

A Comprehensive Review of Deep Learning Techniques for Remote Sensing Image Analysis: Methods, Applications, Challenges, and Future Directions

Ledya Safwat^{*,1} and Mohamed Abd Elaziz²

¹Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, 62511, Egypt

²Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, UAE

Abstract Deep learning has rapidly advanced remote sensing image analysis. However, existing review papers typically focus on a single type of deep learning architecture or lack a structured comparative evaluation across tasks. Therefore, this paper provides a structured and critical review of deep learning for remote sensing image analysis on five major model families— Convolutional Neural Networks (CNNs), Transformer-based architectures, Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and hybrid CNN-Transformer models — and evaluates their performance across four core remote sensing tasks: semantic segmentation, object detection, image classification, and change detection. Our work provides a unified comparative framework that highlights architectural trade-offs, dataset-specific performance trends, and task suitability — a perspective not provided by prior surveys. We also highlight several key challenges (e.g., data scarcity, domain shift, class imbalance, and computational complexity) alongside recent solutions (e.g., self-supervised learning, multimodal fusion, lightweight networks, and explainable AI). We hope this review can provide a structured roadmap for developing scalable, efficient, and interpretable remote sensing systems.

Keywords: Remote Sensing, Deep Learning, Semantic Segmentation, Convolutional Neural Networks, Vision Transformers, Change Detection, Image Classification, Data Fusion, Explainable AI, Self-Supervised Learning.

1 Introduction

Remote sensing has become an essential tool for large-scale Earth observation, with applications including environmental monitoring, urban planning, precision agriculture, and disaster management. The rapid development of satellite and aerial platforms has produced high-resolution multi-spectral and multi-temporal image data at an unprecedented rate, necessitating accurate and automated image analysis systems [1, 2].

*Corresponding author: ls.sobhi023@science.zu.edu.eg

Traditionally, remote sensing analysis relied on hand-crafted feature engineering followed by conventional classifiers such as support vector machines (SVMs), k-nearest neighbors (k-NN) or random forests. While these methods have demonstrated reasonable performance in constrained scenarios, they do not generalize well to larger, more complex environments and are sensitive to low-level feature design [3].

Convolutional Neural Networks (CNNs) have since become the de facto standard for semantic segmentation and image classification problems in remote sensing applications, owing to their capacity for automatic hierarchical feature extraction [1]. However, CNNs cannot model long-range spatial information due to the inductive bias of local convolutional filters, and this is essential in most applications.

To remedy this, Transformer-based models have been proposed such as The Vision Transformer (ViT) [4] and its hierarchical variant, the Swin Transformer [5], have also obtained state-of-the-art results on a variety of scene understanding tasks. Furthermore, hybrid architectures combining CNNs and Transformers have also emerged as a leading direction, combining the spatial precision of convolutional features with the global modeling capacity of attention mechanisms [6, 7]. Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, handle the temporal dimension of remote sensing by enabling multi-temporal change detection and time-series analysis [8], while Generative Adversarial Networks (GANs) further assist via data augmentation, image-to-image translation, and super-resolution tasks [9–11].

Despite these advances, important gaps remain in the literature. First, most surveys focus on a single model family, or merely provide a high-level overview without structured comparative evaluation of model architectures or tasks. Second, emerging topics — including self-supervised learning, multimodal data-fusion, light-weight model design, and explainable AI (XAI) encounter difficulties in their common implementation within the remote sensing domain. Third, domain-specific challenges such as data scarcity, domain shift, class imbalance and high computational complexity continue to obstruct the transition from research to operational deployment.

To fill these gaps, this paper makes the following contributions:

- The well-structured and critical analysis of CNNs, Transformers, RNNs, GANs, and hybrid models, to perform four tasks (semantic segmentation, Object Detection, image classification, and change detection tasks) using consistent benchmark metrics. Unlike prior surveys, our framework explicitly maps each architecture to task suitability and dataset-specific performance trends
- A unified analysis of architectural trade-offs (global context, temporal modeling, data efficiency, and computational cost), providing practitioners with clear selection guidance for real-world deployment scenarios.
- A survey of open research challenges. Forward-looking discussion of new directions for scalable, efficient and interpretable remote sensing systems.

The remainder of this paper is organized as follows. Section 2 reviews related survey works. Section 3 presents a detailed analysis of deep learning techniques. Section 4 discusses key applications. Section 5 identifies challenges and outlines future directions. Section 6 concludes the paper.

2 RELATED WORK

The rapid growth of deep learning applications in remote sensing has inspired numerous surveys. To contextualise the contributions of this paper, we organize existing reviews into four thematic categories and identify the specific gaps this work fills.

2.1 CNN-Focused Surveys

In early surveys, most attention was given to CNNs. Zhu et al. [1] provided a foundational review highlighting CNN effectiveness in spatial feature extraction for classification and segmentation. However, their scope excludes Transformers, hybrid models, or self-supervised learning. Similarly, Rewehel et al. [3] assessed practical applications including land cover classification and environmental monitoring but offered limited discussion of attention mechanisms and temporal modeling.

2.2 Transformer-Oriented Surveys

Khan et al. [12] provided a thorough comparison of Vision Transformers for computer vision tasks, including self-attention mechanisms and specific architectures such as ViT [4] and the Swin Transformer [5]. However, it is not specific to remote sensing, and omits remote sensing-specific tasks such as multi-modal data fusion and large-scale geographic coverage.

