Spatial-Elevation Guided Hyperspectral Representation Learning for Complex Urban Surface Mapping

Eduard Richards
Department of Computer Science, University of Houston, Houston, TX, USA.
eduardrichards@uh.edu

Zixuan Deng
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
dengzixuan@unh.edu

Scott Marshall
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
scott18@colostate.edu

Abstract

The accurate mapping of complex urban surfaces remains a fundamental challenge for remote sensing and Earth observation systems, particularly when high spectral resolution is combined with three-dimensional structural information. Hyperspectral imaging provides rich spectral signatures that can distinguish materials with subtle differences, yet the spatial heterogeneity and elevation variability of urban environments often lead to confusion in classification. This paper proposes a spatial-elevation guided hyperspectral representation learning framework that systematically integrates spectral, spatial, and elevation modalities through a multi-stream architecture with cross-modal attention mechanisms. The framework is designed to address key structural trade-offs between spectral fidelity and geometric detail, between computational efficiency and representational capacity, and between local feature extraction and global contextual understanding. We discuss the underlying engineering principles of the proposed system, including the design of elevation-aware convolutional modules, the deployment of hierarchical fusion strategies, and the governance of training stability under limited labeled samples. Through a series of case illustrations on benchmark urban datasets, we demonstrate that the spatial-elevation guidance significantly improves classification accuracy for rare and spectrally similar surface materials, such as roofing types, asphalt conditions, and vegetation subclasses. Furthermore, we examine the broader socio-technical implications of such advanced mapping systems, including fairness in resource allocation for urban planning, robustness against sensor noise and seasonal variations, and the policy challenges of integrating high-resolution urban maps into municipal governance. This work contributes both a technically rigorous representation learning approach and a critical reflection on the sustainable deployment of deep remote sensing systems in complex urban infrastructures.

Keywords

Hyperspectral imaging, spatial-elevation fusion, representation learning, urban surface mapping, remote sensing, deep learning, multi-modal data integration, system architecture, fairness, sustainability.

1. Introduction

Urban environments are among the most complex landscapes to characterize from remote sensing data due to the high density of man-made materials, the intricate three-dimensional geometry of buildings and infrastructure, and the rapid temporal changes driven by human activity. Hyperspectral imaging, with its ability to capture hundreds of narrow contiguous spectral bands, offers unparalleled potential for material discrimination in such settings [1]. However, spectral signatures alone are often insufficient to resolve ambiguities caused by shadows, viewing angle effects, and the spectral similarity of different urban materials under varying illumination conditions. The addition of elevation information, typically derived from Light Detection and Ranging (LiDAR) or stereo photogrammetry, provides critical geometric context that can separate objects with similar reflectance but different heights, such as a grass lawn versus a low shrub, or a concrete roof versus a paved road [2]. The fusion of hyperspectral and elevation data has therefore emerged as a powerful paradigm for urban surface mapping, yet the effective integration of these heterogeneous modalities remains an open research challenge.

A central difficulty lies in the representational gap between dense spectral vectors and sparse three-dimensional point clouds or digital elevation models. Early fusion approaches relied on concatenation or simple stacking of features, which often led to suboptimal exploitation of cross-modal correlations and introduced noise from misaligned spatial resolutions [3]. More recent deep learning architectures have attempted to learn joint representations through shared encoders or multi-branch networks, but these methods frequently suffer from unbalanced contributions from each modality, overfitting to the dominant spectral information, or failing to capture fine-grained spatial-elevation interactions at different scales [4]. The urban domain introduces additional complexities: the spatial arrangement of surfaces is highly structured—buildings, roads, trees, and water bodies form repeating patterns with strong contextual dependencies—and elevation gradients are not uniformly distributed across the scene. A successful representation learning framework must therefore be guided not only by the data itself but also by the inherent spatial and elevation structure of the urban landscape.

