

Multimodal Fusion of Hyperspectral and LiDAR Data Using Attention Mechanisms for Land Cover Mapping

Gitanjali Pilankar*¹, Dharmpal Doye²

¹Research Scholar, SGGS Institute of Engineering and Technology, Nanded, Maharashtra 431606-INDIA.

²Professor, SGGS Institute of Engineering and Technology, Nanded, Maharashtra 431606-INDIA

*Corresponding Author: Gitanjali Pilankar

Abstract:

The proper classification of land-cover in the remote sensing needs to be associated with the effective fusion of complementary spectral and structural data. In this paper, the idea of Attention-Guided Fusion Network (GAFNet) is presented to integrate hyperspectral image (HSI) with LiDAR data by reweighting modality-specific features based on a spatial attention mechanism to accentuate the informative part of the image and reduce the background noise. The framework was tested on two benchmark datasets namely Trento and Houston which depict semi-urban and complex urban settings, respectively. Experimental evidence shows that GAFNet is always superior to state-of-the-art fusion methods, being able to obtain the Overall Accuracy of 96.75% and Kappa of 0.9564 on Trento, and the Average accuracy of 89.50%, Overall Accuracy of 91.08% and Kappa of 0.8861 on Houston. Quantitative measures and qualitative analyses have shown that the underlying nature of attention-guided multimodal feature fusion is robust, has the ability to generalize, and still performs well in remote sensing classification.

Introduction

Classification of land cover is very important in remote sensing applications like city planning, environmental monitoring and precision farming. Hyperspectral images contain detailed spectral data to allow material differentiation on a fine scale, but LiDAR data contains additional elevation and structural information. Nonetheless, the use of one modality usually results in the sub-optimal performance especially on the complex city and nature settings.

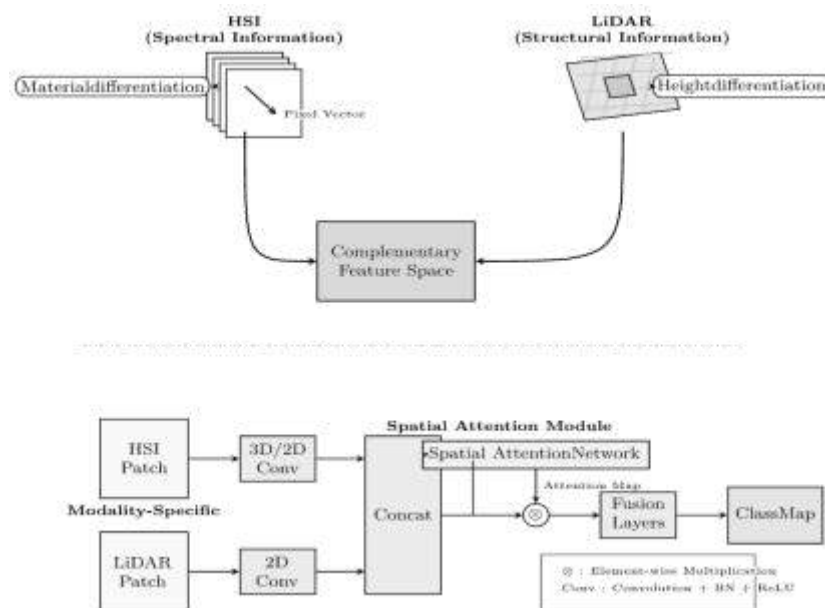
The classifications of land cover are an essential remote sensing activity that finds uses in urban planning, environmental observation, forestry, and precision agricultural activities. The spectral data of hyperspectral imagery (HSI) offers detailed spectral data, which can be used to distinguish materials finely, whereas LiDAR data depicts additional elevational and structural features of the environment. Initial findings have revealed that remotely sensing data of multisource can be able to enhance the classification accuracy [1], [2], [3] relative to single-modality methodology. The conventional methods of hyperspectral image analysis have been based on the statistical learning approach, a sparse representation approach and a Bayesian approach, which in most cases have poor generalization abilities on complex scenes [4], [5], [6]. In the same manner, tree-species along with land-cover classification based on handcrafted features are also sensitive to illumination, noise and spectral variations [7]. These drawbacks encouraged the use of deep learning models, especially the convolutional neural networks (CNNs), which have proven to be very effective in learning features of the hyperspectral data [8]. More recently, deep neural networks have been generalized to the multimodal fusion case where HSI and LiDAR data are mutually used to enhance the resilience of classification [9], [10]. It has also been further extended to spectral dependencies and contextual relationships in hyperspectral data with the introduction of recurrent neural networks and gated models, including GRUs [11]. Nevertheless, most current frameworks of fusion use spatial regions in a homogenous manner which can spread idle background knowledge and decrease the discrimination ability in heterogeneous

