Antony Daniel Rex. J, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science,
Engineering and Technology
An Open Access Journal

Deep Learning Based Semantic Segmentation of Satellite Imagery for Land Cover Classification

Antony Daniel Rex. J[1], Dr. R. Vidya[2], Dr. A. Martin[3]

[1](Research Scholar, Assistant Professor
St. Joseph's College of Arts and Science (Autonomous)
Cuddalore-1
Affiliated to Annamalai University, Chidambaram, Tamil Nadu, India.

[2](Research Supervisor, Associate Professor
St. Joseph's College of Arts and Science (Autonomous)
Cuddalore-1
Affiliated to Annamalai University, Chidambaram, Tamil Nadu, India.

[3](Assistant Professor
Department of Computer Science, Central University of Tamil Nadu
Thiruvarur, India.

Abstract- Land cover classification from remote sensing imagery underpins agricultural monitoring, urban planning, disaster response, and environmental assessment, and deep learning has become the dominant paradigm for extracting land cover information from satellite scenes at pixel-level granularity. This survey reviews recent advances in deep learning-based land cover classification and semantic segmentation, organizing the literature into convolutional neural network (CNN)-based encoder–decoder architectures, attention and self-attention mechanisms, transformer and hybrid CNN–Transformer architectures, and optimization-augmented and foundation-model-adapted approaches. We summarize the datasets, evaluation metrics, and comparative performance reported across the reviewed studies, and use a representative hybrid EfficientNet–Vision Transformer framework as a case study to illustrate current design trends. Persistent challenges are identified, including class imbalance for narrow structures such as roads, the computational cost of hybrid transformer-based models, limited cross-dataset generalization, and the scarcity of densely annotated data. The survey concludes by outlining promising future directions, including lightweight hybrid architectures, self-supervised and weakly supervised learning, adaptation of large vision foundation models, multi-sensor data fusion, and explainable land cover mapping.

Keywords— Land Cover Classification, Semantic Segmentation, Deep Learning Survey, Convolutional Neural Networks, Vision Transformer, Remote Sensing, Hybrid Architectures

I. INTRODUCTION

Land cover classification is the task of assigning each region or pixel of a remotely sensed image to a semantic category such as urban, forest, water, agriculture, or road. It underpins a wide range of applications, including agricultural monitoring, urban planning, deforestation tracking, disaster response, and climate-change assessment. The increasing availability of high-resolution satellite and aerial imagery, together with rapid progress in deep learning, has driven a shift away from manual and pixel-based statistical classification toward automated, data-driven semantic segmentation.

Deep learning models learn discriminative spatial and spectral features directly from raw imagery, removing much of the hand-crafted feature engineering that constrained earlier remote sensing pipelines. Convolutional neural networks (CNNs) were the first architectures to demonstrate strong performance on this task, and subsequent research has explored attention mechanisms, transformer-based global context modelling, hybrid CNN–Transformer designs, and the adaptation of large vision foundation models to the land cover domain. This diversity of architectural strategies, combined with a fragmented landscape of datasets and evaluation protocols, motivates a structured review of the field.

1. Scope and Contributions of this Survey

This survey focuses on deep learning-based approaches to land cover classification and semantic segmentation published in the recent literature. Its contributions are as follows.

- A taxonomy of deep learning approaches for land cover classification, organized into CNN-based encoder–decoder architectures, attention and self-attention-based methods, transformer and hybrid CNN–Transformer architectures, and optimization-augmented or foundation-model-adapted approaches.
- A structured summary of the datasets and evaluation metrics commonly used to benchmark land cover classification models.
- A comparative analysis of representative methods in terms of architecture, dataset, and reported performance.
- A case study of a representative hybrid EfficientNet–Vision Transformer architecture (HCNNT-MLCSS) that illustrates current design trends toward combining local and global feature representations.
- A discussion of open challenges and promising future research directions in the field.

