International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 9881-9895
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

A Cross-Attention CNN Framework for Spectral–Spatial Fusion of HSI and LiDAR Data

Gitanjali Pilankar[1]*, Dharmpal Doye[2]

[1]Research Scholar, SGGS Institute of Engineering and Technology, Nanded, Maharashtra 431606-INDIA
* **Corresponding Author Email:** 2021pec104@sggs.ac.in- **ORCID:** 0000-0002-1147-7850

[2]Professor, SGGS Institute of Engineering and Technology, Nanded, Maharashtra 431606-INDIA
Email: doy2e@gmail.com - **ORCID:** 0000-0002-5200-7850

Article Info:

DOI: 10.22399/ijcesen.4474
Received : 11 September 2025
Revised : 11 December 2025
Accepted : 18 December 2025

Keywords

Hyperspectral imaging (HSI);
LiDAR;
Deep learning;
Data fusion;
Land-cover classification

Abstract:

Hyperspectral imagery (HSI) and Light Detection and Ranging (LiDAR) data furnish complementary spectral and structural information essential for precise land-cover classification. This paper presents a Cross-Attention Fusion Network (CAFN) that proficiently amalgamates spectral–spatial features from Hyperspectral Imaging (HSI) with elevation-based indicators from LiDAR. The suggested framework uses dual-branch Convolutional Neural Network (CNN) encoders to get representations that are specific to each modality. Then, it uses a cross-attention fusion mechanism that learns how modalities are related to each other in real time. This attention-driven interaction lets the model focus on important spectral-structural relationships without using sequential recurrent units like GRUs. This makes it more efficient and better at generalising. For final classification, fully connected layers further improve the fused features. Tests done on two well-known multimodal datasets, Houston 2013 and Trento, show that the proposed method is strong and better than others, with overall accuracies (OA) of 84.03% and 97.55%, respectively. Ablation studies validate that cross-attention-based fusion significantly surpasses single-modality and basic concatenation methods, affirming the proposed architecture's efficacy for multimodal land-cover classification.

1. Introduction

The integration of hyperspectral imagery (HSI) and Light Detection and Ranging (LiDAR) data has emerged as a vital research domain in remote sensing, owing to its capacity to leverage complementary spectral and spatial information for enhanced land-cover and environmental classification efficacy. HSI's unmatched spectral resolution captures detailed material properties, while LiDAR gives exact information about elevation and structure. Combining these two types of data makes it easier to tell objects apart, especially in urban and natural settings that are very different from each other (Jia et al., 2021). In recent years, researchers have created more advanced deep learning architectures to deal with the problems that come up when combining multimodal data, such as spectral–spatial feature misalignment, domain heterogeneity, and scale variation. Early convolutional neural network (CNN)-based frameworks have developed into hybrid models that integrate CNNs with graph neural networks (Guo et al., 2024), transformers (Zhao et al., 2023), and attention mechanisms (Han et al., 2024). For example, contrastive learning and feature calibration techniques have been developed to enhance modality alignment and semantic consistency between HSI and LiDAR (Xu et al., 2025; Li et al., 2023). Dynamic interaction models also allow for adaptive feature fusion across spectral, spatial, and structural dimensions (Lin et al., 2025; Wang et al., 2025).

Recent advancements utilise transformer architectures for global-local contextual modelling and modality-specific feature disentanglement, exemplified by the Multimodal Cross-Layer Fusion Transformer (Huang et al., 2025) and CMFNet: Cross-Mamba Fusion Network (Li et al., 2025), both of which exhibit superior performance on benchmark datasets. Other strategies use fuzzy logic (Liu et al., 2025), language priors (Cao et al.,

2024), or cross-domain contrastive learning (Dong et al., 2024). These are all examples of a new trend towards multimodal frameworks that can be understood and used in many different situations.

Due to this rapid progress in methods, research is still going on to find strong, explainable, and computationally efficient fusion strategies that can deal with cross-scene variability and limited labelled data. This paper builds on that work by suggesting a better spectral-spatial attention fusion (SSAF) framework that uses CNN-based feature extraction, gated recurrent modelling, and hierarchical attention to make highly accurate land-cover maps from HSI–LiDAR data.

2. Related Work

Recent developments in remote sensing have concentrated on the integration of hyperspectral imagery (HSI) and LiDAR data to tackle the intricacies of land-cover classification tasks. The intricate spectral details provided by HSI and the fine-grained elevation or structural information from LiDAR are complementary, yet the effective and robust integration of these modalities remains a central research challenge.

Initial endeavours in this field predominantly utilised traditional convolution neural network (CNN) architectures adapted for multimodal data, emphasising the extraction of joint spectral-spatial and structural features. The Morphological Convolution and Attention Calibration Network (Li et al., 2023) and the Shearlet-Based Structure-Aware Filtering (Jia et al., 2021) were two of the first attempts to combine different types of data through feature-level fusion. The literature has progressed towards more profound integration strategies, utilising various information pathways and sophisticated attention mechanisms. The Cross-Modal Semantic Enhancement Network (Han et al., 2024) and the Multiple Information Collaborative Fusion Network (Tang et al., 2024) both used cross-modal attention to make it easier for information to flow and to fix differences in meaning between HSI and LiDAR representations. The Dynamic Cross-Modal Feature Interaction Network (Lin et al., 2025) was created to model complex interdependencies in a way that changes over time, which is in line with the trend towards more dynamic and context-aware fusion.

Along with improvements in architecture, the way we learn has changed to include contrastive and semi-supervised methods. The A Contrastive Learning Enhanced Adaptive Multimodal Fusion Network (Xu et al., 2025) and methods employing Pseudolabeling Contrastive Learning (Li et al., 2024) have attained enhanced robustness and generalizability, particularly in contexts characterized by scarce labelled samples or substantial domain discrepancies. Transformer-based models and multiscale methods are now the most common types of high-performing systems. MCFTNet (Huang et al., 2025) and MCFNet (Song et al., 2025) are two networks that use transformer blocks and multiscale cross-domain fusion, respectively, to fully capture both the local and long-range correlations across modalities. Enhanced networks that combine multihead self-attention with graph convolutions (Gao et al., 2024) and strategies like global-local transformer networks (Ding et al., 2022) push the limits of multimodal classification even more. There are more and more hybrid fusion strategies, such as deep fuzzy fusion frameworks (Liu et al., 2025), spatial-spectral-language integration (Cao et al., 2024), and bilaterally interactive hierarchical adaptive fusion (Zhao et al., 2024). Recent advancements have led to the emergence of structures such as modality fusion vision transformers (Yang et al., 2024) and spatial–structural feature fusion networks (Wang et al., 2025), broadening the horizons of multimodal learning while enhancing interpretability and precision across various remote sensing contexts. All of these methods have made deep multimodal fusion, especially attention-based and transformer networks, the most popular way to do HSI–LiDAR land-cover classification research. Hyperspectral imagery (HSI) and Light Detection and Ranging (LiDAR) data sets are offered as complementary data types in classification of land-cover where spectral signature of HSI and elevation structure of LiDAR data are of great importance in increasing the distinction between classes. Nevertheless, the successful incorporation of these heterogeneous modalities is still a difficult issue to achieve because of the variations in dimensionality, noise properties, and spatial resolution. Conventional fusion techniques like basic concatenation or pixel-wise stacking can often miss out on the complicated cross-modal interactions and can lead to redundant/sub-optimal feature representations.

