbox Cls/Reg; Prototype constraint for Mask segmentation	[61]
SAM (Shape-Aware Model)	✓	✓	✓	✓	✓	Shape-aware Conv for large-aspect-ratio ships; Dynamic anchor adjustment; RoI sampling optimization for deformed ship parts; Shape loss for Bbox fitting; Shape-constrained Mask for ship-background distinction	[69]
HCP-Mask-RCNN	✓	✓	✓	✓	✓	Frequency-aware (FARS) module for detail features; Fine-grained proposal prioritization; RoI alignment with frequency features; Coarse-fine hierarchy (HCP) for Cls; Frequency-guided Mask for fine structures	[70]
Oriented R-CNN	✓	✓	×	✓	×	Oriented feature enhancement via FPN; Oriented proposal generation; Geospatial object localization/Cls; Serve as teacher network for knowledge distillation	[71]
ISCL-Mask-RCNN	✓	×	×	✓	×	Contrastive learning (CLM) to widen interclass distance; Refined instance switching (ReIS) for class imbalance; Improve airplane fine-grained detection (HBB/OBB)	[72]
PETDet	✓	✓	✓	✓	×	Anchor-free QOPN for high-quality proposals; Bilinear channel fusion (BCFN) for RoI features; Adaptive recognition loss (ARL) for Cls/Reg; Focus on fine-grained target distinction	[73]
SFRNet	✓	✓	✓	✓	✓	SC-Former for spatial-channel interaction; OR-Former for rotation-sensitive features; Multi-RoI loss (MRL) for Cls; Separate feature refinement for Cls/segmentation	[74]
MGANet	✓	✓	✓	✓	×	Local-global alignment (LAM) for ship features; Multigranularity self-attention (MSM) for fusion; RoIAlign optimization for local ship parts; Improve dense ship fine-grained Cls/Reg	[75]
FineShipNet	✓	✓	✓	✓	×	Blend synchronization module for feature reuse; Polarized feature focusing for task decoupling; Adaptive harmony anchor labeling; RoIAlign for ship discriminative features (Cls/Reg)	[76]
HMS-Net	✓	✓	✓	✓	×	Multiscale region feature re-extraction; Top-down feature fusion with guidance; Hierarchical loss for interclass relations (Cls); RoIAlign for ship fine-grained features	[77]
GFA-Net	✓	✓	✓	✓	✓	Graph focusing process (GFP) for structural features; Graph aggregation network (GAN) for node weight; RoIAlign for invariant structure features (Cls/Reg); Mask segmentation for object structure preservation	[78]
DIMA	✓	✓	✓	✓	✓	FARS module for frequency-domain features; Hierarchical classification (HCP) for Cls; RoI alignment with frequency details; Mask refinement for fine target structures	[70]

Note: ✓ = The method involves improvements in this stage; × = The method does not involve improvements in this stage. Abbreviations: Bbox Cls/Reg = Bounding Box Classification and Regression; Cls = Classification; Reg = Regression; FPN = Feature Pyramid Network; RPN = Region Proposal Network; RoIAlign = Region of Interest Align; MRAN = Multisource Region Attention Network; SAM = Shape-Aware Model; QOPN = Quality-Oriented Proposal Network; BCFN = Bilinear Channel Fusion Network; FARS = Frequency-Aware Representation Supplement.

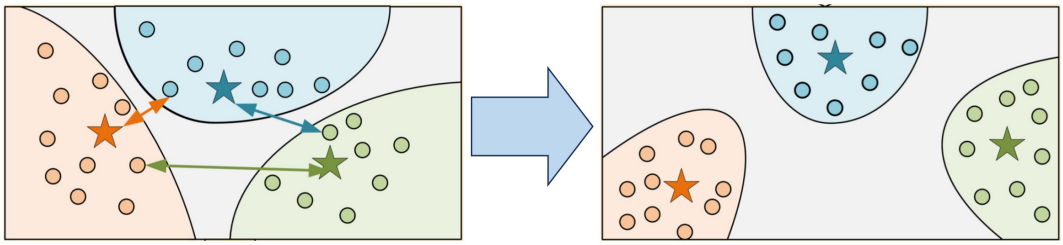


Figure 8. Prototypical Contrast Learning [79].

 doi:10.20944/preprints202509.2113.v1

Contrastive Learning. This subcategory focuses on optimizing the feature space of highly similar targets through inter-sample contrast to amplify inter-class differences and reduce intra-class variations. Its core logic is to construct positive/negative sample pairs and use contrastive loss to guide the model in learning discriminative features, which is particularly effective for scenarios where visual similarity leads to feature confusion.

Existing studies in this subcategory (Contrastive Learning) mainly include: To address insufficient feature discrimination caused by long-tailed distributions, [61] proposed PCLDet, which builds a category prototype library to store feature centers of targets (e.g., ships, aircraft) and introduces Prototypical Contrastive Loss (ProtoCL) to maximize inter-class distances while minimizing intra-class distances. A Class-Balanced Sampler (CBS) further balances sample distribution, ensuring that rare subtypes receive sufficient attention. For the problem of intra-class diversity in fine-grained aircraft detection, [72] designed an Instance Switching-Based Contrastive Learning method. The Contrastive Learning Module (CLM) uses InfoNCE+ loss to expand the feature gap between aircraft subtypes (e.g., passenger aircraft models), while the Refined Instance Switching (ReIS) module mitigates class imbalance and iteratively optimizes features of discriminative regions (e.g., wings, engines). For oriented highly similar targets (e.g., ships), [79] combined Oriented R-CNN (ORCNN) with Adaptive Prototypical Contrastive Learning (APCL). The Spatial-Aligned FPN (SAFPN) solves the spatial misalignment issue of traditional FPN, providing high-quality feature inputs for contrastive learning, and significantly improves the separability of features for ship subtypes (e.g., frigates vs. destroyers) on datasets such as FGSD and ShipRSImageNet. Unknown ship detection via memory bank and uncertainty reduction [80] proposed a method that uses a Class-Balanced Proposal Sampler (CBPS) to balance sample learning and a Fine-Grained memory bank-based Contrastive Learning (FGCL) strategy to separate known/unknown ships. The Uncertainty-Aware Unknown Learner (UAUL) module reduces prediction uncertainty, solving the misjudgment of unknown highly similar ships (e.g., new military ships).

Knowledge Distillation. This subcategory aims to balance detection accuracy and model efficiency by transferring fine-grained knowledge from complex "teacher models" to lightweight "student models." It has expanded from traditional multi-model distillation to self-distillation, enabling knowledge reuse within a single model and adapting to scenarios such as lightweight deployment and few-shot learning.

The technical evolution of this subcategory (Knowledge Distillation) is reflected in three directions: Multi-teacher knowledge distillation for accuracy-efficiency balance [71] used oriented R-CNN as the first teacher to locate vehicles/ships and Coarse-to-Fine Object Recognition Network (CF-ORNet) as the second teacher for fine-grained recognition. By distilling knowledge from both teachers into a student model and combining filter grafting, the model achieves high accuracy on high-resolution remote sensing images while reducing computational costs. Decoupled distillation for lightweight underwater detection [81] proposed the Prototypical Contrastive Distillation (PCD) framework, which uses R-CNN as the teacher model to transfer fine-grained knowledge of underwater targets (e.g., submersibles) via prototypical contrastive learning. The decoupled distillation mechanism allows the student model to focus on discriminative features, and contrastive loss enhances semantic structural attributes, improving the robustness of lightweight models in underwater environments. Self-distillation for few-shot scenarios [82] proposed Decoupled Self-Distillation for fine-grained few-shot detection. The model uses its "high-confidence branch" as an implicit teacher and "low-confidence branch" as a student to transfer knowledge of rare highly similar subtypes (e.g., rare aircraft models). Combined with progressive prototype calibration, this method addresses the problem of insufficient knowledge transfer due to limited data in few-shot scenarios.

Hierarchical Feature Optimization and Highly Similar Feature Mining (HFOSFM). This subcategory follows the logic of "from low-level feature purification to high-level feature fusion" to iteratively improve feature quality, with the ultimate goal of mining subtle discriminative features of highly similar targets. Low-level optimization focuses on eliminating noise (e.g., background interference,

 doi:10.20944/preprints202509.2113.v1

posture misalignment), while high-level optimization emphasizes integrating semantic information to enhance feature completeness.

Key innovations across these HFOSFM studies include: Low-level noise filtering and high-level feature matching [73] proposed PETDet, which uses the Quality-Oriented Proposal Network (QOPN) to generate high-quality oriented proposals (low-level purification) and the Bilinear Channel Fusion Network (BCFN) to extract independent discriminative features for proposals (high-level refinement). Adaptive Recognition Loss (ARL) further guides the R-CNN head to focus on high-quality proposals, solving the mismatch between proposals and features for highly similar targets. Multi-domain feature fusion and semantic association construction [70] proposed DIMA, which synchronously learns image and frequency-domain features via the Frequency-Aware Representation Supplement (FARS) mechanism (low-level detail enhancement) and builds coarse-fine feature relationships using the Hierarchical Classification Paradigm (HCP) (high-level semantic integration). This approach effectively amplifies structural differences between highly similar samples (e.g., ships of different tonnages). For oriented targets (e.g., rotating ships), [74] proposed SFRNet, which uses the Spatial-Channel Transformer (SC-Former) to correct feature misalignment caused by posture variations (low-level spatial interaction) and the Oriented Transformer (OR-Former) to encode rotation angles (high-level semantic supplementation). This ensures that local differences (e.g., wing angles of tilted aircraft) are fully captured.

Category Relationship Modeling and Similarity Measurement Optimization (CRMSMO). This subcategory explicitly models intrinsic relationships between categories (e.g., hierarchical, structural, or functional relationships) to optimize similarity measurement logic, addressing the issue where traditional methods fail to distinguish highly similar targets due to over-reliance on visual features.

Representative studies of CRMSMO are: Semantic decoupling and anchor matching optimization [76] proposed a method for fine-grained ship detection that decouples classification and regression features using a polarized feature focusing module and selects high-quality anchors via adaptive harmony anchor labeling. By optimizing the matching between anchors and category features, it improves the localization accuracy of highly similar ships. Hierarchical relationship constraint and feature distance expansion [77] proposed HMS-Net, which reinforces features at different semantic levels (e.g., ship contours vs. local components) and uses hierarchical relationship constraint loss to model the semantic hierarchy of ship subtypes (e.g., destroyer models). This explicitly expands the feature distance between highly similar subcategories. Invariant structural feature extraction via graph modeling [78] proposed Invariant Structure Representation, which uses the Graph Focusing Process (GFP) module to extract invariant structural features (e.g., cross-shaped aircraft, rectangular vehicles) based on graph convolution. The Graph Aggregation Network (GAN) updates node weights to enhance structural feature expression, enabling the model to distinguish visually similar targets by their inherent structural relationships. Shape-aware modeling for large aspect ratio targets [69] addressed the high similarity and large aspect ratio of ships in high-resolution satellite images by designing a Shape-Aware Feature Learning module to alleviate feature alignment bias and a Shape-Aware Instance Switching module to balance category distribution. This ensures sufficient learning of rare ship subtypes (e.g., special operation ships).

Multi-Source Feature Fusion and Context Utilization. This subcategory compensates for the lack of discriminative information caused by visual similarity by fusing multi-modal data (e.g., RGB, multispectral, LiDAR) and leveraging contextual relationships. It is particularly effective for scenarios where single-modal features are insufficient to distinguish highly similar targets (e.g., street tree subtypes). For example, in [68], it proposed a multisource region attention network that fuses RGB, multispectral, and LiDAR data. A multisource region attention module assigns weights to features of highly similar street tree subtypes, using multi-modal differences (e.g., spectral reflectance, elevation information) to supplement the information gap caused by visual similarity. This approach significantly improves the fine-grained classification accuracy of street trees in remote sensing imagery. Few-shot aircraft detection via cross-modal knowledge guidance [83] proposed the TEMO method, which introduced text-modal descriptions of aircraft and fused text-visual features via a cross-modal assembly

 doi:10.20944/preprints202509.2113.v1

module. This reduces confusion between new categories and known similar aircraft, enabling fine-grained recognition in few-shot scenarios based on the R-CNN two-stage framework.

