Global NEST Journal, **28** (XX), pp 1-11
Available online: 09/04/2026

GMVG: A ViT Embedded Graph-Mamba Network for HSI/LiDAR Land Cover Classification

Lirong Yin[1], Lei Wang[2], Siyu Lu[3], Guangyu Xu[4], Junmin Lyu[5] and Wenfeng Zheng[3]*

[1]Department of Hydrology and Atmospheric Sciences, University of Arizona, 85721, Tucson, AZ, USA
[2]Department of Geography & Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA
[3]School of Automation, University of Electronic Science and Technology of China, 611731, Chengdu, China
[4]School of the Environment, The University of Queensland, Brisbane St Lucia, QLD 4072 Australia
[5]School of Artificial Intelligence, Guangzhou Huashang University, Guangzhou 511300, China

*Correspondence e-mail: winfirms@ieee.org

Abstract

In recent years, the rapid development of geospatial intelligence and remote sensing technologies has brought new ways to study land cover. By combining different data sources, such as hyperspectral and LiDAR data, the AI model's land cover classification accuracy reaches new heights. Yet, due to the difference in data extraction and processing, the fusion of two data types is challenging. This paper presents a novel model that combines two input data sources, hyperspectral and LiDAR, to improve the application of machine learning (ML) or artificial intelligence (AI) in land cover classification. This is a new network based on GNNs and Mamba, leveraging multi-source HSI and LiDAR data to enhance classification results. In this paper, the Gated Recurrent Units (GRUs) and Vision Transformers (ViTs) were selected to enhance performance. By integrating GRUs and ViTs into GNN and Mamba frameworks, the proposed networks aim to leverage the strengths of these components to address challenges in multi-source HSI and LiDAR data classification. All 3 models show outstanding performance across the 3 datasets (MUFFL, Trento, and Houston). By introducing cutting-edge and diverse ML/AI models and components tailored to different tasks, this paper aims to explore the application prospects of ML/AI in land cover studies that could benefit the wider community.

Keywords: Vision Transformers, Gated Recurrent Units, GNN, Mamba, LCLU, LiDAR, RS, Hyperspectral image

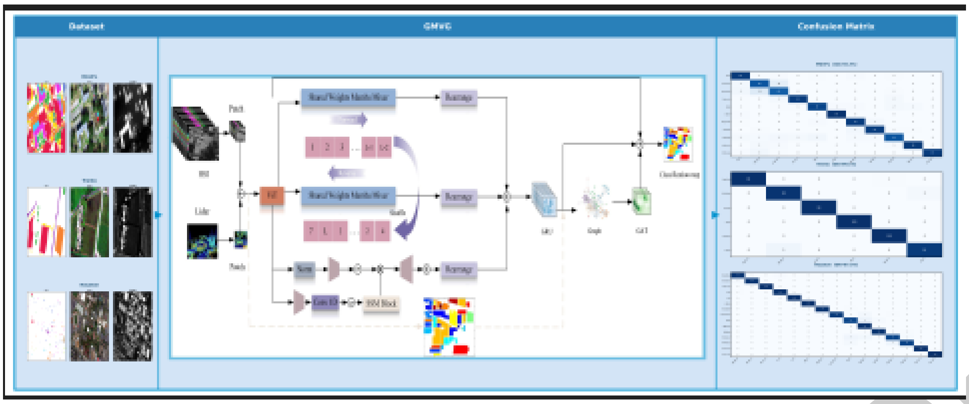
Received: 22/10/2025,

Accepted: 27/03/2026,

Available online: 09/04/2026

https://doi.org/10.30955/gnj.08129

Graphical abstract



1. Introduction

Over the past century, advances in remote sensing have significantly enhanced geospatial analysis by providing abundant and diverse Earth observation data (Zhang *et al.* 2022; Macarringue *et al.* 2022; Jaelani *et al.* 2025). The increasing availability of multi-source remote sensing datasets has opened new opportunities for comprehensive environmental monitoring while simultaneously introducing substantial analytical challenges (Kadhim *et al.* 2025; Madroñero Palacios and Muñoz Guerrero 2025). Among these data types, hyperspectral imagery (HSI) plays a critical role in environmental monitoring and mineral exploration due to its rich spectral resolution, which enables detailed characterization of surface materials (Zhang *et al.* 2024; Sun *et al.* 2022).

