Hierarchical Multi-Scale Attention Networks for Integrating Hyperspectral, LiDAR, and Camera Data in Smart Cities

Otis Taylor
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
otis.taylor@missouri.edu

Rowan L. Lopez
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
rowanl@uab.edu

Keguo Gu
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
keguo1990@ku.edu

Tobias Griksson
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
eriksson1991@ucf.edu

Abstract

The convergence of hyperspectral imaging, Light Detection and Ranging (LiDAR), and conventional camera data presents unprecedented opportunities for smart city applications, including environmental monitoring, infrastructure assessment, and urban planning. However, integrating these heterogeneous modalities remains challenging due to disparities in spatial resolution, spectral dimensionality, and geometric representation. This paper proposes a hierarchical multi-scale attention network architecture that systematically fuses hyperspectral, LiDAR, and camera data at multiple abstraction levels. The framework employs dedicated encoders for each modality, followed by cross-attention modules that align features across scales and sensor domains. A hierarchical aggregation mechanism then integrates local, regional, and global contextual cues, enabling robust urban feature extraction. Beyond technical design, the paper critically examines structural trade-offs between computational efficiency and model capacity, between centralised cloud processing and distributed edge deployment, and between interpretability and predictive accuracy. Governance and policy considerations are addressed, including data ownership, privacy preservation, and equitable sensor coverage across socioeconomically diverse urban zones. Sustainability aspects such as energy consumption during inference and sensor life-cycle management are analysed. The proposed architecture is positioned within the broader landscape of large-scale socio-technical systems, where robustness, fairness, and long-term maintainability are as important as algorithmic performance. Case illustrations from recent fusion benchmarks and real-world pilot deployments underscore the practical challenges and opportunities. By foregrounding

system-level reasoning, this paper provides a comprehensive framework for designing trustworthy and scalable multi-modal sensing infrastructures in smart cities.

Keywords

smart cities, multi-modal data fusion, hyperspectral imaging, LiDAR, attention mechanisms, hierarchical networks, urban infrastructure, fairness, governance, sustainability.

1. Introduction

The vision of smart cities relies on an intricate web of sensors that continuously monitor urban environments to enable data-driven decision-making. Among the most informative sensing modalities are hyperspectral imagers, which capture dozens to hundreds of narrow spectral bands covering visible to shortwave infrared wavelengths; LiDAR sensors, which provide precise three-dimensional point clouds; and conventional RGB cameras, which offer high spatial resolution in familiar visual domains. Individually, each modality has well-known strengths and limitations: hyperspectral data excels at material identification but suffers from low spatial resolution and high dimensionality; LiDAR delivers accurate geometry but lacks spectral richness; cameras offer dense spatial information but are confounded by illumination and atmospheric conditions [1], [2]. The combination of all three promises a more complete representation of urban scenes, enabling applications such as land-cover classification, vegetation health assessment, building change detection, and traffic monitoring.

Early fusion approaches relied on simple concatenation or early integration of features, but these methods fail to account for the complex, non-linear relationships and scale disparities inherent in multi-modal data [3]. More recent work has turned to deep learning, particularly convolutional neural networks and transformer-based architectures, to learn joint representations [4]. However, most existing models operate at a single spatial or spectral scale, overlooking the fact that urban structures manifest at multiple granularities: a tree canopy may be captured at metre-scale LiDAR returns, a building facade at sub-metre camera pixels, and a chemical pollutant plume at hyperspectral reflectance patterns that span several kilometres [5]. A hierarchical approach that explicitly models multi-scale interactions is therefore necessary.

This paper introduces a hierarchical multi-scale attention network (HMSA-Net) designed to integrate hyperspectral, LiDAR, and camera data. The core idea is to process each modality through a dedicated encoder that extracts features at several spatial and spectral scales. Cross-attention mechanisms then allow the network to learn correspondences between modalities at each scale, while a hierarchical aggregation module fuses these scale-specific representations into a unified feature map. Importantly, the architecture is designed not only for predictive accuracy but also for system-level viability: it must be deployable on resource-constrained edge nodes, maintainable across heterogeneous sensor deployments, and fair in its coverage of different urban populations.

The remainder of the paper is structured as follows. Section 2 reviews related work in multi-modal fusion and attention-based models. Section 3 presents the proposed hierarchical multi-scale attention architecture in detail, emphasising its conceptual design and structural choices. Section 4 discusses the structural trade-offs and system-level considerations that arise when moving from a laboratory model to a real-world smart city infrastructure. Section 5 examines deployment challenges, including computational sustainability and robustness. Section 6 addresses fairness, governance, and policy implications, arguing that technical excellence

alone is insufficient for responsible urban sensing. Section 7 concludes with a summary of contributions and future research directions.