The remainder of this survey is organized as follows. Section II presents background on the land cover classification problem and a taxonomy of deep learning approaches. Section III reviews CNN-based methods. Section IV reviews attention-based, transformer-based, and hybrid CNN–Transformer methods, including a case study architecture. Section V reviews optimization-augmented and foundation-model-based approaches. Section VI summarizes datasets and evaluation metrics. Section VII presents a comparative analysis. Section VIII discusses open challenges, Section IX outlines future directions, and Section X concludes the survey.

II. BACKGROUND AND TAXONOMY

Land cover classification can be formulated at three levels of granularity: scene-level classification, which assigns a single label to an entire image tile; object-based classification, which first segments an image into homogeneous regions and then labels each region; and pixel-level semantic segmentation, which assigns a class label to every pixel in the scene. Modern deep learning research on land cover mapping is dominated by the pixel-level formulation, since it provides the spatial precision required for applications such as boundary delineation and change detection.

1. Taxonomy of Deep Learning Approaches

The reviewed literature can be organized into four broad categories based on the core architectural mechanism employed.

- CNN-based encoder–decoder architectures, which rely on convolutional feature extraction combined with an upsampling decoder to recover pixel-wise predictions (e.g., U-Net variants, ResNeXt-based encoders).
- Attention and self-attention augmented CNNs, which incorporate channel or spatial attention modules, including cross-module attention and self-attention pipelines, to emphasize informative features.
- Transformer and hybrid CNN–Transformer architectures, which combine convolutional local feature extraction with transformer-based global context modelling, often through a feature pyramid or multi-scale fusion module.
- Optimization-augmented and foundation-model-adapted approaches, which apply metaheuristic optimization algorithms to refine network design or training, or adapt large pretrained vision foundation models such as the Segment Anything Model to the land cover domain.

Fig. 1 illustrates a representative multi-stage processing pipeline — image and mask preprocessing, dataset preparation, model training, and evaluation — that is broadly characteristic of the deep learning workflows adopted across the surveyed literature.

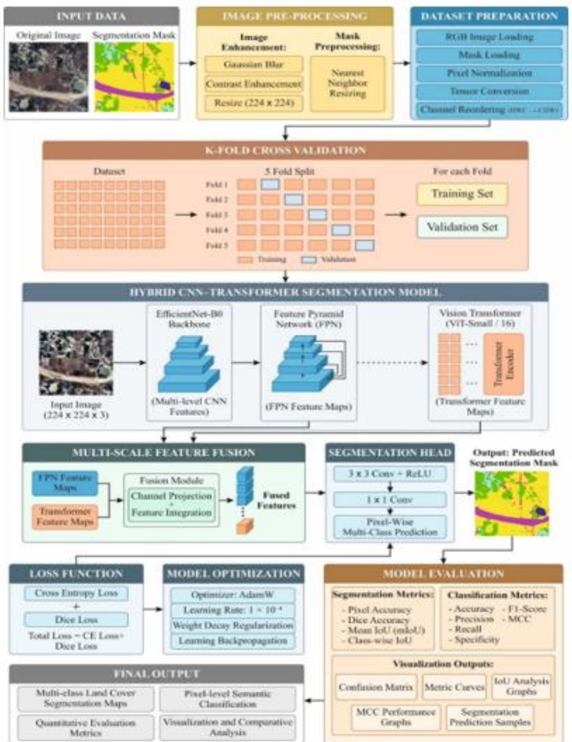

Fig. 1. Representative deep learning processing pipeline for land cover classification

III. CNN-BASED APPROACHES

CNN-based encoder–decoder architectures remain a widely used foundation for land cover segmentation because of their strong local feature extraction capability and comparatively low computational cost relative to transformer-based alternatives.

