

Diffusion-Based Multimodal Remote Sensing Segmentation for Land-Cover Mapping and Environmental Monitoring

Rosaura Espejo

Department of Computer Science, George Mason University, Fairfax, VA, USA.
rosauramail@gmu.edu

Birgitt Klotz

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
birgitt.klotz225@ku.edu

Matthijs van Engelen

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
matthijs550@uab.edu

Marjo Muhonen

Department of Computer Science, University of North Texas, Denton, TX, USA.
hellomarjo@unt.edu

Abstract

Land-cover mapping and environmental monitoring increasingly depend on the integration of heterogeneous remote sensing modalities, including optical imagery, synthetic aperture radar, hyperspectral cubes, thermal bands, and light detection and ranging products. Although deep learning has substantially advanced remote sensing segmentation, conventional discriminative models often struggle with missing modalities, domain shift, class imbalance, and the need for calibrated uncertainty in operational decision-making. Diffusion-based generative models offer a promising alternative because they learn complex data distributions through iterative denoising and can be conditioned on multiple sensor streams, semantic prompts, and temporal context. This paper examines diffusion-based multimodal remote sensing segmentation as a socio-technical system rather than a narrow model architecture. It analyzes fusion strategies, deployment infrastructure, robustness, fairness, governance, and sustainability implications for land-cover mapping and environmental monitoring. The discussion emphasizes structural trade-offs among accuracy, computational cost, latency, interpretability, and institutional trust. It also considers how cloud-edge orchestration, data provenance, model lifecycle management, and policy frameworks shape whether diffusion segmentation can move from benchmark performance to reliable public and private environmental intelligence. The paper concludes that diffusion-based multimodal segmentation is best understood as an infrastructural capability whose value depends on careful alignment among sensing systems, computational resources, regulatory expectations, and the communities affected by land-cover decisions.

Keywords

diffusion models, multimodal remote sensing, semantic segmentation, land-cover mapping, environmental monitoring, geospatial governance, robustness, sustainability.

1. Introduction

Remote sensing segmentation has become a foundational computational layer for land-cover mapping, agricultural assessment, forest monitoring, urban planning, disaster response, and climate-related environmental observation. The shift from pixel-wise classification to semantically rich segmentation has been driven by deep learning, which can learn hierarchical spatial and spectral representations from large Earth observation archives [1]. However, real-world land-cover mapping rarely depends on a single sensor modality. Optical imagery is intuitive but cloud-sensitive; synthetic aperture radar provides structural and dielectric information under adverse weather; hyperspectral data offers fine spectral discrimination; thermal imagery reveals surface energy patterns; and light detection and ranging supplies precise vertical structure. Multimodal fusion is therefore not merely an accuracy strategy but an operational necessity for continuous environmental monitoring [4], [5].

Despite progress, existing segmentation systems face persistent constraints. Domain adaptation remains difficult because models trained in one biome, season, or sensor configuration often degrade when transferred to another [2]. Class imbalance is severe because rare habitats, small water bodies, and transitional land-cover classes occupy limited area but carry high ecological or regulatory significance. Missing modalities are common in practice, especially when cloud cover, sensor failure, or downlink limitations disrupt one stream. Furthermore, operational users require uncertainty estimates, not only hard labels, because land-cover decisions can affect subsidies, conservation enforcement, infrastructure permitting, and disaster relief. These requirements expose the limits of purely discriminative segmentation pipelines, which typically output deterministic masks without an explicit generative account of ambiguity.

Diffusion-based models have emerged as a compelling response to these limitations. By learning to reverse a progressive noising process, diffusion models can approximate complex conditional distributions and generate or refine structured outputs under uncertainty [6], [7]. In segmentation, diffusion has been adapted to produce masks, boundaries, and panoptic scene interpretations, sometimes with strong gains in detail and stability relative to adversarial or purely feed-forward approaches [10], [11]. When applied to remote sensing, diffusion offers a mechanism for integrating multimodal evidence during iterative refinement, allowing the model to reconcile conflicting cues from optical, radar, thermal, and elevation data [14]. Yet the same properties that make diffusion attractive also introduce system-level challenges: high inference cost, sensitivity to conditioning design, opaque uncertainty, and substantial data and compute requirements.

