Attention-Based Multisource Remote Sensing Fusion Using Hyperspectral and LiDAR Observations

Bruce Salonen
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
bruce.work@buffalo.edu

Ananya A. Chandra
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
ananyachandra73@binghamton.edu

Malcolm Neal
Department of Computer Science, George Mason University, Fairfax, VA, USA.
malcolm1977@gmu.edu

Abstract

The integration of hyperspectral imaging and Light Detection and Ranging observations represents a transformative capability in modern remote sensing, enabling simultaneous acquisition of high-dimensional spectral signatures and precise three-dimensional structural information. While conventional fusion methods have relied primarily on stacking feature vectors or applying linear transformations, these approaches often fail to capture the complex, non-linear interactions between spectral and spatial modalities. This paper presents a comprehensive analysis of attention-based architectures for multisource remote sensing fusion, examining their capacity to model long-range dependencies, prioritize salient features, and dynamically weight contributions from heterogeneous data streams. We systematically explore the architectural trade-offs associated with attention mechanisms, including self-attention, cross-attention, and multi-head configurations, within the context of hyperspectral and LiDAR data fusion. The study further investigates the implications of such systems for large-scale infrastructure deployment, emphasizing considerations of computational efficiency, energy consumption, data governance, and model robustness across varying geographic and operational conditions. Through a cross-domain analytical lens, we examine how attention-based fusion frameworks can be designed to support sustainable and equitable deployment in environmental monitoring, urban planning, and disaster response. The paper also addresses challenges related to spectral redundancy, spatial resolution disparities, and label scarcity, proposing governance-oriented strategies for training data curation and model validation. By situating technical innovations within broader socio-technical infrastructure systems, the analysis underscores the need for transparent, auditable, and fair fusion methodologies. We conclude that attention-based multisource fusion, when informed by rigorous structural and policy considerations, offers a robust pathway toward more accurate, resilient, and interpretable remote sensing systems.

Keywords

attention mechanisms; hyperspectral imaging; LiDAR; multisource fusion; remote sensing infrastructure; socio-technical systems; deep learning governance.

1. Introduction

The growing availability of multi-modal Earth observation data has fundamentally altered the landscape of environmental sensing, moving the field beyond single-sensor paradigms toward integrated analytical frameworks that can jointly interpret complementary data streams. Among the most promising combinations for land cover classification, vegetation analysis, and urban mapping is the fusion of hyperspectral imagery with LiDAR point cloud observations [1]. Hyperspectral sensors capture reflectance values across hundreds of narrow contiguous spectral bands, providing rich material composition information [2]. LiDAR, by contrast, delivers high-resolution three-dimensional point clouds that encode surface elevation, canopy height, and structural geometry [3]. When fused, these modalities can overcome the limitations inherent to each individually: hyperspectral data alone often suffers from spectral ambiguity in structurally complex scenes, while LiDAR alone lacks the spectral discrimination necessary for detailed material classification [4]. However, the effective fusion of these disparate data sources presents significant technical and conceptual challenges, ranging from spatial registration and resolution alignment to the modeling of cross-modal feature dependencies [5].

Traditional approaches to multisource data fusion have employed straightforward concatenation of feature vectors extracted from each modality, followed by classification or regression using conventional machine learning models [6]. More recent methods have applied convolutional neural networks to extract hierarchical spatial features from both hyperspectral patches and LiDAR-derived digital surface models [7]. While these techniques have demonstrated measurable improvements over single-modality baselines, they are fundamentally limited by their inability to capture the long-range contextual relationships that span across both spectral and spatial dimensions. The spatial structure of LiDAR data, for instance, is inherently three-dimensional and irregularly sampled, whereas hyperspectral data is organized as dense two-dimensional arrays with deep spectral channels. Aligning these distinct data geometries within a unified representation remains an open area of investigation [8].

Attention mechanisms, originally developed within the domain of natural language processing, have recently emerged as a powerful tool for multi-modal fusion in remote sensing [9]. The core intuition behind attention is conceptually straightforward: rather than treating all input features equally, the model learns to assign differential weights to different regions, spectral bands, or spatial locations based on their relevance to the task at hand. This capacity for dynamic feature selection is particularly well-suited to hyperspectral and LiDAR fusion, where the importance of specific spectral signatures may depend on the structural context provided by elevation data, and vice versa [10]. In a typical attention-based fusion pipeline, the model computes a set of attention scores that reflect the compatibility between query vectors from one modality and key-value pairs from another, effectively constructing a context-aware weighted combination of the two data sources [11]. The study by Yang et al. provides an important baseline for understanding how band ordering strategies interact with fusion architectures, though the present analysis extends this inquiry into broader system-level considerations.