Despite their potential, HSIs face inherent challenges. Due to its passive nature, HSIs are sensitive to interference from atmospheric conditions and other physical factors, including instrumental limitations. These factors have a huge impact on the results in particle applications. For example, there is high variability in response in the "Stressed Grass" class in the Houston2013 dataset. This variability is also found in other HSI datasets (Zhang *et al.* 2024; Sun *et al.* 2022; Xie *et al.* 2021; Du *et al.* 2021). These challenges highlight the need for complementary data sources to enhance robustness.

In order to overcome the limitations that come with the use of hyperspectral imaging (HSI)-based classification, the synergistic approach of Light Detection and Ranging (LiDAR) combined with HSI has become a practical and complementary solution. Unlike HSI, LiDAR is an active remote sensing method, which emits laser pulses and observes the reflected pulses. It is capable of retrieving high-resolution elevation data and the spatial structure data, which are extremely beneficial for terrain classification and the process of defining objects (Zhang *et al.* 2024). This complementary spatial–structural information is highly valuable for terrain analysis and object delineation. The fusion of HSI and LiDAR data has been widely studied, and currently used methods can be categorized into three broad categories: data fusion, feature fusion, and decision fusion strategies (Hussain *et al.* 2024; Ignatious 2024; Qureshi *et al.* 2023). The most common of these is feature-level fusion, which gives a good trade-off between the computation complexity and classification performance by fusing multiple bits of information into one feature representation.

Traditional feature-level fusion methods include filter-based algorithms such as morphological profiles (MPs), attribute profiles (APs), and extinction profiles (EPs), which effectively capture shallow spatial morphological structures while preserving geometric information (Zhao and Zhang 2024; Kumar *et al.* 2024). Their simplicity and efficiency make them attractive to many applications; however, they are susceptible to the Hughes phenomenon because of high-dimensional feature spaces and an additional need for processing to enhance the outputs.

To mitigate the curse of dimensionality, low-rank representation models have been developed to project high-dimensional data into compact subspaces while retaining discriminative characteristics (Butt *et al.* 2023). Composite kernel techniques are another widely used technique that splits spectral, spatial, and elevation data and then assembles them with different kernel learning models. These are effective strategies that leverage the localized methods of filtering and have proven quite precise in challenging classification problems. Nevertheless, they are limited in their generalizability and adaptability to a variety of different or new settings as they are dependent on manually designed feature extractors, thus restricting their application in a more dynamic or broader context.

With the rapid advancement of deep learning, Convolutional Neural Networks (CNNs) have become dominant in HSI–LiDAR classification tasks due to their hierarchical feature extraction capabilities and strong representational power (Alzubaidi *et al.* 2021; Kattenborn *et al.* 2021). CNN-based models excel at feature fusion, a process that extracts features from multiple modalities and combines them to improve classification performance. For example, Chen *et al.* used CNNs to get different types of data and then used deep network layers to improve discriminative and invariant representations. This led to classification by logistic regression (Wang *et al.* 2021). Zeng *et al.* suggested a cross-modal hierarchical frequency network for frequency domain classification, and Han *et al.* proposed a model with multi-scale convolutional kernels and Gaussian-weighted matrices to capture both spatial and spectral features at the same time, resulting in strong classification results (Zeng *et al.* 2024). These methods have demonstrated strong performance; however, CNNs inherently emphasize local receptive fields and may struggle to capture long-range contextual dependencies.

Transformer-based architecture has recently emerged as a powerful alternative due to its self-attention mechanisms, which effectively model global dependencies. Vision Transformers (ViTs) treat images as sequences of patches, enabling direct modeling of long-range spatial relationships. In remote sensing applications, transformer-based models have shown superior performance in handling complex, high-dimensional, and multi-source data. Xue *et al.* put forward a deep hierarchical vision transformer that uses multi-head attention processes to improve domain generalization across a variety of landscapes (Xue *et al.* 2022). Wang *et al.* created a dual-branch hybrid network that uses both CNNs and transformers to take advantage of the different features in data from several sources (Wang *et al.* 2023). Yao *et al.* also created a vision transformer framework that included a cross-modality attention module to find fine-grained spatial correlations at the pixel level (Yao *et al.* 2024).