This paper argues that diffusion-based multimodal remote sensing segmentation should be evaluated as a socio-technical infrastructure. Its success depends not only on architectural novelty but also on deployment pipelines, data governance, energy sustainability, fairness across regions, and policy compatibility. The following sections examine the conceptual foundations, fusion architectures, diffusion segmentation methods, operational systems, governance concerns, evaluation practices, and future trajectories of this emerging field. The objective is to provide an interdisciplinary account that connects machine learning innovation with the institutional realities of environmental monitoring.

2. Background and Conceptual Foundations

Deep learning in remote sensing has evolved through convolutional architectures, recurrent models, graph-based methods, attention mechanisms, and transformer-based designs. Early reviews established that representation learning could outperform handcrafted features in classification, detection, and segmentation tasks [1], [13]. Subsequent work extended these methods to hyperspectral imagery, where spectral-spatial dependencies require careful modeling of both local texture and narrow-band signatures [3], [20]. Domain adaptation and transfer learning became central because Earth observation data are geographically, temporally, and sensor-dependent, making generalization a first-order concern rather than an afterthought [2]. These foundations matter for diffusion segmentation because generative refinement must operate on representations that preserve semantic and physical meaning.

Diffusion models originate from probabilistic generative modeling. Denoising diffusion probabilistic models learn to transform random noise into structured data through a sequence of denoising steps [6]. Score-based formulations generalize this process through stochastic differential equations and provide a unified view of noise-conditioned score matching [7]. Latent diffusion reduces computational burden by performing the generative process in a compressed latent space, enabling high-resolution synthesis with manageable cost [8]. Diffusion models have also been shown to rival or surpass generative adversarial networks in image synthesis quality and mode coverage, although they often require more sampling steps [9]. These properties are relevant to remote sensing because land-cover patterns are multimodal: a single pixel or patch may be consistent with multiple plausible classes depending on season, illumination, soil moisture, and human activity.

Segmentation with diffusion can be framed as conditional generation of a label field or mask. Instead of directly predicting a class map, the model iteratively denoises a mask conditioned on image features, textual prompts, or auxiliary modalities. SegDiff demonstrated that diffusion probabilistic models can produce competitive segmentation masks by learning a conditional reverse process [11]. Generalist panoptic frameworks have extended diffusion-like generative segmentation to images and videos, suggesting that unified perception tasks can be addressed within a generative paradigm [10]. In medical imaging, diffusion has been used for anomaly detection and uncertainty-aware segmentation, where the generative model can highlight deviations from learned anatomical distributions [12]. These cross-domain developments inform remote sensing, where anomalies such as illegal logging, unpermitted construction, or flood extent changes are often more important than average-case accuracy.

The conceptual appeal of diffusion for land-cover mapping lies in its ability to represent ambiguity. A cloud-contaminated optical pixel may be classified as water, shadow, or wetland depending on radar backscatter and elevation. A discriminative model may collapse these possibilities into a single label, whereas a conditional diffusion model can represent a distribution over plausible masks. This is valuable for environmental monitoring because decision-makers often need to know whether a classification is stable or highly sensitive to missing evidence. At the same time, generative flexibility creates governance challenges: if multiple plausible masks exist, which one becomes authoritative for regulatory purposes? The answer depends on evaluation protocols, uncertainty communication, and institutional rules for resolving conflicting evidence.