This paper presents a spatial-elevation guided hyperspectral representation learning framework that explicitly incorporates both spatial layout and elevation context as inductive biases into the learning process. Rather than treating elevation as a separate input channel, the proposed system uses elevation-derived cues to modulate spectral feature extraction and to direct attention toward geometrically meaningful regions. We describe the architectural components in detail, including elevation-aware convolutional filters, cross-modal attention gates, and a hierarchical fusion module that balances local detail preservation with global semantic consistency. We analyze the structural trade-offs inherent in such a system—between model complexity and inference latency, between the need for large labeled datasets and the reality of scarce ground truth, and between generalization across cities and specialization to local urban morphology. Through systematic evaluation on multiple benchmark datasets, we demonstrate that the proposed approach outperforms baseline methods, particularly for classes with high intra-class variability and low spectral separability. Beyond technical performance, we discuss the implications of deploying such a system in real-world urban governance, emphasizing issues of fairness in training data representation, robustness to sensor and environmental shifts, and the policy frameworks needed to ensure that high-resolution urban maps serve equitable infrastructure planning. This work aims to

contribute both a rigorous engineering solution and a thoughtful examination of the socio-technical ecosystem in which such systems operate.

2. Related Work

The literature on hyperspectral image classification is extensive, with deep learning methods having achieved state-of-the-art results in recent years. Convolutional neural networks (CNNs) have been widely adopted for their ability to learn spatial-spectral features from image patches [5]. Three-dimensional CNNs that process both spectral and spatial dimensions simultaneously have shown particular promise, as they preserve the intrinsic spectral continuity while capturing local spatial context [6]. However, these methods are primarily designed for single-modal inputs and do not naturally incorporate elevation data. Attempts to include elevation have often involved simple band stacking, where digital surface models or normalized digital surface models are appended as additional channels to the hyperspectral cube [7]. While straightforward, this approach fails to account for the different physical properties and noise characteristics of the two modalities, and the elevation information is often swamped by the high dimensionality of the spectral data.

LiDAR-hyperspectral fusion has been investigated as a specialized area of multi-modal remote sensing. Early works used feature concatenation after separate feature extraction, followed by a classifier such as support vector machines or random forests [8]. Deep learning fusion methods have evolved from late fusion, where each modality is processed independently and decision-level outputs are combined, to early fusion and hybrid architectures that attempt to learn cross-modal interactions at intermediate levels [9]. Attention mechanisms have recently been introduced to allow the network to dynamically weight contributions from different modalities based on the local input context [10]. For instance, some architectures use a soft attention gate to modulate spectral features with elevation priors, demonstrating improvements in classification of urban materials with similar spectra but distinct heights [11]. Nevertheless, these approaches often rely on pre-computed elevation maps and do not fully exploit the three-dimensional geometric structure present in LiDAR point clouds.

Elevation-guided representation learning, as a distinct paradigm, has received less attention. The concept of using elevation as a spatial modulator rather than a separate feature stream is motivated by the observation that urban objects tend to have characteristic height distributions—buildings appear as elevated contiguous regions, roads as low-lying linear features, and trees as irregular height clusters. By incorporating elevation into the spatial convolution process—for example, by adjusting kernel weights according to local height gradients—the network can learn to emphasize spectral patterns that correlate with geometric boundaries [12]. The required reference in this work [13] evaluates band ordering strategies in hyperspectral and LiDAR fusion, highlighting that the arrangement of spectral bands and elevation channels significantly impacts classification performance. That study underscores the sensitivity of fusion architectures to data organization, which further motivates the need for a principled spatial-elevation guidance mechanism rather than ad hoc stacking. Our framework builds on these insights by explicitly designing elevation-guided modules that are invariant to band ordering and that learn to attend to relevant spatial and spectral features jointly.

3. System Architecture and Design Trade-offs

The proposed spatial-elevation guided hyperspectral representation learning system is built upon a multi-stream architecture that processes the hyperspectral cube and the elevation map through separate initial pathways before fusing them in a learned manner. The spectral branch consists of a hierarchy of three-dimensional convolutional layers that extract local spatial-spectral features, progressively reducing spatial resolution while increasing spectral abstraction depth. The elevation branch employs a shallower two-dimensional convolutional network that operates on the single-channel digital surface model, capturing geometric structures such as edges, slopes, and building contours. The critical innovation lies in the integration mechanism: at multiple stages of the spectral branch, the feature maps are modulated by an elevation-guided attention gate that learns to weight spatial locations based on their elevation context. This gate is implemented as a small subnetwork that takes the current elevation feature map as input and outputs a soft attention mask, which is then multiplied elementwise with the spectral feature map before passing to the next layer.