Ramos and Sappa proposed an enhanced U-Net that adopts SK-ResNeXt as its encoder for multispectral land cover analysis [1]. The adaptive-kernel, multi-cardinality design of SK-ResNeXt allows the network to capture multi-scale detail and adapt to varying spatial resolution, improving segmentation of difficult land cover categories, though the CNN-centric design offers limited global contextual reasoning.

Muhammad et al. introduced CALCNet, a purpose-built encoder–decoder network in which a restoration path recovers fine spatial detail through skip and upsampling connections while a contracting path progressively extracts higher-level semantic features [2]. Despite its cross-module attention design for improved feature fusion, CALCNet does not efficiently combine features across multiple scales and carries considerable computational overhead.

Albarakati et al. examined a fully automated, optimizer-driven, self-attention CNN pipeline for land use and land cover mapping, introducing a contrast-enhancement scheme for data augmentation built around DenseNet-64 and an IBN-based backbone [3]. Boah et al. proposed a deep hybrid UNet-ResNet-50 network for multi-class segmentation of DigitalGlobe satellite imagery, optimized using a metaheuristic particle swarm algorithm [17]. Mahmoud et al. enhanced land use and land cover mapping using an uncertainty-aware 3D-UNet combined with a squeeze-and-excite attention mechanism to strengthen feature discrimination [18]. Sharifi and Safari evaluated U-Net++ architectures for high-resolution image classification in urban planning contexts, demonstrating the benefit of nested skip connections for capturing fine urban structures [19]. KS et al. proposed an attention-based deep learning approach for satellite image segmentation aimed specifically at improving land use and land cover analysis accuracy [20]. Collectively, these CNN-based studies demonstrate steady improvements through deeper encoders, richer skip-connection designs, and lightweight attention modules, but they remain fundamentally constrained by the limited receptive field of convolutional operations, which restricts their ability to model long-range spatial dependencies across large or irregularly shaped land cover regions.

IV. ATTENTION-BASED, TRANSFORMER-BASED, AND HYBRID ARCHITECTURES

To overcome the limited receptive field of pure CNN architectures, recent research has incorporated self-attention and transformer-based global context modelling, either as an add-on to CNN backbones or as a parallel branch fused with convolutional features.

Wang et al. proposed LMFNet, a lightweight multimodal fusion network designed for high-resolution remote sensing image segmentation, balancing multimodal feature integration against computational efficiency [21]. Wei et al. proposed a method to retain and enhance modality-specific information for multimodal remote sensing land use/land cover classification, addressing the information loss that commonly occurs when fusing heterogeneous sensor modalities [10]. Huang et al. introduced a knowledge-graph-enhanced class representation scheme to reduce semantic ambiguity in open-vocabulary remote sensing image segmentation, extending segmentation beyond a fixed label set [11].

1. Case Study: Hybrid EfficientNet–Vision Transformer Architecture (HCNNT-MLCSS)

As a representative example of the hybrid CNN–Transformer design trend, a Hybrid CNN–Transformer network for Multi-Class Land Cover Semantic Segmentation (HCNNT-MLCSS) combines an EfficientNet-B0 encoder for local spatial feature extraction with a Feature Pyramid Network for multi-scale representation and a ViT-Small branch for global contextual reasoning.

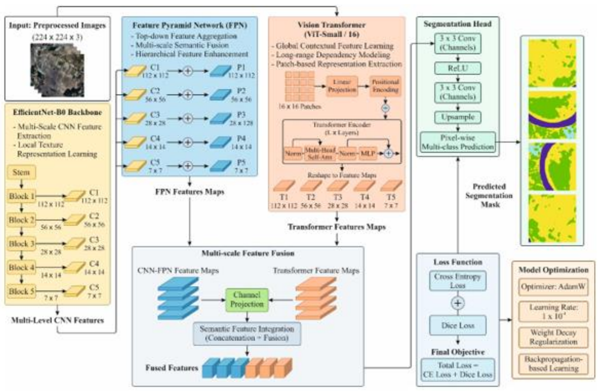

Fig. 2. Representative hybrid CNN–Transformer segmentation architecture (HCNNT-MLCSS case study)