4.2.2. One-Stage Detectors

Single-stage detectors (such as the YOLO series) omit the separation steps of candidate region generation and subsequent classification, and directly perform category prediction and bounding box regression on the feature map. This end-to-end structure significantly reduces model complexity and inference latency, thereby enabling real-time detection capabilities. Although early single-stage methods generally lagged behind two-stage detectors in terms of accuracy, YOLO et al. have made many improvements and achieved significant enhancements in aspects such as network structure optimization, loss function improvement, and the introduction of feature enhancement modules.

YOLO (You Only Look Once) [84] pioneered one-stage detection by treating object detection as a regression task. It divides the image into a grid, with each grid cell predicting bounding boxes and class probabilities, enabling real-time performance. SSD (Single Shot MultiBox Detector) [85] introduced multi-scale feature maps to detect objects of varying sizes, using default bounding boxes (anchors) at different layers to improve small object detection. RetinaNet [86] addressed the class imbalance issue in one-stage detectors with Focal Loss, a modified cross-entropy loss that down-weights easy background examples. This closed the accuracy gap with two-stage methods. YOLOv3 [87] enhanced the original YOLO with multi-scale prediction, a more efficient backbone (Darknet-53), and better class prediction, balancing speed and accuracy. EfficientDet [88] optimized both accuracy and efficiency through compound scaling (co-scaling depth, width, and resolution) and a weighted bi-directional feature pyramid network (BiFPN), achieving state-of-the-art results on COCO. YOLOv7 [89] introduced trainable bag-of-freebies (e.g., ELAN architecture, model scaling) and bag-of-specials (e.g., reparameterization) to boost performance, outperforming previous YOLO variants and other one-stage detectors on speed-accuracy curves.

Methods based on YOLO can be structurally decomposed into Backbone, Neck, and Head. Table 3 summarizes the improvements of existing fine-grained object detection approaches with respect to these decomposed components.

These methods can also be broadly categorized into four groups: data and input augmentation-driven, attention and feature fusion-driven, discriminative learning and task design-driven, and optimization and post-processing-driven. Each direction addresses different technical aspects, yet they share the common goal of enhancing the ability to distinguish visually similar targets and to improve the detection of small objects in complex remote sensing scenes.

Data and Input Augmentation-Driven Methods. This category mainly focuses on enriching input data and sample representation, alleviating challenges of limited training samples and class imbalance in remote sensing. For instance, the improved YOLOv7-Tiny [90] applies multi-scale/rotation augmentation to expand input sample diversity; Lightweight FE-YOLO [91] optimizes input by preprocessing input data to highlight fine-grained features of small targets; YOLOv8 (G-HG) [92], adjusts input feature resolution to match multi-scale remote sensing targets; YOLO-RS [93] adopts context-aware input sampling to focus on crop fine-grained regions; YOLOX-DW [94] applies adaptive sampling to balance the distribution of fine-grained classes in input data . Moreover, DETet [95] and MFL [96] explore image degradation recovery and super-resolution enhancement, offering new approaches to restore fine details in low-quality remote sensing images. These studies highlight that input-level improvements not only enhance robustness but also provide stronger foundations for fine-grained discrimination.

Attention and Feature Fusion-Driven Methods. Methods in this category emphasize enhancing discriminative feature representations by leveraging attention mechanisms and multi-scale fusion. For example, FGA-YOLO [97] and SR-YOLO [98] combine global multi-scale modules, bidirectional FPNs, and super-resolution convolutions to strengthen fine-grained representation of aircraft and UAV targets. WDFA-YOLOX [99] and YOLOv5+CAM [100] address SAR feature loss and wide-area vehicle detection through wavelet-based compensation and attention mechanisms. IF-YOLO [101] and FiFoNet [102] improve feature pyramid and fusion strategies to preserve small-object features and

 doi:10.20944/preprints202509.2113.v1

suppress background noise. These works demonstrate that precise feature modeling under complex backgrounds and scale variations is crucial for fine-grained detection.

Table 3. Summary of One-Stage Methods (YOLO) for Fine-Grained Object Detection

Method	Input Stage	Backbone	Neck	Head	Purpose	Reference
FGA-YOLO	×	✓	✓	✓	Aggregate multi-layer features to enhance multi-scale information; Extract key discriminative features to improve fine-grained recognition; Alleviate imbalance between easy/hard samples via EMA Slide Loss	[97]
SR-YOLO	×	✓	✓	✓	Extract small-target fine-grained features via SR-Conv module; Enhance small-target feature fusion with bidirectional FPN; Improve detection accuracy via Normalized Wasserstein Distance Loss	[98]
IF-YOLO	×	✓	✓	×	Preserve small-target intrinsic features via IPFA module; Suppress conflicting information with CSFM; Fuse multi-scale features via FGAFPN	[101]
WDFA-YOLOX	×	✓	✓	✓	Compensate SAR fine-grained feature loss via WSPP module; Enhance small-ship features with GLFAE; Improve bounding-box regression via Chebyshev distance-GIoU Loss	[99]
Related-YOLO	×	✓	✓	×	Model ship component geometric relationships via relational attention; Adapt to rotated ships with deformable convolution; Optimize anchors via hierarchical clustering	[103]
YOLOv5+CAM	×	✓	✓	✓	Capture key regions via CAM attention module; Fuse multi-scale features with CAM-FPN; Enhance training via coarse-grained judgment + background supervision	[100]
FiFoNet	×	✓	✓	×	Capture global-local context via GLCC module; Select valid multi-scale features to block redundant information; Improve small-target detection in UAV images	[102]
FD-YOLOv8	×	✓	✓	×	Preserve aircraft local details via local feature module; Enhance local-global interaction via focus modulation; Improve fine-grained accuracy in complex backgrounds	[104]
YOLOX (GTDet)	×	✓	✓	✓	Adapt to oriented targets via GCOTA label assignment; Improve angle prediction via DLAAH; Enhance localization via anchor-free detection	[105]
DEDet	×	✓	✓	×	Restore nighttime details via FPP module; Filter background interference via progressive filtering; Improve nighttime UAV target detection	[95]
MFL	×	✓	✓	×	Realize SR-OD mutual feedback via MFL closed-loop; Focus on ROI details via FROI module; Narrow target feature differences via MSOI	[96]
InterMamba	×	✓	✓	✓	Capture long-range dependencies via VMamba backbone; Fuse multi-scale features via cross-VSSM; Optimize dense detection via UIL loss	[106]
Improved YOLOv7-Tiny	✓	✓	×	×	Construct diverse remote sensing aircraft dataset; Apply multi-scale/rotation augmentation to enrich input samples	[90]
Lightweight FE-YOLO	✓	✓	✓	×	Preprocess input data to highlight small-target fine-grained features; Reduce input noise interference via similarity-based channel screening; Optimize input feature distribution for remote sensing scenarios	[91]
YOLOv8 (G-HG)	✓	✓	✓	×	Adjust input feature resolution to match multi-scale remote sensing targets; Retain fine-grained details in input via redundant feature map sampling; Optimize input data utilization for complex background scenarios	[92]
YOLO-RS	✓	✓	✓	✓	Adopt context-aware input sampling to focus on crop fine-grained regions; Balance input class distribution via AC mix module	[93]
YOLOX-DW	✓	✓	×	×	Apply adaptive sampling to balance fine-grained class distribution in input; Optimize input sample selection to avoid rare class underrepresentation	[94]

Note: ✓ = The method involves improvements in this YOLO stage; × = The method does not involve improvements in this YOLO stage.

Discriminative Learning and Task Design-Driven Methods. This research line emphasizes introducing additional discriminative constraints or multi-task mechanisms to improve the separation of visually similar categories. FD-YOLOv8 [104] captures subtle differences in aircraft through local detail modules and focus modulation mechanisms. Related-YOLO [103] leverages relational attention, hierarchical clustering, and deformable convolutions to model structural relations between ship components. GTDet [105] enhances classification-regression consistency for oriented objects using optimal transport-based label assignment and decoupled angle prediction. MFL [96] builds a closed-loop between detection and super-resolution, guiding degraded images to recover discriminative details. Overall, these methods contribute discriminative signals by focusing on local part modeling, relational learning, and multi-task integration.

Optimization and Post-Processing-Driven Methods. This category centers on loss function design and post-processing optimization, improving adaptation to fine-grained targets during both training and inference. WDFA-YOLOX [99] and SR-YOLO [98] introduce novel regression losses (Chebyshev distance-IoU and normalized Wasserstein distance) to improve bounding box localization for small objects. GTDet [105] applies optimal transport-based assignment to address the scarcity of positive samples for oriented objects with large aspect ratios. SA-YOLO [107] dynamically adjusts

 doi:10.20944/preprints202509.2113.v1

class weights with adaptive loss functions to mitigate bias from data imbalance. DETet [95] employs iterative filtering during post-processing to suppress noise and false positives in night-time UAV imagery. These strategies demonstrate that careful optimization and post-processing not only stabilize training but also ensure the preservation of small and fine-grained targets during inference.

Overall, YOLO-based fine-grained detection research in remote sensing has established a comprehensive improvement pathway spanning input augmentation, feature modeling, discriminative learning, and optimization strategies. Data and input enhancements improve baseline robustness, attention and feature fusion strengthen discriminative representations, discriminative learning and task design introduce novel supervision signals, and optimization and post-processing ensure stability and reliability across stages. Future trends are expected to further integrate these directions, such as combining input augmentation with discriminative learning, or unifying feature modeling and optimization strategies into an end-to-end framework, to comprehensively improve fine-grained detection performance in remote sensing.

4.2.3. Other Methods for Fine-Grained Object Detection

In the field of fine-grained object detection in remote sensing, aside from YOLO and RCNN-based methods, existing studies can be broadly categorized into four classes: methods based on Transformer/DETR, classification/recognition networks, customized approaches for special modalities or scenarios, and graph-based or structural feature modeling methods. These approaches address challenges such as category ambiguity, feature indistinctness, and complex scene conditions from different perspectives, including global feature modeling, fine-grained feature optimization, environment-specific adaptation, and structural information exploitation.

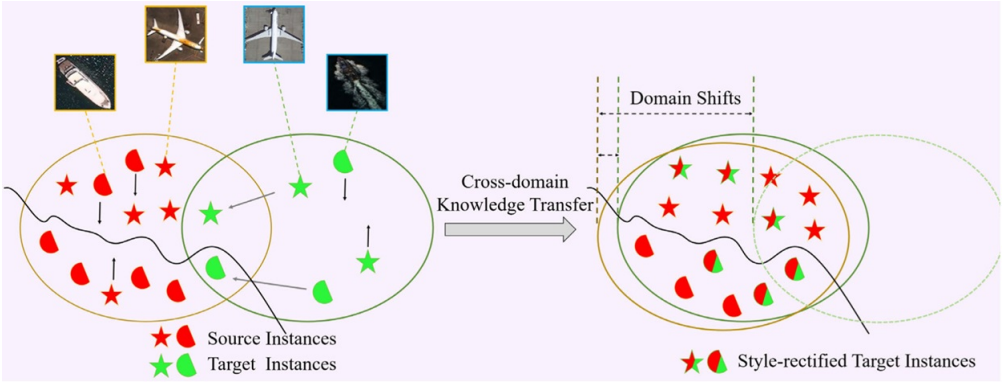


Figure 9. FSDA-DETR [108]. Domain shift is observed between the source and target domains when the given target-domain data is scarce.