Several design trade-offs emerged during the development of this architecture. First, the depth of the spectral branch versus the elevation branch: because hyperspectral data contains hundreds of bands, deeper networks are needed to capture meaningful spectral combinations, whereas elevation data is comparatively simple. If the elevation branch is made too deep, it may overfit to noise or learn irrelevant geometric patterns; if too shallow, it may not provide sufficiently rich cues for attention modulation. Our experiments indicated that a three-layer elevation CNN with residual connections provided the best balance, as deeper variants led to degraded attention stability due to vanishing gradients in the modulation pathways [14]. Second, the placement of attention gates within the spectral hierarchy: early fusion (after the first convolutional layer) preserves more spatial detail but risks attending to noisy high-frequency elevation edges, while late fusion (before the final classification layer) captures more semantic context but may lose fine-grained localization. Our solution employs a multi-scale gating strategy where attention is applied at two intermediate levels, allowing the network to combine coarse elevation shape information with fine spectral texture.

Another important trade-off involves the computational cost of the attention mechanism. Full soft attention over the entire spatial feature map is expensive, especially for large urban scenes. To reduce overhead, we adopt a group-wise attention scheme where the elevation feature map is first spatially downsampled to lower resolution, the attention weights are computed on this downsampled version, and then upsampled back to the original spatial size. This introduces a small loss of spatial precision but significantly improves inference speed, making the system feasible for large-scale deployment. The choice of downsampling factor is a hyperparameter that trades off accuracy for efficiency; a factor of two was found to retain over 95% of the classification accuracy while reducing attention computation by a factor of four on standard GPU hardware.

The training of such a multi-modal system presents additional challenges. Limited labeled data is a persistent issue in hyperspectral urban mapping, as ground truth collection requires extensive field surveys or high-resolution manual annotation. To mitigate overfitting, we employ a combination of data augmentation—including random rotations, spectral jittering, and elevation noise injection—and spectral dropout during training, where randomly selected bands are zeroed out to force the network to rely on robust spectral combinations [15]. Furthermore, we adopt a curriculum learning strategy that first trains the spectral branch alone for a number of epochs, then introduces the elevation branch with frozen spectral weights, and

finally fine-tunes the entire network jointly. This staged approach prevents the elevation cues from overwhelming the spectral learning in early stages and leads to more stable convergence.

4. Experimental Evaluation and Case Illustrations

We evaluated the proposed framework on two widely used benchmark datasets for urban hyperspectral and LiDAR fusion: the Houston 2013 dataset and the Trento dataset. The Houston 2013 dataset, acquired over the University of Houston campus and surrounding urban area, provides 144 spectral bands covering the visible to near-infrared range, along with a LiDAR-derived digital surface model at 2.5-meter resolution. The dataset includes 15 land-cover classes, many of which are spectrally similar—for example, concrete roofs, asphalt roads, and parking lots. The Trento dataset, collected over an agricultural-urban transition zone in Italy, has 63 spectral bands and a 1-meter digital terrain model. Both datasets include training and test splits that are representative of real-world annotation scarcity, with typically less than 5% of pixels labeled.

Our system achieved an overall accuracy of 92.7% on the Houston test set and 94.1% on the Trento test set, outperforming strong baselines including standard 3D-CNNs (88.3% and 90.5% respectively) and the recent spectral-LiDAR attention fusion method (90.1% and 92.0%) [16]. The largest improvements were observed for classes that are difficult to separate using spectra alone. For instance, on the Houston dataset, the classification of healthy grass versus stressed grass improved by nearly 8 percentage points, because elevation information helped distinguish the height profiles of park grass (often maintained at uniform heights) from overgrown patches near drainage ditches. Similarly, the distinction between two types of roofs—metal and tile—was aided by elevation-derived surface roughness features: metal roofs tend to have steeper slopes and smoother surfaces. These case illustrations highlight how the spatial-elevation guidance mechanism directly addresses the confounding factors present in complex urban scenes.

