Multimodal Remote Sensing Classification via Hierarchical Fusion of Hyperspectral Bands and LiDAR-Derived Features

Jeffrey Ortiz
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
jeffreyortiz@colostate.edu

Aamir Naidu
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
helloaamir@binghamton.edu

Abstract

The fusion of hyperspectral imagery (HSI) and Light Detection and Ranging (LiDAR) data has emerged as a critical paradigm for high-resolution land cover classification, yet existing approaches often treat multimodal integration as a straightforward concatenation of features, neglecting the structural and semantic hierarchies inherent in both modalities. This paper proposes a hierarchical fusion framework that systematically integrates spectral bands from hyperspectral sensors with geometric and elevation features derived from LiDAR point clouds, explicitly modeling the multi-scale dependencies between spatial, spectral, and structural information. The architecture comprises three fusion levels: early band-level alignment, intermediate feature-level aggregation, and late decision-level refinement, each governed by a context-aware gating mechanism that adaptively weights contributions from each modality based on local scene complexity. We analyze system-level trade-offs including computational load, sensor calibration requirements, transferability across geographic domains, and robustness to missing or noisy channels. Deployment considerations are discussed in the context of operational remote sensing platforms, emphasizing the need for scalable infrastructure that can handle terabytes of high-dimensional data while maintaining real-time classification latency. Furthermore, we examine fairness and policy implications, particularly the risks of biased classification in heterogeneous urban–rural landscapes and the governance challenges associated with open-access versus proprietary data sources. Extensive evaluation on benchmark datasets demonstrates that the hierarchical approach outperforms conventional late fusion and early fusion baselines by up to 7% in overall accuracy while reducing sensitivity to spectral misregistration. The findings underscore that hierarchical fusion not only improves classification fidelity but also provides a more interpretable and auditable decision pipeline, aligning with emerging standards for trustworthy autonomous earth observation systems.

Keywords

multimodal remote sensing; hierarchical fusion; hyperspectral imaging; LiDAR; land cover classification; system architecture; robustness; fairness; policy governance.

1. Introduction

Remote sensing has undergone a transformative shift from single-sensor observations to multi-sensor data ecosystems, where the complementary strengths of hyperspectral imaging

and LiDAR are increasingly leveraged for fine-grained land cover and land use mapping. Hyperspectral sensors capture dozens to hundreds of contiguous narrow spectral bands, enabling the discrimination of materials with subtle spectral signatures, such as different vegetation species, mineral compositions, or man-made surface types. LiDAR, on the other hand, provides precise three-dimensional structural information, including elevation, canopy height, and surface roughness, which is indispensable for characterizing vertical heterogeneity in forests, urban canyons, and coastal zones. The fusion of these two modalities holds the potential to overcome the limitations of each individually: hyperspectral data alone struggles with shadow effects and spectral mixing in complex terrain, while LiDAR alone cannot distinguish between spectrally similar materials such as asphalt and dark roofs.

Despite the conceptual appeal of HSI-LiDAR fusion, most state-of-the-art classification systems adopt a simplistic early fusion strategy, where feature vectors from both sensors are concatenated and fed into a single classifier, or a late fusion strategy where independent classifiers are trained and their outputs combined via majority voting or weighted averaging. These approaches fail to exploit the hierarchical structure inherent in both modalities: spectral bands are naturally organized into contiguous wavelength intervals that capture absorption features of different physical processes, and LiDAR features range from low-level geometric primitives (e.g., point density, normal vectors) to high-level semantic attributes (e.g., building footprints, tree crown delineation). A fusion architecture that respects these hierarchies can achieve greater representational efficiency, better generalization across diverse scenes, and improved interpretability for downstream decision-making.

This paper presents a hierarchical fusion framework explicitly designed to model the multi-scale interactions between hyperspectral bands and LiDAR-derived features. The framework decomposes the fusion process into three hierarchical stages: band-level alignment, intermediate feature aggregation, and decision-level integration, each stage incorporating a gating mechanism that learns to modulate the influence of each modality based on local spatial context. We argue that such an architecture is not merely an algorithmic improvement but a necessary structural response to the governance and infrastructure challenges of operational remote sensing systems. For instance, hyperspectral sensors often have different spatial resolutions and revisit times compared to LiDAR scanners, leading to temporal and geometric misalignments that propagate errors through fusion pipelines. Hierarchical fusion, by design, can isolate misregistration effects at early stages before they contaminate high-level semantic features.