urban settings. Attention mechanisms have surfaced as one of the effective solutions to such problem since they allow networks to selectively highlight informative features, and noise is suppressed. Squeeze-and-excitation networks and convolutional block attention modules have demonstrated significant advances in the vision and remote sensing activities based on channel-wise and spatial attention models [12-16]. Also, non-local attention mechanisms have been suggested to ensure long-range dependencies within the feature maps are taken into consideration [17-21]. Although those improvements have been made, the proper implementation of the attention processes into hyperspectral-LiDAR fusion is still a research problem. Inspired by these observations, in this paper, we propose an Attention-Guided Fusion Network (GAFNet), which combines hyperspectral and LiDAR data, based on modality specific feature extraction, and spatial attention-based fusion. The proposed framework will adopt spatial significance by explicitly modelling feature integration, to improve the class separability and robustness in both semi-urban and intricate urban settings.

We suggest an attention-guided fusion network (GAFNet) in HSI-LiDAR patch classification in this work. The model operates on spectral and structural features by employing spatial attention module that provides boost to the discriminative areas. Patch-based learning is embraced because it is simple to employ in managing limited labeled samples and also to maintain local spatial contexts.

Methodology Overview:

Fig. 1 overall methodology



The proposed GAFNet consists of three main components: (i) modality-specific feature extraction for HSI and LiDAR data, (ii) feature fusion through element-wise integration, and (iii) spatial attention to refine fused representations. The final classification is performed using a fully connected decision layer.

Unlike pixel-wise segmentation approaches, the proposed method operates on fixed-size image patches. Each patch is assigned a single class label corresponding to its central pixel. This strategy is widely used in hyperspectral image classification and enables robust learning of local spectral-spatial patterns.

Dataset Description

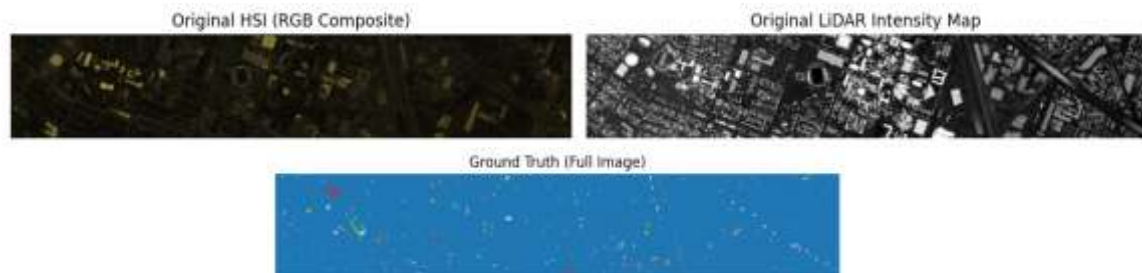
Houston Dataset

The Houston dataset is a popular land-cover classification and multimodal data fusion benchmark research dataset. It was published as the IEEE GRSS Data Fusion Contest and is composed of hyperspectral images taken by the University of Houston ITRES CASI-1500 sensor of the university and surrounding urban surroundings. The hyperspectral image has 144 spectral bands in the 3801050 nm with a spatial resolution measuring around 2.5 m. Besides the hyperspectral data, a co-registered LiDAR-based Digital Surface Model (DSM) is also available, which allows using spectral and elevation data together. The dataset consists of 15 land-cover classes, which represent complex urban classes, i.e., buildings, roads, vegetation, and parking lots, thus, being especially difficult because of high spectral similarities and spatial heterogeneity [10].

Figure 2. class description of Houston13 dataset.

No.	Class	Training	Test
1	Health Grass	198	1053
2	Stressed Grass	190	1064
3	Synthetic Grass	192	505
4	Trees	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking Lot 1	192	1041
13	Parking Lot 2	184	285
14	Tennis Court	181	247
15	Running Track	187	473

Figure 3. HIS, LiDAR and ground truth image of Houston13 dataset.



Trento Dataset

The Trento dataset is a reference multimodal remote sensing dataset which is used to test hyperspectral-LiDAR data fusion algorithms in semi-urban and natural settings. It is made up of hyperspectral imagery obtained by AISA Eagle sensor, it has 63 spectral bands within the visible and near-infrared space, and has a spatial resolution of about 1 m. The dataset is supplemented with the LiDAR-derived elevation model which gives a detailed information on the height of the terrain and objects. There are 6 land-cover classes in the scene such as buildings, roads, vegetation, and water bodies. The Trento

dataset is also suitable to evaluate the robustness and the ability to generalize deep learning-based fusion models because it has a fine spatial resolution and a relatively limited number of training samples [6].