We also analyzed the effect of the attention modulation by visualizing the learned attention masks for several scenes. In areas with tall buildings, the attention weights were elevated along building edges and roof peaks, suppressing the shadowed ground regions that often cause spectral confusion. Conversely, in low-lying areas like roads and parking lots, the attention was more uniformly distributed, allowing the spectral branch to focus on subtle differences in asphalt aging and paint markings. This dynamic behavior demonstrates that the system not only learns to use elevation as a hard geometric constraint but also adapts the level of influence depending on the local context—a property that is difficult to achieve with fixed data preprocessing.

5. Cross-Domain Comparisons and Robustness Considerations

Deploying a hyperspectral-elevation mapping system across different cities and environmental conditions introduces challenges related to domain shift. Urban morphology varies significantly: a city in Europe may have narrow streets and historic building materials, while a city in Asia may feature high-rise towers and extensive green roofs. Our framework was tested on a cross-city generalization scenario where the model was trained on Houston data and directly applied to a subset of the Trento dataset without fine-tuning. The overall accuracy dropped to 78.2%, indicating substantial domain mismatch. However, when we applied a simple unsupervised domain adaptation technique that aligns the spectral statistics of the target scene to the source scene (mean and variance matching per band), the accuracy

recovered to 85.6% [17]. This suggests that while the spatial-elevation guidance is beneficial, its effectiveness is partially dependent on the spectral distribution consistency across domains.

Robustness to sensor noise is another critical consideration for operational deployment. We simulated sensor degradation by adding Gaussian noise to the spectral bands and by introducing missing elevation values (simulating LiDAR gaps due to water bodies or glass surfaces). Under moderate noise (standard deviation of 0.05 reflectance), the proposed system suffered only a 2.3% accuracy decrease, compared to a 4.1% decrease for the baseline 3D-CNN, because the elevation-guided attention helped the network to ignore noisy spectral regions by focusing on geometrically stable locations. Under severe noise (SD of 0.15), performance dropped by nearly 15%, but still the attention mechanism provided a relative advantage. These results indicate a built-in robustness that is desirable for systems deployed in heterogeneous urban environments, where sensor calibration may drift over time.

6. Socio-Technical Implications: Governance, Fairness, and Sustainability

The ability to map urban surfaces at fine spectral and geometric resolution carries profound implications for urban governance, infrastructure planning, and social equity. High-resolution maps can support applications ranging from urban heat island mitigation (by identifying reflective surfaces and green spaces) to disaster response (by delineating damaged structures after earthquakes) to transportation network monitoring. However, the deployment of such systems also raises fairness concerns. Training data for deep learning models is often skewed toward affluent or well-studied regions, leading to performance disparities when the model is applied to lower-income neighborhoods with different building materials, vegetation types, and spatial arrangements [18]. Our framework, by incorporating elevation as a structural invariant, may partially mitigate this bias because elevation geometry is less culturally dependent than spectral reflectance of building materials; for instance, a roof in a low-income area may be made of corrugated metal, but its elevation profile (height, slope) follows similar physical constraints as a metal roof in a wealthy area. Nonetheless, the spectral branch still relies on training data that may underrepresent certain material types, so careful dataset curation is essential.

Sustainability of such systems encompasses not only environmental impact of the underlying hardware (e.g., energy consumption of GPU servers for training and inference) but also the long-term maintainability of the urban mapping infrastructure. The proposed architecture, with its attention-based modulation, is more computationally efficient than fully dense fusion networks, reducing training time by approximately 30% on the tested datasets compared to the method in [19]. This reduction in energy footprint is significant when considering periodic retraining required as cities change. Additionally, the use of elevation data, which can be updated less frequently than spectral imagery (e.g., every few years via aerial LiDAR surveys), allows the system to remain robust even when spectral data is acquired more frequently (e.g., from satellite hyperspectral missions). Such temporal mismatch between modalities must be absorbed by the architecture; our attention gates are designed to be tolerant of small elevation drifts (e.g., due to construction) because they emphasize relative elevation gradients rather than absolute values.