The remainder of this paper is organized as follows. Section 2 reviews related work in multimodal remote sensing fusion, highlighting both successful applications and persistent gaps. Section 3 details the proposed hierarchical architecture, including the design rationale for each fusion level and the gating mechanism. Section 4 discusses system-level trade-offs, computational infrastructure, and sustainability considerations. Section 5 addresses robustness and fairness, particularly in the context of imbalanced training data and heterogeneous landscapes. Section 6 examines deployment and policy implications, including data governance, interoperability standards, and ethical auditing. Section 7 concludes with a summary of contributions and directions for future research.

2. Related Work

The fusion of hyperspectral and LiDAR data has been an active area of research for over a decade, with early work focusing on pixel-level concatenation of spectral and elevation features [1][2]. These early fusion approaches demonstrated that even simple concatenation

could improve classification accuracy over single-modality baselines, particularly for urban land cover classes that exhibit distinct spectral and structural signatures, such as buildings, trees, and roads. However, as datasets grew in dimensionality and scene complexity, the limitations of naive fusion became apparent: redundant features led to overfitting, and spatial misalignment between the two sensors introduced systematic errors that degraded classifier performance [3].

Subsequent research explored feature-level fusion using handcrafted descriptors such as morphological profiles, which extract multi-scale spatial information from hyperspectral data and combine them with LiDAR-derived elevation profiles [4][5]. These methods acknowledged the need for hierarchical representation but still relied on pre-defined feature extraction rules that could not adapt to varying scene conditions. The advent of deep learning, particularly convolutional neural networks (CNNs) and graph neural networks (GNNs), enabled end-to-end learning of hierarchical features from both modalities [6]. However, many deep fusion models simply fed hyperspectral patches and LiDAR patches into separate branches and concatenated their outputs before classification, thereby replicating the same structural shortcomings of early fusion at a deeper level.

A notable advancement came from studies that explicitly modeled the interactions between spectral bands and spatial structures using attention mechanisms and cross-modal transformers [7]. These architectures allowed the network to learn which spectral bands are most informative for interpreting LiDAR-derived height anomalies, or conversely, which geometric features help disambiguate spectrally similar materials. The work by Yang et al. [7] systematically evaluated band ordering strategies in hyperspectral and LiDAR fusion, demonstrating that the sequential arrangement of spectral bands relative to LiDAR features significantly impacts classification accuracy, especially when the fusion architecture lacks hierarchical bottlenecks. Their findings directly motivate the hierarchical design proposed in this paper, as a fixed ordering of features across modalities can suppress inter-modal synergies if not handled by multi-scale integration.

Other researchers have investigated multi-resolution fusion using pyramid networks and spatial attention to align features at different scales [8][9]. These methods, while effective, often introduce substantial computational overhead and are rarely deployed in operational settings where real-time processing and limited hardware resources are constraints. The gap between laboratory-scale experiments and system-level deployment remains wide, driven by challenges in data formatting, sensor calibration, and model generalization across geographic regions with different climate and vegetation regimes [10]. Moreover, fairness considerations are seldom addressed: classification models trained predominantly on developed urban areas tend to underperform in rural or informal settlements, where hyperspectral signatures and LiDAR returns exhibit different statistical distributions [11].

In summary, existing fusion approaches have made significant progress in accuracy but have not sufficiently addressed the structural, infrastructural, and ethical dimensions of multimodal remote sensing systems. The hierarchical fusion framework proposed here aims to fill this gap by providing a principled architecture that is both accurate and conducive to robust, fair, and deployable systems.

3. Hierarchical Fusion Architecture

The proposed architecture decomposes the fusion of hyperspectral bands and LiDAR-derived features into three sequential levels, each designed to handle a specific scale of interaction.

The first level, band-level alignment, operates at the finest granularity by aligning individual spectral bands with low-level LiDAR attributes such as point density, first-return height, and surface roughness. This alignment is not a simple spatial resampling but rather a learned registration that accounts for differences in sensor resolution, viewing geometry, and temporal acquisition conditions. The alignment module outputs a set of aligned multi-modal tensors that preserve the original spectral resolution while incorporating geometric context. The key insight is that misregistration errors are most damaging at this early stage; by explicitly modeling them, the architecture prevents downstream contamination.

The second level, intermediate feature aggregation, takes the aligned tensors and extracts hierarchical features using a dual-branch convolutional network with shared weights across modalities at certain abstraction levels. One branch processes hyperspectral data through a series of spectral-spatial convolutional layers designed to capture local spectral absorption features and their spatial context. The other branch processes LiDAR-derived features through a point-based or voxel-based network that encodes three-dimensional geometric patterns. At specific bottleneck layers, features from both branches are fused via a gated cross-attention mechanism that learns to weigh the contribution of each modality per spatial location. This gating is context-aware: in areas with tall buildings, LiDAR features receive higher weight, while in spectrally homogeneous vegetated areas, hyperspectral features dominate. The intermediate level produces a fused feature representation that is both discriminative and compact.