A dedicated multi-scale feature fusion module merges the convolutional and transformer-derived representations before a lightweight convolutional segmentation head produces the final pixel-wise class map, as illustrated in Fig. 2. Training relies on a hybrid cross-entropy and Dice loss optimized with AdamW, following a 5-fold cross-validation protocol. On the Bhuvan Satellite Land Cover dataset, this hybrid design achieved a pixel accuracy of 92.96%, a Dice score of 79.45%, and a mean IoU of 69.58%, outperforming CNN-only baselines such as U-Net, ConvNeXt, and UResNet-50, which illustrates the practical benefit of combining local and global feature representations for multi-class land cover segmentation.

Tu et al. investigated VCNet, a weakly supervised deep learning strategy for large-scale, high-resolution land cover mapping that automatically refines coarse, low-resolution label products without manual annotation, reducing the annotation burden associated with transformer-based training [6]. Together, these studies indicate that hybrid CNN–Transformer designs consistently improve segmentation accuracy relative to CNN-only baselines, at the cost of increased computational complexity and training resource requirements.

V. OPTIMIZATION-AUGMENTED AND FOUNDATION-MODEL-BASED APPROACHES

A further category of methods augments deep learning pipelines with metaheuristic optimization or adapts large pretrained vision foundation models to the land cover domain, aiming to reduce manual architecture and hyperparameter tuning.

Aljebreen et al. proposed a land use and land cover classification approach built on the River Formation Dynamics optimization algorithm combined with deep learning, automatically distinguishing between different surface categories in remote sensing scenes [4]. Deressu et al. similarly enhanced land use and land cover classification through deep learning-based satellite imagery segmentation refined by systematic hyperparameter and architecture search [23].

Yang et al. developed MLFA-SAM, which applies a three-stage refinement procedure to adapt the Segment Anything Model foundation architecture to land use and land cover classification, reporting notable accuracy gains across varied high-resolution remote sensing conditions [5]. Moon and Cho proposed a resource-efficient architecture that reuses a pretrained land cover model without additional fine-tuning, combining a FLAIR U-Net with MiT-B5 and ResNet-34 backbones across a four-stage pipeline, illustrating the practical value of transfer learning for reducing training cost in operational settings [7].

Papadopoulos et al. analysed fully polarimetric SAR imagery, drawing on scattering-intensity components from the Krogager, Cloude–Pottier, and Pauli decomposition models to strengthen feature identification for land cover discrimination, illustrating how domain-specific signal decomposition can complement learned deep features [8][9]. Nguyen et al. integrated object-based image analysis (OBIA) with the tidymodels framework for a multi-metric evaluation of machine learning algorithms in land use/land cover classification, highlighting that classical object-based approaches remain competitive benchmarks in specific application settings [12]. Kumar et al. presented a comparative evaluation of pixel-based and object-based classification approaches using deep learning on satellite data, reporting that the optimal choice between paradigms depends strongly on scene heterogeneity and spatial resolution [13].

Explainability has also begun to receive attention in this domain: Gavade and Gavade examined the role of explainable AI in transforming land use land cover classification, arguing that model transparency is increasingly important for adoption in operational and policy-relevant mapping contexts [14].

VI. DATASETS AND EVALUATION METRICS

1. Benchmark Datasets

A variety of benchmark datasets are used across the reviewed literature, differing in spatial resolution, sensor modality, geographic coverage, and class taxonomy. Table 1 summarizes representative datasets referenced across the surveyed studies.