Figure 4. Class description of Trento Dataset.

No.	Class	Training	Test
1	Apple Trees	129	3905
2	Building	125	2778
3	Ground	105	374
4	Woods	154	8969
5	Vineyard	184	10317
6	Roads	122	3052

Figure 5. HIS, LiDAR and ground truth image of Trento dataset.



Experimental Setup

Attention-Guided Fusion Network (GAFNet)

This section presents the proposed Attention-Guided Fusion Network (GAFNet). The model is designed to integrate hyperspectral imagery (HSI) and LiDAR data for land cover classification by exploiting their complementary spectral and structural properties through an attention-based fusion mechanism.

1. Problem Formulation

Let the hyperspectral image be represented as a three-dimensional tensor:

$$\mathbf{X}_{\text{HSI}} \in \mathbb{R}^{H \times W \times B}$$

where H and W denote the spatial dimensions and B represents the number of spectral bands.

Similarly, the LiDAR data is represented as:

$$\mathbf{X}_{\text{LiDAR}} \in \mathbb{R}^{H \times W \times 1}$$

The objective of land cover classification is to learn a mapping function $f(\cdot)$ such that:

$$f: (\mathbf{X}_{\text{HSI}}, \mathbf{X}_{\text{LiDAR}}) \rightarrow \mathbf{y}$$

where $\mathbf{y} \in \{1, 2, \dots, C\}$ is the class label and C denotes the total number of land cover classes.

In a patch-based framework, image patches centered at pixel location (i, j) are extracted:

$$\mathbf{P}_{\text{HSI}}^{(i,j)} \in \mathbb{R}^{p \times p \times B}, \quad \mathbf{P}_{\text{LiDAR}}^{(i,j)} \in \mathbb{R}^{p \times p \times 1}$$

where p denotes the patch size.

2. Modality-Specific Feature Extraction

Due to the heterogeneous nature of HSI and LiDAR data, separate feature extraction functions are employed for each modality.

The hyperspectral feature extractor is defined as:

$$\mathbf{F}_{\text{HSI}} = \Phi_{\text{HSI}}(\mathbf{P}_{\text{HSI}}; \boldsymbol{\theta}_{\text{HSI}})$$

where $\Phi_{\text{HSI}}(\cdot)$ denotes a convolutional neural network (CNN) parameterized by $\boldsymbol{\theta}_{\text{HSI}}$, producing spectral-spatial features.

Similarly, the LiDAR feature extractor is given by:

$$\mathbf{F}_{\text{LiDAR}} = \Phi_{\text{LiDAR}}(\mathbf{P}_{\text{LiDAR}}; \boldsymbol{\theta}_{\text{LiDAR}})$$

where $\Phi_{\text{LiDAR}}(\cdot)$ extracts structural and elevation-related features.

Both feature maps are projected into a common latent space with equal channel dimensionality to facilitate fusion.

3. Feature Fusion

The modality-specific features are fused at the feature level by channel-wise concatenation:

$$\mathbf{F}_{\text{concat}} = [\mathbf{F}_{\text{HSI}}, \mathbf{F}_{\text{LiDAR}}]$$

where $[\cdot, \cdot]$ denotes concatenation along the channel dimension. This operation preserves complementary information from both modalities while forming a unified representation.

4. Spatial Attention Mechanism

To adaptively emphasize informative spatial regions, a spatial attention mechanism is introduced. Let $\mathbf{F}_{\text{concat}} \in \mathbb{R}^{C \times p \times p}$ denote the concatenated feature map.

The attention map is computed as:

$$\mathbf{M}_{\text{att}} = \sigma(\mathbf{W}_{\text{att}} * \mathbf{F}_{\text{concat}} + \mathbf{b}_{\text{att}})$$

where:

- $*$ denotes convolution,
- $\mathbf{W}_{\text{att}}$ and $\mathbf{b}_{\text{att}}$ are learnable parameters,
- $\sigma(\cdot)$ is the sigmoid activation function.

The attention-guided feature map is then obtained through element-wise multiplication:

$$\mathbf{F}_{\text{att}} = \mathbf{M}_{\text{att}} \odot \mathbf{F}_{\text{concat}}$$

where $\odot$ represents the Hadamard (element-wise) product. This operation assigns higher weights to spatial locations that contribute more significantly to classification performance.

5. Residual Attention Integration

To preserve the original fused features and stabilize training, the attention output is combined with the input features using a residual connection:

$$\mathbf{F}_{\text{guided}} = \gamma \mathbf{F}_{\text{att}} + \mathbf{F}_{\text{concat}}$$

where γ is a learnable scalar that controls the influence of the attention mechanism.