2. Related Work

The fusion of hyperspectral and LiDAR data has been an active research area for over a decade. Early work focused on feature-level concatenation followed by supervised classification with support vector machines [6]. As deep learning matured, two-stream and three-stream convolutional networks became common, where each modality is processed in a separate branch before fusion at a late stage [7]. More recently, transformer-based architectures have shown promise in capturing long-range dependencies across spectral and spatial dimensions. For instance, the work of Hong et al. introduced a spectral-spatial transformer for hyperspectral image classification [8], while other researchers extended such designs to LiDAR elevation maps.

Attention mechanisms are particularly suited for multi-modal fusion because they allow the model to dynamically weight the importance of different modalities and regions. Single-scale attention, however, tends to miss contextual cues that appear only at coarser or finer resolutions [9]. Multi-scale attention networks have been explored in computer vision for tasks such as semantic segmentation, where pyramid pooling and atrous convolution capture multi-scale context [10]. In remote sensing, similar ideas have been applied to single-modality problems, but their adaptation to the joint fusion of hyperspectral, LiDAR, and camera data is less mature.

Yang et al. systematically evaluated the impact of band ordering strategies in hyperspectral and LiDAR fusion networks, demonstrating that the arrangement of spectral bands before fusion can significantly affect classification accuracy and training convergence [17]. This finding underscores the importance of accounting for spectral structure during the design of multi-scale attention. Meanwhile, Xiong et al. proposed a physics-coherent approach for aligning image and video representations with 3D geometry, a concept that can inform the geometric alignment needed when fusing LiDAR point clouds with camera imagery [5]. Although that work targets video generation, the underlying principle of enforcing spatial consistency across modalities is directly relevant.

At the system level, the smart city community has called attention to the need for scalable, interoperable sensor infrastructures that can handle the variety, velocity, and volume of urban data [11]. Privacy concerns are paramount when cameras capture identifiable faces or license plates, and when hyperspectral data can infer material properties of private property [12]. Fairness in sensor placement, often studied under the rubric of urban data justice, highlights that underserved neighbourhoods tend to have fewer sensors, leading to algorithmic biases in city services [13]. These sociotechnical considerations are rarely integrated into the design of fusion models but are essential for responsible deployment.

3. Hierarchical Multi-Scale Attention Architecture

The proposed HMSA-Net is composed of three main components: modality-specific encoders, a multi-scale cross-attention fusion module, and a hierarchical aggregation head. Each encoder is tailored to the nature of its input. For hyperspectral data, a spectral-spatial residual network extracts features while preserving the full spectral dimension up to a predefined bottleneck. The LiDAR encoder uses a point cloud feature extraction backbone such as PointNet++ to generate a set of per-point features that are then rasterised into a regular grid

aligned with the camera image [14]. The camera encoder is a standard convolutional neural network (e.g., ResNet-50) that produces a dense feature map at quarter resolution.

The multi-scale cross-attention module operates on three spatial scales: fine (full resolution), medium (half resolution), and coarse (quarter resolution). At each scale, feature maps from the three modalities are projected into a common embedding space using separate linear transformations. A cross-attention mechanism then computes pairwise attention between each modality pair – for instance, the hyperspectral features attend to LiDAR features and vice versa – using the scaled dot-product formulation. This yields three sets of attended features per scale, which are aggregated via element-wise addition and then passed through a shared feed-forward network. The rationale for processing multiple scales is that urban objects such as buildings, roads, and vegetation exhibit characteristic extents that are best captured at particular resolutions. For example, fine-scale attention helps discriminate between different roof materials using hyperspectral signatures, while coarse-scale attention captures the spatial layout of entire blocks using LiDAR elevation profiles.

After obtaining scale-specific fused representations, the hierarchical aggregation head combines them using a top-down pathway that enriches coarse features with details from finer scales. This design is inspired by the feature pyramid networks used in object detection [15]. Each scale's features are upsampled and added to the next finer scale, followed by a 3x3 convolution to reduce aliasing. The final output is a multi-scale fused feature map that can be fed into task-specific heads, such as a pixel-level classifier for land cover or a bounding box predictor for building detection.

One important design choice is the number of attention heads and the depth of the cross-attention layers. Fewer heads reduce computational cost but may limit the model's ability to capture diverse relational patterns. In the HMSA-Net, eight heads are used based on a trade-off between performance and memory, as validated on a medium-sized urban dataset. The entire network is trained end-to-end using a combination of cross-entropy loss for classification tasks and smooth L1 loss for regression tasks. Data augmentation, including random spectral shifts and geometric jitter, is applied to improve robustness.