2.3 Application-Specific Surveys

In their review of deep learning-based change detection, Lei et al. [13] focus on temporal modeling of RNN- and Siamese network-based architectures in detail, but did not cover segmentation or object detection approaches. Haut et al. [14] surveyed deep learning approaches for remote sensing image super-resolution, covering CNN-based and GAN-based enhancement methods, but did not address broader architectural comparisons or downstream task efficiency.

2.4 Broad Multi-Architecture Surveys

As concluded by Paheding et al. [2] surveyed CNNs, Transformers, and GANs across various remote sensing tasks — one of the most comprehensive recent reviews. Nevertheless, it misses a structured comparative framework evaluating architectural trade-offs across specific tasks and benchmark datasets, and emerging topics such as self-supervised learning and XAI are discussed only superficially. Asif et al. [6] explored hybrid CNN-Transformer architectures but did not provide a unified analysis spanning all five major model families.

2.5 Positioning of This Survey

No existing survey provides a unified comparative evaluation of all five major model families across the four core remote sensing tasks, nor sufficiently addresses the interplay between architectural design and task-specific performance. This paper directly addresses these gaps. Table 1 below summarizes the positioning of this work relative to existing surveys; the final column entries confirm that only this paper covers all five model families

and all four tasks simultaneously, while also including XAI and self-supervised learning (SSL).

Table 1: Positioning of This Survey Relative to Existing Works

Survey	Year	CNN	Trans.	RNN	GAN	Hybrid	Seg.	Det.	Cls.	CD	XAI	SSL
Zhu et al. [1]	2022	✓	×	×	×	×	✓	✓	✓	×	×	×
Paheding et al. [2]	2024	✓	✓	×	✓	×	✓	✓	✓	×	×	×
Rewehel et al. [3]	2023	✓	×	×	×	×	✓	×	✓	×	×	×
Khan et al. [12]	2022	×	✓	×	×	×	✓	✓	✓	×	×	×
Asif et al. [6]	2025	✓	✓	×	×	✓	✓	✓	×	×	×	×
Lei et al. [13]	2025	✓	×	✓	×	×	×	×	×	✓	×	×
This Paper	2026	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

3 DEEP LEARNING TECHNIQUES FOR REMOTE SENSING IMAGE ANALYSIS

Deep learning has fundamentally transformed remote sensing image analysis by enabling automatic hierarchical feature learning from high-dimensional data. This section provides a structured and critical analysis of five major deep learning model families. A unified comparative analysis is presented in Table 2.

3.1 Convolutional Neural Networks (CNNs)

CNNs remain widely used as the backbone architecture in remote sensing due to their strong capacity for local spatial feature extraction. The encoder-decoder architecture presented by U-Net [15] has achieved great success in many pixel-wise segmentation by combining high-level semantic features with low-level spatial details through skip connections. DeepLabV3+ [16] improved segmentation with atrous (dilated) convolutions and Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context aggregation. Feature Pyramid Networks were used for detection (FPN) [17] introduce scale variation by constructing multi-scale feature hierarchies, this allows robust detection of both large and small objects within high-resolution remote sensing imagery.

Despite their effectiveness, CNN-based models are fundamentally constrained by local receptive fields, limit the model’s ability to capture long-range spatial dependencies [1]. The over all performance U-Net obtains an overall accuracy of approximately 85.9% and mIoU of 73.9% on the ISPRS Potsdam dataset [15], while DeepLabV3+ achieves an overall accuracy of approximately 87.3% and mIoU of 82.7% on the same benchmark [16].

3.2 Transformer-Based Models

The Vision Transformer (ViT) [4] denoted a paradigm shift by substituting convolutional operations with self-attention mechanisms. ViT processes an input image by splitting it into fixed-size patches and applying multi-head self-attention across all patch tokens, allowing global context modeling from the earliest layers.

The Swin Transformer [5] tackles ViT’s quadratic computational complexity by introducing shifted window attention, achieving linear complexity and producing multi-scale features compatible with dense prediction tasks. In remote sensing applications, the Swin

Transformer along with a UPerNet decoder achieves mIoU scores of 83–89% on the ISPRS Potsdam benchmark, outperforming CNN baselines by 3–6% [18]. For scene classification on the AID dataset [19], ViT-based models achieve 95–97% accuracy, compared to 92–94% for ResNet-based CNNs [20]. These gains are most pronounced when large-scale pretraining data are available; performance advantages narrow considerably in low-data regimes.

3.3 Recurrent Neural Networks (RNNs)

Long Short-Term Memory (LSTM) [8] and Gated Recurrent Unit (GRU) networks apply recurrent hidden state propagation to model time-dependent relationships. In remote sensing, RNNs have been used mainly for multi-temporal analysis in a time series of images for land-surface change detection and dynamic processes, such as crop growth cycles or urbanization. For change detection of the LEVIR-CD dataset [21], CNN-LSTM hybrid models achieve F1 scores of 87–91% [22].