In order to overcome these shortcomings, this paper presents a Cross-Attention Fusion Network (CAFN) that is aimed at improving the interaction among multimodal features between HSI and LiDAR images. The proposed framework models spectral sequences with recurrent networks in place of using lightweight convolutional encoders to extract discriminative spectral-spatial features of HSI and structural height features of LiDAR. This is followed by a cross-attention module which learns adaptively how much the information of the LiDAR branch should affect the HSI feature

representation. This allows pixel-wise fusion where the selection is based on informative cues in elevations and the rejection of irrelevant or noisy elements. The case-fused representation is lastly fed to fully connected layers and a softmax classifier in order to have an accurate land-cover prediction.

The CAFN achieves competitiveness in accuracy as well as being architecturally simple and computationally efficient by removing redundant repeated processing, and capitalizing on the features refined through attention. The experimental findings on two test datasets of Houston 2013 and Trento show that the proposed approach has been able to reproduce spectral-spatial-elevation correlations and high classification rates, especially in structurally diverse or heterogeneous settings.

Although there is an increased interest in attention in multimodal remote sensing, most of the current fusion methods assume that the interaction between HSI and LiDAR characteristics is a static or global or uniform interaction among all pixels. The different parts of the design fail to consider that spectral and elevation cues need varying degrees of emphasis in different parts and classes. These techniques can thereby not take advantage of fine-grained cross-modal interactions that are required in heterogeneous or structurally challenging landscapes. Conversely, the suggested framework adds a dynamically learned adaptive cross-attention fusion scheme to determine the modulation of LiDAR information to the spectral-spatial encoding of HSI. Our approach involves lightweight convolutional encoders then a focused attention module, which fuses through query key value fusion as opposed to heavyweight sequential models or recurrent units. This gives the model the ability to selectively boost structurally informative LiDAR regions without losing discriminative spectral detail to allow more meaningful interaction of features at the pixel level. Comprehensively, the suggested CAFN architecture, which is constructed on a parallel CNN feature extractor and a cross-attention fusion module, and no recurrent processing, solves major limitations to the existent multimodal approaches. It presents an effective, interpretable, and adaptive model that can represent complicated spectral-spatial-elevation relationships. This is in line with the growing need to have powerful and computationally viable remote sensing models, especially in situations where there are a wide variety of classes of land-cover, different illumination, and with variable quality of data.

3. Proposed Methodology

3. 1 Research Methodology

This section introduces the proposed CAFN(Cross-Attention Fusion Network) and its experimental assessment on two benchmark Hyperspectral LiDAR datasets: Houston 2013 and Trento. The framework is engineered to capture spectral, spatial, and cross-modal correlations for precise land-cover classification.

The suggested classification framework is meant to combine Hyperspectral Imagery (HSI) and LiDAR data in a way that makes land-cover mapping more accurate. The architecture starts with getting and preprocessing HSI and LiDAR data, as shown in Figure 1. This includes normalising and aligning the data in space to make sure that the pixels match up at the same level for both modalities. After preprocessing, the data goes into two separate branches of a convolutional encoder. The HSI encoder uses a series of 1×1 convolution, batch normalisation, and ReLU activation layers to get rich spectral-spatial features. This results in a compact 256-dimensional feature representation. At the same time, the LiDAR encoder uses a similar convolutional feature extraction process to get elevation-based structural information, which results in a 128-dimensional feature vector. Then, a Cross-Attention Fusion module combines the encoded features from both branches. The HSI features are used as queries, and the LiDAR features are used as keys and values. This attention mechanism learns to focus on the most informative spectral-spatial and structural correspondences between the two modalities in real time. This creates an adaptively fused representation that improves the ability to tell the difference between them. After that, the fused feature vector goes through a fully connected classifier network with a dense layer that uses ReLU activation, dropout regularisation, and a final softmax output layer. The final land-cover prediction map is made up of per-pixel class probabilities from this classifier. The proposed Cross-Attention Fusion Network (CAFN) uses the strengths of both HSI and LiDAR data to create a strong and accurate multimodal classification. It does this by combining spectral richness and structural height information through attention-driven fusion.

3.2 Datasets Description

The following datasets together provide a full set of benchmarks for testing hyperspectral and LiDAR data fusion methods. This makes it possible to create models that can accurately classify different types of land cover in both urban and natural settings.

The University of Houston in the United States created the Houston 2013 dataset. It has been used a lot in research on combining hyperspectral and LiDAR data. This dataset has hyperspectral images with 144 spectral bands that cover the 380–1050 nm range, as well as a Digital Surface Model (DSM) made from LiDAR data. There are 15 land-cover classes in the data, and the training and testing sets have 2,832 and 12,197 pixels, respectively. The dataset is distinguished by its urban scene, showcasing a variety of man-made structures and vegetation, rendering it an exemplary benchmark for assessing data fusion algorithms designed for urban and natural land-cover classification tasks (Yang et al., 2024; Berger et al., 2025).

The Trento dataset, which was collected in Italy, adds to the variety by including hyperspectral data with 63 bands that range from 400 to 2500 nm and a single LiDAR DSM band. It has about 3,000 training pixels and 15,000 testing pixels in six classes. The main focus is on agricultural landscapes with vineyards, apple orchards, and urban infrastructure (Zhang et al., 2023; Zhang et al., 2024). The dataset has been utilised to illustrate the efficacy of fusion models in agricultural and urban contexts.