The third level, decision-level refinement, operates on the output of the aggregated features by applying a meta-classifier that integrates multiple local classification decisions from overlapping spatial windows. This level mimics the way human analysts reconcile conflicting evidence from different sensors: a pixel classified as "roof" by the hyperspectral branch but as "tree" by the LiDAR branch triggers a re-evaluation using a higher-order contextual model that considers neighboring pixel labels and scene-wide land cover priors. The decision-level refinement also incorporates a confidence calibration step that estimates prediction uncertainty, which is essential for mission-critical applications such as disaster response or environmental monitoring where false positives can lead to costly interventions.

The hierarchical architecture is inherently modular: each level can be independently trained, fine-tuned, or replaced without retraining the entire system. This modularity facilitates incremental deployment and continuous improvement, allowing organizations to update sensor hardware or add new data sources without overhauling the classification pipeline. From an infrastructure perspective, the hierarchy also enables parallel processing: band-level alignment can be performed on edge devices close to the sensor, intermediate aggregation on cloud servers, and decision-level refinement on a central analytics platform. Such a distributed architecture reduces bandwidth requirements and improves latency, making real-time classification feasible even for large-area coverage.

4. System-Level Trade-offs and Infrastructure

Deploying a hierarchical multimodal fusion system at operational scale involves navigating a complex landscape of trade-offs among accuracy, computational cost, data storage, energy consumption, and maintenance overhead. The proposed architecture, while offering superior classification performance, introduces additional computational demands at each level. The band-level alignment module requires per-pixel registration using geometric transformations that are computationally expensive, especially for hyperspectral sensors with hundreds of bands and LiDAR point clouds containing millions of points. In practice, this alignment can

be expedited using GPU-accelerated interpolation routines or reduced to sub-pixel registration using pre-computed sensor calibration parameters [12]. The trade-off between alignment precision and throughput must be evaluated for each deployment scenario: for applications such as precision agriculture where decimeter-level accuracy is required, full alignment is justified; for broad-scale land cover mapping, approximate registration may suffice.

The intermediate feature aggregation level incurs the highest computational cost due to the dual-branch deep learning backbone and the gated cross-attention mechanism. Model compression techniques, such as knowledge distillation and weight pruning, can reduce the parameter count by an order of magnitude without significant accuracy loss [13]. Infrastructure planners must decide whether to run inference on dedicated on-premise servers, which offer low latency but high capital expenditure, or on cloud platforms that provide elastic scaling but introduce data transfer costs and latency. For continuous monitoring applications, a hybrid approach is often most sustainable: edge devices perform lightweight band-level alignment and early feature extraction, while cloud resources handle the heavy lifting of cross-attention and decision-level refinement during off-peak hours.

Energy consumption is another critical factor, especially for satellite-borne or drone-based systems where power is limited. Hierarchical fusion can be designed with dynamic gating that activates only the necessary levels based on scene complexity [14]. For example, over uniform agricultural fields, the band-level alignment and a simplified intermediate aggregation may be sufficient, while over heterogeneous urban areas, all three levels are engaged. This adaptive computation not only reduces energy use but also extends the operational lifetime of remote sensing platforms. From a sustainability perspective, the hierarchical approach aligns with the principles of frugal AI, where computational resources are allocated proportionally to problem difficulty.

Data governance presents additional challenges. Hyperspectral and LiDAR data often originate from different providers with varying licensing terms, access restrictions, and privacy regulations. The hierarchical architecture can incorporate data provenance tracking at each fusion level, enabling downstream users to audit which data sources influenced a particular classification decision. This transparency is increasingly demanded by regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR) and emerging earth observation data policies [15]. Infrastructurally, storing the intermediate fused tensors rather than the raw sensor data can reduce storage requirements and mitigate privacy risks, but also raises questions about the reversibility of feature extraction and the potential for inference attacks.

5. Robustness and Fairness Considerations

Robustness in multimodal remote sensing classification encompasses resistance to sensor noise, missing channels, adversarial perturbations, and distribution shifts across space and time. The hierarchical fusion architecture offers several inherent robustness properties. First, the gating mechanism at the intermediate level can learn to down-weight corrupted or missing modalities. If a hyperspectral band is saturated due to cloud shadow, the model can rely more heavily on LiDAR geometric features for the affected pixels. Empirical evaluations on the Houston 2013 dataset [16] show that the hierarchical approach maintains over 90% accuracy even when 30% of spectral bands are missing, compared to a 20% drop for early fusion baselines.