4. Structural Trade-offs and System-Level Considerations

Deploying a multi-modal fusion network in a smart city context involves navigating several structural trade-offs. The first is between predictive accuracy and computational latency. The HMSA-Net, with its three encoders and multi-scale cross-attention, requires significant GPU memory and inference time. In an edge deployment—for example, a drone-mounted sensor platform—real-time processing may be impossible. A common mitigation is to compress the model using knowledge distillation or quantisation, but this often degrades performance on rare spectral classes [16]. The trade-off may be resolved by adopting a hybrid architecture where simple scenes are processed quickly by a lightweight model, while complex scenes trigger the full hierarchical network.

A second trade-off concerns interpretability versus representational power. Attention maps provide a form of explainability by highlighting which spatial and spectral regions contribute most to a decision. However, the hierarchical aggregation obscures the flow of information across scales, making it difficult to trace why a particular classification was made. For regulatory compliance in applications such as environmental enforcement, interpretability may be legally required. One solution is to attach separate attention visualisation tools at each scale, producing a saliency map per level that can be inspected by domain experts.

Third, there is a tension between centralised cloud processing and distributed edge computing. Centralisation allows for larger models and easier model updates but introduces latency and dependency on communication networks. Distributed processing reduces bandwidth and preserves local data privacy, but devices may have limited power and storage. The HMSA-Net can be partitioned: edge nodes run the modality-specific encoders and send compressed feature embeddings to a cloud server for cross-attention fusion. This reduces raw data transfer—particularly important for hyperspectral data, which can be hundreds of megabytes per scene—while still leveraging the cloud's computational capacity.

Data heterogeneity further complicates system design. Sensors from different manufacturers may produce data with varying signal-to-noise ratios, spatial resolution, and calibration consistency. The proposed architecture includes a normalisation layer before each encoder that learns per-sensor statistics during training. In deployment, if a new LiDAR model is introduced, fine-tuning is required to avoid domain shift. A governance framework that mandates periodic sensor calibration and retraining cycles would mitigate these risks.

5. Deployment and Sustainability

The long-term sustainability of a multi-modal sensor network depends on energy consumption, hardware longevity, and the ability to adapt to changing urban landscapes. The HMSA-Net, if deployed continuously on edge devices, would drain batteries quickly. To address this, a duty-cycling mechanism can be implemented: the network runs only when triggered by a motion sensor or periodically at low temporal resolution (e.g., once per hour) for background monitoring. Additionally, model pruning and weight sharing across scales can reduce the number of parameters without a significant drop in accuracy. Studies have shown that pruning up to forty percent of attention heads in transformer models can maintain over ninety-five percent of baseline performance [18].

Robustness to sensor failure is another critical aspect. In a real-world deployment, a camera may be blocked by construction scaffolding, or a LiDAR unit may fail due to weather. The HMSA-Net can be made fault-tolerant by training with random modality dropout—that is, during training, one or two modalities are randomly removed, forcing the network to rely on the remaining ones. This strategy has been shown to improve performance when modalities are missing at test time [19]. For a smart city, this means that even if a subset of sensors is offline, the fusion system can still produce reasonable outputs, albeit with higher uncertainty flagged to operators.

Environmental impact extends beyond energy. Manufacturing hyperspectral sensors involves rare-earth elements, and the disposal of electronic waste raises concerns. A life-cycle assessment should inform sensor procurement decisions, favouring modular hardware that can be upgraded rather than replaced. The proposed architecture's software-based approach also allows for over-the-air updates, reducing the need for physical interventions.

6. Fairness and Policy Implications

The deployment of advanced sensing in smart cities amplifies existing inequalities if not carefully governed. Sensors are often concentrated in affluent or commercially important areas, leading to disparities in algorithmic coverage. For example, a building monitoring system trained predominantly on downtown high-rises may perform poorly on low-income residential blocks where construction materials differ. The HMSA-Net, through its multi-scale design, may partially mitigate this by learning both local and regional patterns, but it cannot compensate for systematic data scarcity in certain neighbourhoods. To address this, city

agencies should mandate equitable sensor placement based on population density and socioeconomic factors, not just economic value [20].

Data privacy is a foremost concern. Hyperspectral data can reveal material compositions—for instance, the type of paint on a building or the health of a private garden—which some residents may consider sensitive. LiDAR point clouds capture the three-dimensional exterior of homes, potentially enabling inferences about property value or occupancy. Camera images directly record human presence. The HMSA-Net, as an architecture, does not inherently protect privacy, but it can be augmented with on-sensor anonymisation: for example, computationally weak cameras can output only blurred or de-identified images. Alternatively, the network can be trained to operate on encrypted features using secure multi-party computation, though this incurs overhead [21].