3.4 Generative Adversarial Networks (GANs)

GANs are widely applied for image generation, transformation, and optimization in remote sensing. Pix2Pix [9] proposes a supervised image-to-image translation, applicable to SAR-to-optical conversion under cloud-cover conditions. CycleGAN [10] extends this strategy to unpaired image translation through cycle-consistency constraints. For super-resolution, SRGAN [11] and its variants obtain PSNR improvements of 2–4 dB over bicubic upsampling on satellite imagery benchmarks [23]. Additionally, the use of GANs for data augmentation has led to a 2 to 5% improvement in downstream classification accuracy in low-data scenarios [24].

3.5 Hybrid CNN-Transformer Architectures

More general hybrid architectures that combine CNNs with Transformers allow models to make use of CNNs for local fine-grained feature extraction and Transformers for global long-range dependencies. TransUNet [25] was originally developed for medical image segmentation and led to a variety of hybrid designs in remote sensing; it should not be treated as a direct remote sensing contribution but as a foundational reference for subsequent adaptations in this domain. On the DOTA object detection benchmark [26], hybrid CNN-Transformer detectors achieve mAP scores of 78–83%, compared to 72–76% for CNN-only baselines [27]. On ISPRS segmentation benchmarks, hybrid models regularly achieve mIoU scores above 87% [7, 18].

3.6 Comparative Analysis of Techniques

Table 2 presents a unified comparison of the five model families. CNN-based models offer the best data efficiency and lowest computational cost, making them suitable for resource-constrained deployment. Transformer-based models achieved superior global context modeling and state-of-the-art effectiveness in scene classification, but require large-scale training data. RNNs remain the top choice for temporal and sequential analysis tasks. GANs obtain unique capabilities for image synthesis and data augmentation. The hybrid of CNN-Transformer architectures currently represent state of the art across

most tasks, combining local and global feature learning at the cost of very high computational complexity. Reported performance ranges reflect variability across datasets and experimental settings; results on individual benchmarks should be interpreted accordingly. In the next section, several applications will be explained.

Table 2: Comparison of Deep Learning Techniques for Remote Sensing

Model	Common Models	Global Context	Temp.	Data Eff.	Comp. Cost	Best Tasks	Typical Perf.
CNN [15–17]	U-Net, DeepLabV3+, FPN	Low	×	High	Low	Segmentation, Classification	mIoU: 73–87%
Transformer [4, 5]	ViT, Swin	High	×	Low	High	Scene Understanding, Classification	OA: 91–97%
RNN [8]	LSTM, GRU	Low	✓	Medium	Medium	Change Detection, Time-series	F1: 87–91%
GAN [9–11]	Pix2Pix, CycleGAN, SRGAN	—	×	Medium	High	Super-resolution, Augmentation	PSNR: +2–4 dB
Hybrid [7, 18, 25]	TransUNet, CNN+Swin	High	×	Medium	Very High	All Tasks	mIoU: 87–89%

4 APPLICATIONS OF DEEP LEARNING IN REMOTE SENSING

Deep learning techniques have been effectively applied across a broad spectrum of remote sensing applications. This section reviews the four primary application domains — semantic segmentation, image classification, object detection, and change detection along benchmark performance, examining representative methods, and key challenges specific to each task. Figure 1 illustrates examples of all applications and Table 3 presents the definitions of the evaluation metrics used.

4.1 Semantic Segmentation

Semantic segmentation demands allocation of semantic class label to every pixel in an image, and is essential for land use, land cover mapping, urban infrastructure delineation, vegetation monitoring, and water body extraction [1]. The field has originated from early Fully Convolutional Networks (FCNs) [28] through encoder–decoder architectures such as U-Net [15] to modern Transformer-based and hybrid models integrating global context modeling [7, 29].

Representative results on the ISPRS Potsdam benchmark are summarized in Table 4, CNN-based models such as DeepLabV3+ [16] achieve overall accuracy (OA) values of 86–87%, while Transformer-based models such as SegFormer [29] reach 91–92% OA. Hybrid CNN–Transformer architectures currently represent the state of the art, achieving OA values above 91% and mIoU scores exceeding 84%.

4.2 Image Classification

Image classification assigns a single semantic label to an entire image or scene. Transfer learning has emerged as the dominant strategy, with pretrained models fine-tuned on remote sensing datasets achieving strong performance even with limited labeled data [1, 32]. On the AID dataset [19], fine-tuned ResNet-50 [20] achieves approximately 93.97%