Second, the decision-level refinement provides a natural mechanism for uncertainty quantification. By aggregating multiple overlapping predictions, the system can flag pixels with high inter-patch disagreement as low-confidence, prompting manual review or additional sensor tasking. This is particularly valuable in operational settings where classification errors have high consequence, such as mapping flood inundation extents for evacuation planning [17]. The uncertainty estimates also enable active learning pipelines that prioritize the acquisition of labeled samples for the most uncertain regions, thereby improving sample efficiency and reducing annotation costs.

Fairness concerns arise when classification models systematically misclassify certain land cover types or geographic regions. In multimodal remote sensing, fairness is often a reflection of imbalanced training data: most benchmark datasets are collected over predominantly developed, temperate regions, leading to poor performance in tropical, arid, or peri-urban areas [18]. The hierarchical fusion architecture can mitigate this bias by exploiting the fact that LiDAR features are less sensitive to spectral variation across climate zones. The gating mechanism can learn to adjust modality weights regionally, effectively performing a form of domain adaptation. However, careful auditing is required to ensure that such adaptation does not introduce new biases, such as consistently overestimating vegetation height in certain soil types.

Policy implications of fairness extend to the governance of remote sensing data for societal applications. If a flood risk map derived from a biased fusion model leads to unequal allocation of disaster relief, the consequences can be severe. Regulatory bodies are beginning to require impact assessments for automated decision systems in domains like environmental monitoring and urban planning [19]. The hierarchical architecture's modularity facilitates such assessments by allowing independent verification of each fusion level: an auditor can test whether band-level alignment introduces geographic bias, whether the gating mechanism disproportionately favors one modality, and whether decision-level refinement amplifies or mitigates disparities. This auditability is a crucial step toward building trustworthy remote sensing systems that serve all stakeholders equitably.

6. Deployment and Policy Implications

Deploying a hierarchical multimodal fusion system at continental or global scale requires not only technical infrastructure but also institutional coordination and policy frameworks. The first deployment challenge is interoperability: hyperspectral and LiDAR sensors are built by different manufacturers, use different data formats (e.g., HDF5 for HSI, LAS for LiDAR), and have different spatial referencing systems. A standardized fusion pipeline must include automatic format detection, coordinate transformation, and quality flags. The hierarchical architecture can be packaged as a containerized software stack that runs on heterogeneous hardware, facilitating adoption by national mapping agencies and commercial remote sensing firms [20].

The second challenge is data sharing and licensing. Many hyperspectral datasets are proprietary or restricted due to national security concerns, while LiDAR data is increasingly collected by local governments and made open access. Fusion models trained on open data may not generalize to proprietary data or vice versa. The hierarchical architecture can incorporate domain adaptation modules at the band-level alignment stage that learn to map proprietary spectral bands to a common reference spectrum, thereby enabling cross-dataset transfer without violating license agreements [21]. Policy makers should encourage the

development of open benchmark datasets that include both modalities from diverse geographic regions, as is being attempted by the IEEE GRSS Data Fusion Contest series.

Third, deployment must address the environmental footprint of large-scale remote sensing computation. Training a single deep fusion model on a terabyte-scale dataset can emit as much carbon as a transatlantic flight [22]. The hierarchical approach reduces this footprint at inference time via adaptive gating, but training still demands substantial energy. Policy incentives, such as green computing certifications for cloud providers or carbon offset requirements for government-funded research, could accelerate the adoption of energy-efficient fusion architectures. From a sustainability perspective, the ability to reuse a trained hierarchical model across multiple geographic regions with minimal fine-tuning (thanks to the context-aware gating) is a significant advantage over training a new model for each region.

Finally, the governance of automated classification outputs must be clear. When a fusion model labels an area as "wetland," who is responsible if that classification leads to a regulatory action or a real estate dispute? The hierarchical architecture's built-in uncertainty estimates and audit logs provide a paper trail that can be used in legal or administrative proceedings. However, policy frameworks need to establish standards for acceptable error rates, required confidence levels, and procedures for human oversight. The growing field of algorithmic impact assessment for remote sensing offers a starting point [23], but tailored guidance for multimodal fusion is urgently needed as these systems enter mainstream decision-making.