The present paper offers a long-form analytical discussion of attention-based fusion for hyperspectral and LiDAR data, with an emphasis on architectural design, infrastructure deployment, governance, and sustainability. Rather than presenting a single experimental benchmark, we aim to map the design space of attention-based fusion systems and to critically evaluate their potential and limitations when deployed at scale. This perspective is

informed by the recognition that remote sensing technologies increasingly function as components of larger socio-technical infrastructures, embedded within policy frameworks, economic incentives, and community decision-making processes. As such, the technical evaluation of attention mechanisms must be complemented by an analysis of their implications for data equity, algorithmic transparency, and environmental sustainability.

2. Architectural Foundations of Attention-Based Fusion

The design of attention mechanisms for multisource remote sensing fusion must contend with the fundamental heterogeneity of the input data. Hyperspectral imagery is typically represented as a three-dimensional tensor with spatial dimensions in the horizontal and vertical axes and a spectral dimension along the third axis, while LiDAR data is most commonly represented as an unordered point cloud with spatial coordinates and, optionally, intensity or return number attributes [12]. To fuse these representations, it is necessary to project them into a common embedding space where attention can operate across modalities. This embedding process introduces several architectural choices that carry significant implications for model capacity, computational cost, and generalization performance.

One widely adopted architecture for attention-based fusion involves the use of separate encoder branches for each modality, followed by a cross-attention module that computes the interactions between the two encoded representations [13]. In this design, the hyperspectral encoder might consist of a series of three-dimensional convolutional layers that capture spectral-spatial features, while the LiDAR encoder employs either a voxel-based convolutional network or a point-based neural network such as PointNet++. The outputs of these encoders are then fed into a cross-attention mechanism that allows the model to selectively attend to regions in one modality based on information from the other [14]. The attended features are subsequently aggregated through a pooling operation or a transformer block before being passed to a classification head. This modular architecture facilitates independent pretraining of the encoders on large corpora of single-modality data, which can be beneficial when paired multi-modal training data is scarce.

An alternative architectural paradigm is the unified transformer, wherein both modalities are tokenized and concatenated into a single sequence that is processed by a shared transformer backbone [15]. In such a framework, each patch of hyperspectral image and each LiDAR point is treated as a token, with positional encodings capturing both spatial coordinates and, in the case of hyperspectral data, spectral band indices. The self-attention layers within the transformer then compute pairwise relationships among all tokens, regardless of their source modality. This approach has the advantage of allowing the model to learn cross-modal interactions from the outset, without the need for a separate fusion module. However, it also introduces substantial computational overhead, as the attention matrix scales quadratically with the total number of tokens. For large-area remote sensing scenes containing millions of pixels and points, this scaling behavior can render the unified transformer impractical without aggressive token reduction strategies.

Multi-head attention represents a further refinement that allows the model to attend to information from different representation subspaces simultaneously [16]. In a multi-head configuration, the input embeddings are projected multiple times into lower-dimensional spaces, each of which is processed by an independent attention head. The outputs of all heads are then concatenated and projected back to the original dimensionality. For hyperspectral and LiDAR fusion, multi-head attention is particularly valuable because different heads can specialize in capturing distinct types of cross-modal relationships. One head might focus on

spectral similarities between surface materials identified by LiDAR-derived elevation changes, while another might attend to textural patterns in the hyperspectral data that correlate with the vertical structure of vegetation canopies. The use of multi-head attention thereby increases the representational capacity of the fusion model without requiring an exponential increase in parameters.

A critical architectural consideration that arises in attention-based fusion is the handling of scale and resolution disparities. Hyperspectral sensors typically have a spatial resolution on the order of meters to tens of meters, while LiDAR point clouds can achieve sub-meter resolution along the vertical axis. When fusing these data sources, it is often necessary to downsample the LiDAR point cloud to match the grid of the hyperspectral image, or alternatively to upsample the hyperspectral image to the native resolution of the LiDAR data. Both strategies involve loss of information. Attention mechanisms can partially mitigate these losses by learning to weigh the contributions of different points or pixels based on their spatial alignment, but the fundamental resolution mismatch remains a structural constraint that must be addressed through careful preprocessing and data governance protocols [17].