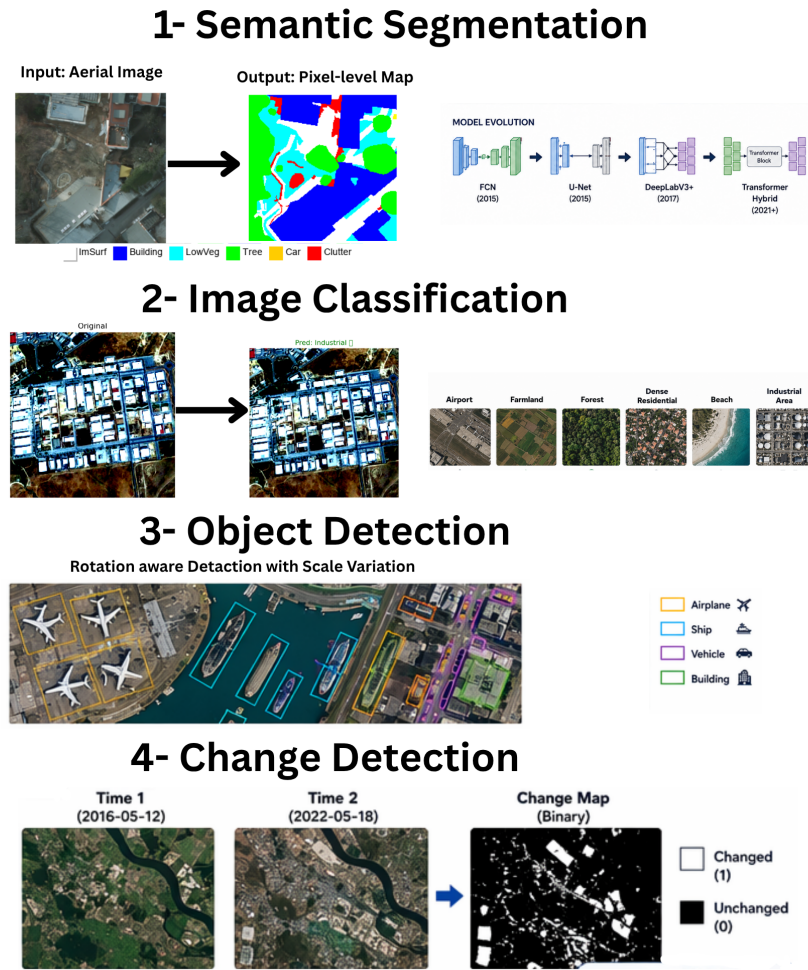


Figure 1: Application overview

accuracy [33], while ViT-based models [4] reach 95.8%. On NWPU-RESISC45 [32], state-of-the-art hybrid models achieve accuracies exceeding 97%, as shown in Table 5.

4.3 Object Detection

Object detection in remote sensing involves simultaneous localization and classification of objects within an image. This task is considerably more challenging due to the top-down viewing perspective, large variation in object scale and orientation, extreme object density in urban scenes, and complex backgrounds [26]. On the DOTA-v1.0 dataset [26] with 15 object categories, state-of-the-art oriented object detectors achieve mAP scores of 68–82% across different architectures. Hybrid CNN–Transformer detectors achieve consistent mAP gains of 4–6% over CNN-only baselines, as shown in Table 6.

4.4 Change Detection

Change detection identifies significant differences in land surface between images acquired at different time points. Applications include urban expansion monitoring, deforestation detection, agricultural dynamics, and disaster assessment [13, 21]. For binary change detection on the LEVIR-CD dataset [21], methods achieve F1 scores ranging from 83.4%

Table 3: Summary of Evaluation Metrics Used in This Paper

Metric	Full Name	Applies To	Interpretation
OA	Overall Accuracy	Segmentation, Classification	% of correctly labeled pixels/samples. Higher = better.
mIoU	Mean Intersection over Union	Semantic Segmentation	Mean class overlap ratio (0–1). Higher = better per-class accuracy.
mAP	Mean Average Precision	Object Detection	Mean AP across all categories & IoU thresholds. Higher = better.
F1	F1-Score	Change Detection	Harmonic mean of Precision & Recall (0–1). Higher = better balance.
IoU	Intersection over Union (binary)	Change Detection	Overlap ratio of predicted vs. true change mask. Higher = better.
PSNR	Peak Signal-to-Noise Ratio	Super-Resolution	Reconstruction fidelity in dB. Higher = less distortion.
SSIM	Structural Similarity Index Measure	Super-Resolution	Perceptual quality (0–1). Higher = more visually faithful output.

Note: All metrics are reported as percentages (%)

(CNN baselines) to above 91% (hybrid Transformer models), as summarised in Table 7. The improvement from CNN to hybrid architectures reflects the benefit of capturing long range spatial dependencies when comparing temporally separated image pairs.

4.5 Task-Specific Deep Learning Paradigms in Remote Sensing

4.5.1 Super-Resolution and Image Enhancement

Super-resolution reconstructs high-resolution imagery from low-resolution inputs, directly benefiting downstream tasks such as object detection and semantic segmentation [14]. GAN-based models such as SRGAN [11] achieve PSNR improvements of 2–4 dB and SSIM gains of 0.05–0.12 over bicubic interpolation at $4\times$ upscaling on satellite imagery benchmarks [23]. More recent diffusion model-based approaches further improve perceptual quality by generating sharper high-frequency details compared to GAN-based methods, though at higher computational cost [42].