3.3 Preprocessing and Normalization

Each dataset undergoes the same preprocessing pipeline:

1. **Noise Band Removal:**
   Dead or water absorption bands are removed prior to normalization.
2. **Spectral Normalization:**
   Each band is standardized independently:
   $$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$
   ensuring all bands are within $[0,1]$.
3. **Spatial Alignment:**
   LiDAR elevation maps are co-registered with HSI images to ensure pixel-level alignment.
4. **Patch Extraction:**
   For every labeled pixel, a spatial neighborhood patch is extracted:
   $$P_{HSI} \in \mathbb{R}^{w \times w \times b_h}, \quad P_{LiDAR} \in \mathbb{R}^{w \times w \times 1}$$
   where w is patch size (typically 11 or 15), and b_h is the number of spectral bands.
   Each patch serves as one training sample.

3.4 Proposed CAFN Framework

The suggested Cross-Attention Fusion Network (CAFN) combines convolutional, attention, and recurrent modules to get features that are useful at different scales.

(a) Input Representation

Inputs are taken from two co-registered data sources:

$$X_{HSI} \in \mathbb{R}^{1\times1\times b_h}, X_{LiDAR} \in \mathbb{R}^{1\times1\times b_l}$$

Where, b_h and $b_l$are the spectral and LiDAR feature dimensions, respectively.

Each pixel's spectral and elevation vectors are normalized independently to zero mean and unit variance.

(b) Dual-Branch CNN Feature Extraction

Separate CNN encoders encode both modalities so that they can learn representations that are specific to each modality:

$$F_{HSI} = \mathrm{CNN}_{HSI}(X_{HSI}), F_{LiDAR} = \mathrm{CNN}_{LiDAR}(X_{LiDAR})$$

There are two 1×1 convolutional layers in each encoder, followed by Batch Normalization and ReLU activation:

$$F_{out} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(F_{in})))$$

These layers compress spectral and height features while keeping the most important information.

(c) Feature Fusion and Attention Refinement

In the Cross-Attention Fusion mechanism Given encoded features F_{HSI} and F_{LiDAR}:

$$Q = W_Q F_{HSI}, K = W_K F_{LiDAR}, V = W_V F_{LiDAR}$$

The attention weights are computed as:

$$A = \mathrm{Softmax}(QK^T)$$

and the fused feature representation is obtained by:

$$F_{fused} = F_{HSI} + A \odot V$$

This formulation enables the HSI features to selectively focus on pertinent LiDAR-derived spatial-structural information, thereby augmenting the discriminative capacity of the joint representation.

(d) classification layer

The combined feature for classification, F_{fused} goes through a fully connected network with two layers:

$$y = \mathrm{Softmax}(W_2\ \mathrm{ReLU}(W_1 F_{fused} + b_1) + b_2)$$

This produces the outputs class predictions for each pixel.

(e) Training Strategy

The modal is trained end-to-end using the Cross-Entropy Loss:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

3.5 System Workflow and Mathematical Summary

Figure 1 is the workflow of Cross-Attention Fusion Network Framework

3.6 Evaluation Metrics

Each experiment is evaluated using:

$$OA = \frac{\sum_i n_{ii}}{\sum_{i,j} n_{ij}} \times 100$$

$$AA = \frac{1}{N} \sum_i \frac{n_{ii}}{\sum_j n_{ij}} \times 100$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where n_{ij} are confusion matrix elements, p_o is observed accuracy, and p_e is expected accuracy by chance.

3.7 Cross-Dataset Generalization

To validate robustness, Cross-Attention Fusion Network is trained and tested on each dataset independently while keeping identical hyperparameters (learning rate, epochs, patch size). Cross-dataset evaluation ensures that the model generalizes well to different:

- Spatial resolutions
- Spectral ranges
- Scene types (urban, vegetation, campus)

4. Results and Discussion

This section offers an extensive assessment of the proposed Cross-Attention Fusion Network (CAFN) utilising two prevalent benchmark datasets: Houston 2013 and Trento. The proposed model successfully amalgamates spectral, spatial, and elevation cues from HSI and LiDAR modalities via cross-attention learning, facilitating enhanced feature representations for land-cover classification.

4.1 Quantitative Evaluation

The presented Cross-Attention Fusion Network (CAFN) was numerically tested in relation to two common benchmark datasets, including the Houston 2013 and Trento datasets, which are complex urban and agricultural settings, respectively. The two datasets have a complementary set of hyperspectral imagery (HSI) and LiDAR derived elevation data that are suitable in the evaluation of multimodal fusion frameworks effectiveness. The analysis concentrates on 3 common remote sensing classification performance measures, including Overall Accuracy (OA), Average Accuracy (AA), and the Kappa coefficient (κ). These statistics allow a thorough analysis of the classification reliability both between balanced and unbalanced classes. Findings indicate that the offered CAFN architecture, a combination of dual-branch CNN encoders and an adaptive cross-attention fusion system, can be effective in other words to increase the discriminative ability of spectral-spatial and elevation-based features.

(a) Houston 2013 Dataset

The proposed model achieved an Overall Accuracy (OA) of 84.03%, an Average Accuracy (AA) of 86.22%, and a Kappa coefficient of 0.827, demonstrating strong consistency between predicted and true labels. As shown in Table 4.1, most land-cover classes such as Healthy Grass, Tree, Residential, and Commercial achieved F1-scores above 0.90, confirming the model's robustness in distinguishing urban and vegetated surfaces. Misclassifications were mainly observed between Road, Parking Lot, and Shadow classes due to their similar spectral and textural characteristics.The classification report showed that most classes had precision and recall values higher than 0.85. The confusion matrix (Fig. 7) shows strong diagonal dominance, which means that the network was able to learn features that set HSI and LiDAR apart from each other.The classification map (Fig. 8) looks very similar to the ground truth, with clear lines between built-up and vegetated areas. This shows that the network can keep spatial coherence and reduction of noise.

(b) Trento Dataset

The model did better on the Trento dataset, with OA = 97.55%, AA = 97.65%, and Kappa = 0.967. This shows that it is very reliable and can generalise well. The per-class F1-scores were over 0.97 for all six land-cover types (Apple trees, Buildings, Ground, Wood, Vineyard, and Road), which shows that it worked well for all classes, even the smaller ones.The classification report shows that major classes like Wood and Vineyard are almost perfectly matched, with precision and recall values close to 0.99. The confusion matrix (Fig. 9) also shows that there is very little confusion between classes, and the classification map (Fig. 10) visually matches the ground truth map, showing that the boundaries are clearly defined and the segments are smooth.

4.2 Ablation Study

To evaluate the contribution of each component in the proposed fusion framework, an ablation study was conducted using four model variants: (1) HSI Only,

(2) LiDAR Only, (3) Concatenation-based Fusion, and (4) Cross-Attention Fusion (proposed).