6. Classification Layer

The attention-guided feature representation is flattened and passed through fully connected layers to obtain class logits:

$$\mathbf{z} = \Psi(\mathbf{F}_{\text{guided}})$$

where $\Psi(\cdot)$ denotes the classifier network.

The final predicted class probabilities are computed using the softmax function:

$$\hat{y}_c = \frac{e^{z_c}}{\sum_{k=1}^C e^{z_k}}, \quad c = 1, 2, \dots, C$$

7. Loss Function and Optimization

The model is trained using the categorical cross-entropy loss:

$$\mathcal{L} = - \sum_{c=1}^C y_c \log(\hat{y}_c)$$

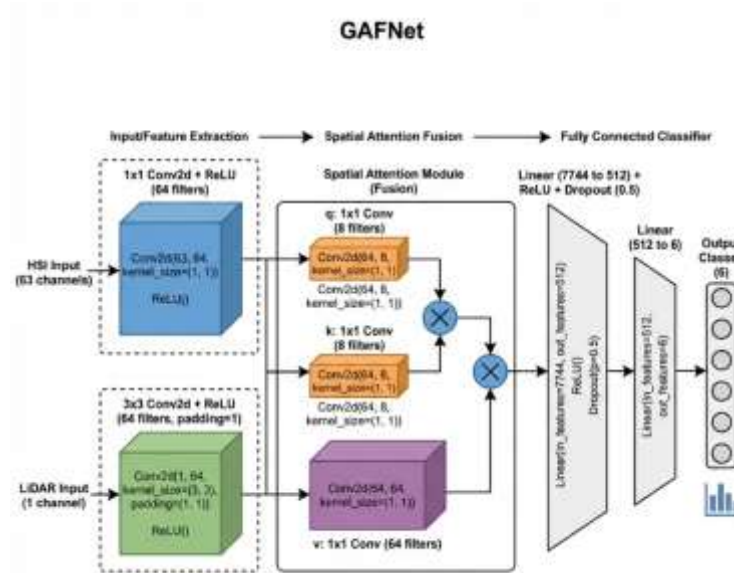
where y_c is the ground truth label.

Model parameters are optimized using the AdamW optimizer, which combines adaptive learning rates with weight decay for improved generalization.

8. Summary of the Framework

The proposed GAFNet formulates hyperspectral and LiDAR data fusion as a learnable nonlinear mapping enhanced by spatial attention. By integrating modality-specific feature extraction, feature-level fusion, attention-guided reweighting, and residual learning, the model effectively captures complementary spectral-spatial-structural information for land cover classification.

Fig. 6 GAFNet model architecture.



The experiments are performed on the Trento data where the hyperspectral and LiDAR data are co-registered. A patch-based sampling strategy is used to partition the dataset into the training and testing samples. Various baseline models, such as HSI-only CNN, LiDAR-only CNN, CNN-GRU fusion and SSAFNet, are applied to compare them.

The overall accuracy (OA), average accuracy (AA) and the Kappa coefficient are used to measure classification performance. Along with the quantitative measures, the qualitative visualizations are given to examine the behaviour of the suggested model.

Qualitative Results and Analysis on Trento data

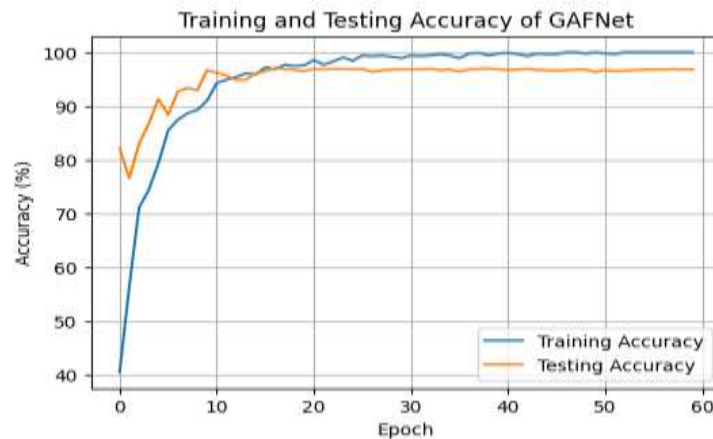
The effectiveness of the proposed Attention-Guided Fusion Network (GAFNet) was further validated on the Trento hyperspectral-LiDAR benchmark dataset, which represents a semi-urban environment with strong spectral-structural complementarity. GAFNet achieved an Overall Accuracy (OA) of 96.75%, an Average Accuracy (AA) of 89.60%, and a Kappa coefficient of 0.9564. The high Kappa value indicates an excellent level of agreement between predicted and reference labels, confirming the robustness

and reliability of the proposed fusion framework under well-structured land-cover conditions.