3. Multimodal Remote Sensing Data and Fusion Architectures

Multimodal remote sensing fusion can be organized by the stage at which modalities interact. Early fusion concatenates or co-registers raw or minimally processed data, enabling the model to learn cross-modal interactions from the outset. This approach can capture fine-grained

complementarity, but it is sensitive to spatial misalignment, radiometric inconsistency, and missing channels. Late fusion combines independent predictions from per-modality models, which improves robustness to missing data but may lose subtle cross-modal dependencies. Intermediate or hybrid fusion integrates features at multiple depths, often using attention, gating, or graph-based mechanisms to weight modalities according to reliability. Reviews of multimodal remote sensing fusion show that no single strategy dominates; performance depends on sensor characteristics, task granularity, label availability, and computational constraints [4]. Data fusion more broadly has long been recognized as a relationship between sensing physics, information theory, and application requirements [5].

For land-cover mapping, modality selection is shaped by environmental conditions and institutional capacity. Optical Sentinel-2 and Landsat imagery provide broad spectral coverage and long temporal archives, but persistent cloud cover limits tropical and high-latitude monitoring. Synthetic aperture radar, such as Sentinel-1, penetrates clouds and provides sensitivity to surface roughness and moisture, yet its interpretation is less intuitive and affected by speckle. Hyperspectral sensors offer diagnostic spectral signatures for vegetation species, minerals, and water quality, but their spatial resolution and data volume impose trade-offs [3], [20]. Thermal infrared supports urban heat, drought, and evapotranspiration monitoring. Light detection and ranging provides canopy height and building morphology but is expensive and spatially sparse. A diffusion-based segmentation system must therefore be designed around a modality portfolio rather than a fixed input vector.

Fusion architectures for diffusion segmentation face a distinctive conditioning problem. The generative denoising process needs to be guided by multimodal evidence without allowing any single modality to dominate arbitrarily. Attention-based conditioning can learn to attend to reliable regions, but it requires representative training data and may fail under distribution shift. Cross-attention between latent image features and modality-specific tokens is promising, but its memory footprint grows with resolution and sequence length. Gating mechanisms can explicitly model missing modalities by routing only available streams, yet they may introduce discontinuities when modality availability changes across a scene. A robust system might combine early fusion for spatially aligned optical and thermal data, intermediate fusion for radar and hyperspectral features, and late fusion for elevation products, with uncertainty estimates controlling how each stream influences the final mask.

The infrastructural implications of fusion are substantial. Multimodal archives require co-registration, atmospheric correction, radiometric normalization, and quality masking. These preprocessing steps are often more expensive and error-prone than the segmentation model itself. In many operational settings, data arrive from different agencies with different update cycles, licenses, and spatial reference systems. A diffusion segmentation system must therefore include provenance tracking, versioned preprocessing, and reproducible geospatial transformations. Without such infrastructure, multimodal fusion becomes a laboratory exercise rather than a monitoring capability. The choice of fusion architecture is thus inseparable from data engineering, institutional coordination, and long-term archive maintenance.

4. Diffusion Models for Segmentation and Land-Cover Mapping

Diffusion-based segmentation adapts generative denoising to produce structured label fields. The model is typically conditioned on an input image, a feature pyramid, or a multimodal embedding, and it learns to iteratively refine a noisy mask toward a semantically coherent output. This process can improve boundary adherence because the model can revise uncertain

regions across multiple steps. It can also model multimodal uncertainty by generating multiple plausible masks from different noise initializations. For remote sensing, this is particularly relevant in transitional zones such as wetlands, riparian buffers, and urban fringes, where class definitions are inherently fuzzy and spectral mixtures are common. A recent diffusion-driven semantic segmentation framework with multimodal image fusion for enhanced perception illustrates how generative refinement can be coupled with multiple sensor streams to improve segmentation under complex conditions [14].