4.5.2 Multimodal Data Fusion

Remote sensing platforms acquire data across multiple modalities including optical, Synthetic Aperture Radar (SAR), hyperspectral, and LiDAR sensors, each providing complementary surface information [43, 44]. SAR–optical fusion is particularly valuable under cloud-cover conditions where optical imagery is unavailable; deep learning fusion models have demonstrated overall accuracy improvements of up to 5.7% over optical-only baselines on land cover classification benchmarks [45]. Hyperspectral–LiDAR fusion achieves land cover classification accuracies ranging from 94.4% to 96.0% on the IEEE GRSS

Table 4: Representative Semantic Segmentation Results on ISPRS Benchmarks

Method	Architecture	Dataset	OA (%)	mIoU (%)	Year
U-Net [15]	CNN	ISPRS Potsdam	85.90	73.90	2015
DeepLabV3+ [16]	CNN	ISPRS Potsdam	87.30	82.70	2018
SegFormer [29]	Transformer	ISPRS Potsdam	91.60	84.74	2021
ST-UNet [18]	Hybrid (CNN+Swin)	ISPRS Vaihingen	—	84.10	2022
UNetFormer [7]	Hybrid (CNN+Trans.)	ISPRS Potsdam	91.56	84.74	2022

*Baseline results reproduced from [30]. OA and mIoU from [31].

Table 5: Representative Remote Sensing Image Classification Results

Method	Architecture	Dataset	OA (%)	Year
ResNet-50 [20]	CNN	AID [19]	93.97	2018
EfficientNet-B3 [34]	CNN	NWPU-RESISC45 [32]	92.12	2021
ViT-B/16 [4]	Transformer	AID [19]	95.78	2021
Swin-B [5]	Transformer	NWPU-RESISC45 [32]	94.70	2021
ST-UNet [18]	Hybrid (CNN+Swin)	AID [19]	97.20	2022

* Results reproduced on AID/NWPU in [18].

Houston et al. [46, 47].

Fusion strategies are broadly categorized into three paradigms. Early fusion concatenates multimodal inputs at the data level prior to feature extraction [42]. Late fusion independently processes each modality and combines their prediction outputs [42]. Intermediate fusion integrates complementary features at multiple network layers and has consistently demonstrated superior performance compared to early or late fusion strategies, particularly for complex urban land cover classification [43, 45]. Key challenges include spatial misalignment between sensors, temporal asynchrony between acquisitions, and increased computational complexity introduced by processing multiple input streams. These challenges are discussed further in next section from an architectural perspective

4.5.3 Disaster Management and Environmental Monitoring

Deep learning allows rapid, accurate, and large-scale analysis in time-sensitive disaster scenarios [21, 48]. For flood detection, semantic segmentation models applied to Sentinel-1 SAR imagery have been validated on the Sen1Floods11 benchmark dataset [49], demon-

Table 6: Object Detection Results on the DOTA-v1.0 Benchmark (OBB Task)

Method	Architecture	mAP (%)	Year
Faster R-CNN [35]	CNN	68.16	2018
YOLOv5 [36]	CNN	72.30	2020
Oriented R-CNN [37]	CNN	75.87	2021
AO2-DETR [38]	Transformer	79.22	2022
Oriented DINO [27]	Hybrid (CNN+Trans.)	81.36	2023

* Faster R-CNN baseline from the original DOTA paper [26].

YOLOv5 results reproduced from the Ultralytics repository [36].

Table 7: Change Detection Results on the LEVIR-CD Benchmark

Method	Architecture	Dataset	F1 (%)	IoU (%)	Year
FC-EF [39]	CNN	LEVIR-CD [21]	83.40	71.54	2018
BIT [22]	Transformer	LEVIR-CD [21]	89.31	80.68	2021
ChangeFormer [40]	Hybrid (Trans.+MLP)	LEVIR-CD [21]	90.40	82.48	2022
FTN [41]	Hybrid (Swin+FPN)	LEVIR-CD [21]	91.01	83.51	2022

strating the all-weather capability of SAR for reliable flood extent delineation [50]. Post-earthquake building damage assessment using very high-resolution optical imagery has been extensively benchmarked on the xBD dataset [48], the largest publicly available building damage assessment dataset, which contains 850,736 annotated building instances across 45,362 km² of imagery covering 19 disaster types. Transformer-based hierarchical models applied to xBD achieve competitive F1 scores for damage classification compared to CNN baselines [51]. The critical requirements for disaster management applications are processing speed, model reliability under domain shift, and the ability to generalize to unseen geographic regions and disaster types. In the next section defining this challenges and the next future work.

5 CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Despite remarkable progress, the transition from research-oriented models to robust real-world deployment remains constrained by fundamental challenges spanning data, architecture, computational, and interpretability dimensions. This section provides a structured critical analysis of each challenge category, reviews existing mitigation strategies, and identifies the most promising future research directions.

5.1 Data-Related Challenges

5.1.1 Scarcity of Labeled Data

The lack of high-quality annotated data is the most fundamental bottleneck of DL for remote sensing. Unlike general natural images in datasets such as ImageNet, remote sensing data annotation is labor-intensive and requires domain expertise especially for dense pixel-wise prediction tasks such as semantic segmentation [1]. This is particularly exacerbated for change detection and disaster assessment problems where annotating the temporal pairs is expensive and time-consuming. Current mitigation strategies include transfer learning from large-scale pretrained models [4, 20], semi-supervised learning with consistency regularization [2], and self-supervised pretraining methods through contrastive learning and masked image modeling. These techniques have consistently reduced the need for labeled data on various remote sensing datasets in both classification and segmentation benchmarks [1, 2].