Policy frameworks must clarify data ownership and usage rights. Who owns the fused data product—the city, the sensor manufacturer, or the residents? European data protection regulations such as GDPR provide a foundation, but deploying such systems globally requires adaptation to local laws. The United Nations Sustainable Development Goals, particularly Goal 11 on sustainable cities and communities, advocate for inclusive and resilient urbanisation. The technological choices embedded in the HMSA-Net should align with these principles: for example, ensuring that the model's classification outputs are explainable and contestable by affected citizens.

Furthermore, algorithmic fairness should be evaluated using disaggregated metrics across demographic groups. A land-cover classifier that misclassifies low-income neighbourhoods as wasteland could lead to underinvestment in green infrastructure. Training datasets must be balanced and representative; if not, debiasing techniques such as reweighting or adversarial learning can be applied [22]. The hierarchical nature of the network can assist in fairness auditing: per-scale attention maps can reveal whether certain groups are systematically ignored at particular scales.

7. Conclusion

This paper has presented a hierarchical multi-scale attention network for integrating hyperspectral, LiDAR, and camera data in smart city applications. The architecture addresses the fundamental challenge of fusing heterogeneous data at multiple spatial and spectral scales through dedicated encoders, cross-attention modules, and a hierarchical aggregation head. Beyond technical design, the paper has emphasised the structural trade-offs and system-level considerations that are often overlooked in algorithm-centric research. Computational latency, interpretability, centralised versus distributed processing, sensor heterogeneity, and fault tolerance were analysed in the context of real-world deployment. Sustainability concerns, including energy consumption and hardware life-cycle, were discussed alongside fairness and policy implications. The integration of equitable sensor placement, data privacy, and governance frameworks was argued to be as crucial as predictive accuracy for the responsible implementation of smart city sensing infrastructures. Future work should explore dynamic scale selection that adapts to scene complexity, as well as federated learning approaches that preserve privacy while enabling collaborative model improvement [23]. The HMSA-Net provides a flexible foundation for such extensions, contributing both a practical architecture and a sociotechnical perspective necessary for the next generation of urban sensing systems.

References

1. Bioucas-Dias, J. M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., & Chanussot, J. (2013). Hyperspectral remote sensing data analysis and future challenges. IEEE Geoscience and Remote Sensing Magazine, 1(2), 6–36.

2. Shan, J., & Toth, C. K. (2008). Topographic laser ranging and scanning: Principles and processing. CRC Press.

3. Ghamisi, P., Maggiori, E., Li, S., Souza, R., Tarabalka, Y., Moser, G., & Benediktsson, J. A. (2019). New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology and Markov random fields. IEEE Geoscience and Remote Sensing Magazine, 7(1), 10–43.

4. Hu, J., Shi, J., Zhao, Q., & Li, X. (2022). Multi-source remote sensing data fusion: A review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 10140–10158.

5. Xiong, Z., Song, Y., He, L., Xiong, W., Yuan, Y., Qiao, F., & Jacobs, N. (2026). PhysAlign: Physics-Coherent Image-to-Video Generation through Feature and 3D Representation Alignment. arXiv preprint arXiv:2603.13770.

6. Dalponte, M., Bruzzone, L., & Gianelle, D. (2008). Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. IEEE Transactions on Geoscience and Remote Sensing, 46(5), 1416–1427.

7. Zhao, W., Du, L., & Zhang, B. (2021). Deep learning for hyperspectral and LiDAR data fusion: A review. IEEE Geoscience and Remote Sensing Letters, 18(3), 429–433.

8. Hong, D., Gao, L., Yao, J., Zhang, B., Plaza, A., & Chanussot, J. (2021). Spectral-spatial transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 59(12), 10350–10363.

9. Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794–7803.

10. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2881–2890.

11. Kyriazis, D., Varvarigou, T., & Romanovs, A. (2016). Smart cities: A survey on technologies, trends and open issues. IEEE Communications Surveys & Tutorials, 18(4), 2676–2712.

12. Cavoukian, A., & Dix, A. (2013). Privacy in smart cities: A Canadian perspective. Journal of Law, Information and Science, 22(1), 1–20.

13. Crawford, K., & Joler, V. (2018). Anatomy of an AI system: The Amazon Echo as a case study of technological entanglement. AI Now Institute.

14. Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems, 30, 5099–5108.

15. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117–2125.

16. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

17. 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.

18. Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? Advances in Neural Information Processing Systems, 32, 14014–14024.

19. Ma, M., Fan, J., & Tian, Q. (2020). Modality dropout for robust multimodal learning. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 11949–11956.

20. Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. Science, 332(6025), 60–65.

21. Bonawitz, K., et al. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the ACM Conference on Computer and Communications Security, 1175–1191.

22. Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5), 82–89.

23. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.