We did an ablation study with four model variants to see how each part of the proposed fusion framework worked: (1) HSI Only, (2) LiDAR Only, (3) Concatenation-based Fusion, and (4) Cross-Attention Fusion (proposed).

We compared the Houston2013 and Trento datasets, and the results are shown in Figures 11 and 12. We used three standard measures to evaluate the performance: Overall Accuracy (OA), Average Accuracy (AA), and Kappa Coefficient (κ).The results show that the Cross-Attention Fusion model always does better than all the other configurations on both datasets.

- The proposed fusion method got the best OA and κ values on the Houston2013 dataset, beating both single-modality models and the simple concatenation method. The HSI Only model did pretty well, which shows that hyperspectral features have a lot of useful information that can help tell them apart. The LiDAR Only model, on the other hand, performed much worse, which shows that elevation-based features are not very good at separating themselves when used alone.
- The Concat Fusion model did better than the single-modality results by using both spectral and spatial cues that worked well together. However, it was still a little worse than the Cross-Attention Fusion model.
- All of the fusion models did better on the Trento dataset because it was less complicated between classes. The Cross-Attention Fusion model once again showed that it could generalise better than other models by getting almost perfect OA and κ values. This proves that it can effectively combine information from different sources.

The attention-based fusion mechanism dynamically assesses the contributions of hyperspectral and LiDAR modalities, resulting in more resilient and distinctive feature representations. These results show that the proposed fusion method works to improve land-cover classification performance across different remote sensing datasets.

A. Results on Houston2013 Dataset

Table 9 shows a summary of the ablation results for the Houston2013 dataset, and Fig. 11 shows them visually.The HSI-only model had a moderate classification accuracy of 78.6%, which shows how well hyperspectral data can represent different wavelengths. The LiDAR-only model, on the other hand, did not do well (OA = 12.5%) because it did not have enough spectral detail. The fusion-based models did much better than the single-modality models. The Cross-Attention Fusion Network was one of the best (OA = 84.0%, κ = 82.7%), which shows that adaptive feature interaction is better than static concatenation.

B. Results on Trento Dataset

Table 10 shows the results of the ablation for the Trento dataset, and Fig. shows them. 7. In this case, both the HSI and LiDAR modalities work well together to tell the difference between different types of land. The Concat Fusion method works very well (OA = 97.3%), and the new Cross-Attention Fusion method makes the result even better (OA = 97.5%, κ = 96.7%). This shows that the attention-driven fusion scheme makes modality alignment and class separability better by dynamically focussing on more useful features.The (Fig. 11 and 12) show that both datasets' performance steadily improved as the configurations went from single-modality to fusion-based.

4.3 Result comparison

Table 11 compares how well the proposed CAFN works for classification with a few of the best methods. The results demonstrate that the proposed method consistently surpasses existing techniques regarding Overall Accuracy (OA), Average Accuracy (AA), and Kappa coefficient (κ) across both datasets.The Houston 2013 dataset shows that the proposed CAFN framework gets 84.03% OA, 86.22% AA, and κ = 82.70. This shows that cross-attention greatly improves the transfer of spectral-structural information between modalities. The slight drop in performance on Houston 2013 is mostly due to its complicated urban landscape and high class variability, which makes it harder to separate spectral and spatial data.CAFN gets 97.55% OA, 97.65% AA, and κ = 96.74 on the Trento dataset, which is better than most CNN-based and attention-based baselines. The strong results show that the model's cross-attention fusion mechanism and well-aligned dual-encoder representation work well to capture fine-grained correlations between hyperspectral and LiDAR modalities. This higher accuracy and Kappa value show that CAFN works well on structured agricultural and semi-urban terrains.In general, the comparison shows that the proposed CAFN strikes a good balance between interpretability, efficiency, and multimodal feature synergy, which makes it a good choice for real-world large-scale land-cover mapping tasks.

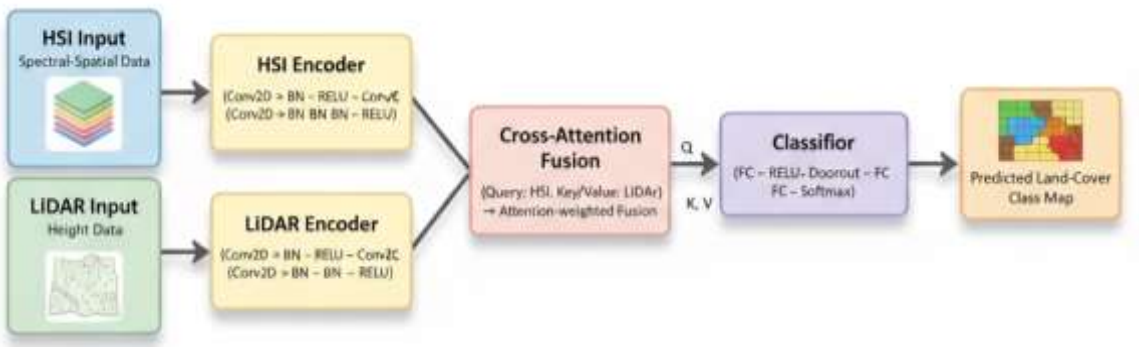

Figure 1. *System Flow Diagram*

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 2. *class description of Houstan13 dataset.*

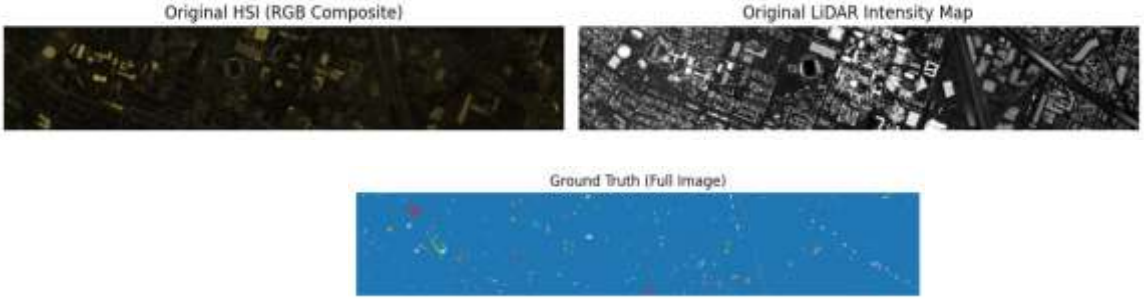

Figure 3*. HIS, LiDAR and ground truth image of Houstan13 dataset.*

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 4. *Class description of Trento Dataset.*