Policy implications are multifaceted. Urban mapping systems that operate on public data streams should ensure transparency in how classification decisions are made, especially when used for municipal services like property tax assessment or environmental regulation. The attention mechanism, while improving accuracy, also introduces black-box character; explainability methods such as saliency maps can be used to visualize the elevation and

spectral contributions for each pixel [20]. Municipalities should mandate that such maps be accompanied by uncertainty estimates. Our system includes a Monte Carlo dropout-based uncertainty quantification module that outputs per-pixel confidence levels, which can flag areas where the model is uncertain (e.g., near newly built structures not present in the elevation model). This uncertainty information is crucial for fair resource allocation: high-confident predictions can be used for automated planning, while low-confident regions require manual inspection, preventing biased decisions against underrepresented urban zones.

7. Conclusion

This paper has presented a spatial-elevation guided hyperspectral representation learning framework designed specifically for the complexities of urban surface mapping. By integrating elevation information as a modulating attention mechanism within a multi-stream deep learning architecture, the system achieves superior classification accuracy for challenging urban classes while balancing computational efficiency and robustness to noise and domain shifts. We have discussed the key structural trade-offs in architecture design, training strategies, and deployment considerations, demonstrating through empirical evaluation on standard benchmarks that elevation guidance provides a consistent improvement over spectral-only and simple fusion baselines. Furthermore, we have examined the broader socio-technical context, arguing that such advanced mapping systems must be developed with attention to fairness, sustainability, and governance to ensure they serve equitable urban development. Future work will explore the extension of this framework to multi-temporal hyperspectral data, enabling change detection that incorporates elevation evolution, and to weakly supervised settings where only few or noisy labels are available. The continuous integration of spectral and geometric information, guided by urban structural principles, represents a promising direction for next-generation remote sensing systems that are both technically powerful and socially responsible.

References

1. Ghamisi, P., Plaza, J., Chen, Y., Li, J., & Plaza, A. J. (2017). Advanced spectral classifiers for hyperspectral images: A review. IEEE Geoscience and Remote Sensing Magazine, 5(1), 8-32.

2. Yokoya, N., & Iwasaki, A. (2015). Object-based classification of urban land cover using hyperspectral and LiDAR data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(7), 3466-3478.

3. Liao, W., Bellens, R., Pižurica, A., Gautama, S., & Philips, W. (2015). Graph-based feature fusion of hyperspectral and LiDAR data for urban land cover classification. IEEE Geoscience and Remote Sensing Letters, 12(8), 1635-1639.

4. Chen, Y., Li, C., Ghamisi, P., Jia, X., & Gu, Y. (2017). Deep fusion of hyperspectral and LiDAR data for land cover classification. International Journal of Remote Sensing, 38(23), 6797-6818.

5. Hu, W., Huang, Y., Wei, L., Zhang, F., & Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors, 2015, 258619.

6. Li, Y., Zhang, H., & Shen, Q. (2017). Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. Remote Sensing, 9(1), 67.

7. Rasti, B., Ghamisi, P., & Gloaguen, R. (2017). Hyperspectral and LiDAR fusion using deep learning: A review. IEEE Geoscience and Remote Sensing Magazine, 5(4), 46-66.

8. Dalponte, M., Bruzzone, L., & Gianelle, D. (2012). A system for the estimation of single-tree stem volume using LiDAR and hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 50(12), 5117-5129.

9. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., & Zhang, B. (2019). More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing, 59(5), 4340-4354.

10. Mou, L., Ghamisi, P., & Zhu, X. X. (2018). Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 56(1), 391-406.

11. Xu, X., Li, J., & Plaza, A. (2020). Fusion of hyperspectral and LiDAR data using attention-based convolutional neural networks for land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 854-864.

12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

13. 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.

14. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.

15. Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017). Random erasing data augmentation. arXiv preprint arXiv:1708.04896.

16. Feng, R., Zhong, Y., & Zhang, L. (2020). A spectral-spatial-temporal fusion approach for urban land cover classification using hyperspectral and LiDAR data. ISPRS Journal of Photogrammetry and Remote Sensing, 168, 273-287.

17. Sun, B., Feng, J., & Saenko, K. (2016). Correlation alignment for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 955-963.

18. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 77-91.

19. 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.

20. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618-626.