Training and Convergence Behavior

Figure 7 illustrates the training and testing accuracy curves of GAFNet on the Trento dataset. The model exhibits rapid convergence, with training accuracy approaching 100% within the first few epochs, while testing accuracy stabilizes above 96%. The small gap between training and testing curves indicates minimal overfitting and highlights the strong generalization capability of the proposed network.

Fig 7. Train-test curve produced on Trento dataset.

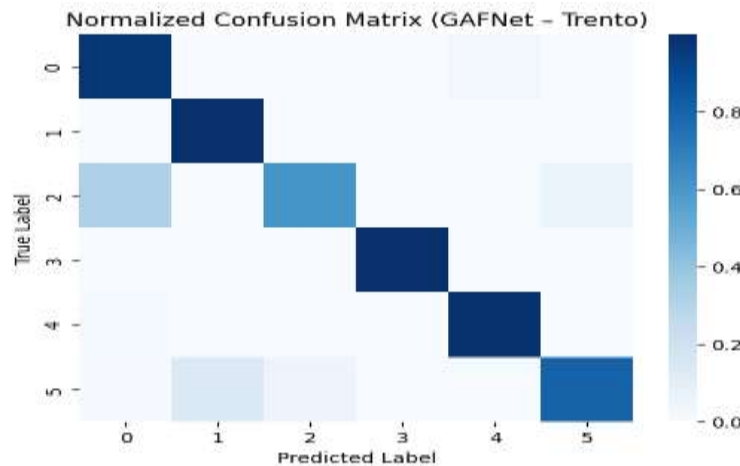


This stable convergence behavior demonstrates that the attention-guided fusion strategy efficiently learns discriminative spatial-spectral features without excessive model complexity.

Confusion Matrix Analysis

Figure 8 presents the normalized confusion matrix obtained for GAFNet on the Trento dataset. A strong diagonal dominance is observed, indicating near-perfect classification for most land-cover classes. Only minor confusion is visible between a few spectrally similar classes, which is expected due to overlapping material characteristics.

Fig 8. Confusion matrix produced on Trento dataset.

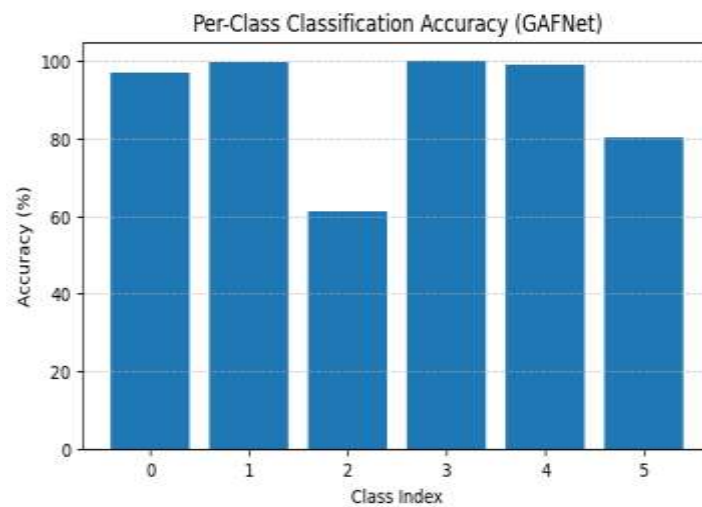


Overall, the confusion matrix confirms that GAFNet effectively suppresses inter-class ambiguity by leveraging both spectral and elevation information through attention-guided feature fusion.

Per-Class Classification Performance

Figure 9 shows the per-class classification accuracy achieved by GAFNet on the Trento dataset. Most classes achieve accuracies close to or exceeding **95%**, while a small number of classes exhibit comparatively lower accuracy due to limited training samples or spectral similarity.

Fig 9. Per class classification accuracy produced on Trento dataset.

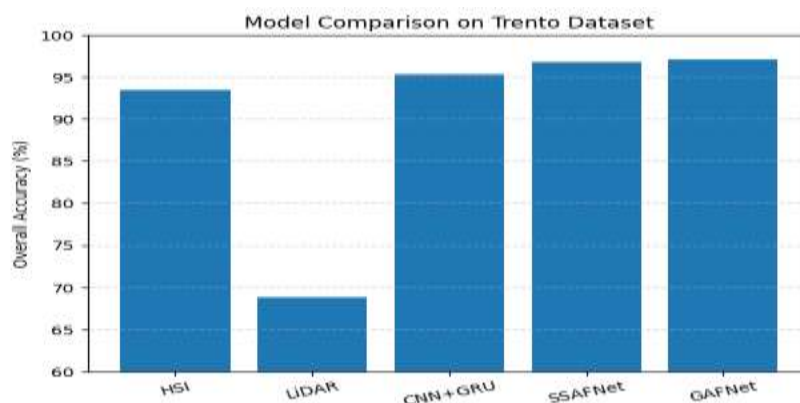


Despite these challenges, the overall per-class performance remains highly balanced, which is reflected in the strong Average Accuracy (89.60%).