Transformer/DETR-Based Methods. Transformer and DETR-based methods leverage self-attention mechanisms to model global feature dependencies, enabling the capture of fine-grained target relationships across the entire image and supporting end-to-end detection. Typical studies include: FSDA-DETR [108] employs cross-domain style alignment and category-aware feature calibration to achieve effective adaptation in optical-SAR cross-domain few-shot scenarios; GMODet [109] integrates region-aware and semantic-spatial progressive interaction modules within the DETR framework to capture spatio-temporal correlations of ground-moving targets, enabling efficient detection in large-scale remote sensing images; InterMamba [106] combines cross-visual selective scanning with global attention and user interaction feedback to optimize detection and annotation in dense scenes, enhancing discriminability in crowded environments.

CNN-Based Feature Interaction and Classification (Non-YOLO/RCNN). This category primarily focuses on fine-grained object classification or recognition. Most methods do not involve explicit detection or YOLO/RCNN structures(without localization modules), but rely on feature optimization, data augmentation, and feature purification to improve category discriminability. A few approaches (Context-Aware method [110]) incorporate lightweight localization modules in addition to classifi-

 doi:10.20944/preprints202509.2113.v1

cation. Representative works include [111], which combines CNN features with natural language attributes for zero-shot recognition; [112], which uses region-aware instance modeling and adversarial generation to mitigate inter-class similarity; EFM-Net [113], which leverages feature purification and data augmentation to enhance fine-grained characteristics; [114], integrating weak and strong features to iteratively optimize discriminative regions in low-resolution images; [115], proposing a coarse-to-fine hierarchical framework for urban village classification; and [116], which uses feature decoupling and pyramid transformer encoding to distinguish visually similar targets in UAV videos. Overall, these methods emphasize enhancing classification capability under limited or ambiguous feature conditions.

Customized Methods for Special Modalities or Scenarios. These methods target fine-grained object detection under specific modalities (e.g., thermal infrared, underwater) or challenging scenarios (e.g., low-light, night-time), optimizing feature extraction and localization through specialized modules. Typical studies include: U-MATIR [117] constructs a multi-angle thermal infrared dataset and leverages heterogeneous label spaces with hybrid view cascade modules to enable efficient detection of thermal infrared targets; DEDet [95] employs pixel-level exposure correction and background noise filtering to improve feature quality and detection performance under low-light UAV imagery; PCD method [81] uses prototype contrastive learning and decoupled distillation to transfer features and lighten models for underwater fine-grained targets, enhancing overall detection performance.

Graph-Based or Structural Feature Modeling Methods. Graph-based methods model structural relationships among target components, reinforcing classification and localization through structural consistency.Typical studies include: GFA-Net [78] employs a graph-focused aggregation network to model structural features and node relations, achieving precise detection of structurally deformed targets; In [118], it integrates geospatial priors with frequency-domain analysis to infer the distribution and class relationships of aircraft in large-scale SAR images, enabling efficient localization.

Overall, the three categories of fine-grained object detection methods form a complementary technical system targeting the core challenge of "high inter-class similarity": R-CNN-based methods mainly achieve high precision through specialized technical paths (contrastive learning, knowledge distillation, hierarchical feature optimization) and are suitable for complex scenarios (few-shot, unknown categories); YOLO-based methods mainly prioritize efficiency via multi-scale fusion and attention mechanisms, making them ideal for real-time scenarios (UAV, SAR); Other methods break through traditional frameworks to address special scenarios (cross-domain, nighttime, zero-shot), providing innovative supplements.

4.3. Fine-Grained Scene-Level Recognition

Fine-grained scene-level recognition is playing an increasingly important role in remote sensing applications, where distinguishing subtle differences between visually similar scenes has become more complex and challenging. Many studies on scene understanding in remote sensing images draw on methods and models from the field of computer vision. In the field of computer vision, image scene understanding generally follows two fundamental paradigms: bottom-up and top-down approaches.

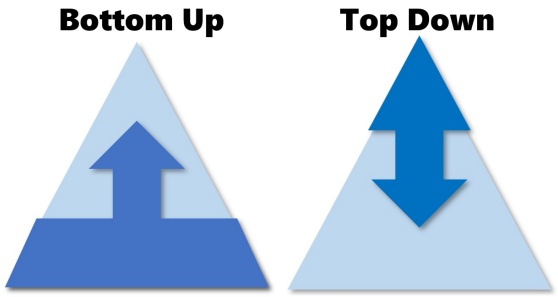
Bottom-up methods start from pixels and low-level features, progressively extracting textures, shapes, and spectral information, and then aggregating them into high-level semantics through deep neural networks, as in Figure 10. Their advantages lie in being data-driven, well-suited for large-scale imagery, and capable of automatic feature learning with good transferability. However, they often lack high-level semantic constraints, making them vulnerable to intra-class variability and complex backgrounds, which may lead to insufficient semantic interpretability.

Top-down methods, in contrast, begin with task objectives or prior knowledge, employing geographic knowledge graphs, ontologies, or semantic rules to guide and constrain the interpretation of low-level features, as in Figure 10. These approaches have the strengths of semantic clarity and interpretability, aligning more closely with human cognition. Their limitations, however, include dependence on high-quality prior knowledge, high construction costs, and limited scalability in large-scale automated tasks.

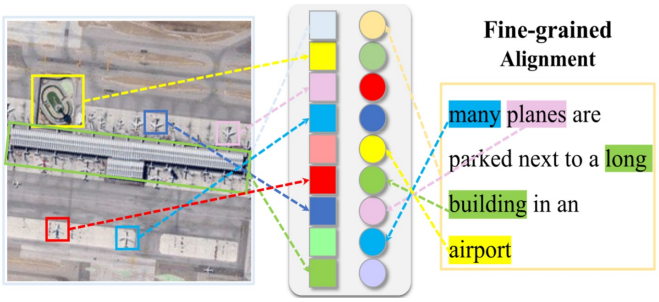
 doi:10.20944/preprints202509.2113.v1

In terms of research trends, bottom-up methods dominate the current literature. In fine-grained remote sensing scene understanding in particular, most studies rely on multi-scale feature modeling, attention mechanisms, convolutional neural networks, and Transformer architectures to capture subtle inter-class differences through hierarchical abstraction. These approaches are well-suited to large-scale data-driven training and have therefore become the mainstream. By contrast, top-down methods are mainly explored in knowledge-based scene parsing, cross-modal alignment, and zero-shot learning, and remain relatively limited in number, though they show promise for enhancing semantic interpretability and cross-domain generalization.

In summary, fine-grained remote sensing scene understanding is currently almost exclusively driven by bottom-up feature learning approaches, while top-down methods remain at an exploratory stage. This paper mainly reviews two studies on the scene level of remote sensing images: scene classification and image retrieval. Most of them are bottom-up fine-grained image recognition or understanding.



(**a**) Bottom-Up vs. Top-Down



(**b**) Example of Scene-level Recognition (Bottom-Up) [119]

Figure 10. Paradigms and Example of Scene-Level Recognition.

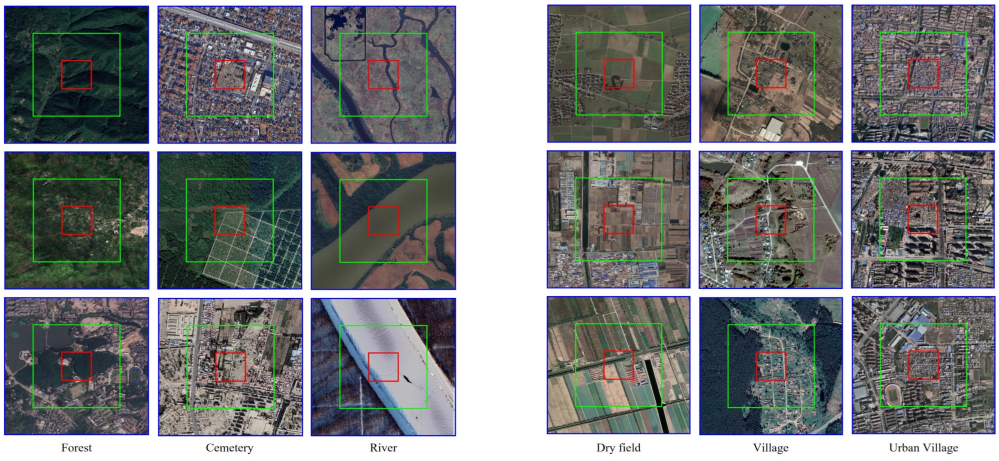
4.3.1. Scene Classification

The core challenges of fine-grained remote sensing image scene classification converge on four dimensions: feature confusion caused by "large intra-class variation and high inter-class similarity", data constraints from "high annotation costs and scarce samples", modeling imbalance between "local details and global semantics", and domain shift across "sensors and regions". The studies can be categorized into four core classes and one category of scattered research according to technical objectives and methodological logic.

Multi-Granularity Feature Modeling. It is one of the core approaches to resolving "Intra-Class Variation-Inter-Class Similarity". This category represents the fundamental technical direction for fine-grained classification. Its core logic involves mining multi-dimensional features (e.g., "local-global", "low-high resolution", "high-low frequency") to capture subtle discriminative information between subclasses, thereby addressing the pain points of "large intra-class variation and high inter-class similarity" in remote sensing scenes. Its technical evolution has progressed from single-granularity enhancement to multi-granularity collaborative decoupling, which can be further divided into two technical branches: **1) Multi-Level Feature Fusion and Semantic Collaboration.** This branch strengthens the transmission and discriminability of fine-grained semantics through feature interaction across

 doi:10.20944/preprints202509.2113.v1

different network levels. Typical studies include: [120] proposed the MGML-FENet framework, innovatively designing a Channel-Separate Feature Generator (CS-FG) to extract multi-granularity local features (e.g., building edge textures, crop ridge structures) at different network levels; [121] proposed MGSN, pioneering a coarse-grained guiding fine-grained bidirectional mechanism. Its MGSL module enables simultaneous learning of global scene structures and local details; [122] proposed the MG-CAP framework: it generates multi-granularity features via progressive image cropping, and uses Gaussian covariance matrices (replacing traditional CNN features) to capture feature high-order correlations. **2) Frequency/Scale** Decoupling and Enhancement. Typical studies include: Targeting the characteristic of remote sensing images where "high-frequency details are separated from low-frequency structures", this branch strengthens the independence and discriminability of fine-grained features through frequency decomposition or multi-scale modeling; [123] proposed MF²CNet: it realizes parallel extraction and decoupling of high/low-frequency features (high-frequency for fine-grained details like road markings, low-frequency for global structures like road orientation); [124] designed a Multi-Granularity Decoupling Network (MGDNet), focusing on "fine-grained feature learning under class-imbalanced scenarios". The network is guided to focus on subclass differences using region-level supervision; [125] proposed the ECA-MSDWNet, which integrates "multi-scale feature extraction" with incremental learning: the Efficient Channel Attention module focuses on key fine-grained features, while the multi-scale depthwise convolution reduces computational costs.



(**a**) Illustration of Intra-class Variation (**b**) Illustration of Inter-class Similarity

Figure 11. Illustration of Intra-Class Variation and Inter-Class Similarity in Scene-Level Recognition in MEET [22].

Cross-Domain and Domain Adaptation Learning. It is one of key technologies for addressing "sensor-region" Shift. In practical applications of fine-grained remote sensing classification, distribution shifts between training data (source domain) and test data (target domain) (e.g., optical-SAR sensor differences, regional differences between southern and northern farmlands) drastically reduce model generalization. This category mainly includes: **1) Open-Set Domain Adaptation**: Addressing Real-World Scenarios with "Unknown Subclasses in the Target Domain" Traditional domain adaptation assumes "complete category overlap between source and target domains", while Open-Set Domain Adaptation (OSDA) is more aligned with remote sensing reality (e.g., unknown subclasses such as "new artificial islands" or "special crops" appearing in the target domain). Its core lies in "separating unknown classes and aligning fine-grained features of known classes". [126] proposed IAFAN, which innovatively designs a USS mechanism (calculating sample semantic correlations via instance affinity matrix to identify unknown classes) and uses SDE loss to expand fine-grained differences between known classes (e.g., parking lot-industrial park vehicle arrangement density differences). **2) Multi-Source Domain Adaptation.** It achieves fine-grained alignment with limited annotations. For scenarios where the target domain contains only a small number of annotations, this branch improves the domain adaptability of fine-grained features through "pseudo-label optimization" and

 doi:10.20944/preprints202509.2113.v1

"multi-source subdomain modeling". [127] uses a bidirectional prototype module for source-target category/pseudo-label alignment, introduces adversarial training to optimize pseudo-labels (reducing mislabeling like "elevated roads as bridges"). In [128] for multi-source-single-target domain shifts, it adopts a "shared + dual-domain feature extractor" architecture (learning multi-source shared features like water bodies low reflectance first, then fine-grained subdomain alignment per source-target pair). In [129] to address fine-grained domain shift via frequency dimension, its HFE module aligns source-target fine-grained details (e.g., road marking edge intensity), LFE module aligns global structures (e.g., road network topology).