3. Structural Trade-Offs and System Robustness

The deployment of attention-based fusion systems for operational remote sensing tasks introduces a series of structural trade-offs that extend beyond pure predictive accuracy. These trade-offs encompass computational and memory requirements, inference latency, model interpretability, and robustness to distributional shifts. Understanding and managing these trade-offs is essential for the design of fusion systems that are not only accurate but also practical, sustainable, and trustworthy.

The computational cost of attention mechanisms is dominated by the pairwise similarity computation between query and key vectors, which scales as the product of the number of tokens. For hyperspectral data, each pixel may be considered a token, leading to tens of thousands of tokens for a single image patch. When LiDAR tokens are added to the same sequence, the total token count can increase substantially. While sparse attention patterns and linear attention approximations have been proposed to reduce this complexity, these methods often sacrifice the ability to capture long-range global dependencies, which are precisely the strengths that motivate the use of attention in the first place [18]. From an infrastructure perspective, the choice between full attention and its approximations should be informed by the specific deployment context: high-throughput, near-real-time applications such as disaster monitoring may require linear attention, while offline, high-accuracy mapping campaigns may justify the computational expense of full attention.

Model interpretability is another dimension where attention mechanisms offer both opportunities and challenges in the remote sensing fusion context. Attention weights can be visualized as heatmaps that indicate which regions of the input were most influential in producing a given output. This can be valuable for verifying that the model is focusing on physically meaningful features, such as vegetation boundaries or building edges, rather than spurious correlations [19]. However, recent research has shown that attention weights are not always reliable indicators of feature importance, particularly when the attention heads interact in complex ways or when the model is overparameterized. In the regulatory and governance context of remote sensing, where decisions about land use classification or natural resource management may have significant socio-economic consequences, the opacity of attention-based models poses a challenge to accountability. Development of attention-based fusion systems should therefore be accompanied by complementary interpretability tools, such as

gradient-based attribution methods or concept-based explanations, to ensure that model outputs can be rigorously audited.

Robustness to domain shift is a particularly pressing concern for multisource fusion systems. Hyperspectral sensors are sensitive to atmospheric conditions, illumination angles, and seasonal vegetation changes, while LiDAR returns can be affected by surface reflectivity, scan angle, and flying altitude. When an attention-based fusion model is trained on data from one geographic region or acquisition campaign and deployed in a different setting, its performance may degrade significantly if the joint distribution of spectral and structural features shifts [20]. Attention mechanisms, by virtue of their capacity for dynamic weighting, can theoretically adapt to some forms of domain shift by reweighting features based on their local context. In practice, however, this adaptation is limited by the extent to which the training data covers the space of possible cross-modal variations. Strategies such as domain adversarial training, self-supervised pretraining on large multi-modal corpora, and test-time adaptation have been proposed to enhance the robustness of fusion models, but each of these approaches introduces additional complexity and computational overhead.

4. Infrastructure, Deployment, and Sustainability

The operational deployment of attention-based fusion systems for hyperspectral and LiDAR data is not merely a technical problem but an infrastructural one. Airborne and spaceborne hyperspectral sensors are expensive to develop and operate, and LiDAR surveys are similarly resource-intensive, typically requiring dedicated aircraft or satellite platforms. The data volumes generated by modern hyperspectral sensors routinely exceed tens of gigabytes per flight line, and LiDAR point clouds can add similar or larger data volumes. Attention-based neural networks, particularly those employing transformer architectures, require substantial computational resources for both training and inference, often involving multiple graphics processing units operating over extended periods. The energy consumption associated with training large deep learning models has been estimated to produce carbon emissions comparable to several transatlantic flights, raising sustainability concerns for the widespread adoption of these technologies [21].

From a sustainability perspective, the deployment of attention-based fusion systems must therefore be evaluated within a lifecycle assessment framework that accounts not only for the accuracy of land cover maps but also for the environmental cost of producing them. This includes the energy consumed during sensor operation, data transmission, storage, preprocessing, model training, and inference. Attention-based methods, while potentially more accurate than simpler alternatives, may not always be the most resource-efficient choice for applications where moderate accuracy is sufficient. Decision-makers in government agencies and environmental monitoring organizations should consider tiered deployment strategies, where lightweight models are used for routine monitoring and attention-based systems are reserved for high-stakes or high-complexity tasks.