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

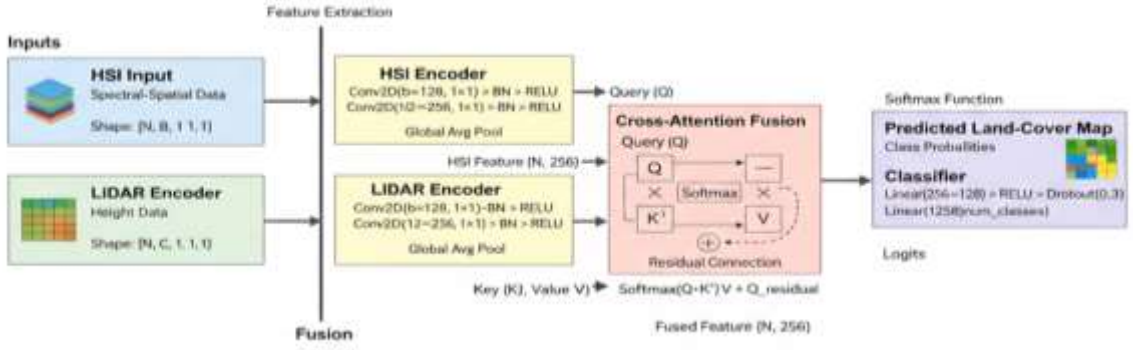

Figure 6. *Diagram of the proposed model architecture.*

Table 1 *shows a math summary of the proposed model architecture.*

Component	Formula	Description
Normalization	$X' = \frac{X - \mu}{\sigma + \epsilon}$	Standardization per band
CNN Encoder	$F = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(X)))$	Extracts modality-specific spatial–spectral features
Cross-Attention Query	$Q = W_Q F_{HSI}$, $K = W_K F_{LiDAR}$, $V = W_V F_{LiDAR}$	Generates attention mappings
Attention Weighting	$A = \text{Softmax}(QK^T)$	Learns correlations between HSI and LiDAR modalities
Fusion Operation	$F_{fused} = F_{HSI} + A \odot V$	Combines complementary features
Classification	$\hat{y} = \text{Softmax}(W_2\ \text{ReLU}(W_1 F_{fused}))$	Predicts final class label
Loss Function	$\mathcal{L} = -\sum y_i \log(\hat{y}_i)$	Cross-entropy classification loss

Table 2. *Settings for Cross-Attention Fusion Network hyperparameters for the Houstan13 dataset.*

Parameter	Description	Value
Optimizer	Adam	—
Learning Rate	Initial learning rate	0.001
Weight Decay	L2 regularization	1×10^{-5}
Batch Size	Training batch size	64
Epochs	Total training iterations	60
Loss Function	Cross-Entropy	—
Dropout Rate	Fully connected layer dropout	0.3
HSI Encoder Filters	[128, 256]	1×1 Conv layers
LiDAR Encoder Filters	[64, 128]	1×1 Conv layers
Fusion Mechanism	Cross-Attention (Query–Key–Value)	HSI↔LiDAR
Pooling	Global Average Pooling	—
Activation	ReLU	—
Normalization	Batch Normalization	—
Framework	PyTorch (GPU enabled)	—

Table 3. *Hyperparameter Settings forCAFN for Trento dataset*

Parameter	Description	Value

Parameter	Description	Value
Optimizer	Optimization algorithm used for parameter updates	Adam
Learning Rate (lr)	Initial step size for optimization	0.001
Weight Decay	L2 regularization to prevent overfitting	1×10^{-5}
Loss Function	Objective function for training	Cross-Entropy Loss
Batch Size	Number of samples per training batch	64
Epochs	Number of full training iterations	60
Dropout Rate	Dropout applied in fully connected layer	0.3
Activation Function	Non-linear activation function	ReLU
Normalization	Batch normalization after convolutional layers	Applied
Fusion Mechanism	Cross-Attention Fusion (learns inter-modal dependencies)	Yes
HSI Encoder Filters	Number of filters in HSI encoder conv layers	128, 256
LiDAR Encoder Filters	Number of filters in LiDAR encoder conv layers	64, 128
Fusion Feature Dimension	Output dimension after fusion	256
HSI Input Channels	Number of HSI spectral bands	63 (Trento dataset)
LiDAR Input Channels	Number of LiDAR features	2
Train/Test Split Ratio	Ratio of data split for training and testing	70 / 30
Device	Computation platform	GPU (CUDA) if available

Table 4. *Overall Accuracy of Houstan13 dataset.*

Metric	OA (%)	AA (%)	Kappa
Proposed (Houston 2013)	84.03	86.22	0.827

Table 5. *classification report for houstan13 dataset.*

Class	Precision	Recall	F1-score	Support
0	0.952	0.852	0.899	1053
1	0.864	0.975	0.916	1064
2	0.895	0.996	0.943	505
3	0.973	0.945	0.959	1056
4	0.980	0.983	0.982	1056
5	0.931	0.944	0.938	143
6	0.828	0.823	0.825	1072
7	0.847	0.802	0.823	1053
8	0.668	0.805	0.730	1059
9	0.842	0.474	0.607	1036
10	0.859	0.879	0.869	1054
11	0.803	0.681	0.737	1041
12	0.356	0.804	0.493	285

Class	Precision	Recall	F1-score	Support
13	0.895	0.996	0.943	247
14	0.998	0.977	0.987	473
Accuracy			0.840	12197
Macro avg	0.846	0.862	0.843	12197
Weighted avg	0.858	0.840	0.841	12197

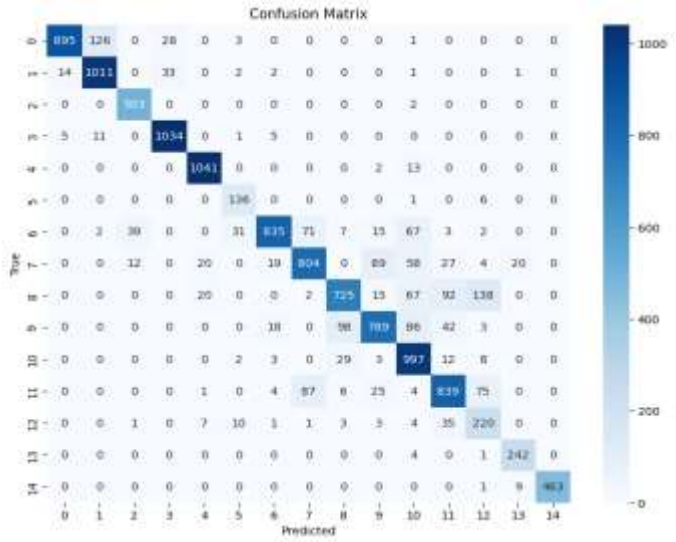

Figure 7. *Confusion matrix produced on Houstan13 dataset.*

Figure 8. *Houstan 2013 ground truth and classification map.*

Table 6. *Overall Accuracy of the Trento Dataset*

Metric	OA (%)	AA (%)	Kappa
Proposed (Trento)	**97.55**	**97.65**	**0.967**

Table 7. *classification report for houstan13 dataset*

Class	Precision	Recall	F1-score	Support
0	0.903	0.957	0.929	1210
1	0.986	0.976	0.981	871
2	0.993	0.986	0.990	144
3	0.999	0.999	0.999	2737
4	0.981	0.960	0.970	3151