5.1.2 Class Imbalance

The class imbalance between the classes that occupy a majority of most remote sensing datasets (the background classes: vegetation, bare soil, etc.) and the semantically rich objects (the foreground classes: buildings, roads, vehicles, etc.) is very high [1, 32]. This poses a challenge to deep learning methods that tend to focus on the majority classes. Existing approaches include weighted cross-entropy loss functions [15], focal loss [52], oversampling and data augmentation approaches, such as Generative Adversarial Network [24] based approaches, are standard, and more principled solutions to the statistical imbalance of the data observed in remote sensing remain open research questions.

5.1.3 Domain Shift

Models trained on one geographic region or sensor type frequently degrade when applied to different domains, due to the diversity in satellite platforms, sensor specifications, atmospheric conditions, seasonal variations, and geographic landscapes encountered in real-world deployment [2]. Adversarial domain adaptation and domain generalization approaches have shown promise in cross-sensor transfer scenarios [1]. However, achieving reliable generalization across the full diversity of real-world remote sensing conditions remains an open challenge.

5.2 Architectural and Computational Challenges

5.2.1 Computational Complexity

High computational requirements of state-of-the-art deep learning models, particularly Transformer-based architectures, represent a significant barrier to practical deployment in remote sensing applications [4, 5]. Model compression techniques such as knowledge distillation — which transfers the predictive behaviour of a large teacher model to a compact student model [53] — quantization, which reduces the numerical precision of model weights from 32-bit to 8-bit representations [54], and Neural Architecture Search (NAS) have been applied to discover efficient remote sensing architectures [55]. However, models achieving both state-of-the-art accuracy and real-time inference on standard hardware remain largely unavailable for high-resolution remote sensing imagery.

5.2.2 Multi-Scale and Fine-Detail Modeling

Remote sensing images contain objects spanning multiple orders of magnitude in spatial scale. While multi-scale architectures such as FPN [17] and hierarchical Transformers such as the Swin Transformer [5] have made significant progress in handling scale variation, accurately detecting and segmenting very small objects within complex and cluttered backgrounds continues to challenge even the most advanced models [7, 26].

5.2.3 Multimodal Data Fusion

Whereas Section 4.5.2 discussed multimodal fusion from an applications perspective, we focus here on the architectural challenges it poses. The heterogeneous nature of remote sensing data — encompassing optical, SAR, hyperspectral, and LiDAR modalities with fundamentally different physical measurement principles, spatial resolutions, and noise characteristics — introduces significant challenges for unified deep learning frameworks [43, 44]. Developing robust, generalizable, and computationally efficient multimodal fusion frameworks that accommodate variable modality availability at inference time, and handle spatial misalignment and temporal asynchrony between sensors, represents an important open research direction.

5.3 Interpretability and Reliability Challenges

5.3.1 Lack of Model Interpretability

Deep learning models are widely characterized as black-box systems whose internal decision-making processes are opaque. In high-stakes remote sensing applications such as disaster damage assessment and environmental law enforcement, the inability to explain model predictions limits operational trust and regulatory acceptance [56, 57]. Explainable AI (XAI) techniques including gradient-based saliency maps [58], Class Activation Mapping (CAM) [59], Gradient-weighted CAM (Grad-CAM) [60], and attention visualization [4] have been applied to provide post-hoc explanations of model decisions in remote sensing contexts [61]. However, current XAI methods remain limited in providing faithful, semantically meaningful explanations that align with expert domain knowledge, and a comprehensive evaluation framework for XAI in remote sensing remains an open research problem [56, 61].

5.3.2 Uncertainty Quantification

Overconfident incorrect predictions can have serious consequences in applications such as flood mapping, building damage assessment, and ecological monitoring [48]. Bayesian deep learning approaches and ensemble methods have been explored for uncertainty estimation in remote sensing, but their high computational cost limits practical deployment at scale [1]. Lightweight uncertainty quantification methods suitable for large-scale remote sensing analysis, and evaluation frameworks for quantifying prediction robustness in operational settings, represent an important gap in the current literature.

Table 8: Summary of Open Challenges, Current Solutions, and Future Research Directions in Deep Learning for Remote Sensing

Challenge	Impact Level	Current Solutions	Future Direction
Labeled data scarcity [1]	Very High	Transfer learning [20], SSL [62], semi-supervised learning [2]	Foundation models [63], self-supervised pretraining [64]
Class imbalance [32]	High	Focal loss [52], GAN augmentation [9], weighted cross-entropy [15]	Principled imbalance modeling
Domain shift [2]	Very High	Adversarial adaptation, domain generalisation	Meta-learning, diverse pretraining [65]
Computational complexity [5]	High	Knowledge distillation [53], quantization [54], NAS	Lightweight architectures [55], edge deployment
Multi-scale modeling [26]	Medium	FPN [17], hierarchical Transformers [5]	Adaptive scale-aware architectures
Multimodal fusion [44]	High	Early/late/intermediate fusion [42], cross-modal attention [43]	Flexible multimodal Transformers
Interpretability [56]	High	Grad-CAM [60], saliency maps, attention visualisation [4]	Remote sensing XAI frameworks [61]
Uncertainty quantification [1]	Medium	Bayesian deep learning, ensembles	Lightweight uncertainty estimation

5.4 Future Research Directions

5.4.1 Self-Supervised and Foundation Models

Large-scale self-supervised pretraining frameworks specifically designed for remote sensing data represent perhaps the most impactful near-term research direction. Geospatial foundation models such as SatMAE [62], SSL4EO-S12 [64], and Prithvi [63] show few-shot and zero-shot generalization, following pretraining on large-scale multispectral satellite collections; these models improve upon ImageNet-pretrained models by approximately 4–5% [65]. Extension of these methods to cover multi-spectral, SAR and multi-temporal data modalities has shown promise in reducing the amounts of labeled data required for model training.