Latent diffusion architectures are especially relevant for high-resolution land-cover mapping because they decouple perceptual quality from pixel-level computation [8]. By operating in a compressed latent space, they reduce the cost of iterative sampling, making diffusion more feasible for large geospatial tiles. Transformer-based diffusion backbones further improve scalability by replacing convolutional inductive biases with attention mechanisms that can model long-range dependencies [15]. Video and temporal diffusion models extend this logic to time-series remote sensing, where land-cover change, crop phenology, and disaster progression require consistent segmentation across frames [16]. These developments suggest a trajectory toward multimodal, temporal, and spatially aware diffusion systems that treat segmentation as a dynamic inference process rather than a one-shot classification.

However, diffusion segmentation is not automatically superior to discriminative segmentation. Iterative sampling increases latency and energy consumption, which can be problematic for near-real-time monitoring. The quality of generated masks depends heavily on conditioning design, noise schedule, and the balance between generative prior and input evidence. If the generative prior is too strong, the model may hallucinate plausible but incorrect land-cover patterns; if it is too weak, diffusion offers little advantage over conventional refinement. This trade-off is especially sensitive in environmental monitoring, where false positives can trigger unnecessary enforcement or misallocate conservation resources. Evaluation must therefore assess not only average accuracy but also hallucination rates, calibration, and performance on rare classes.

Diffusion also changes the interpretability profile of segmentation. Discriminative models can sometimes be explained through saliency maps or attention weights, although these explanations are often unreliable. Diffusion models introduce an additional temporal dimension: the denoising trajectory. Analysts can inspect how a mask evolves across steps, which may reveal regions of high uncertainty or conflicting multimodal evidence. At the same time, the generative process is complex, and naive visualizations can create false confidence. Effective governance requires that uncertainty communication be designed with users, not merely appended as a post hoc heatmap. Land-cover maps are political and economic artifacts as well as technical outputs, so the interpretability of diffusion segmentation must be evaluated in relation to decision contexts.

5. System Architecture, Deployment, and Infrastructure

Deploying diffusion-based multimodal segmentation at scale requires an architecture that spans data ingestion, preprocessing, model training, inference, post-processing, and dissemination. A typical pipeline begins with acquisition from multiple satellite and aerial platforms, followed by co-registration, cloud masking, atmospheric correction, and modality-specific normalization. These steps must be reproducible and auditable. The segmentation service then conditions a diffusion model on the fused representation and produces one or more candidate masks with uncertainty estimates. Post-processing may enforce topological constraints, smooth boundaries, or align outputs with administrative units. Finally, results are

distributed through geospatial APIs, dashboards, or decision-support tools. Each stage introduces potential failure points that can propagate into environmental decisions [1], [4].

Compute infrastructure is a central constraint. Training diffusion models for high-resolution multimodal segmentation demands substantial graphics processing unit or tensor processing unit capacity, high-bandwidth storage, and efficient data loaders. Inference is also expensive because iterative denoising multiplies the cost of each prediction. Cloud platforms offer elasticity, but they raise concerns about data sovereignty, vendor lock-in, and recurring operational expenditure. Edge deployment can reduce latency and bandwidth use, but edge devices rarely have the memory or power budget for full diffusion sampling. Hybrid cloud-edge architectures may distribute work: lightweight modality-specific encoders run near sensors, while heavier generative refinement occurs in regional data centers. This division must account for intermittent connectivity and security risks.

Model lifecycle management is equally important. Remote sensing environments change through seasonal cycles, land-use transitions, sensor degradation, and policy-driven shifts in class definitions. A diffusion segmentation model trained on historical data may become stale or biased. Continuous monitoring, drift detection, and periodic retraining are necessary, but retraining diffusion models is costly and can introduce instability if not carefully validated. Version control must cover not only model weights but also preprocessing rules, class taxonomies, and fusion configurations. Without such discipline, operational maps may become irreproducible, undermining legal and scientific credibility. Infrastructure therefore includes organizational routines as much as hardware and software.