Table 1: Representative datasets used in the surveyed land cover classification literature

Dataset	Typical Use	Notes
Bhuvan Satellite Land Cover	5-class pixel-level segmentation	Urban, water, forest, agriculture, road classes
Five-Billion-Pixels	Large-scale multispectral segmentation	Used to benchmark U-Net/SK-ResNeXt variants
EuroSAT	Scene-level land use classification	Widely used for CNN and hybrid model benchmarking
AID / UCMerced-LandUse	Aerial scene classification	Common cross-attention CNN benchmark
NWPU-RESISC45	Scene classification	Large class-count remote sensing benchmark
SIRI-WHU	Urban land use classification	High-resolution urban imagery
ISPRS Potsdam / HRLC	High-resolution segmentation	Used to benchmark SAM-adapted models
Chesapeake Bay / Tokyo	Weakly supervised label refinement	Coarse-to-fine label mapping tasks
FLAIR / AI-HUB Land Cover	Transfer-learning benchmarking	Multi-source pretrained model reuse

2. Evaluation Metrics

Segmentation performance is most commonly reported using pixel accuracy, the Dice similarity coefficient, and mean Intersection-over-Union (mIoU), while classification-style performance is reported using accuracy, precision, recall, specificity, F1-score, and the Matthews Correlation Coefficient. Overall accuracy (OA) is frequently used in scene-classification benchmarks such as EuroSAT and AID, while R^2 is occasionally reported for pixel-level fused regression-style mapping tasks.

VII. COMPARATIVE ANALYSIS

Table 2 presents a comparative summary of representative methods discussed in this survey, organized by their primary architectural category, together with the dataset used and the headline performance figure reported in the original study.

Table 2: Comparative summary of representative land cover classification methods

Method	Category	Dataset	Reported Performance
Ramos & Sappa [1]	CNN encoder-decoder	Five-Billion-Pixels	mIoU 8.906%
CALCNet [2]	CNN + cross-module attention	AID/UCMerced/NWPU/EuroSAT	Acc. up to 99.47%
Albarakati et al. [3]	Self-attention CNN	SIRI-WHU/EuroSAT/NWPU	Acc. up to 98.20%
Aljebreen et al. [4]	Optimization-augmented CNN	EuroSAT	Acc. 98.15%
MLFA-SAM [5]	Foundation-model adaptation	HRLC, ISPRS Potsdam	OA 87.91%
VCNet [6]	Weakly supervised hybrid	Chesapeake Bay, Tokyo	mIoU 65.53%
Moon & Cho [7]	Transfer-learning U-Net	FLAIR, AI-HUB	R^2 0.706–0.999
Papadopoulos et al. [8]	Decomposition-based fusion	EuroSAT	OA 95.78%
LMFNet [21]	Lightweight multimodal fusion	High-res RS imagery	Reported gains vs. baselines
HCNNT-MLCSS (case study)	Hybrid CNN–Transformer	Bhuvan Satellite	Pixel Acc. 92.96%, mIoU 69.58%

This comparison highlights a general trend: pure CNN encoder–decoder methods achieve strong accuracy at low computational cost but plateau on mIoU for spatially complex classes; optimization-augmented and self-attention CNNs improve robustness at moderate additional cost; and hybrid CNN–Transformer and foundation-model-adapted approaches achieve the strongest overall segmentation quality, at the expense of higher computational and training resource requirements.

Open Challenges

- Class imbalance: narrow, linear structures such as roads are consistently the weakest-performing class across the surveyed studies, since they occupy a small proportion of scene pixels relative to broad classes such as agriculture or forest.
- Computational cost: hybrid CNN–Transformer and foundation-model-adapted architectures deliver the strongest accuracy but require substantially more compute and memory than CNN-only baselines, complicating deployment on resource-constrained or edge platforms.
- Cross-dataset generalization: most reviewed studies are evaluated on a single dataset or sensor modality, and performance often degrades when models are applied to imagery from different sensors, resolutions, or geographic regions.
- Annotation scarcity: densely labelled pixel-level ground truth is expensive to produce at scale, motivating continued interest in weakly supervised and label-refinement strategies such as VCNet.
- Multi-sensor fusion complexity: combining optical, SAR, and multispectral modalities can improve discrimination but introduces alignment, calibration, and modality-specific noise challenges that are not yet fully resolved.
- Interpretability: as land cover maps increasingly inform policy and regulatory decisions, the

limited interpretability of deep, hybrid architectures remains an obstacle to trust and adoption.