5.4.2 Lightweight and Edge-Deployable Architectures

Many applications of remote sensing increasingly demand near-real-time processing for disaster response and environmental monitoring, motivating the development of lightweight network models that can be deployed on edge computing devices and onboard satellite processors [50]. This direction encompasses neural architecture search for remote sensing, structured pruning of Transformer-based models, knowledge distillation [53, 55], and hardware-aware training objectives that optimise for inference efficiency on resource-constrained platforms.

5.4.3 Robust Domain Adaptation and Generalization

Developing models that generalize robustly across the full diversity of remote sensing domains — encompassing sensor types, geographic regions, seasonal conditions, and imaging configurations — without requiring domain-specific retraining is essential for scalable deployment [2]. Promising directions include meta-learning approaches that enable rapid adaptation to new domains with minimal labeled data, and large-scale pretraining on geographically and temporally diverse datasets [64, 65] to build broadly generalizable representations.

5.4.4 Advanced Multimodal Fusion

Future work should prioritize flexible multimodal fusion frameworks that accommodate variable modality availability, handle spatial and temporal misalignment, and adaptively weight modality contributions based on data quality and the context of scenes [44]. Attention-based cross-modal Transformer architectures [43] and unified multimodal pre-training strategies are particularly promising directions to consider.

5.4.5 Explainable and Trustworthy AI

Closing the interpretability, uncertainty and reliability gaps is important to adopt remote sensing deep learning systems in operational or high-stakes usages [56, 57]. In the upcoming research should develop XAI methods tailored to the spectral and spatial characteristics of remote sensing data, uncertainty-adaptive training objectives, and evaluation frameworks that not only the model’s predictive performance but also the quality of explanations and reliability of predictions [61].

5.4.6 Temporal and Dynamic Modeling

Recent years have seen an explosion of detailed time-series satellite data from platforms such as Sentinel-1/2 and Planet Labs, evolving deep learning architectures capable of effectively modeling complex temporal dynamics — including missing observations, irregular time sampling, and long-range temporal dependencies — signifies a significant and underexplored opportunity [13,22]. Transformer-based temporal attention mechanisms [4] and state space models offer innovative architectural foundations for this direction.

Table 8 offers an overview of the key challenges and potential future directions

6 CONCLUSION

In this survey, we provide a structured, critical, and comparative review of deep learning for remote sensing image analysis based on the five models CNNs, Transformers, RNNs, GANs, and hybrid methods across four tasks: semantic segmentation, object detection, image classification, and change detection. The apparent redundancy among network architectures is addressed through a critical assessment of the available literature. Although CNN based methods are powerful and data-efficient, are being consistently surpassed by Transformer-based and hybrid architectures as dataset sizes grow. Hybrid CNN–Transformer models have achieved state-of-the-art performance with mIoU exceeding 87% on ISPRS benchmarks [7, 18] and mAP above 80% on DOTA [26]. RNNs have been used multi-temporal analysis [8] and GANs have been used for data augmentation

and super-resolution [9, 10]. The field is obviously changing toward two complementary goals: universality through geospatial foundation models and self-supervised pretraining that decrease reliance on labeled data; and efficiency through lightweight architectures and edge-deployable models that bring state-of-the-art efficiency to resource-constrained platforms. Fundamental barriers nonetheless remain, including data scarcity [1], domain shift [2], computational complexity [48], and LOW interpretability [56], and the comparative framework presented here is intended as a practical roadmap toward remote sensing models that are accurate, efficient, and trustworthy.

References

- [1] X. X. Zhu et al. Deep learning in remote sensing: A comprehensive review. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):8–36, 2022.
- [2] S. Paheding et al. Advancing horizons in remote sensing: A comprehensive survey of deep learning models. *Neural Computing and Applications*, 2024.
- [3] E. Rewehel et al. Deep learning methods used in remote sensing images: A review. *Journal of Environmental and Earth Sciences*, 2023.
- [4] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [5] Z. Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [6] M. Asif et al. Hybrid deep learning models for remote sensing applications. *Scientific Reports*, 2025.
- [7] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022.
- [8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [9] P. Isola et al. Image-to-image translation with conditional adversarial networks (pix2pix). In *CVPR*, 2017.
- [10] J.-Y. Zhu et al. Unpaired image-to-image translation using cycle-consistent adversarial networks (cyclegan). In *ICCV*, 2017.
- [11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017.
- [12] S. Khan et al. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022.