Deployment also intersects with data governance. Multimodal remote sensing often combines commercial imagery, public satellite data, and sensitive ancillary layers such as cadastral boundaries or critical infrastructure. Licensing terms may restrict redistribution or derivative use. Privacy concerns arise when very high-resolution imagery reveals individual properties or activities. A responsible deployment architecture must implement access control, provenance metadata, retention policies, and audit trails. These requirements can conflict with open-science norms, especially when environmental monitoring aims to empower civil society. The design challenge is to enable transparency and reproducibility while respecting legitimate restrictions and security concerns.

6. Robustness, Fairness, Governance, and Policy

Robustness in diffusion-based segmentation encompasses sensor noise, adversarial perturbations, missing modalities, and domain shift. Because diffusion models learn a data distribution, they may be vulnerable to out-of-distribution inputs that produce confident but implausible masks. Multimodal conditioning can improve robustness when modalities are complementary, but it can also create new dependencies. If a model learns to rely on thermal data that is unavailable in some regions, performance may drop sharply. If radar preprocessing differs between archives, the model may confuse speckle patterns with land-cover texture. Robustness testing should therefore include modality dropout, sensor degradation, seasonal extremes, and geographic transfer. Medical anomaly detection research offers a useful analogy: generative models can detect deviations, but they require careful calibration to avoid false alarms [12].

Fairness in land-cover mapping is often overlooked because segmentation appears technical and apolitical. Yet classification schemes, training data, and evaluation metrics embed normative choices. A model trained predominantly on temperate forests may perform poorly

on tropical agroforestry or arid rangelands. Cloud masking algorithms can systematically obscure regions with persistent cloud cover, leading to lower map quality in humid tropics. Urban informal settlements may be misclassified if training data overrepresents formal housing. Diffusion models can reproduce or amplify these biases because they learn from historical data and generate outputs consistent with dominant patterns. Fairness thus requires geographic representation, participatory class definition, and disaggregated evaluation across regions and communities [19].

Governance frameworks must address accountability when segmentation informs environmental regulation. If a diffusion model produces a land-cover map used to detect deforestation, wetland loss, or illegal dumping, affected parties need mechanisms to contest errors. This requires documentation of training data, model versions, confidence thresholds, and human review processes. Policy instruments such as the European Union's Copernicus program, national remote sensing strategies, and environmental regulations shape what counts as acceptable evidence. Governance must also consider dual-use risks: the same multimodal segmentation capability that monitors conservation can support surveillance or resource extraction. Responsible innovation requires transparency about intended use, restrictions on harmful applications, and independent oversight.

Policy implications extend to procurement and public capacity. Governments and NGOs may lack the computational resources to train diffusion models but still depend on their outputs. Public-private partnerships, open foundation models, and shared benchmarking infrastructure can reduce barriers, but they also create dependencies on private vendors. Procurement rules should require model cards, reproducibility, and performance reporting across affected regions. Data-sharing agreements should clarify rights and responsibilities. More fundamentally, policy should treat environmental monitoring as a public information infrastructure, not merely a market for analytics. Diffusion segmentation can support that vision only if governance ensures that benefits and risks are distributed fairly.

7. Evaluation, Case Illustrations, and Operational Trade-offs

Evaluation of diffusion-based multimodal segmentation must go beyond overall accuracy and intersection-over-union. Land-cover mapping requires class-wise performance, boundary quality, temporal consistency, uncertainty calibration, and computational cost. Rare classes such as wetlands, small water bodies, and endangered habitats deserve separate reporting because aggregate metrics can hide failure. Evaluations should also measure robustness under modality dropout and domain shift. The remote sensing literature has repeatedly shown that deep learning models can achieve impressive benchmark results while failing in operational conditions [17], [18]. Diffusion models are no exception; their generative flexibility makes rigorous evaluation more, not less, important.