Future Research Directions

- Lightweight hybrid architectures that retain the accuracy benefits of CNN–Transformer fusion while reducing parameter count and inference latency for real-time or edge deployment.
- Broader adoption of self-supervised and weakly supervised learning to reduce dependence on densely annotated training data.
- Continued adaptation of large vision foundation models, such as the Segment Anything Model, to the land cover domain through efficient fine-tuning or prompt-based refinement strategies.
- Multi-source and multi-sensor data fusion frameworks capable of robustly combining optical, SAR, and multispectral imagery.
- Standardized, geographically diverse benchmark datasets and evaluation protocols to enable fairer cross-study comparison.
- Integration of explainability techniques into land cover segmentation pipelines to support operational and policy-relevant adoption.

VIII. CONCLUSION

This survey has reviewed recent deep learning approaches to land cover classification and semantic segmentation, organizing the literature into CNN-based encoder–decoder architectures, attention and self-attention augmented CNNs, transformer and hybrid CNN–Transformer architectures, and optimization-augmented or foundation-model-adapted approaches. A case study of a representative hybrid EfficientNet–Vision Transformer architecture illustrated the practical benefit of combining local convolutional features with global transformer-based context. Across the reviewed studies, hybrid and transformer-augmented designs consistently achieve the strongest segmentation performance, though at increased computational cost, while class imbalance, limited cross-dataset generalization, and annotation scarcity remain open challenges. Continued progress in lightweight hybrid design, weak supervision, foundation-model adaptation, multi-sensor fusion, and explainability is expected to shape the next generation of land cover classification research.

REFERENCES

1. L. T. Ramos and A. D. Sappa, "Leveraging U-Net and selective feature extraction for land cover classification using remote sensing imagery," Sci. Rep., vol. 15, p. 784, 2025.
2. F. Muhammad, Y. Hikmat, J. Weiwei, I. A. Hassan, I. Muhammad, and D. L. Minh, "CALCNet: A novel cross-module attention network for efficient land cover classification," Remote Sens., vol. 18, no. 8, p. 1218, 2026.
3. H. M. Albarakati et al., "A novel deep learning architecture for agriculture land cover and land use classification from remote sensing images based on network-level fusion of self-attention architecture," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 17, pp. 6338–6353, 2024.
4. M. Aljebreen, H. A. Mengash, M. Alamgeer, S. S. Alotaibi, A. S. Salama, and M. A. Hamza, "Land use and land cover classification using river formation dynamics algorithm with deep learning on remote sensing images," IEEE Access, vol. 12, pp. 11147–11156, 2024.
5. H. Yang et al., "A high-resolution remote sensing land use/land cover classification method based on multi-level features adaptation of segment anything model," Int. J. Appl. Earth Obs. Geoinf., vol. 141, p. 104659, 2025.
6. S. Tu, Q. Cheng, Q. Zhao, Y. Du, and H. Cheng, "Weakly supervised segmentation with cross-feature learning for fine-scale land cover map updating," Geo-spatial Inf. Sci., pp. 1–26, 2026.
7. H. J. Moon and N. W. Cho, "Training-free lightweight transfer learning for land cover segmentation using multispectral calibration," Remote Sens., vol. 18, no. 2, 2026.
8. S. Papadopoulos, G. Koukiou, and V. Anastassopoulos, "A multi-class Bahadur–Lazarsfeld expansion framework for pixel-level fusion in multi-sensor land cover classification," Remote Sens., vol. 18, no. 3, p. 399, 2026.
9. X. Sun, J. Gao, and Y. Yuan, "Fully PolSAR image reconstruction for enhanced land cover mapping," Pattern Recognit., vol. 169, p. 111895, 2026.