- [13] T. Lei et al. Deep learning for remote sensing image change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [14] Juan Mario Haut, Mercedes Eugenia Paoletti, Javier Plaza, and Antonio Plaza. Remote sensing image superresolution using deep residual channel attention networks. *IEEE Geoscience and Remote Sensing Letters*, 16(11):1670–1674, 2019.
- [15] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.
- [16] L.-C. Chen et al. Encoder-decoder with atrous separable convolution for semantic image segmentation (deeplabv3+). In *ECCV*, 2018.
- [17] T.-Y. Lin et al. Feature pyramid networks for object detection. In *CVPR*, 2017.
- [18] Xin He, Yong Zhou, Jiewen Zhao, Di Zhang, Rui Yao, and Yang Xue. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.
- [19] G.-S. Xia et al. Aid: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [21] H. Chen et al. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection (levir-cd). *Remote Sensing*, 2020.
- [22] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.
- [23] Shenglong Lei, Zhenwei Shi, and Zhengxia Zou. Super-resolution for remote sensing images via local–global combined network. *IEEE Geoscience and Remote Sensing Letters*, 14(8):1243–1247, 2017.
- [24] Xuejie Hao et al. A review of data augmentation methods of remote sensing image target recognition. *Remote Sensing*, 15(3):827, 2023.
- [25] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [26] G.-S. Xia et al. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018.
- [27] Jian Cao, Qilong Yao, and Junnan Liu. Oriented DINO: Towards high-quality oriented object detection using transformer. *arXiv preprint arXiv:2308.04394*, 2023.
- [28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.
- [29] E. Xie et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

- [30] Haoran Wang et al. LGD-DeepLabV3+: An enhanced framework for remote sensing semantic segmentation via multi-level feature fusion and global modeling. *Sensors*, 26(3):1008, 2025.
- [31] JiangQing Wang, Ting Chen, Lu Zheng, et al. A multi-scale remote sensing semantic segmentation model with boundary enhancement based on UNetFormer. *Scientific Reports*, 15:14723, 2025.
- [32] G. Cheng et al. Remote sensing image scene classification: Benchmark and state of the art (nwpu-resisc45). *Proceedings of the IEEE*, 2017.
- [33] Gang Li, Chunju Zhang, Mengmeng Wang, Feng Gao, and Xiuping Zhang. Scene classification of high-resolution remote sensing image using transfer learning with multi-model feature extraction framework. In *PRCV*, pages 230–241, 2018.
- [34] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.
- [35] S. Ren et al. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.
- [36] Glenn Jocher et al. YOLOv5 by ultralytics. <https://github.com/ultralytics/yolov5>, 2022. Version 7.0.
- [37] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *ICCV*, pages 3520–3529, 2021.
- [38] Lei Dai, Qian Liu, Jianzhuang Ma, Mingxiang Xia, and Fangqing Chen. AO2-DETR: Arbitrary-oriented object detection transformer. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [39] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, 2018.
- [40] W. Bandara and V. Patel. A transformer-based siamese network for change detection. In *IGARSS*, 2022.
- [41] Qingyun Yan, Jincheng Yang, Liqian Li, and Zening Wang. Fully transformer network for change detection of remote sensing images. In *ACCV*, pages 1691–1707, 2022.
- [42] Rui Zhang et al. Land use and land cover classification with deep learning-based fusion of SAR and optical data. *Remote Sensing*, 17(7):1298, 2025.
- [43] Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):375–395, 2020.
- [44] Pedram Ghamisi, Behnood Rasti, Naoto Yokoya, Qian Wang, Bernhard Hofle, Lorenzo Bruzzone, Francesca Bovolo, Mingmin Chi, Katharina Anders, Richard Gloaguen, Peter M. Atkinson, and Jon Atli Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):6–39, 2019.

- [45] Rui Zhang et al. SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216:260–275, 2024.
- [46] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019.
- [47] Xu Tang et al. Classification of hyperspectral and LiDAR data by transformer-based enhancement. *International Journal of Remote Sensing*, 45(20):7591–7614, 2024.
- [48] Ritwik Gupta, Bryce Goodman, Nirav Patel, Richard Hosfelt, Sandra Sajeew, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xBD: A dataset for assessing building damage from satellite imagery. In *CVPR Workshops*, pages 10–17, 2019.
- [49] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *CVPR Workshops*, pages 835–845, 2020.
- [50] Shan Lv et al. High-performance segmentation for flood mapping of HISEA-1 SAR remote sensing images. *Remote Sensing*, 14(21):5504, 2022.
- [51] Navneet Kaur, Chung-Ching Lee, Ali Mostafavi, and Ali Mahdavi-Amiri. Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. *Computer-Aided Civil and Infrastructure Engineering*, 38(15):2072–2091, 2023.
- [52] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.
- [53] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [54] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, pages 2704–2713, 2018.
- [55] Fang Wang et al. Applications of knowledge distillation in remote sensing: A survey. *Information Fusion*, 110:102453, 2024.
- [56] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.
- [57] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.
- [58] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

- [59] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.
- [60] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.
- [61] Ioannis Kakogeorgiou and Konstantinos Karantzas. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103:102520, 2021.
- [62] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS*, volume 35, pages 197–211, 2022.
- [63] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, et al. Foundation models for generalist geospatial artificial intelligence. In *arXiv preprint arXiv:2310.18660*, 2023.
- [64] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023.
- [65] Matías Mendieta, Boran Han, Xingjian Shi, Hengbo Zhu, Chih-Yao Chen, Jian Li, and Ming-Hsuan Chen. Towards geospatial foundation models via continual pretraining. In *ICCV*, pages 16806–16816, 2023.