Semi-Supervised and Zero-Shot Learning. It is the inevitable path to reducing fine-grained annotation dependence. Annotations for fine-grained remote sensing scene classification require dual expertise in pixel-level labeling and subclass semantics, resulting in extremely high annotation costs. This category overcomes data constraints through "limited annotations + unlabeled sample utilization" or "knowledge transfer", serving as a key enabler for the large-scale application of fine-grained classification. **1) Semi-Supervised Learning.** Pseudo-Label Optimization and Consistency Constraints, mainly focused on enhancing the model's ability to learn fine-grained features through "supervised signals + unsupervised consistency". [130] modeled fine-grained road scene understanding as semi-supervised semantic segmentation. It optimizes supervised loss on annotated samples and consistency loss (e.g., perturbed prediction consistency) on unlabeled ones via ensemble prediction, and cuts annotation cost by using few annotated samples to reach high accuracy of fully supervised models on a self-built dataset. [124] solved fine-grained minority sample mislabeling—evaluating pseudo-label reliability via model confidence and class prior, prioritizing high-reliability samples for updates, and combined DCF loss to improve minority class accuracy in subclass classification. **2) Zero-Shot** Learning. By knowledge graphs or cross-modal transfer, this method targets scenarios with no annotations for target subclasses. This branch achieves fine-grained classification through knowledge transfer, with a core focus on establishing accurate mappings between visual features and semantic descriptions. For example, In [131], this study constructs a Remote Sensing Knowledge Graph (RSKG) for the first time. It generates semantic representations of remote sensing scene categories through graph representation learning, so as to improve the ability of domain semantic expression.

Figure 12. In [132], a hierarchical graph-enhanced Transformer network based on GNN is proposed, which improves the accuracy of remote sensing scene classification through dual attention mechanisms, multi-stage feature extraction, and graph structure modeling.

CNN and Transformer Fusion Modeling. This is one of the technical trend of balancing local details-global semantics. Traditional CNNs excel at extracting local fine-grained features (e.g., edges, textures) but lack strong global semantic modeling capabilities; Transformers excel at capturing global dependencies (e.g., spatial correlations between targets) but have insufficient local detail perception. This category achieves complementary advantages through strategies such as knowledge distillation, feature embedding, and hierarchical fusion, addressing the core contradiction of fragmented local features and missing global semantics in fine-grained classification. **1) Knowledge Distillation.** Lightweight Decoupling and Semantic Transfer With transformers as teachers and CNNs as students,

 doi:10.20944/preprints202509.2113.v1

global semantic knowledge is transferred to lightweight models, balancing fine-grained accuracy and computational efficiency. [133] proposed ET-GSNet, first using ViT as a fine-grained semantic teacher and ResNet18 as a student model; dynamic knowledge distillation lets the student learn CNN local details and ViT global semantics. [134] innovated a multi-path self-distillation mechanism: with ResNet34 as backbone, it fuses multi-scale fine-grained features and builds bidirectional self-distillation to couple global semantics and local details, achieving high accuracy in urban functional zone fine-grained classification. **2) Feature Embedding and Hierarchical Fusion.** Through CNN feature embedding into Transformers or hierarchical Transformers with local modules, in-depth collaboration between local details and global semantics for fine-grained features is achieved. In [135], it embeds CNN-extracted local fine-grained features into ViT's Patch Embedding layer, enabling "local+global" learning with limited data and accelerating convergence/improving accuracy vs. pure ViTs. In [136], for multi-scale fine-grained targets, swin transformer captures global structures, CNN extracts multi-resolution local features, and multi-scale fusion achieves high urban scene classification accuracy with better discrimination than pure Transformers. As Figure 12 which is a good example [132], it introduced GNNs for fine-grained structural features. It designs a dual attention (DA) module to suppress background noise by extracting channel-spatial key features, builds a three-stage hierarchical transformer extractor to capture multiscale global features, develops a pixel-level GNN extractor to distinguish similar scenes via spatial topology.

Some Other Research: [137] systematically reviewed a large number of deep learning methods, clarifying for the first time that "fine-grained classification needs to overcome the local limitations of CNNs and data dependence of Transformers", and summarizing key directions such as "multi-granularity features" and "cross-domain adaptation", providing a framework for subsequent research. [138] conducted a meta-analysis of multiple studies, quantitatively showing the rise of Transformers in fine-grained classification in recent years (with several representative studies), with AID and NWPU-RESISC45 as the most commonly used benchmarks. It also identified unresolved issues such as "fine-grained feature confusion" and "cross-sensor domain shift", providing data support for research directions. For efficiency and robustness optimization, [139,140] proposed a bilinear model based on MobileNetv2, enhancing fine-grained features through "dual convolutional layer feature transformation + Hadamard product". On the UC-Merced dataset, the number of parameters is much lower than that of traditional CNNs, while accuracy is improved, making it suitable for fine-grained classification requirements on edge devices. [141] proposed the Confounder-Free Fusion Network (CFF-NET), eliminating "spurious correlations between background interference and fine-grained features" (e.g., misclassifying "cloud shadows as water textures") through three branches ("global-local-target"). It achieves SOTA performance in fine-grained classification and retrieval tasks for aerial images, providing new ideas for "robust fine-grained modeling".

4.3.2. Image Retrieval

Fine-grained remote sensing image retrieval (FRSIR) has become an active research area, aiming to capture subtle distinctions among highly similar remote sensing (RS) images. The existing methods have conducted fine-grained optimization for remote sensing image retrieval from different perspectives.

Global–local and multi-scale feature fusion methods integrate information across different levels. For instance, GaLR introduces global and local representation with dynamic fusion and relational enhancement [142], while GLISA leverages global–local soft alignment with adaptive local information extraction [143]. FAAMI aggregates multi-scale information with cross-layer connections and consistency enhancement [144], and MSITA learns salient information with multiscale fusion and image-guided text alignment [145]. Related approaches such as AMFMN also focus on multiscale representation and semantic consistency [146].

Fine-grained semantic alignment focuses on aligning visual patches with textual words. FGVLA introduces spatial mask and contrastive losses to enhance patch-to-word correspondence [79]. MAFA-Net combines multi-attention fusion with fine-grained alignment for bidirectional retrieval [147].

 doi:10.20944/preprints202509.2113.v1

JGDN captures intra-modal fine-grained semantics to guide cross-modal learning [148], while CDMAN introduces cues-driven alignment and multi-granularity association learning [119]. Earlier work on semantic alignment networks also paved the way for this line of research [149].

Optimization and training strategies aim to improve generalization and address modality imbalance. RSITR-FFT adapts CLIP to RS domains through fine-grained tuning with consistency regularization [150]. FGIS introduces information supplementation and value-guided learning [151]. SWPE uses strong and weak prompts to capture both global and fine-grained semantics [152], while RDB bridges representation discrepancies with differential and hierarchical attention [153]. SMLGN further integrates multi-subspace joint learning with adversarial training to achieve modality consistency [154].

Other novel feature modeling methods. FRORS integrates fine-grained prototype memory and Gram-based learning to capture intra-class heterogeneity and inter-class commonality [155]. DMFH introduces multiscale fine-grained hashing for efficient retrieval [156]. GNN-based methods explicitly model associations between text and images for fine-grained matching [157]. SWAN employs scene-aware aggregation to mitigate semantic confusion in RS retrieval [158]. Other explorations, such as sketch-based fine-grained retrieval, also extend the paradigm [159,160].

This section mainly reviews two aspects of research on fine-grained scene-level recognition: scene classification and image retrieval. Among them, research on fine-grained scene classification is relatively abundant, while research on fine-grained image retrieval is still in the exploratory stage. From a technical perspective, most studies adopt a bottom-up paradigm. “large intra-class variation and high inter-class similarity” remain core challenges in fine-grained scene recognition. Compared with pixel-level and object-level fine-grained recognition, scene-level fine-grained recognition employs more diverse and comprehensive technologies, and thus is more challenging.

4.4. Summary of Methods

In this chapter, with the three core levels of remote sensing image interpretation (pixel-level, object-level, and scene-level) as the framework, fine-grained interpretation is taken as a technical deepening centered on the demand for “subclass distinction” at each level, rather than an independent paradigm. Centering on the core issues of fine-grained interpretation such as small inter-class differences and large intra-class variations, we systematically sort out the representative methods of fine-grained interpretation under the three levels. we also find that some methods appear repeatedly in fine-grained interpretation at different levels, such as Utilizing MultiSource Data, Modeling Relationships Between Classes, Advanced Data Annotation Strategies, Knowledge Distillation, Novel Data Representation, Component Relationship Learning, Enhanced Attention Mechanism, Few shot or Zero shot, Prototypical Contrastive Learning etc. The advantages and disadvantages of these methods were summarized. Meanwhile, this chapter analyzes the technical logic and applicable scenarios of different methods, and finally constructs a fine-grained remote sensing image interpretation method classification system covering core tasks, providing systematic reference for method research in this field.

5. Discussion

5.1. Challenge

Although fine-grained interpretation of remote sensing images has made significant progress in recent years, the field still faces many unresolved challenges before it can achieve reliable and scalable applications in real-world scenarios. These challenges span conceptual definitions, methodological limitation, dataset construction, and relation to human cognition etc. Below, we discuss several key challenges in depth.

1. Lack of a unified definition of fine granularity. One of the most fundamental problems is the absence of a unified definition of “fine-grained” in the context of remote sensing. At the pixel level, fine granularity may refer to distinguishing subtle spectral variations; at the object level, it often refers to recognizing subcategories of the same class, such as different ship types or aircraft models; at the scene level, it may mean subdividing complex environments into finer categories, such as different types of

 doi:10.20944/preprints202509.2113.v1

residential or agricultural areas. These different perspectives are not hierarchical and are often defined independently, which creates inconsistencies across studies. As a result, it is difficult to fairly compare methods or transfer models between datasets. The lack of a common framework also complicates the creation of benchmarks, since each dataset may adopt its own class definitions and granularity standards. Establishing a standardized and hierarchical definition of fine granularity would provide a foundation for dataset construction, model evaluation, and cross-domain generalization.

2. Heavy reliance on computer vision methods but with domain gaps. Fine-grained remote sensing interpretation has borrowed heavily from advances in computer vision, including deep convolutional neural networks, attention mechanisms, and transformer-based architectures. These methods have enabled rapid progress, but the direct transfer of computer vision techniques has limitations. Remote sensing images differ fundamentally from natural images in several ways: they are top views and cover much larger spatial extents, often include multiple modalities (optical, SAR, LiDAR, hyperspectral), and are acquired under widely varying conditions such as seasons, illumination, and sensor platforms. For instance, models that perform well on ImageNet-style natural images may fail to handle the scale variation and sensor noise inherent in remote sensing. Furthermore, remote sensing tasks often demand recognition of subtle differences across highly similar categories, which is less common in computer vision benchmarks. This domain gap indicates the need to adapt or redesign methods specifically for remote sensing, rather than relying solely on transferring computer vision techniques.