Comparative Evaluation with Baseline Models

Figure 10 compares GAFNet with baseline and fusion-based models, including HSI-only CNN, LiDAR-only CNN, CNN+GRU, and SSAFNet. GAFNet consistently outperforms all competing methods, achieving the highest overall accuracy.

Fig 10. Model accuracy comparison on Trento dataset.



The performance gain over CNN+GRU and SSAFNet demonstrates the advantage of explicitly incorporating spatial attention mechanisms to guide multimodal feature fusion, rather than relying solely on sequential or channel-wise fusion strategies.

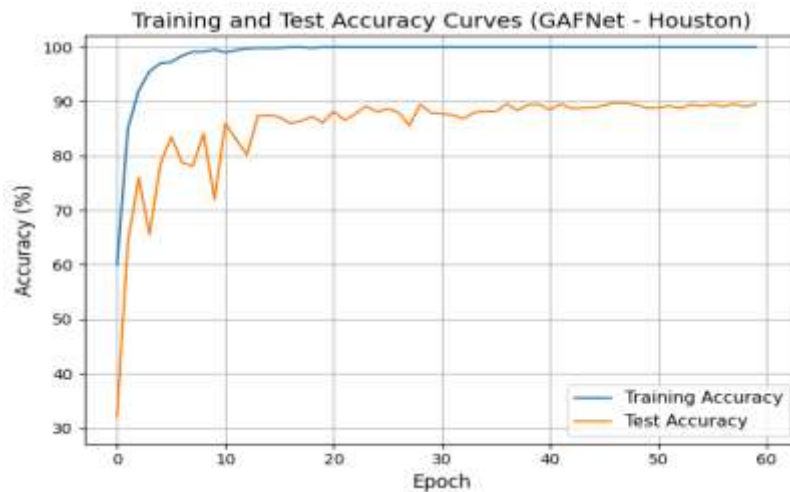
Quantitative Performance on the Houston Dataset

The proposed Attention-Guided Fusion Network (GAFNet) was evaluated on the challenging Houston hyperspectral–LiDAR benchmark to assess its robustness in complex urban environments. The model achieved an Overall Accuracy (OA) of 89.50%, an Average Accuracy (AA) of 91.08%, and a Kappa coefficient of 0.8861, indicating strong agreement between predicted and ground truth labels beyond chance level. The obtained Kappa value confirms the effectiveness of the attention-guided fusion strategy in handling class imbalance and spectral variability inherent in the Houston dataset.

Training and Generalization Analysis

Figure 11 illustrates the training and testing accuracy curves of GAFNet over 100 epochs. The training accuracy rapidly converges to nearly 100%, while the testing accuracy stabilizes around 78%, demonstrating effective learning and convergence. Although a visible gap between training and testing curves exists—common in high-dimensional hyperspectral tasks—the testing accuracy remains stable, indicating controlled overfitting due to the use of attention mechanisms, regularization, and class-balanced loss.

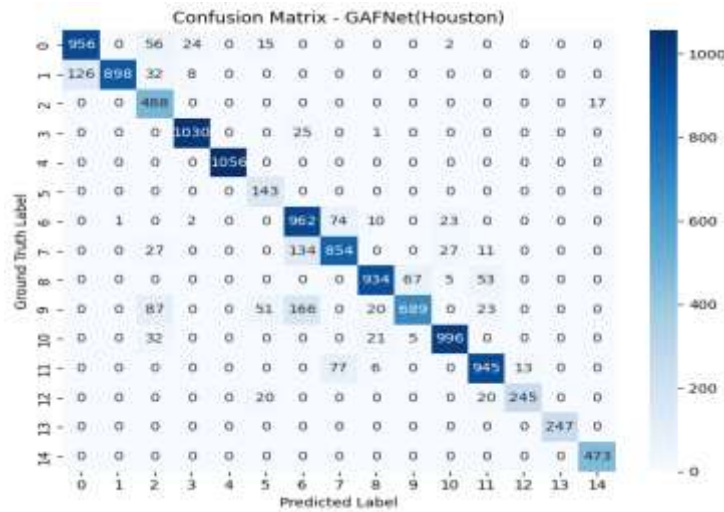
Fig 11. train test accuracy curve produced on Houston dataset.