Nevertheless, transformer-based methods still face limitations, including insufficient cross-modal interaction modeling and inadequate preservation of edge and texture information, both of which are critical for fine-grained land-use and land-cover (LULC) classification.

State Space Models (SSMs) have emerged as a powerful option to all the above problems. They can express long-range dependencies with almost linear computational complexity through convolutional processes. Mamba is an SSM-based architecture, which is characterized by high training and inference efficiency. This is due to the fact that it employs time-varying parameters and maximizes hardware performance. RSMamba is a special SSM model constructed through the Mamba structure, which is developed to be applied in categorizing remote sensing photographs (Chen *et al.* 2024). RSMamba splits input photos into overlapping patch tokens and processes them forwardly, backwardly, and randomly with common parameterized Mamba blocks. This structure is permissible in extracting worldwide contextual interconnections and making computation manageable. This makes it suitable for large-scale pre-training jobs that do not require many resources. The network architecture used in this study is based on RSMamba, as it is capable of being modified and expanded.

In parallel, Graph Neural Networks (GNNs) have gained prominence for modeling relational structures in graph-formatted data. Through message-passing mechanisms, GNNs aggregate information from neighboring nodes to encode both local and global contextual relationships. Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) further enhance representation learning via spectral filtering and attention-based aggregation (Munir *et al.* 2024; Li *et al.* 2024; Han *et al.* 2022, 2023). Despite challenges such as over-smoothing and scalability, recent advancements in sparse computation and adaptive attention mechanisms have significantly improved their applicability in geospatial analysis.

Recurrent architectures, particularly Gated Recurrent Units (GRUs), provide efficient mechanisms for modeling sequential dependencies (Chung *et al.* 2014). GRUs are ideal for modeling time series data (i.e., sequential data) since information flow in the network is governed by their gating mechanisms. The update gate controls the intensity of the retention and transmission of old information, which simplifies the learning of long-term dependencies. The reset gate, conversely, determines the amount of the past information to be added to the present input. This allows the model to concentrate on short-term trends as required. These dynamic gates enable GRUs to easily adapt to different time dependencies, improving the model's ability to learn from sequences of varying lengths. GRUs are simpler in comparison to Long Short-Term Memory (LSTM) networks. LSTMs contain three gates—input, forget, and output—and discrete memory cells. GRUs, conversely, incorporate these functions into two gates and do not use separate memory cells. This makes GRUs more suitable for training and inference as it reduces the number of parameters and decreases computational cost (Chung *et al.* 2014; Emmert-Streib *et al.* 2022; Sherstinsky 2020; DiPietro *et al.* 2020). GRUs are relatively effective in capturing short-term and long-term dependencies in spite of being structurally simple. They achieve similar or even higher performance in most situations than LSTMs. Their efficiency makes them well-suited for temporal modeling and sequential feature integration in multi-modal fusion frameworks.

Another architecture, called Vision Transformer or ViT, was proposed by Dosovitskiy *et al.* in 2020, extending transformer architectures to image processing by representing images as sequences of patches enriched with positional encodings. Through multi-head self-attention mechanisms, ViTs effectively model global spatial interactions. Although they require large-scale pre-training due to limited inductive biases, variants such as DeiT (Data-efficient Image Transformer) enable data-efficient training (Yang *et al.* 2024; Anzum *et al.* 2024; Touvron *et al.* 2021). ViT has already established itself as a significant aspect of enhancing deep learning for computer vision by providing a novel method of image representation and being more effective in large-scale settings.

To address the limitations of existing approaches, this paper proposes a novel hybrid architecture that integrates

Graph Neural Networks (GNNs) and Mamba-based State Space Models for multi-source HSI–LiDAR classification. To further enhance performance, the framework incorporates GRUs for sequential feature refinement and Vision Transformers for global attention modeling.