Case illustrations clarify the trade-offs. In agricultural monitoring, multimodal diffusion segmentation could fuse optical phenology, radar backscatter, and thermal stress indicators to map crop types and irrigation status. The generative model might better handle cloud gaps, but it could also hallucinate crop patterns in fallow fields if the prior is overly strong. In forest monitoring, radar and optical fusion can detect degradation and deforestation, yet diffusion sampling may be too slow for daily alerting without substantial optimization. In urban heat mapping, thermal and optical fusion can identify vulnerable neighborhoods, but fairness concerns arise if model accuracy varies with housing density or vegetation cover. In disaster response, rapid flood mapping benefits from radar-optical fusion, but uncertainty must be communicated clearly because false negatives can delay relief.

Operational trade-offs often favor hybrid approaches. A discriminative segmentation model can provide fast initial masks, while a diffusion model refines uncertain regions or generates alternative hypotheses. This cascaded architecture reduces average inference cost and allows human reviewers to focus on ambiguous areas. Alternatively, a diffusion model can be distilled into a faster student network for routine monitoring, with the full model reserved for high-stakes analysis. Such designs acknowledge that environmental monitoring includes both routine surveillance and episodic decision-making. The appropriate architecture depends on latency requirements, budget, regulatory tolerance for error, and the availability of human expertise. No single diffusion configuration is universally optimal.

Benchmarking infrastructure should reflect these operational realities. Static datasets with fixed train-test splits are useful for comparison but insufficient for deployment. Evaluation should include temporal holdouts, geographic cross-validation, sensor-transfer tests, and stress tests with degraded inputs. Reporting should document energy use, inference time, and hardware requirements. Open benchmarks can accelerate progress, but they must be designed with privacy, licensing, and local knowledge in mind. Otherwise, they may reinforce the dominance of well-resourced regions and marginalize the very environments where monitoring is most needed [19]. Evaluation is therefore both a technical and an ethical practice.

8. Sustainability and Future Perspectives

The sustainability of diffusion-based remote sensing segmentation has two meanings. The first concerns environmental sustainability as an application domain: these models can support climate mitigation, biodiversity protection, and sustainable land management. The second concerns the computational sustainability of the models themselves. Iterative denoising, large transformer backbones, and multimodal training can consume significant energy. Latent diffusion and efficient samplers reduce cost, but the overall system footprint includes data storage, preprocessing, network transfer, and cooling [8], [15]. As environmental monitoring scales to global coverage and high revisit rates, efficiency becomes an environmental imperative, not merely an engineering preference.

Future systems are likely to move toward foundation models for Earth observation. A shared multimodal backbone could be pretrained on optical, radar, hyperspectral, and elevation data, then adapted to segmentation tasks through lightweight conditioning. Diffusion could serve as a generative decoder that produces masks, change maps, or uncertainty ensembles. Temporal diffusion models could learn land-cover dynamics and generate consistent sequences rather than independent frames [16]. Such models could improve change detection, phenology tracking, and disaster monitoring. However, foundation models concentrate power and resources, raising questions about who controls the pretrained weights and who can fine-tune them. Public investment, open licensing, and federated learning may be necessary to ensure broad participation.

Another direction is uncertainty-aware decision support. Rather than delivering a single map, diffusion systems could provide ensembles of plausible land-cover configurations, each with a likelihood or confidence score. Decision-makers could then apply rules based on risk tolerance: for example, requiring human review when a wetland boundary varies across samples. This approach aligns with environmental governance, where precaution and contestability are important. It also creates new design challenges in visualization, explanation, and auditing. A probabilistic map is only useful if users understand what it represents and if institutions can act on uncertainty without paralysis.

Finally, diffusion-based multimodal segmentation should be integrated with participatory monitoring and local knowledge. Satellite and aerial data cannot capture all dimensions of land cover, especially customary tenure, seasonal grazing, or culturally significant landscapes. Community mapping, ground sensors, and citizen science can provide complementary evidence. A diffusion model conditioned on both remote sensing and local observations could produce more contextually valid maps, but it must avoid extractive data practices. Benefit-sharing, consent, and capacity building are essential. The future of environmental monitoring is not only more generative; it is also more pluralistic, requiring technical systems that can accommodate multiple ways of knowing and multiple claims to authority.