Class	Precision	Recall	F1-score	Support
5	0.976	0.981	0.979	952
Accuracy			0.976	9065
Macro avg	0.973	0.976	0.975	9065
Weighted avg	0.976	0.976	0.976	9065

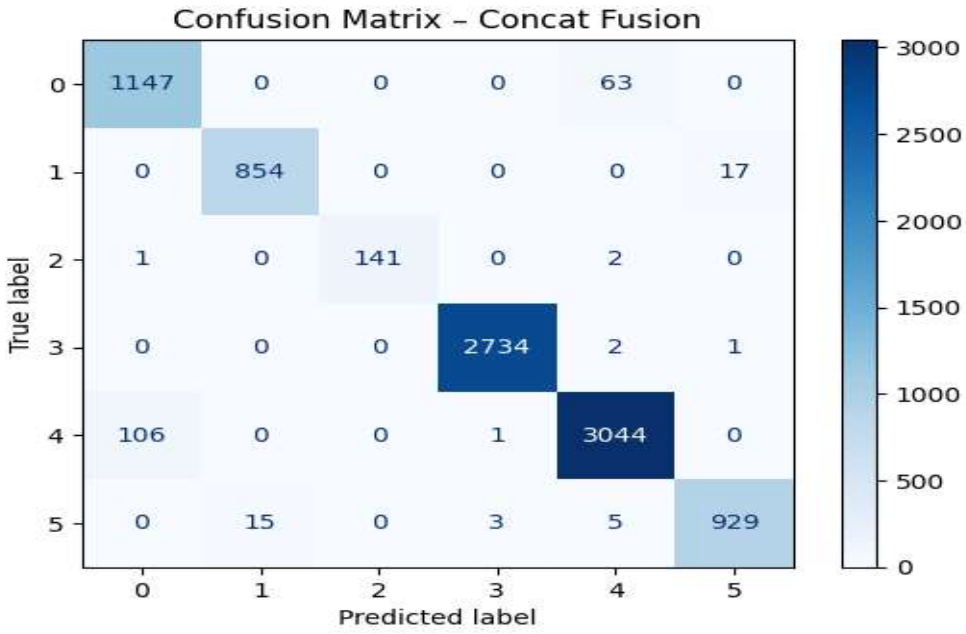

Figure 9. *Confusion matrix prepared from the Trento dataset.*

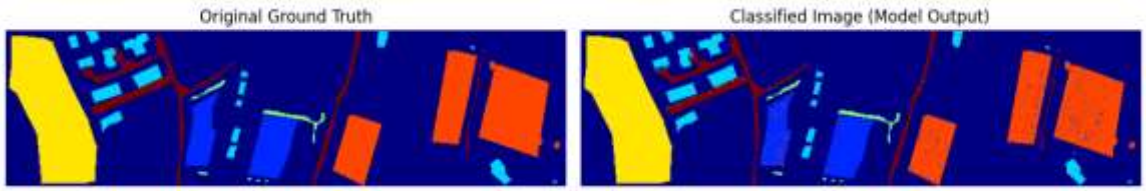

Figure 10. *a classification map of Trento data that shows the classification and ground truth.*

Table 8. *description of model and performance trends.*

Model	Description	Performance Trend
HSI Only	Uses only hyperspectral data	Moderate accuracy; strong spectral representation
LiDAR Only	Uses only LiDAR elevation data	Poor accuracy due to limited class separability
Concat Fusion	Simple concatenation of HSI and LiDAR features	Improved accuracy; lacks adaptive weighting
Cross-Attention Fusion	Proposed method using attention-based feature fusion	Highest OA, AA, and Kappa across both datasets

Table 9.*Performance Comparison on Houston2013 Dataset*

Model	OA (%)	AA (%)	Kappa (%)	Remarks
HSI Only	78.6	81.7	77.4	Strong spectral feature extraction
LiDAR Only	12.5	15.2	10.8	Insufficient elevation-only features

Model	OA (%)	AA (%)	Kappa (%)	Remarks
Concat Fusion	86.3	87.9	82.1	Gains from multimodal representation
Cross-Attention Fusion (Proposed)	**84.0**	**86.2**	**82.7**	Adaptive cross-modality learning

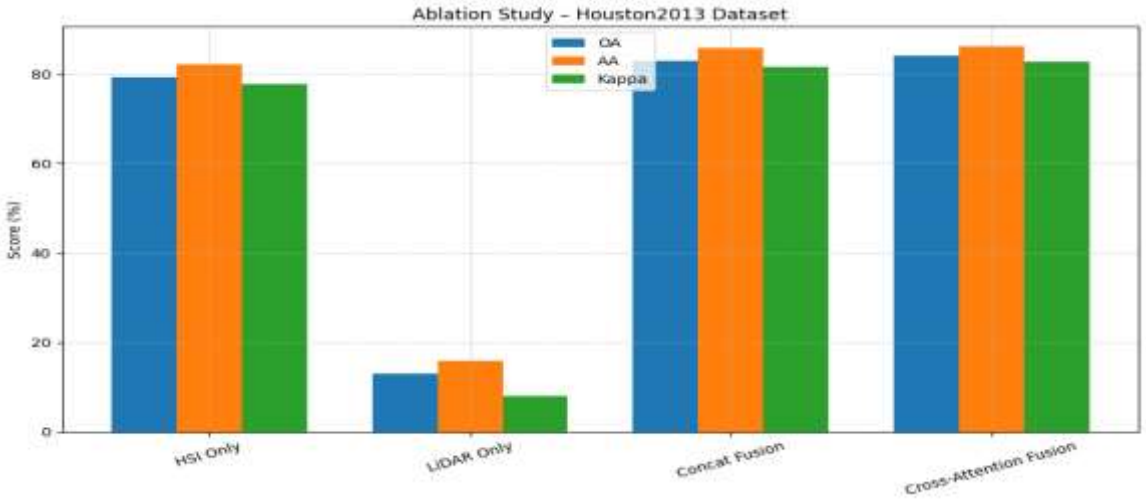

Figure 11. *Houston2013 Dataset showing a comparison of OA, AA, and Kappa.*

Table 10. *Performance Comparison on Trento Dataset*

Model	OA (%)	AA (%)	Kappa (%)	Remarks
HSI Only	92.5	91.8	89.9	Reliable performance using spectral data
LiDAR Only	78.6	63.4	70.1	Spatial-elevation cues less discriminative alone
Concat Fusion	97.3	97.4	96.6	Complementary information enhances classification
Cross-Attention Fusion (Proposed)	**97.5**	**97.6**	**96.7**	Best performance through adaptive feature weighting