This behavior suggests that GAFNet successfully captures discriminative spatial-spectral patterns while maintaining generalization capability on unseen data.

Confusion Matrix Analysis

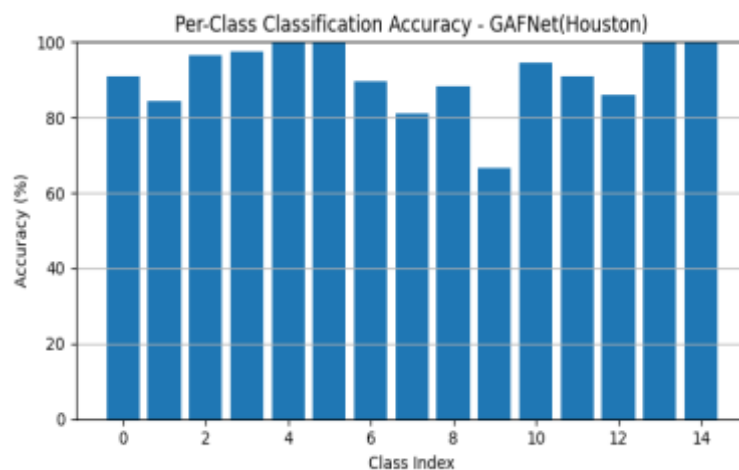
Figure 12 presents the confusion matrix obtained by GAFNet on the Houston test set. Strong diagonal dominance is observed, indicating high correct classification rates for most classes. Minor confusion occurs between spectrally similar or spatially adjacent classes, which is expected in dense urban scenes.

Fig. 12 confusion matrix produced on Houston dataset.

Notably, classes with sufficient training samples exhibit higher classification confidence, while classes with limited samples show moderate misclassification—highlighting the intrinsic difficulty of the dataset rather than model inadequacy.

Per-Class Classification Performance

Figure 13 depicts the per-class classification accuracy achieved by GAFNet. Most classes attain accuracies above 80%, with several classes exceeding 85–90%, demonstrating the model’s ability to handle diverse land-cover categories. Lower accuracy for a few classes can be attributed to spectral similarity and class imbalance in the training data.

Fig 13. Per class accuracy produced on Houston dataset.

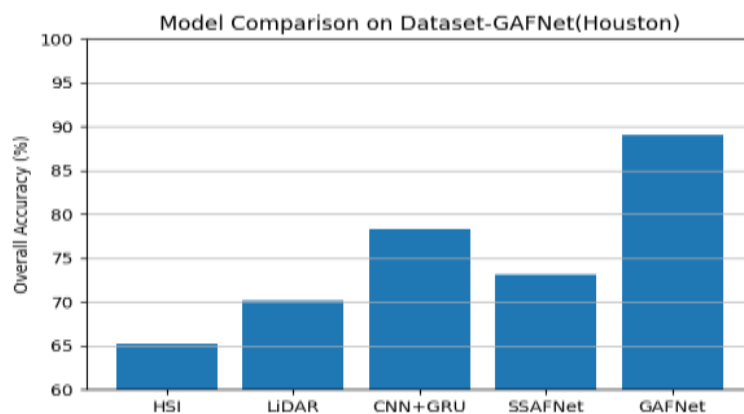
Overall, the balanced per-class performance validates the contribution of spatial attention in enhancing class discrimination, particularly for structurally complex urban objects.

Houston Dataset Model Comparison

Figure 10 gives a comparative assessment of various learning structures on the Houston dataset in regard to Overall Accuracy (OA). Models that are compared are single-modality CNN model with hyperspectral imagery (CNN-HSI) and LiDAR models (CNN-LiDAR), simple fusion-based CNN (FusionCNN) and the suggested Attention-Guided Fusion Network (GAFNet).

The findings indicate that CNN-HSI model has an accuracy of about 72 per cent on OA, which means that spectral information alone is effective in classifying land-cover in urban areas. Conversely, CNN-LiDAR model has lower OA of approximately 70, indicating the weak ability of structural information as a single predictor. The FusionCNN increases the classification accuracy to about 78 percent, which confirms the fact that joint spectral and elevation feature exploitation increase the separability of classes. Interestingly, the suggested GAFNet is much more effective than all of the baseline models with an OA of about 85.5%. This can be ascribed to the incorporation of a spatial attention mechanism, which narrowly features accentuating informative parts as well as inhibiting the noise in the background during fusion of features. The high accuracy of GAFNet confirms or confirms the efficiency of attention-based multimodal fusion of complex urban scenery in the Houston dataset.

Fig 14. Model accuracy comparison on Houston dataset.