By synergistically combining relational modeling (GNN), long-range efficient dependency learning (Mamba), sequential adaptation (GRU), and global contextual awareness (ViT), the proposed network effectively addresses the complexity and heterogeneity of multi-source remote sensing data. This integrated framework significantly improves classification accuracy and robustness for fine-grained land cover mapping. The remainder of this paper is organized as follows: Section 2 introduces the datasets and model architecture; Section 3 presents experimental results; Sections 4 and 5 provide discussion and conclusions.

2. Dataset and Method

2.1. Data description

For the testing of the model, 3 datasets, Trento, MUUFL, and Houston 2013 (**Figure 1**), are used for comparison with other methods (Zhang *et al.* 2024) since some of the existing models are only trained and tested with certain datasets. The datasets are all well-established HIS/LiDAR datasets for training and testing models for their ability to identify and analyze land cover.

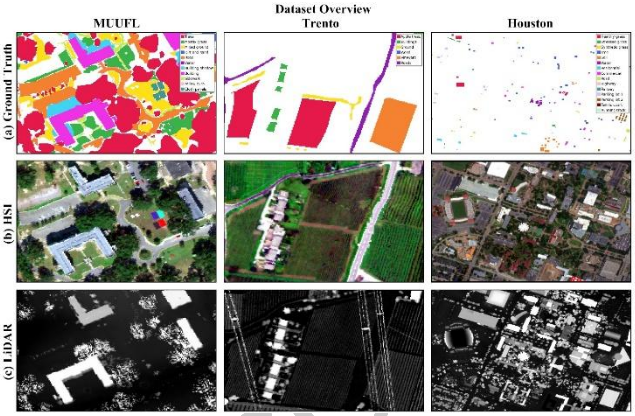


Figure 1. Datasets visualization of 3 datasets (a) Ground Truth; (b) Hyper-Spectral Image; (c) LiDAR Image.

The Trento dataset was collected over a rural area south of Trento, Italy. It comprises 600 × 166 pixels and includes LiDAR DSM data acquired using the Optech ALTM 3100EA sensor, alongside hyperspectral data captured by the AISA Eagle sensor, both with a spatial resolution of 1 meter. The hyperspectral data consists of 63 bands, covering wavelengths from 402.89 to 989.09 nm with a spectral resolution of 9.2 nm. Six classes of interest were identified in this dataset: Building, Woods, Apple Trees, Roads, Vineyard, and Ground.

The MUUFL dataset was collected near the University of Southern Mississippi Gulf Park in Long Beach, Mississippi (2010), using the Reflective Optics System Imaging Spectrometer (ROSIS) sensor. It consists of 325 × 220 pixels with 72 spectral bands, accompanied by LiDAR data containing elevation information from two raster sets. Due to noise, the first and last eight spectral bands were removed, leaving 64 bands. The dataset includes 53,687 ground-truth pixels, classified into 11 distinct urban land-cover categories.

The Houston 2013 dataset, supplied by the Hyperspectral Image Analysis Group and the NSF-funded Airborne Laser Mapping Center (NCALM) at the University of Houston, was originally developed for scientific research and featured in the IEEE GRSS Data Fusion Competition 2013. It consists of 144 spectral bands covering wavelengths from 0.38 to 1.05 μm. The dataset includes 349 × 1,905 pixels with a spatial resolution of 2.5 meters and is categorized into 15 distinct classes.

The detailed sample size for each dataset is listed in **Table 1**. In the MUFFL dataset, the number of samples is below 150, except for the Tree class. This choice is made based on the total number of samples, which are very small for some classes (Water, Yellow curb, and Cloth panels) but by comparison very large for the Tree class. The Trento dataset has a 150 train sample per class due to the more balanced total sample size for all the classes. The Houston dataset has 12 classes with 120 training samples and 3 classes with 75 training samples, also selected based on the total number of samples per class. The overall training size is kept low to reduce the overfitting issue and reduce the learning process. Since the dataset contains many bands (63, 72, and 144), we reduce the number of training samples to reduce the learning time.

2.2. Model structure

The proposed models (Graph Mamba with ViT-only, Graph Mamba with GRU-only, and GMVG-Graph Mamba with ViT and GRU) are multi-module network that utilizes Vision Transformer (ViT) (Liu *et al.* 2023), Graph Attention network(GNN-GAT) (Munir *et al.* 2024; Li *et al.* 2024; Han *et al.* 2023), and Gated Recurrent Units (GRUs) for feature extraction and classification.