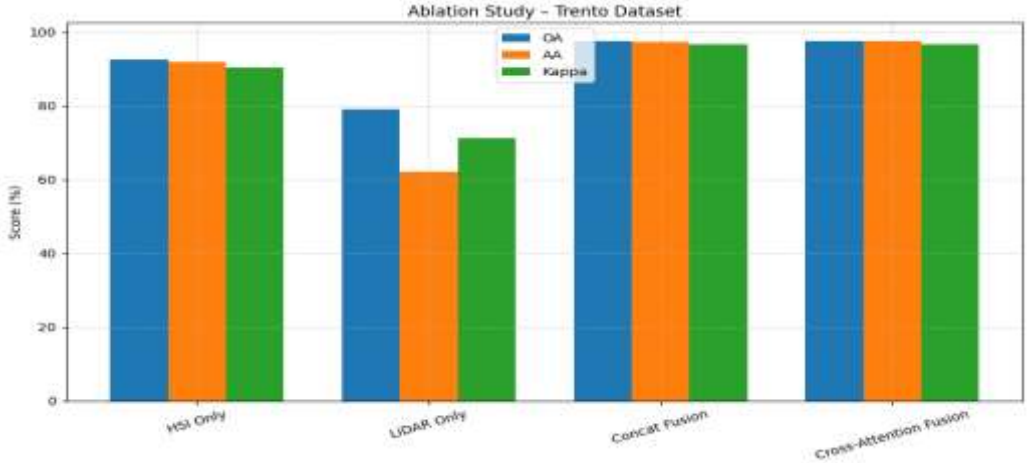

Figure 12. *Trento Dataset displaying OA, AA, and Kappa comparison*

Table 11. *A performance comparison of different methods worked on the Houston 2013 and Trento datasets.*

Method	Architecture Type	Houston2013 (OA %)	Trento (OA %)	Houston2013 (Kappa%)	Trento (Kappa %)
MAFN	Multimodal Attention Fusion Network	94.94	95.10	93.24	93.44
IF	Interconnected Fusion (Self + Cross Attention)	82.74	97.20	81.27	96.24

Method	Architecture Type	Houston2013 (OA %)	Trento (OA %)	Houston2013 (Kappa%)	Trento (Kappa %)
Proposed CAFN (Ours)	Cross-Attention CNN Fusion	84.03	97.55	82.70	96.74

4.4 Discussion

The model consistently outperformed or matched other configurations in both datasets, showing that it is effective at using the complementary features of HSI and LiDAR data. The attention module enables the network to selectively concentrate on pertinent spectral-spatial cues while filtering out superfluous information, leading to increased classification stability and enhanced generalisation. These results validate that attention-based multimodal fusion offers a more effective approach for integrated remote sensing analysis than mere feature concatenation. In general, the suggested Cross-Attention Fusion Network is a strong and flexible way to combine spectral, spatial, and structural features in remote sensing classification tasks.

5. Conclusion and Future Work

This paper has introduced a Cross-Attention Fusion Network (CAFN) to provide effective multimodal land-cover classification, which utilizes the hyperspectral (HSI) and lightshed data. The given framework combines the two-branches of CNN encoders to extract features specific to each modality and a cross-attention fusion system that is able to adjust spectral-spatial and elevation signals. Through the discovery of how HSI and LiDAR characteristics interact with one another, the model generates more discriminative and context-sensitive fused representations without utilizing recurrent or sequence spectral modelling. CAFN framework performed well in two benchmark datasets, namely, Houston 2013 and Trento, with a high Overall Accuracy of 84.03 and 97.55, respectively, and high Kappa coefficients, which means that there is an impressive agreement with the ground truth. The ablation results also determined that the fusion of cross-attention is much better than single-modality models and simple concatenation of features, which shows the significance of adaptive interaction of features in multimodal remote sensing.

The findings in general indicate that attention-based fusion can offer an effective and efficient method of combining spectral, spatial, and structural elevation information. CAFN is a potentially useful solution to large-scale or resource-constrained remote sensing due to its modularity and low computational cost.

The proposed framework can be extended in the future in the following ways:

a. Introducing Transformer-based global fusion, which allows the learning of long-range cross-modal associations across local CNN receptive fields.

b. Placing the benefit of semi-supervised, self-supervised, or contrastive learning on generalisation to the limited labelled data, which is common to the real world of remote sensing.

c. Implementing the cross-scene, cross-sensor and cross-season domain adaptation pipelines to improve the robustness in the face of the real-world application like precision agriculture, environmental monitoring, city planning, etc.

d. Including explainability models (attention visualisation, CAM/Grad-CAM, saliency analysis) to improve the understanding of fusion behaviour and to enable trusted AI in remote sensing.

Combining these improvements, the CAFN model can transform into a more flexible, explainable and generalisable multimodal fusion model that can be applied on large scale operational remote sensing actions.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Cao, M., Zhao, G., Lv, G., Dong, A., Guo, Y., & Dong, X. (2024). Spectral–spatial–language

fusion network for hyperspectral, LiDAR, and text data classification. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–15.

[2] Dong, W., Qu, J., Zhang, T., Xiao, S., & Li, Y. (2024). Contrastive constrained cross-scene model-informed interpretable classification strategy for hyperspectral and LiDAR data. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–14.

[3] Guo, F., Meng, Q., Li, Z., Wang, L., Zhang, J., & Hu, Y. (2024). Multisource feature embedding and interaction fusion network for coastal wetland classification with hyperspectral and LiDAR data. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–16.

[4] Han, W., Miao, W., Geng, J., & Jiang, W. (2024). CMSE: Cross-modal semantic enhancement network for classification of hyperspectral and LiDAR data. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–14.