3. Intrinsic difficulty of small inter-class differences and large intra-class variations. The co-existence of small inter-class differences and large intra-class variations is an intrinsic difficulty of fine-grained tasks. Many fine-grained remote sensing categories are visually similar, such as different models of aircraft, ships, or ecologically related tree species. At the same time, instances of the same class can appear very different depending on seasonality, vegetation phenology, viewing angle, or illumination. For example, the same type of crop field can exhibit drastic changes in spectral characteristics over its growth cycle, while the same residential area may look very different under varying imaging resolutions or atmospheric conditions. This dual challenge requires models to learn highly discriminative features while also being robust to intra-class variability. Existing approaches, such as attention mechanisms or part-based feature learning, partially address these issues, but their effectiveness is still limited in real-world, large-scale scenarios. Tackling this challenge will likely require more sophisticated strategies, including dynamic feature adaptation, multi-scale representation learning, and domain-aware training.

4. Insufficient multi-modal and multi-source data fusion. Remote sensing inherently involves multiple modalities of data, each capturing complementary information. Optical images provide rich texture and color cues, SAR images capture structural and backscatter properties, LiDAR data reveals 3D geometry, and hyperspectral sensors provide dense spectral signatures that can separate visually similar categories. Despite this, most fine-grained interpretation research still relies almost exclusively on optical images. This limitation arises partly from the lack of multi-modal benchmark datasets with aligned and consistent annotations, and partly from the difficulty of designing models that can effectively integrate heterogeneous modalities. Without multi-source fusion, models may fail to fully exploit the complementary strengths of different sensors, leading to suboptimal performance in complex or ambiguous scenarios. Future progress will depend on building large-scale, well-annotated multi-modal datasets and developing fusion strategies that can dynamically balance modality contributions under varying conditions.

5. High annotation cost and lack of consistency. Fine-grained datasets require extensive high-quality annotations, which are both costly and difficult to obtain. Unlike coarse-grained classification, fine-grained labeling often requires domain expertise—for example, identifying tree species or aircraft variants demands specialist knowledge that crowdsourcing cannot easily provide. Furthermore, fine-grained annotation is prone to inconsistencies, as different annotators may disagree on subtle distinctions. This variability undermines dataset reliability and can introduce noise into model training

 doi:10.20944/preprints202509.2113.v1

and evaluation. While some efforts have been made to reduce costs using weakly supervised learning, semi-supervised learning, or automatic labeling techniques, these methods are still insufficient for achieving the level of accuracy required for fine-grained interpretation. Without scalable annotation solutions, progress in this field will remain constrained. Future research should explore integrating expert knowledge with automatic or semi-automatic annotation strategies, as well as leveraging knowledge graphs and large language models to support label consistency.

6. Limited cross-domain generalization and open-world recognition. Most fine-grained models perform well on specific benchmark datasets but generalize poorly when applied to new regions, sensors, or time periods. This performance drop is largely due to domain shifts caused by differences in geography, climate, imaging conditions, and sensor characteristics. In real-world applications, remote sensing imagery frequently contains novel classes not seen during training, making the traditional closed-set assumption unrealistic. Addressing this challenge requires developing models with stronger cross-domain adaptability and the ability to detect and learn from unseen categories. Open-world recognition and zero-shot learning have been proposed as potential solutions, but current work in these areas is still at an exploratory stage. Progress will require not only new methods but also large-scale open-world benchmarks to evaluate and train models under realistic conditions.

7. Limited integration with cognitive theories. Humans are adept at fine-grained recognition because of cognitive mechanisms such as selective attention, hierarchical semantic reasoning, and multi-modal integration. However, current fine-grained remote sensing methods are largely disconnected from cognitive science. Most rely on purely data-driven deep learning approaches without incorporating principles inspired by human perception. Questions such as how to replicate human-like selective attention to key image regions, how to leverage semantic hierarchies for more interpretable reasoning, and how to dynamically integrate multiple information sources remain underexplored. Bridging remote sensing interpretation with cognitive theories could lead to new paradigms in feature learning and model design, potentially making models both more accurate and more interpretable.

5.2. Future Directions

Looking ahead, fine-grained interpretation of remote sensing images is expected to evolve along several promising directions, many of which directly correspond to the challenges outlined above.

1. Establishing unified definitions and hierarchical frameworks. A first priority is to standardize the definition of fine granularity across pixel-level, object-level, and scene-level tasks. Developing a unified and hierarchical framework will enable more consistent dataset construction, fairer benchmarking, and clearer methodological comparisons. Such a framework could also serve as a foundation for cross-task transfer and multi-level integration.

2. Designing domain-specific methods beyond computer vision transfer. While computer vision remains an important source of inspiration, the remote sensing community needs to design methods tailored to the unique properties of remote sensing imagery. This includes accounting for large-scale spatial structures, multi-resolution data, and sensor-specific characteristics. Hybrid models that integrate domain knowledge with advanced architectures such as transformers may offer a way forward.

3. Addressing the dual challenge of inter-class similarity and intra-class variability. Future models must be capable of simultaneously handling subtle inter-class differences and large intra-class variations. Promising directions include multi-scale representation learning, part-based feature modeling, and dynamic adaptation mechanisms that adjust feature extraction based on environmental conditions. Incorporating temporal and contextual cues may also help reduce ambiguity in challenging cases.

4. Building and exploiting multi-modal and multi-source datasets. The integration of optical, SAR, LiDAR, hyperspectral, and temporal data will be crucial for fine-grained interpretation. Future efforts should focus on constructing large-scale, well-aligned multi-modal datasets and developing fusion strategies that adaptively balance complementary modalities. Advances in cross-modal representation learning and self-supervised pretraining are expected to play an important role.

 doi:10.20944/preprints202509.2113.v1

5. Reducing annotation cost and improving label consistency. Novel strategies are needed to address the bottleneck of high annotation cost. These include active learning to prioritize informative samples, semi- and weakly supervised learning to leverage partially labeled data, and automated annotation assisted by knowledge graphs or large foundation models. Such approaches can reduce reliance on expensive expert labeling while maintaining high label quality and consistency.

6. Enhancing cross-domain generalization and supporting open-world recognition. Developing robust models that can generalize across regions, sensors, and time periods will be a key research focus. Domain adaptation, domain generalization, and meta-learning provide potential solutions. At the same time, constructing open-world benchmarks and incorporating zero-shot and few-shot learning methods will allow models to better handle previously unseen categories, bringing remote sensing closer to real-world requirements.

7. Bridging remote sensing interpretation with cognitive science. Finally, future research may benefit from stronger integration with cognitive theories. Models inspired by human perception—such as attention-guided reasoning, semantic hierarchy utilization, and multi-modal integration—can potentially achieve both higher accuracy and interpretability. This interdisciplinary approach could open up new directions for fine-grained interpretation, enabling systems that not only match but also mimic human analytical capabilities.

6. Conclusions

In this survey, we have provided a comprehensive review of fine-grained interpretation of remote sensing images across three levels: pixel-level classification and segmentation, object-level detection and recognition, and scene-level understanding. We systematically examined existing benchmark datasets, summarized representative methods, and analyzed their strengths and limitations. The review highlights that, although deep learning has driven substantial advances in accuracy and applicability, fine-grained interpretation remains an inherently challenging problem due to issues such as inconsistent definitions, small inter-class differences with large intra-class variations, high annotation costs, and limited generalization across domains.

We further identified seven key challenges that constrain the progress of this field, ranging from the lack of a unified definition of fine granularity to the insufficient integration with cognitive theories. Addressing these challenges will require joint efforts in dataset construction, algorithmic innovation, and interdisciplinary collaboration.

Looking forward, the future of fine-grained remote sensing interpretation will likely move toward unified frameworks, multi-modal integration, scalable annotation strategies, and robust open-world recognition. At the same time, bridging remote sensing research with cognitive science and harnessing advances in foundation models and self-supervised learning are expected to open new possibilities. With continued progress along these directions, fine-grained interpretation has the potential to significantly enhance applications in environmental monitoring, agriculture, urban planning, and disaster management, ultimately contributing to a deeper and more actionable understanding of our planet.

Author Contributions: Conceptualization, Dongbo Wang, Zedong Yan and Peng Liu; formal analysis, Dongbo Wang, Zedong Yan and Peng Liu; investigation, Peng Liu; resources, Peng Liu; writing—original draft preparation, Dongbo Wang, Zedong Yan and Peng Liu; writing—review and editing, Dongbo Wang, Zedong Yan and Peng Liu; supervision, Peng Liu.

Funding: This research was funded by Comprehensive Site Selection System Project grant number E5E2180501.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections.

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

 doi:10.20944/preprints202509.2113.v1

References

1. Wan, M.; Zhong, G.; Wu, Q.; Zhao, X.; Lin, Y.; Lu, Y. CR-Mask RCNN: An Improved Mask RCNN Method for Airport Runway Detection and Segmentation in Remote Sensing Images. *Sensors* **2025**, *25*, 657. https://doi.org/10.3390/s25030657.
2. Li, N.; et al. Airport Detection in Remote Sensing Real-Open World Using Deep Learning and Geographical Analysis. *Engineering Applications of Artificial Intelligence* **2023**, *120*, 106083. https://doi.org/10.1016/j.engappai.2023.106083.
3. Chen, F.; et al. HRTBDA: a network for post-disaster building damage assessment. *Natural Hazards Review* **2024**. https://doi.org/10.1080/17538947.2024.2418880.
4. Wu, Z.; et al. A Hybrid YOLO-E and SAM2 Approach for Damaged Building Extraction Using Multi-Source Remote Sensing Images. *Sensors* **2025**, *25*, 4375. https://doi.org/10.3390/s25144375.
5. Yan, J.; Gu, X.; Chen, Y. CropSTS: A Remote Sensing Foundation Model for Cropland Classification with Decoupled Spatiotemporal Attention. *Remote Sensing* **2025**, *17*, 2481. https://doi.org/10.3390/rs17142481.
6. Xia, L.; et al. A precise spatiotemporal fusion crop classification framework for smallholder agricultural systems: PITT (Parcel-level Integration of Time series and Texture). *Scientific Reports* **2025**, *15*, 33351. https://doi.org/10.1038/s41598-025-03351-7.
7. Zhang, S.; Cao, Y.; Bai, L.; Wu, Z. Research on Camouflage Target Classification and Recognition Based on Mid-Wave Infrared Hyperspectral Imaging. *Remote Sensing* **2025**, *17*, 1475. https://doi.org/10.3390/rs17081475.
8. Zhang, T.; Zhang, D.; Liu, Y. Research on Camouflage Target Detection Method Based on Dual Band Optics and SAR Image Fusion. In Proceedings of the Proceedings of International Conference on Image, Vision and Intelligent Systems (ICIVIS 2023). Springer, 2024, pp. 320–335. https://doi.org/10.1007/978-981-97-0855-0_31.
9. Chen, F.; Ren, R.; Van de Voorde, T.; Xu, W.; Zhou, G.; Zhou, Y. Fast Automatic Airport Detection in Remote Sensing Images Using Convolutional Neural Networks. *Remote Sensing* **2018**, *10*, 443. https://doi.org/10.3390/rs10030443.
10. Zhang, B.; Zhao, L.; Zhang, X. Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images. *Remote Sensing of Environment* **2020**, *247*, 111938.
11. Zhu, Y.; Li, W.; Zhang, M.; Pang, Y.; Tao, R.; Du, Q. Joint feature extraction for multi-source data using similar double-concentrated network. *450*, 70–79. https://doi.org/https://doi.org/10.1016/j.neucom.2021.03.088.
12. Yuan, S.; Lin, G.; Zhang, L.; Dong, R.; Zhang, J.; Chen, S.; Zheng, J.; Wang, J.; Fu, H. FUSU: A multi-temporal-source land use change segmentation dataset for fine-grained urban semantic understanding. *Advances in Neural Information Processing Systems* **2024**, *37*, 132417–132439.
13. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods. SciTePress, 2017, Vol. 2, pp. 324–331.
14. Di, Y.; Jiang, Z.; Zhang, H. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing* **2021**, *13*, 747.
15. Zhang, Z.; Zhang, L.; Wang, Y.; Feng, P.; He, R. Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *14*, 8458–8472.
16. Wang, Z.; Zhou, Y.; Wang, F.; Wang, S.; Gao, G.; Zhu, J.; Wang, P.; Hu, K. Mfbfs: High-resolution multispectral remote sensing image fine-grained building feature set. *Journal of Remote Sensing* **2024**, *28*.
17. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3965–3981.
18. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE* **2017**, *105*, 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.
19. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing* **2018**, *145*, 197–209.
20. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. Mlrsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing* **2020**, *169*, 337–350.
21. Long, Y.; Xia, G.S.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.X.; Zhang, L.; Li, D. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2021**, *14*, 4205–4230.