9. Conclusion

Diffusion-based multimodal remote sensing segmentation offers a powerful approach to land-cover mapping and environmental monitoring by combining generative refinement with heterogeneous sensor evidence. It can represent ambiguity, integrate missing modalities, and produce uncertainty-aware outputs that are valuable in complex environmental decisions. Yet its benefits are not automatic. Diffusion models introduce high computational costs, complex conditioning requirements, and new risks of hallucination and opacity. Their performance depends on data infrastructure, fusion architecture, deployment strategy, and governance. Treating diffusion segmentation as a socio-technical system reveals that accuracy alone is an insufficient measure of success.

The field should prioritize robust evaluation, geographic fairness, energy efficiency, and accountable deployment. Hybrid architectures that combine discriminative speed with generative refinement may offer practical near-term value. Latent and temporal diffusion, transformer backbones, and foundation models point toward scalable future systems, but they also concentrate resources and require public oversight. Policy frameworks must ensure that land-cover maps are contestable, reproducible, and aligned with environmental justice. If these conditions are met, diffusion-based multimodal segmentation can become a durable infrastructure for environmental intelligence rather than a transient benchmark achievement.

References

1. Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4), 8-36. <https://doi.org/10.1109/MGRS.2017.2762307>
2. Tuia, D., Persello, C., & Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 41-57. <https://doi.org/10.1109/MGRS.2016.2548504>
3. Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., & Chanussot, J. (2021). Graph convolutional networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7), 5966-5978. <https://doi.org/10.1109/TGRS.2020.3015157>
4. Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102926. <https://doi.org/10.1016/j.jag.2022.102926>

5. Schmitt, M., & Zhu, X. X. (2016). Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4), 6-23. <https://doi.org/10.1109/MGRS.2016.2561021>
6. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840-6851.
7. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*.
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684-10695.
9. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780-8794.
10. Chen, T., Li, L., Saxena, S., Hinton, G., & Fleet, D. J. (2023). A generalist framework for panoptic segmentation of images and videos. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 909-919.
11. Amit, T., Nachmani, E., Shaharbany, T., & Wolf, L. (2021). SegDiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.
12. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., & Cattin, P. C. (2022). Diffusion models for medical anomaly detection. *Medical Image Computing and Computer Assisted Intervention*, 35-45.
13. Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 22-40. <https://doi.org/10.1109/MGRS.2016.2540798>
14. Wang Z, He D, Zhao L, et al. Difusionseg: Diffusion-driven semantic segmentation with multi-modal image fusion for enhanced perception[J]. *Knowledge-Based Systems*, 2025: 114481.
15. Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195-4205.
16. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023). Align your latents: High-resolution video synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563-22575.
17. Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(6), e1264. <https://doi.org/10.1002/widm.1264>
18. Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152, 166-177. <https://doi.org/10.1016/j.isprsjprs.2019.04.015>
19. Persello, C., Wegner, J. D., Hänsch, R., Tuia, D., Ghamisi, P., Kooistra, L., & Herold, M. (2022). Deep learning and earth observation to support the sustainable development

goals. IEEE Geoscience and Remote Sensing Magazine, 10(2), 176-198.
<https://doi.org/10.1109/MGRS.2021.3136100>

20. Ghamisi, P., Maggiori, E., Li, S., Souza, R., Tarablaka, Y., Moser, G., De Giorgi, A., Fang, L., Chen, Y., Chi, M., Serpico, S. B., & Benediktsson, J. A. (2018). New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning. IEEE Geoscience and Remote Sensing Magazine, 6(3), 10-43.
<https://doi.org/10.1109/MGRS.2018.2854840>