10. T. Wei, H. Chen, W. Liu, L. Chen, and J. Wang, "Retain and enhance modality-specific information for multimodal remote sensing image land use/land cover classification," IEEE Trans. Geosci. Remote Sens., 2025.
11. W. Huang, H. Li, S. Zhang, and F. Deng, "Reducing semantic ambiguity in open-vocabulary remote sensing image segmentation via knowledge graph-enhanced class representations," ISPRS J. Photogramm. Remote Sens., vol. 231, pp. 837–853, 2026.
12. N. T. Nguyen, L. C. K. Pham, K. V. Pham, and T. H. Nguyen, "Integrating OBIA and tidymodels for land use/land cover classification: a multi-metric evaluation of machine learning algorithms," Int. J. Environ. Sci. Technol., vol. 23, no. 1, p. 25, 2026.
13. A. Kumar, A. K. Ranjan, P. Paygude, and R. Daimary, "Comparative evaluation of pixel-based and object-based classification approaches for land use/land cover mapping using deep learning on satellite data," Appl. Geomat., vol. 18, no. 1, p. 16, 2026.
14. A. B. Gavade and P. A. Gavade, "Explainable AI in transforming land use land cover classification," in Mitigation and Adaptation Strategies Against Climate Change in Natural Systems, Cham: Springer Nature Switzerland, 2025, pp. 343–356.
15. D. K. Behera, G. S. Pujar, R. Kumar, and S. K. Singh, "A comprehensive approach towards enhancing land use land cover classification through machine learning and object-based image analysis," J. Indian Soc. Remote Sens., vol. 53, no. 3, pp. 731–749, 2025.
16. A. Putty, B. Annappa, and S. Pariserum Perumal, "Semantic segmentation of remotely sensed images for land-use and land-cover classification: a comprehensive review," IETE Tech. Rev., vol. 42, no. 2, pp. 222–237, 2025.
17. E. A. Boah, Y. Issaka, R. Arhinful, A. A. Borsah, W. S. Aklamati, and F. Tabase, "Multi-class segmentation of land cover types for DigitalGlobe satellite imagery using deep hybrid UNet-ResNet-50 network optimised with metaheuristic particle swarm algorithm," Urban Inform., vol. 5, no. 1, p. 8, 2026.
18. A. S. Mahmoud, N. M. Afify, E. Farg, S. M. Arafat, and M. S. Moustafa, "Enhancing land use and land cover mapping with uncertainty-aware 3D-UNet and squeeze and excite attention mechanism," Int. J. Image Data Fusion, vol. 17, no. 1, 2026.
19. A. Sharifi and M. M. Safari, "Evaluation of U-Net++ architectures in high resolution image classification for urban planning," Appl. Geomat., vol. 18, no. 1, pp. 1–15, 2026.
20. G. KS, R. K. R, V. Ramesh, and S. Natarajan, "Enhancing satellite image segmentation for land use and land cover analysis using attention-based deep learning approach," Int. J. Syst. Assur. Eng. Manage., 2026.
21. T. Wang, G. Chen, X. Zhang, C. Liu, J. Wang, X. Tan, W. Zhou, and C. He, "LMFNet: Lightweight multimodal fusion network for high-resolution remote sensing image segmentation," Pattern Recognit., vol. 164, p. 111579, 2025.
22. "Satellite Image and Mask dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/khushiipatni/satellite-image-and-mask/data.
23. T. F. Deressu, A. K. Bojer, T. G. Debelee, W. G. Negera, S. Nadarajah, and K. W. Gebissa, "Enhancing land use and land cover classification with deep learning-based satellite imagery segmentation," Int. J. Appl. Earth Obs. Geoinf., vol. 144, p. 104839, 2025.