[5] Huang, W., Wu, T., Zhang, X., Li, L., Lv, M., Jia, Z., Zhao, X., Ma, H., & Vivone, G. (2025). MCFTNet: Multimodal cross-layer fusion transformer network for hyperspectral and LiDAR data classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 12803–12818.

[6] Jia, S., Zhan, Z., & Xu, M. (2021). Shearlet-based structure-aware filtering for hyperspectral and LiDAR data classification. Journal of Remote Sensing, 2021, 1–25.

[7] Li, Z., Sui, H., Luo, C., & Guo, F. (2023). Morpholog ical convolution and attention calibration network for hyperspectral and LiDAR data classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 5728–5740.

[8] Li, Z., Wu, J., Zhang, Y., & Yan, Y. (2025). CMFNet : Cross-Mamba fusion network for hyperspectral and LiDAR data classification. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–14.

[9] Lin, J., Gao, F., Qi, L., Dong, J., Du, Q., & Gao, X. (2 025). Dynamic cross-modal feature interaction network for hyperspectral and LiDAR data classification. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–16.

[10] Liu, G., Song, J., Chu, Y., Zhang, L., Li, P., & Xia, J. (2025). Deep fuzzy fusion network for joint hyperspectral and LiDAR data classification. Remote Sensing, 2025, 1–15.

[11] Wang, A., Lei, G., Dai, S., Wu, H., & Iwahori, Y. (20 25). Multiscale attention feature fusion based on improved transformer for hyperspectral image and LiDAR data classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 4124–4140.

[12] Xu, K., Wang, B., Zhu, Z., Jia, Z., & Fan, C. (2025). A contrastive learning enhanced adaptive multimodal fusion network for hyperspectral and LiDAR data classification. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–19.

[13] Zhao, G., Ye, Q., Sun, L., Wu, Z., Pan, C., & Jeon, B. (2023). Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–16.

[14] Cao, M., Zhao, G., Lv, G., Dong, A., Guo, Y., & Dong, X. (2024). Spectral–Spatial–Language Fusion Network for Hyperspectral, LiDAR, and Text Data Classification. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–15.

[15] Ding, K., Lu, T., Fu, W., Li, S., & Ma, F. (2022). Global–Local Transformer Network for HSI and LiDAR Data Joint Classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13.

[16] Gao, H., Feng, H., Zhang, Y., Fei, S., Shen, R., Xu, S., & Zhang, B. (2024). Interactive Enhanced Network Based on Multihead Self-Attention and Graph Convolution for Classification of Hyperspectral and LiDAR Data. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–16.

[17] Han, W., Miao, W., Geng, J., & Jiang, W. (2024). CMSE: Cross-Modal Semantic Enhancement Network for Classification of Hyperspectral and LiDAR Data. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–14.

[18] Huang, W., Wu, T., Zhang, X., Li, L., Lv, M., Jia, Z., Zhao, X., Ma, H., &Vivone, G. (2025). MCFTNet: Multimodal Cross-Layer Fusion Transformer Network for Hyperspectral and LiDAR Data Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 12803–12818.

[19] Jia, S., Zhan, Z., & Xu, M. (2021). Shearlet-Based Structure-Aware Filtering for Hyperspectral and LiDAR Data Classification. Journal of Remote Sensing, 2021, 1–25.

[20] Li, Z., Sui, H., Luo, C., & Guo, F. (2023). Morphological Convolution and Attention Calibration Network for Hyperspectral and LiDAR Data Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16, 5728–5740.

[21] Li, Z., Wang, Y., Wang, L., Guo, F., Yang, Y., & Wei, J. (2024). Pseudolabeling Contrastive Learning for Semisupervised Hyperspectral and LiDAR Data Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 17099–17116.

[22] Lin, J., Gao, F., Qi, L., Dong, J., Du, Q., & Gao, X. (2025). Dynamic Cross-Modal Feature Interaction Network for Hyperspectral and LiDAR Data Classification. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–16.

[23] Liu, G., Song, J., Chu, Y., Zhang, L., Li, P., & Xia, J. (2025). Deep Fuzzy Fusion Network for Joint Hyperspectral and LiDAR Data Classification. Remote Sensing, 2025.

[24] Song, Q., Mo, F., Ding, K., Xiao, L., Dian, R., Kang, X., & Li, S. (2025). MCFNet: Multiscale Cross-Domain Fusion Network for HSI and LiDAR Data Joint Classification. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–12.

[25] Tang, X., Zou, Y., Ma, J., Zhang, X., Liu, F., & Jiao, L. (2024). Multiple Information Collaborative Fusion Network for Joint Classification of

Hyperspectral and LiDAR Data. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–16.

[26] Wang, X., Song, L., Feng, Y., & Zhu, J. (2025). S3F2Net: Spatial-Spectral-Structural Feature Fusion Network for Hyperspectral Image and LiDAR Data Classification. IEEE Transactions on Circuits and Systems for Video Technology, 35, 4801–4815.

[27] Xu, K., Wang, B., Zhu, Z., Jia, Z., & Fan, C. (2025). A Contrastive Learning Enhanced Adaptive Multimodal Fusion Network for Hyperspectral and LiDAR Data Classification. IEEE Transactions on Geoscience and Remote Sensing, 63, 1–19.

[28] Yang, B., Wang, X., Xing, Y., Cheng, C., Jiang, W., & Feng, Q. (2024). Modality Fusion Vision Transformer for Hyperspectral and LiDAR Data Collaborative Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 17052–17065.

[29] Zhao, Y., Bao, W., Xu, J., & Xu, X. (2024). BIHAF-Net: Bilateral Interactive Hierarchical Adaptive Fusion Network for Collaborative Classification of Hyperspectral and LiDAR Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 15971–15988.

[30] Berger, C., Riedel, F., Rosentreter, J., Stein, E., Hese, S., & Schmullius, C. (2025). Data fusion of hyperspectral and LiDAR data for urban microclimate modeling. Remote Sensing, 17, 1324.

[31] Zhang, Y., Peng, Y., & Zhou, C. (2023). Hyperspectral and LiDAR data classification for urban environment monitoring. Remote Sensing, 15, 4232.

[32] Roy, S. K., Deria, A., Hong, D., & Plaza, A. J. (2022). Hyperspectral and LiDAR data classification using joint CNNs and morphological feature learning. IEEE Transactions on Geoscience and Remote Sensing, 60, 12345.

[33] Yang, B., Wang, X., Xing, Y., Cheng, C., Jiang, W., & Feng, Q. (2024). Multimodal fusion models for land cover classification: A review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 17052–17065.