Data governance represents another critical dimension of infrastructure for attention-based fusion. Hyperspectral and LiDAR data often contain sensitive information about land ownership, infrastructure, and natural resources. In many jurisdictions, high-resolution remote sensing data is subject to export controls, privacy regulations, or restrictions on use by non-governmental actors. Attention-based models, by learning to attend to specific spatial and spectral features, can inadvertently encode information that is proprietary or confidential. The governance of training datasets must therefore include protocols for data anonymization, access control, and licensing. Furthermore, the outputs of fusion models, such as classified

land cover maps, should be subjected to verification procedures that involve ground-truthing by domain experts, particularly when the results are used to inform policy decisions or resource allocation.

Equity and fairness are additional governance considerations that are often overlooked in the remote sensing literature. Attention-based fusion models trained on data from well-resourced regions with extensive ground truth may perform poorly when applied to underrepresented regions with different ecological or urban characteristics [22]. This can lead to systematic biases in the production of environmental data products, with wealthier regions receiving more accurate and up-to-date maps while poorer regions are left with lower-quality information. Attention mechanisms that dynamically weight features based on the local data distribution may partially compensate for these biases, but structural inequities in data availability and sensor coverage require policy-level interventions as well. International collaborations and open data initiatives, such as the European Space Agency's Copernicus program, represent important steps toward more equitable access to both remote sensing data and the analytical tools needed to interpret them.

5. Cross-Domain Comparisons and Forward-Looking Perspectives

Attention-based fusion of hyperspectral and LiDAR data shares conceptual and architectural similarities with multi-modal fusion problems in other domains, including medical imaging, autonomous driving, and natural language processing. In medical imaging, for instance, the fusion of magnetic resonance imaging and computed tomography scans using attention mechanisms has been explored to combine soft tissue contrast with bone structure information [23]. In autonomous driving, transformer-based architectures have been used to fuse camera imagery with radar and LiDAR point clouds for object detection and path planning [24]. These cross-domain parallels offer valuable lessons for the remote sensing community, particularly regarding the importance of temporal alignment, sensor calibration, and uncertainty quantification.

One notable insight from other domains is the value of incorporating auxiliary tasks into the training of attention-based fusion models. In autonomous driving, for example, multi-task learning frameworks that jointly predict object detections, lane boundaries, and semantic segmentation have been shown to improve the quality of learned representations across all tasks. Applied to remote sensing, an attention-based fusion model could be trained simultaneously on land cover classification, vegetation height estimation, and building footprint delineation, leveraging shared features across the hyperspectral and LiDAR inputs [25]. Such multi-task training can improve sample efficiency and model generalization, while also producing multiple useful outputs from a single inference pass. However, it also requires careful balancing of loss functions and may introduce conflicts between tasks with competing gradient signals.

Looking forward, several emerging trends are likely to shape the evolution of attention-based remote sensing fusion. The advent of foundation models for remote sensing, pretrained on massive corpora of satellite and aerial imagery, offers the potential to provide powerful feature extractors that can be fine-tuned for fusion tasks with limited labeled data [26]. These foundation models, often based on vision transformer architectures, can serve as a drop-in replacement for the spectral and spatial encoders in an attention-based fusion pipeline. Another trend is the increasing availability of onboard processing capabilities on satellite platforms, which would enable real-time inference of attention-based models in orbit, reducing the need for downlinking large volumes of raw data. This aligns with broader

movement toward edge computing in Earth observation, where latency and bandwidth constraints drive the need for efficient on-device models.

The integration of attention-based fusion with physics-based models represents another promising direction. Remote sensing data is governed by well-understood physical processes, including radiative transfer, atmospheric scattering, and geometric optics. Rather than relying solely on learned representations, a hybrid approach could embed physical constraints into the attention mechanism, ensuring that the fusion model respects the physical relationships between spectral reflectance and structural geometry [27]. For example, an attention head could be designed to compute weights based on the bidirectional reflectance distribution function, encouraging the model to attend to viewing and illumination angles that are physically plausible. Such physically informed attention mechanisms could improve the generalization of fusion models to novel acquisition conditions and reduce the data requirements for training.

6. Conclusion

Attention-based mechanisms offer a principled and powerful framework for fusing hyperspectral and LiDAR observations in remote sensing, enabling models to dynamically prioritize the most informative features from each modality and to capture complex cross-modal dependencies that are inaccessible to traditional fusion methods. However, the successful deployment of these systems at scale requires a holistic perspective that extends beyond algorithmic innovation to encompass architectural design trade-offs, computational sustainability, data governance, and social equity. The architectural choices between separate encoder branches with cross-attention, unified transformers, and multi-head configurations carry significant implications for model capacity, interpretability, and robustness. These choices must be evaluated in light of the specific operational requirements of the deployment environment, including latency budgets, energy constraints, and the availability of labeled training data.