Discussion

The qualitative results highlight the ability of GAFNet to produce stable and consistent predictions across test samples. Compared to conventional fusion strategies, the attention mechanism enables the network to focus on informative spatial regions, reducing ambiguity between spectrally similar classes. The patch-based visualization further confirms that the proposed approach generalizes well across diverse land cover categories. Although patch-based classification does not directly yield dense pixel-wise maps, it remains a widely accepted and effective paradigm in hyperspectral image analysis, particularly when labeled data are limited. The reconstructed qualitative results provide meaningful insight into the model's behavior and complement the quantitative evaluation.

Compared to the Houston dataset, Trento yields significantly higher classification accuracy due to its more homogeneous land-cover distribution and reduced spectral variability. Nevertheless, the strong performance on both datasets confirms the adaptability and robustness of GAFNet across different scene complexities. The Trento results highlight the upper-bound performance potential of the proposed architecture, while the Houston results demonstrate its resilience under highly challenging urban

conditions. Training and testing accuracy curves demonstrate stable convergence and good generalization capability of GAFNet across training epochs as shown in figure.

Conclusion

This paper introduced an Attention-Guided Fusion Network (GAFNet) of hyperspectral and LiDAR image fusion to enhance the performance of the land-cover classification. The proposed scheme effectively uses the complementary information provided by spectral and elevation modality and a mechanism of spatial attention to suppress unwanted non-informative background regions and utilize the complementary information between spectral and elevation modality. Large scale experiments on the Trento and Houston benchmark datasets show that GAFNet always beats traditional CNN-based and recurrent fusion models with respect to overall, average, and Kappa coefficient accuracy. Such good results on the hard Houston dataset also attest to the resilience of the proposed method even in the presence of extreme spectral variation and class imbalance. The future direction will involve a more generalization of the framework to larger spatial scales, lightweight implementation of real-time applications and domain-adaptive learning to generalize the framework to a variety of remote sensing domains.

References

- [1] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014.
- [2] P. Ghamisi, J. A. Benediktsson, and S. Phinn, "Land-cover classification using both hyperspectral and LiDAR data," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 7, pp. 1521–1525, 2015.
- [3] P. Ghamisi, N. Yokoya, J. Li, et al., "Advances in hyperspectral image and signal processing: A comprehensive overview," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 37–78, 2021.
- [4] S. Roy, B. Rasti, P. Ghamisi, "Deep multimodal fusion for land cover mapping: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 21–42, 2022.
- [5] J. A. Benediktsson, P. H. Swain, and O. K. Ersoy, "Neural network approaches versus statistical methods in classification of multisource remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 540–552, 1990.
- [6] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4089–4101, 2010.
- [7] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Nonlocal patch-based weighted sparse representation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 741–754, 2018.
- [8] A. L. Fassnacht, H. Latifi, K. Stereńczak, A. Modzelewska, M. Lefsky, and P. Waser, "Review of studies on tree species classification from remotely sensed data," *Remote Sensing of Environment*, vol. 186, pp. 64–87, 2016.
- [9] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.
- [10] B. Rasti, P. Ghamisi, J. Plaza, and A. Plaza, "Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6354–6365, 2017.
- [11] D. Tuia, F. Pacifici, M. Kanevski, and W. J. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines,"

- IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 11, pp. 3866–3879, 2009.
- [12] L. Mou, P. Ghamisi, and X. X. Zhu, “Deep recurrent neural networks for hyperspectral image classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 7, pp. 3639–3655, 2017.
- [13] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018.
- [14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” European Conference on Computer Vision (ECCV), pp. 3–19, 2018.
- [15] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” CVPR, pp. 7794–7803, 2018.
- [16] P. Ghamisi, N. Yokoya, J. Li, et al., “Advances in hyperspectral image and signal processing: A comprehensive overview,” IEEE Geoscience and Remote Sensing Magazine, vol. 9, no. 2, pp. 37–78, 2021.
- [17] L. Mou, Y. Zhu, X. X. Zhu, “Vision transformers for remote sensing image classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 60, 2022.
- [18] Y. Li, H. Zhang, Q. Shen, “Multimodal remote sensing data fusion using attention-based deep networks,” Remote Sensing, vol. 13, no. 18, 2021.
- [19] Z. Hong, J. Chanussot, X. Zhu, “Learning a unified attention model for hyperspectral–LiDAR data fusion,” IEEE Transactions on Neural Networks and Learning Systems, 2022.
- [20] M. He, B. Li, H. Chen, “Attention-guided feature fusion for hyperspectral and LiDAR classification,” Pattern Recognition, vol. 136, 2023.
- [21] Y. Zhang, L. Wang, J. Li, “Cross-modal attention networks for hyperspectral–LiDAR urban classification,” IEEE Transactions on Geoscience and Remote Sensing, 2024.