 doi:10.20944/preprints202509.2113.v1

22. Li, Y.; Wu, Y.; Cheng, G.; Tao, C.; Dang, B.; Wang, Y.; Zhang, J.; Zhang, C.; Liu, Y.; Tang, X.; et al. MEET: A Million-Scale Dataset for Fine-Grained Geospatial Scene Classification with Zoom-Free Remote Sensing Imagery. *arXiv preprint arXiv:2503.11219* **2025**.
23. Chen, K.; Wu, M.; Liu, J.; Zhang, C. Fgsd: A dataset for fine-grained ship detection in high resolution satellite images. *arXiv preprint arXiv:2003.06832* **2020**.
24. Huang, X.; Ren, L.; Liu, C.; Wang, Y.; Yu, H.; Schmitt, M.; Hänsch, R.; Sun, X.; Huang, H.; Mayer, H. Urban building classification (ubc)-a dataset for individual building detection and classification from satellite imagery. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1413–1421.
25. Huang, X.; Chen, K.; Tang, D.; Liu, C.; Ren, L.; Sun, Z.; Hänsch, R.; Schmitt, M.; Sun, X.; Huang, H.; et al. Urban building classification (ubc) v2—a benchmark for global building detection and fine-grained classification from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–16.
26. Liu, G.; Peng, B.; Liu, T.; Zhang, P.; Yuan, M.; Lu, C.; Cao, N.; Zhang, S.; Huang, S.; Wang, T.; et al. Large-scale fine-grained building classification and height estimation for semantic urban reconstruction: Outcome of the 2023 IEEE GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *17*, 11194–11207.
27. Wu, Z.Z.; Wan, S.H.; Wang, X.F.; Tan, M.; Zou, L.; Li, X.L.; Chen, Y. A benchmark data set for aircraft type recognition from remote sensing images. *Applied Soft Computing* **2020**, *89*, 106132.
28. Yu, W.; Cheng, G.; Wang, M.; Yao, Y.; Xie, X.; Yao, X.; Han, J. Mar20: Remote sensing image military aircraft target recognition dataset. *Journal of Remote Sensing* **2023**, *27*, 2688–2696.
29. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* **2022**, *184*, 116–130.
30. Xiang, X.; Xu, Z.; Deng, Y.; Zhou, Q.; Liang, Y.; Chen, K.; Zheng, Q.; Wang, Y.; Chen, X.; Gao, W. Openearthsensing: Large-scale fine-grained benchmark for open-world remote sensing. *arXiv arXiv:2502.20668* **2025**.
31. Xiao, Z.; Long, Y.; Li, D.; Wei, C.; Tang, G.; Liu, J. High-resolution remote sensing image retrieval based on cnns from a dimensional perspective. *Remote Sensing* **2017**, *9*, 725.
32. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *57*, 1155–1167.
33. Li, H.; Dou, X.; Tao, C.; Wu, Z.; Chen, J.; Peng, J.; Deng, M.; Zhao, L. Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors* **2020**, *20*, 1594.
34. Li, Y.; Kong, D.; Zhang, Y.; Tan, Y.; Chen, L. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing* **2021**, *179*, 145–158.
35. Hua, Y.; Mou, L.; Jin, P.; Zhu, X.X. Multiscene: A large-scale dataset and benchmark for multiscene recognition in single aerial images. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–13.
36. Yuan, J.; Ru, L.; Wang, S.; Wu, C. Wh-mavs: A novel dataset and deep learning benchmark for multiple land use and land cover applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 1575–1590.
37. Zhao, D.; Yuan, B.; Chen, Z.; Li, T.; Liu, Z.; Li, W.; Gao, Y. Panoptic Perception: A Novel Task and Fine-Grained Dataset for Universal Remote Sensing Image Interpretation **2024**. *62*, 1–14. https://doi.org/10.1109/TGRS.2024.3392778.
38. Guo, Z.; Zhang, M.; Jia, W.; Zhang, J.; Li, W. Dual-concentrated network with morphological features for tree species classification using hyperspectral image. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2022**, *15*, 7013–7024.
39. Peng, Y.; Zhang, Y.; Tu, B.; Li, Q.; Li, W. Spatial–Spectral Transformer With Cross-Attention for Hyperspectral Image Classification **2022**. *60*, 1–15. https://doi.org/10.1109/TGRS.2022.3203476.
40. Han, Z.; Xu, S.; Gao, L.; Li, Z.; Zhang, B. GRetNet: Gaussian Retentive Network for Hyperspectral Image Classification **2024**. *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3450536.
41. Jia, C.; Zhang, X.; Meng, H.; Xia, S.; Jiao, L. CenterFormer: A Center Spatial–Spectral Attention Transformer Network for Hyperspectral Image Classification **2025**. *18*, 5523–5539. https://doi.org/10.1109/JSTARS.2025.3529985.
42. Zhao, Y.; Bao, W.; Xu, X.; Zhou, Y. E2TNet: Efficient enhancement Transformer network for hyperspectral image classification **2024**. *142*, 105569. https://doi.org/10.1016/j.infrared.2024.105569.

 doi:10.20944/preprints202509.2113.v1

43. Li, Z.; Guo, F.; Li, Q.; Ren, G.; Wang, L. An Encoder–Decoder Convolution Network With Fine-Grained Spatial Information for Hyperspectral Images Classification **2020**. *8*, 33600–33608. https://doi.org/10.1109/ACCESS.2020.2974025.
44. Roy, S.K.; Kar, P.; Hong, D.; Wu, X.; Plaza, A.; Chanussot, J. Revisiting Deep Hyperspectral Feature Extraction Networks via Gradient Centralized Convolution **2022**. *60*, 1–19. https://doi.org/10.1109/TGRS.2021.3120198.
45. Zhang, M.; Li, W.; Zhao, X.; Liu, H.; Tao, R.; Du, Q. Morphological Transformation and Spatial-Logical Aggregation for Tree Species Classification Using Hyperspectral Imagery **2023**. *61*, 1–12. https://doi.org/10.1109/TGRS.2022.3233847.
46. Guo, Z.; Zhang, M.; Jia, W.; Zhang, J.; Li, W. Dual-Concentrated Network With Morphological Features for Tree Species Classification Using Hyperspectral Image **2022**. *15*, 7013–7024. https://doi.org/10.1109/JSTARS.2022.3199618.
47. Roy, S.K.; Deria, A.; Shah, C.; Haut, J.M.; Du, Q.; Plaza, A. Spectral–Spatial Morphological Attention Transformer for Hyperspectral Image Classification **2023**. *61*, 1–15. https://doi.org/10.1109/TGRS.2023.3242346.
48. Ji, R.; Tan, K.; Wang, X.; Tang, S.; Sun, J.; Niu, C.; Pan, C. PatchOut: A novel patch-free approach based on a transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images **2025**. *138*, 104457. https://doi.org/10.1016/j.jag.2025.104457.
49. Yuan, J.; Wang, S.; Wu, C.; Xu, Y. Fine-Grained Classification of Urban Functional Zones and Landscape Pattern Analysis Using Hyperspectral Satellite Imagery: A Case Study of Wuhan **2022**. *15*, 3972–3991. https://doi.org/10.1109/JSTARS.2022.3174412.
50. Chen, Z.; Xu, T.; Pan, Y.; Shen, N.; Chen, H.; Li, J. Edge Feature Enhancement for Fine-Grained Segmentation of Remote Sensing Images **2024**. *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3443247.
51. Chen, Y.; Huang, L.; Zhu, L.; Yokoya, N.; Jia, X. Fine-Grained Classification of Hyperspectral Imagery Based on Deep Learning **2019**. *11*, 2690. https://doi.org/10.3390/rs11222690.
52. Miao, J.; Zhang, B.; Wang, B. Coarse-to-Fine Joint Distribution Alignment for Cross-Domain Hyperspectral Image Classification **2021**. *14*, 12415–12428. https://doi.org/10.1109/JSTARS.2021.3129177.
53. Wu, H.; Xue, Z.; Zhou, S.; Su, H. Overcoming Granularity Mismatch in Knowledge Distillation for Few-Shot Hyperspectral Image Classification **2025**. *63*, 1–17. https://doi.org/10.1109/TGRS.2025.3530614.
54. Huang, Y.; Peng, J.; Zhang, G.; Sun, W.; Chen, N.; Du, Q. Adversarial Domain Adaptation Network With Calibrated Prototype and Dynamic Instance Convolution for Hyperspectral Image Classification **2024**. *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3387990.
55. Ma, Y.; Deng, X.; Wei, J. Land Use Classification of High-Resolution Multispectral Satellite Images With Fine-Grained Multiscale Networks and Superpixel Postprocessing **2023**. *16*, 3264–3278. https://doi.org/10.1109/JSTARS.2023.3260448.
56. Zhao, C.; Chen, M.; Feng, S.; Qin, B.; Zhang, L. A Coarse-to-Fine Semisupervised Learning Method Based on Superpixel Graph and Breaking-Tie Sampling for Hyperspectral Image Classification **2023**. *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3297110.
57. Ni, K.; Xie, Y.; Zhao, G.; Zheng, Z.; Wang, P.; Lu, T. Coarse-to-Fine High-Order Network for Hyperspectral and LiDAR Classification **2025**. *63*, 1–16. https://doi.org/10.1109/TGRS.2025.3554802.
58. Liu, Y.; Ye, Z.; Xi, Y.; Liu, H.; Li, W.; Bai, L. Multiscale and Multidirection Feature Extraction Network for Hyperspectral and LiDAR Classification **2024**. *17*, 9961–9973. https://doi.org/10.1109/JSTARS.2024.3400872.
59. Liu, Z.; Li, J.; Wang, L.; Plaza, A. Integration of Remote Sensing and Crowdsourced Data for Fine-Grained Urban Flood Detection **2024**. *17*, 13523–13532. https://doi.org/10.1109/JSTARS.2024.3433010.
60. Bai, J.; Yuan, A.; Xiao, Z.; Zhou, H.; Wang, D.; Jiang, H.; Jiao, L. Class Incremental Learning With Few-Shots Based on Linear Programming for Hyperspectral Image Classification **2022**. *52*, 5474–5485. https://doi.org/10.1109/TCYB.2020.3032958.
61. Ouyang, L.; Guo, G.; Fang, L.; Ghamisi, P.; Yue, J. PCLDet: Prototypical Contrastive Learning for Fine-Grained Object Detection in Remote Sensing Images **2023**. *61*, 1–11. https://doi.org/10.1109/TGRS.2023.3290091.
62. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
63. Girshick, R. Fast r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
64. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in neural information processing systems, 2015, pp. 91–99.