From an infrastructure standpoint, attention-based fusion systems must be designed and governed with an awareness of their environmental footprint, their potential for bias in data-scarce regions, and their compliance with regulatory frameworks governing the use of high-resolution remote sensing data. Cross-domain comparisons with medical imaging and autonomous driving reveal useful design strategies, including multi-task learning and physically informed attention, that can enhance the utility and reliability of remote sensing fusion models. As foundation models and onboard processing capabilities continue to mature, the integration of attention-based fusion into operational Earth observation pipelines is likely to accelerate. The present analysis underscores the importance of embedding these technical advances within a broader socio-technical infrastructure that prioritizes transparency, fairness, and long-term sustainability. By doing so, the remote sensing community can harness the full potential of attention-based multisource fusion to produce more accurate, equitable, and actionable environmental information for society.

References

1. Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., ... & Benediktsson, J. A. (2019). Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. IEEE Geoscience and Remote Sensing Magazine, 7(1), 6-39.

2. Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., & Chanussot, J. (2012). Hyperspectral remote sensing data analysis and future challenges. IEEE Geoscience and Remote Sensing Magazine, 1(2), 6-36.

3. Hoffer, R. M., & Fleming, M. D. (2020). Airborne laser scanning for vegetation structure assessment: A review of methods and applications. Remote Sensing of Environment, 245, 111852.

4. Cao, X., Xu, L., Meng, D., Zhao, Q., & Xu, Z. (2021). Integration of hyperspectral and LiDAR data for land cover classification using deep learning: A critical review. ISPRS Journal of Photogrammetry and Remote Sensing, 176, 124-140.

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

6. Liao, W., Pizurica, A., Philips, W., & Pi, Y. (2015). A review of fusion frameworks for hyperspectral and LiDAR data integration. IEEE Geoscience and Remote Sensing Magazine, 3(3), 8-24.

7. Ha, V. K., Ho, Q. T., & Jung, H. (2020). Convolutional neural networks for land cover classification using fused hyperspectral and LiDAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 4726-4738.

8. Xu, D., & Ouyang, S. (2021). Spatial-spectral alignment for multimodal remote sensing fusion using graph attention networks. IEEE Transactions on Geoscience and Remote Sensing, 59(11), 9361-9374.

9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

10. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., & Zhang, B. (2021). More diverse means better: Multimodal deep learning meets remote sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing, 59(5), 4340-4354.

11. 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.

12. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 652-660.

13. Chen, C., Li, S., & Qin, J. (2022). Cross-attention fusion network for multisource remote sensing data classification. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.

14. Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3), 331-368.

15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

16. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5797-5808.

17. Zhang, X., Sun, Y., Song, H., & Li, J. (2023). Resolution-adaptive fusion of hyperspectral and LiDAR data using transformer networks. IEEE Transactions on Geoscience and Remote Sensing, 61, 1-15.

18. Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The efficient transformer. International Conference on Learning Representations.

19. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 3543-3556.

20. 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.

21. 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.

22. Rolf, E., Proctor, J., Rodolfi, T., Carleton, T., & Greicius, T. (2021). The unequal distribution of satellite-based environmental data products. Environmental Research Letters, 16(9), 094035.

23. Xu, Y., Zheng, Y., & Yu, J. (2022). Multi-modal medical image fusion using attention-guided deep networks: A review and taxonomy. Information Fusion, 83, 34-52.

24. Huang, Y., Liu, Y., Su, Y., & Zhang, L. (2022). Multi-modal fusion for autonomous driving: A survey of sensor integration and attention mechanisms. IEEE Transactions on Intelligent Transportation Systems, 23(10), 16974-16991.

25. Kampffmeyer, M., Salberg, A. B., & Jenssen, R. (2016). Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. IEEE Geoscience and Remote Sensing Letters, 13(7), 932-936.

26. Cong, Y., Khanna, S., Mok, S., & Zhu, X. X. (2023). Foundation models for Earth observation: A comprehensive survey and outlook. arXiv preprint arXiv:2311.04967.

27. Rasti, B., Ghamisi, P., & Gloaguen, R. (2020). Physics-aware deep learning for hyperspectral image analysis: A review of models, datasets, and future directions. IEEE Geoscience and Remote Sensing Magazine, 8(4), 86-112.