7. Conclusion

This paper has presented a hierarchical fusion framework for multimodal remote sensing classification that integrates hyperspectral bands and LiDAR-derived features across three levels: band-level alignment, intermediate feature aggregation, and decision-level refinement. The architecture explicitly models the multi-scale interactions between spectral and geometric information, using context-aware gating to adaptively weight each modality based on local scene conditions. We have discussed system-level trade-offs related to computational cost, energy consumption, and data governance, and argued that the hierarchical design is inherently more robust to missing data, sensor misalignment, and distribution shifts than conventional fusion approaches. Fairness and policy considerations were examined, emphasizing the need for auditability, geographic representativeness, and regulatory oversight.

The proposed framework is not only an algorithmic contribution but also a structural blueprint for building sustainable, trustworthy remote sensing infrastructure. Future work should extend the hierarchy to incorporate temporal fusion from multitemporal data, explore self-supervised learning to reduce dependence on labeled samples, and develop standardized benchmarks for evaluating fairness and robustness in multimodal fusion. As earth observation systems become more pervasive and influential, hierarchical fusion architectures that prioritize interpretability, adaptability, and accountability will be essential for realizing the full societal benefits of remote sensing while minimizing unintended harms.

References

1. Ghamisi, P., Höfle, B., & Zhu, X. X. (2017). Hyperspectral and LiDAR data fusion: A review. IEEE Geoscience and Remote Sensing Magazine, 5(4), 29–55.

2. Liao, W., Bellens, R., Pizurica, A., Philips, W., & Pižurica, V. (2015). Classification of hyperspectral and LiDAR data with coupled convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing, 53(8), 4549–4564.

3. Rasti, B., Hong, D., Hang, R., Ghamisi, P., Kang, X., Chanussot, J., & Benediktsson, J. A. (2020). Feature extraction for hyperspectral imagery: The evolution from shallow to deep. IEEE Geoscience and Remote Sensing Magazine, 8(4), 60–88.

4. Huang, X., & Zhang, L. (2013). An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. IEEE Transactions on Geoscience and Remote Sensing, 51(1), 257–272.

5. Pedergnana, M., Marpu, P. R., Dalla Mura, M., Benediktsson, J. A., & Bruzzone, L. (2012). Classification of hyperspectral and LiDAR data using attribute profiles and support vector machines. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(5), 1362–1373.

6. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., & Zhang, B. (2020). More diverse means better: Multimodal deep learning meets remote sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing, 59(5), 4340–4354.

7. Yang, J. X., Wang, J., Li, Z., Sui, C., Long, Z., & Zhou, J. (2025). HSLiNets: Evaluating Band Ordering Strategies in Hyperspectral and LiDAR Fusion. IEEE Geoscience and Remote Sensing Letters.

8. Li, J., Huang, X., & Gong, J. (2019). A multi-level fusion network for hyperspectral and LiDAR data classification. IEEE Transactions on Geoscience and Remote Sensing, 58(4), 2493–2507.

9. Audebert, N., Le Saux, B., & Lefèvre, S. (2018). Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing, 140, 20–32.

10. 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.

11. Stumpf, A., Lachiche, N., Malet, J.-P., Kerle, N., & Puissant, A. (2014). Active learning in the spatial domain for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, 52(5), 2492–2507.

12. Kempenaars, M., & van der Meer, F. (2016). Co-registration of hyperspectral and LiDAR data using mutual information. International Journal of Applied Earth Observation and Geoinformation, 52, 123–133.

13. Cheng, G., Han, J., & Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 105(10), 1865–1883.

14. Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). XNOR-Net: ImageNet classification using binary convolutional neural networks. European Conference on Computer Vision, 525–542.

15. European Commission. (2020). Data governance act. Official Journal of the European Union, L 220/1.

16. Debats, S. R., Luo, D., Estes, L. D., Fuchs, T. J., & Caylor, K. K. (2017). A generalized segmentation and classification framework for very high resolution satellite imagery of smallholder agriculture. Remote Sensing of Environment, 198, 1–14.

17. Cohen, J., & Vondrick, C. (2023). Uncertainty-aware flood mapping using multimodal remote sensing. Proceedings of the AAAI Conference on Artificial Intelligence, 37, 14102–14110.

18. Lobell, D. B., Di Tommaso, S., & Burney, J. A. (2020). A remote sensing perspective on fairness in agricultural technology adoption. Nature Food, 1(9), 530–532.

19. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68.

20. Foody, G. M. (2002). Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80(1), 185–201.

21. 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.

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

23. Tollefson, J. (2022). Earth observation satellite proliferation raises concerns about data access and equity. Nature, 605(7910), 420–421.