 doi:10.20944/preprints202509.2113.v1

65. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
66. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.
67. Han, Y.; Yang, X.; Pu, T.; Peng, Z. Fine-Grained Recognition for Oriented Ship Against Complex Scenes in Optical Remote Sensing Images **2022**. *60*, 1–18. https://doi.org/10.1109/TGRS.2021.3123666.
68. Sumbul, G.; Cinbis, R.G.; Aksoy, S. Multisource Region Attention Network for Fine-Grained Object Recognition in Remote Sensing Imagery **2019-07**. *57*, 4929–4937, [1901.06403 [cs]]. https://doi.org/10.1109/TGRS.2019.2894425.
69. Guo, B.; Zhang, R.; Guo, H.; Yang, W.; Yu, H.; Zhang, P.; Zou, T. Fine-Grained Ship Detection in High-Resolution Satellite Images With Shape-Aware Feature Learning **2023**. *16*, 1914–1926. https://doi.org/10.1109/JSTARS.2023.3241969.
70. Cheng, J.; Yao, X.; Yang, X.; Yuan, X.; Feng, X.; Cheng, G.; Huang, X.; Han, J. DIMA: Digging Into Multigranular Archetype for Fine-Grained Object Detection **2024**. *62*, 1–14. https://doi.org/10.1109/TGRS.2024.3415809.
71. Wang, L.; Zhang, J.; Tian, J.; Li, J.; Zhuo, L.; Tian, Q. Efficient Fine-Grained Object Recognition in High-Resolution Remote Sensing Images From Knowledge Distillation to Filter Grafting **2023**. *61*, 1–16. https://doi.org/10.1109/TGRS.2023.3260883.
72. Zeng, L.; Guo, H.; Yang, W.; Yu, H.; Yu, L.; Zhang, P.; Zou, T. Instance Switching-Based Contrastive Learning for Fine-Grained Airplane Detection **2022**. *60*, 1–16. https://doi.org/10.1109/TGRS.2022.3218533.
73. Li, W.; Zhao, D.; Yuan, B.; Gao, Y.; Shi, Z. PETDet: Proposal Enhancement for Two-Stage Fine-Grained Object Detection **2024**. *62*, 1–14. https://doi.org/10.1109/TGRS.2023.3343453.
74. Cheng, G.; Li, Q.; Wang, G.; Xie, X.; Min, L.; Han, J. SFRNet: Fine-Grained Oriented Object Recognition via Separate Feature Refinement **2023**. *61*, 1–10. https://doi.org/10.1109/TGRS.2023.3277626.
75. Ouyang, L.; Fang, L.; Ji, X. Multigranularity Self-Attention Network for Fine-Grained Ship Detection in Remote Sensing Images **2022**. *15*, 9722–9732. https://doi.org/10.1109/JSTARS.2022.3220503.
76. Liu, Y.; Liu, J.; Li, X.; Wei, L.; Wu, Z.; Han, B.; Dai, W. Exploiting Discriminating Features for Fine-Grained Ship Detection in Optical Remote Sensing Images **2024**. *17*, 20098–20115. https://doi.org/10.1109/JSTARS.2024.3486210.
77. Yang, Y.; Zhang, Z.; Feng, P.; Yan, Y.; He, G.; Liu, S.; Zhang, P.; Gao, H. HMS-Net: A Hierarchical Multilabel Fine-Grained Ship Detection Network in Remote Sensing Images **2025**. *18*, 15394–15411. https://doi.org/10.1109/JSTARS.2025.3570872.
78. Zhu, Z.; Sun, X.; Diao, W.; Chen, K.; Xu, G.; Fu, K. Invariant Structure Representation for Remote Sensing Object Detection Based on Graph Modeling **2022**. *60*, 1–17. https://doi.org/10.1109/TGRS.2022.3181686.
79. Li, Y.; Chen, L.; Li, W. Fine-Grained Ship Recognition With Spatial-Aligned Feature Pyramid Network and Adaptive Prototypical Contrastive Learning **2025**. *63*, 1–13. https://doi.org/10.1109/TGRS.2024.3524621.
80. Gong, T.; Cheng, W.; Chen, Y.; Xiong, S.; Lu, X. Discover the Unknown Ones in Fine-Grained Ship Detection **2025**. *63*, 1–14. https://doi.org/10.1109/TGRS.2025.3578908.
81. Chen, X.; Chen, X.; Ge, X.; Chen, J.; Wang, H. Online Decoupled Distillation Based on Prototype Contrastive Learning for Lightweight Underwater Object Detection Models **2025**. *63*, 1–14. https://doi.org/10.1109/TGRS.2025.3549129.
82. Guo, H.; Liu, Y.; Pan, Z.; Hu, Y. Advancing Fine-Grained Few-Shot Object Detection on Remote Sensing Images with Decoupled Self-Distillation and Progressive Prototype Calibration **2025-01**. *17*, 495. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs17030495.
83. Lu, X.; Sun, X.; Diao, W.; Mao, Y.; Li, J.; Zhang, Y.; Wang, P.; Fu, K. Few-Shot Object Detection in Aerial Imagery Guided by Text-Modal Knowledge **2023**. *61*, 1–19. https://doi.org/10.1109/TGRS.2023.3250448.
84. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
85. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European conference on computer vision. Springer, 2016, pp. 21–37.
86. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
87. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.

 doi:10.20944/preprints202509.2113.v1

88. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 2020, pp. 10781–10790.
89. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696* **2022**.
90. Zhang, Y.; Li, S.; Wang, H.; Liu, Y.; Zhang, J. Aircraft Target Detection in Remote Sensing Images Based on Improved YOLOv7-Tiny Network. *IEEE Geoscience and Remote Sensing Letters* **2024**, *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3420642.
91. Chen, Y.; Liu, J.; Zhang, Y.; Li, W.; Wang, H. A Remote Sensing Target Detection Model Based on Lightweight Feature Enhancement and Feature Refinement Extraction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *17*, 5265–5279. https://doi.org/10.1109/JSTARS.2024.3396462.
92. Luo, Y.; Xiong, G.; Li, X.; Wang, Z.; Chen, J. An Improved YOLOv8 Detector for Multi-Scale Target Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2024**, *62*, 1–16. https://doi.org/10.1109/TGRS.2024.3428459.
93. Li, M.; Zhang, W.; Wang, Q.; Zhao, Y. YOLO-RS: Remote Sensing Enhanced Crop Detection Methods, 2025, [arXiv:cs.CV/2504.11165].
94. Wang, C.; Li, J.; Zhang, H.; Liu, X. YOLOX-DW: A Fine-Grained Object Detection Algorithm for Remote Sensing Images. *Remote Sensing* **2024**. https://doi.org/10.21203/rs.3.rs-5122331/v1.
95. Xi, Y.; Jia, W.; Miao, Q.; Feng, J.; Ren, J.; Luo, H. Detection-Driven Exposure-Correction Network for Nighttime Drone-View Object Detection **2024**. *62*, 1–14. https://doi.org/10.1109/TGRS.2024.3351134.
96. Yang, J.; Fu, K.; Wu, Y.; Diao, W.; Dai, W.; Sun, X. Mutual-Feed Learning for Super-Resolution and Object Detection in Degraded Aerial Imagery **2022**. *60*, 1–16. https://doi.org/10.1109/TGRS.2022.3198083.
97. Wu, J.; Zhao, F.; Yao, G.; Jin, Z. FGA-YOLO: A one-stage and high-precision detector designed for fine-grained aircraft recognition **2025-02-14**. *618*, 129067. https://doi.org/10.1016/j.neucom.2024.129067.
98. Zhao, S.; Chen, H.; Zhang, D.; Tao, Y.; Feng, X.; Zhang, D. SR-YOLO: Spatial-to-Depth Enhanced Multi-Scale Attention Network for Small Target Detection in UAV Aerial Imagery **2025-01**. *17*, 2441. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs17142441.
99. Wu, F.; Hu, T.; Xia, Y.; Ma, B.; Sarwar, S.; Zhang, C. WDFA-YOLOX: A Wavelet-Driven and Feature-Enhanced Attention YOLOX Network for Ship Detection in SAR Images **2024-01**. *16*, 1760. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs16101760.
100. Song, Y.; Wang, S.; Li, Q.; Mu, H.; Feng, R.; Tian, T.; Tian, J. Vehicle Target Detection Method for Wide-Area SAR Images Based on Coarse-Grained Judgment and Fine-Grained Detection **2023-01**. *15*, 3242. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs15133242.
101. Zhang, J.; Zhang, Y.; Shi, Z.; Zhang, Y.; Gao, R. Unmanned Aerial Vehicle Object Detection Based on Information-Preserving and Fine-Grained Feature Aggregation **2024-01**. *16*, 2590. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs16142590.
102. Xi, Y.; Jia, W.; Miao, Q.; Liu, X.; Fan, X.; Li, H. FiFoNet: Fine-Grained Target Focusing Network for Object Detection in UAV Images **2022-01**. *14*, 3919. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs14163919.
103. Ma, S.; Wang, W.; Pan, Z.; Hu, Y.; Zhou, G.; Wang, Q. A Recognition Model Incorporating Geometric Relationships of Ship Components **2024-01**. *16*, 130. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs16010130.
104. Jiang, X.N.; Niu, X.Q.; Wu, F.L.; Fu, Y.; Bao, H.; Fan, Y.C.; Zhang, Y.; Pei, J.Y. A Fine-Grained Aircraft Target Recognition Algorithm for Remote Sensing Images Based on YOLOV8 **2025**. *18*, 4060–4073. https://doi.org/10.1109/JSTARS.2025.3526982.
105. Huang, Q.; Yao, R.; Lu, X.; Zhu, J.; Xiong, S.; Chen, Y. Oriented Object Detector With Gaussian Distribution Cost Label Assignment and Task-Decoupled Head **2024**. *62*, 1–16. https://doi.org/10.1109/TGRS.2024.3395440.
106. Liu, S.; Yang, Z.; Li, Q.; Wang, Q. InterMamba: A Visual-Prompted Interactive Framework for Dense Object Detection and Annotation **2025**. *63*, 1–11. https://doi.org/10.1109/TGRS.2025.3559798.
107. Su, Y.; Zhang, T.; Li, F. SA-YOLO: Self-Adaptive Loss Function for Imbalanced Sample Detection. *Journal of Electronics and Information Technology* **2024**, *46*, 123–134.
108. Yang, B.; Han, J.; Hou, X.; Zhou, D.; Liu, W.; Bi, F. FSDA-DETR: Few-Shot Domain-Adaptive Object Detection Transformer in Remote Sensing Imagery **2025**. *63*, 1–16. https://doi.org/10.1109/TGRS.2025.3574245.
109. Wang, B.; Sui, H.; Ma, G.; Zhou, Y.; Zhou, M. GMODet: A Real-Time Detector for Ground-Moving Objects in Optical Remote Sensing Images With Regional Awareness and Semantic–Spatial Progressive Interaction **2025**. *63*, 1–23. https://doi.org/10.1109/TGRS.2025.3526799.

 doi:10.20944/preprints202509.2113.v1

110. Xu, X.; Chen, Z.; Zhang, X.; Wang, G. Context-Aware Content Interaction: Grasp Subtle Clues for Fine-Grained Aircraft Detection **2024**. *62*, 1–19. https://doi.org/10.1109/TGRS.2024.3464851.
111. Sumbul, G.; Cinbis, R.G.; Aksoy, S. Fine-Grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery **2018-02**. *56*, 770–779. https://doi.org/10.1109/TGRS.2017.2754648.
112. Zhang, J.; Zhong, Z.; Wei, X.; Wu, X.; Li, Y. Remote Sensing Image Harmonization Method for Fine-Grained Ship Classification **2024-06-17**. *16*, 2192. https://doi.org/10.3390/rs16122192.
113. Yi, Y.; You, Y.; Li, C.; Zhou, W. EFM-Net: An Essential Feature Mining Network for Target Fine-Grained Classification in Optical Remote Sensing Images **2023**. *61*, 1–16. https://doi.org/10.1109/TGRS.2023.3265669.
114. Zhao, W.; Tong, T.; Yao, L.; Liu, Y.; Xu, C.; He, Y.; Lu, H. Feature Balance for Fine-Grained Object Classification in Aerial Images **2022**. *60*, 1–13. https://doi.org/10.1109/TGRS.2022.3161433.
115. Chen, D.; Tu, W.; Cao, R.; Zhang, Y.; He, B.; Wang, C.; Shi, T.; Li, Q. A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data **2022-02-01**. *106*, 102661. https://doi.org/10.1016/j.jag.2021.102661.
116. Wu, H.; Nie, J.; He, Z.; Zhu, Z.; Gao, M. One-Shot Multiple Object Tracking in UAV Videos Using Task-Specific Fine-Grained Features **2022-01**. *14*, 3853. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs14163853.
117. Jiang, C.; Ren, H.; Li, F.; Hong, Z.; Huo, H.; Zhang, J.; Xin, J. Object detection from aerial multi-angle thermal infrared remote sensing images: Dataset and method **2025-10-01**. *228*, 438–452. https://doi.org/10.1016/j.isprsjprs.2025.07.024.
118. Luo, R.; He, Q.; Zhao, L.; Zhang, S.; Kuang, G.; Ji, K. Geospatial Contextual Prior-Enabled Knowledge Reasoning Framework for Fine-Grained Aircraft Detection in Panoramic SAR Imagery **2024**. *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3487780.
119. Chen, Y.; Huang, J.; Sun, Z.; Xiong, S.; Lu, X. Thread the Needle: Cues-Driven Multiassociation for Remote Sensing Cross-Modal Retrieval **2024**. *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3509639.
120. Zhao, Q.; Lyu, S.; Li, Y.; Ma, Y.; Chen, L. MGML: Multigranularity Multilevel Feature Ensemble Network for Remote Sensing Scene Classification **2023-05**. *34*, 2308–2322. https://doi.org/10.1109/TNNLS.2021.3106391.
121. Guo, W.; Li, S.; Yang, J.; Zhou, Z.; Liu, Y.; Lu, J.; Kou, L.; Zhao, M. Remote Sensing Image Scene Classification by Multiple Granularity Semantic Learning **2022**. *15*, 2546–2562. https://doi.org/10.1109/JSTARS.2022.3158703.
122. Wang, S.; Guan, Y.; Shao, L. Multi-Granularity Canonical Appearance Pooling for Remote Sensing Scene Classification **2020**. *29*, 5396–5407. https://doi.org/10.1109/TIP.2020.2983560.
123. Bai, L.; Liu, Q.; Li, C.; Ye, Z.; Hui, M.; Jia, X. Remote Sensing Image Scene Classification Using Multiscale Feature Fusion Covariance Network With Octave Convolution **2022**. *60*, 1–14. https://doi.org/10.1109/TGRS.2022.3160492.
124. Miao, W.; Geng, J.; Jiang, W. Multigranularity Decoupling Network With Pseudolabel Selection for Remote Sensing Image Scene Classification **2023**. *61*, 1–13. https://doi.org/10.1109/TGRS.2023.3244565.
125. Ye, Z.; Zhang, Y.; Zhang, J.; Li, W.; Bai, L. A Multiscale Incremental Learning Network for Remote Sensing Scene Classification **2024**. *62*, 1–15. https://doi.org/10.1109/TGRS.2024.3353737.
126. Niu, B.; Pan, Z.; Chen, K.; Hu, Y.; Lei, B. Open Set Domain Adaptation via Instance Affinity Metric and Fine-Grained Alignment for Remote Sensing Scene Classification **2023**. *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3276968.
127. Li, Y.; Li, Z.; Su, A.; Wang, K.; Wang, Z.; Yu, Q. Semisupervised Cross-Domain Remote Sensing Scene Classification via Category-Level Feature Alignment Network **2024**. *62*, 1–14. https://doi.org/10.1109/TGRS.2024.3392984.
128. Wang, Y.; Shu, Z.; Feng, Y.; Liu, R.; Cao, Q.; Li, D.; Wang, L. Enhancing Cross-Domain Remote Sensing Scene Classification by Multi-Source Subdomain Distribution Alignment Network **2025-01**. *17*, 1302. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs17071302.
129. Zhu, P.; Zhang, X.; Han, X.; Cheng, X.; Gu, J.; Chen, P.; Jiao, L. Cross-Domain Classification Based on Frequency Component Adaptation for Remote Sensing Images **2024-01**. *16*, 2134. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs16122134.
130. Xiao, R.; Wang, Y.; Tao, C. Fine-Grained Road Scene Understanding From Aerial Images Based on Semisupervised Semantic Segmentation Networks **2022**. *19*, 1–5. https://doi.org/10.1109/LGRS.2021.3059708.
131. Li, Y.; Kong, D.; Zhang, Y.; Tan, Y.; Chen, L. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification **2021-09-01**. *179*, 145–158. https://doi.org/10.1016/j.isprsjprs.2021.08.001.

 doi:10.20944/preprints202509.2113.v1

132. Li, Z.; Xu, W.; Yang, S.; Wang, J.; Su, H.; Huang, Z.; Wu, S. A Hierarchical Graph-Enhanced Transformer Network for Remote Sensing Scene Classification **2024**. *17*, 20315–20330. https://doi.org/10.1109/JSTARS.2024.3491335.
133. Xu, K.; Deng, P.; Huang, H. Vision Transformer: An Excellent Teacher for Guiding Small Networks in Remote Sensing Image Scene Classification **2022**. *60*, 1–15. https://doi.org/10.1109/TGRS.2022.3152566.
134. Shi, C.; Ding, M.; Wang, L.; Pan, H. Learn by Yourself: A Feature-Augmented Self-Distillation Convolutional Neural Network for Remote Sensing Scene Image Classification **2023-01**. *15*, 5620. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs15235620.
135. Wang, G.; Chen, H.; Chen, L.; Zhuang, Y.; Zhang, S.; Zhang, T.; Dong, H.; Gao, P. P2FEViT: Plug-and-Play CNN Feature Embedded Hybrid Vision Transformer for Remote Sensing Image Classification **2023-01**. *15*, 1773. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs15071773.
136. Solomon, A.A.; Agnes, S.A. MSCAC: A Multi-Scale Swin–CNN Framework for Progressive Remote Sensing Scene Classification **2024-09**. *4*, 462–480. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/geographies4030025.
137. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities **2020**. *13*, 3735–3756. https://doi.org/10.1109/JSTARS.2020.3005403.
138. Thapa, A.; Horanont, T.; Neupane, B.; Aryal, J. Deep Learning for Remote Sensing Image Scene Classification: A Review and Meta-Analysis **2023-01**. *15*, 4804. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/rs15194804.
139. Yu, D.; Xu, Q.; Guo, H.; Zhao, C.; Lin, Y.; Li, D. An Efficient and Lightweight Convolutional Neural Network for Remote Sensing Image Scene Classification **2020-01**. *20*, 1999. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/s20071999.
140. Yu, D.; Xu, Q.; Guo, H.; Zhao, C.; Lin, Y.; Li, D. An Efficient and Lightweight Convolutional Neural Network for Remote Sensing Image Scene Classification **2020-01**. *20*, 1999. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/s20071999.
141. Xiong, W.; Xiong, Z.; Cui, Y. A Confounder-Free Fusion Network for Aerial Image Scene Feature Representation **2022**. *15*, 5440–5454. https://doi.org/10.1109/JSTARS.2022.3189052.
142. Yuan, Z.; Zhang, W.; Tian, C.; Rong, X.; Zhang, Z.; Wang, H.; Fu, K.; Sun, X. Remote Sensing Cross-Modal Text-Image Retrieval Based on Global and Local Information **2022**. *60*, 1–16. https://doi.org/10.1109/TGRS.2022.3163706.
143. Hu, G.; Wen, Z.; Lv, Y.; Zhang, J.; Wu, Q. Global–Local Information Soft-Alignment for Cross-Modal Remote-Sensing Image–Text Retrieval **2024**. *62*, 1–15. https://doi.org/10.1109/TGRS.2024.3401031.
144. 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 **2023-01**. *23*, 8437. Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/s23208437.
145. Chen, Y.; Huang, J.; Li, X.; Xiong, S.; Lu, X. Multiscale Salient Alignment Learning for Remote-Sensing Image–Text Retrieval **2024**. *62*, 1–13. https://doi.org/10.1109/TGRS.2023.3340870.
146. 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 **2022**. *60*, 1–19, [2204.09868 [cs]]. https://doi.org/10.1109/TGRS.2021.3078451.
147. Cheng, Q.; Zhou, Y.; Huang, H.; Wang, Z. Multi-Attention Fusion and Fine-Grained Alignment for Bidirectional Image-Sentence Retrieval in Remote Sensing **2022-08**. *9*, 1532–1535. https://doi.org/10.1109/JAS.2022.105773.
148. Yang, L.; Feng, Y.; Zhou, M.; Xiong, X.; Wang, Y.; Qiang, B. A Jointly Guided Deep Network for Fine-Grained Cross-Modal Remote Sensing Text–Image Retrieval **2023-09-15**. *32*, 2350221. Publisher: World Scientific Publishing Co., https://doi.org/10.1142/S0218126623502213.
149. 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 **2021**. *14*, 4284–4297. https://doi.org/10.1109/JSTARS.2021.3070872.
150. Xiu, D.; Ji, L.; Geng, X.; Wu, Y. RSITR-FFT: Efficient Fine-Grained Fine-Tuning Framework With Consistency Regularization for Remote Sensing Image-Text Retrieval **2024**. *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3478176.
151. Zhou, Z.; Feng, Y.; Qiu, A.; Duan, G.; Zhou, M. Fine-Grained Information Supplementation and Value-Guided Learning for Remote Sensing Image-Text Retrieval **2024**. *17*, 19194–19210. https://doi.org/10.1109/JSTARS.2024.3480014.

 doi:10.20944/preprints202509.2113.v1

152. Sun, T.; Zheng, C.; Li, X.; Gao, Y.; Nie, J.; Huang, L.; Wei, Z. Strong and Weak Prompt Engineering for Remote Sensing Image-Text Cross-Modal Retrieval **2025**. *18*, 6968–6980. https://doi.org/10.1109/JSTARS.2025.3534474.
153. Ning, H.; Wang, S.; Lei, T.; Cao, X.; Dou, H.; Zhao, B.; Nandi, A.K.; Radeva, P. Representation discrepancy bridging method for remote sensing image-text retrieval **2025-10-14**. *650*, 130915. https://doi.org/10.1016/j.neucom.2025.130915.
154. Chen, Y.; Huang, J.; Xiong, S.; Lu, X. Integrating Multisubspace Joint Learning With Multilevel Guidance for Cross-Modal Retrieval of Remote Sensing Images **2024**. *62*, 1–17. https://doi.org/10.1109/TGRS.2024.3369042.
155. Mao, Y.Q.; Jiang, Z.; Liu, Y.; Zhang, Y.; Qi, K.; Bi, H.; He, Y. FRORS: An Effective Fine-Grained Retrieval Framework for Optical Remote Sensing Images **2025**. *18*, 7406–7419. https://doi.org/10.1109/JSTARS.2025.3545828.
156. Huang, J.; Feng, Y.; Zhou, M.; Xiong, X.; Wang, Y.; Qiang, B. Deep Multiscale Fine-Grained Hashing for Remote Sensing Cross-Modal Retrieval **2024**. *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3351368.
157. Yu, H.; Yao, F.; Lu, W.; Liu, N.; Li, P.; You, H.; Sun, X. Text-Image Matching for Cross-Modal Remote Sensing Image Retrieval via Graph Neural Network **2023**. *16*, 812–824. https://doi.org/10.1109/JSTARS.2022.3231851.
158. Pan, J.; Ma, Q.; Bai, C. Reducing Semantic Confusion: Scene-aware Aggregation Network for Remote Sensing Cross-modal Retrieval. In Proceedings of the Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. Association for Computing Machinery, 2023-06-12, ICMR '23, pp. 398–406. https://doi.org/10.1145/3591106.3592236.
159. Yang, B.; Wang, C.; Ma, X.; Song, B.; Liu, Z.; Sun, F. Zero-Shot Sketch-Based Remote-Sensing Image Retrieval Based on Multi-Level and Attention-Guided Tokenization. *Remote Sensing* **2024**, *16*, 1653.
160. Liu, Y.; Dang, Y.; Qi, H.; Han, J.; Shao, L. Zero-shot sketch-based remote sensing image retrieval based on cross-modal fusion. *Neural Networks* **2025**, p. 107796.