The model structure is shown in **Figure 2**.

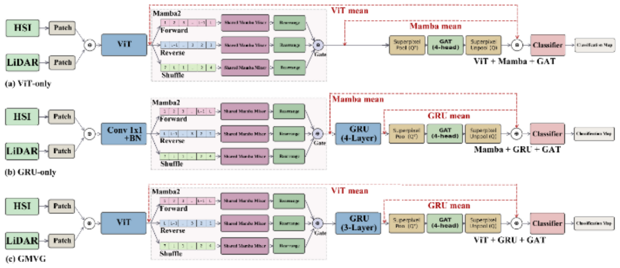


Figure 2. Model Structure of the three networks: (a) Graph-Mamba with ViT-only; (b) Graph-Mamba with GRU-only; (c) Graph-Mamba with ViT and GRU (GMVG)

The three models share a similar structure, where the GMVG has both ViT and GRU added to the process, whereas the ViT-only and GRU-only models contain only one of the modules. The GMVG first extracts the patch, uses ViT (detailed structure shown in **Figure 3**) to extract features (first feature group), and then uses Multipath Mamba (Chen *et al.* 2024) to extract features further. The extracted feature passes through 3 layers of GRU before converting to the linear layer (second feature group). Next, split an image into multiple polygons, each polygon as a node of the graph, forming a graph structure. Then, GAT was used to extract graph features (third feature group), and finally, a linear layer was used to get the final classification result using the 3 feature groups extracted and processed by different modules.

3. Result

Model performance is quantified using four key metrics—overall accuracy (OA), average accuracy (AA), the kappa coefficient, and per-class classification accuracy—with higher scores on each metric denoting superior classification results.

A higher value for each indicator indicates a better classification effect. The calculation functions are listed below:

$$OA = \frac{N_c}{N_a} \tag{1}$$

$$AA = \frac{1}{K}\sum_{j=1}^{K}\frac{N_c^j}{N_a^j} \tag{2}$$

$$\text{Kappa} = \frac{OA - P_e}{1 - P_e} \tag{3}$$

Where N_c and N_a are the number of correctly classified pixels and the total test sample size, and N_c^j and N_a^j are those numbers for class j. The formulation of Kappa addresses class imbalance through the hypothetical probability of chance agreement P_e, are calculated by:

$$P_e = \frac{\sum_{j=1}^{K} N_a^j \times N_p^j}{(N_a)^2} \tag{4}$$

The overall OA, AA, and Kappa of the 4 different models are shown in **Table 2**. Across MUUFL, Trento, and Houston, the confusion matrices and class-wise accuracy tables highlight two consistent themes. The datasets differ in intrinsic difficulty, and methods with stronger multi-modal/structural modeling tend to reduce systematic confusion that affects minority or spatially thin classes. **Table 1** reinforces this observation: MUUFL yields the lowest OA/AA/κ across methods, Trento approaches saturation for all models, and Houston lies in between, where residual errors are concentrated in spectrally similar urban surfaces.

Table 1. Sample size by class of the 3 datasets

MUFFL			Trento			Houston 2013		
Class	**Train**	**All**	**Class**	**Train**	**All**	**Class**	**Train**	**All**
Trees	550	23246	Healthy grass	150	4034	Healthy grass	120	1131
Mostly grass	150	4270	Stressed grass	150	2903	Stressed grass	120	1134
Mixed ground	150	6882	Synthetic grass	150	479	Synthetic grass	120	577
Dirt and sand	150	1826	Tree	150	9123	Tree	120	1124
Road	150	6687	Soil	150	10501	Soil	120	1122
Water	80	466	Water	150	3174	Water	75	250
Building shadow	150	2233				Residential	120	1148
Building	150	6240				Commercial	120	1124
Sidewalk	150	1385				Road	120	1132
Yellow curb	80	183				Highway	120	1107
Cloth panels	80	269				Railway	120	1115
						Parking lot 1	120	1113
						Parking lot 2	75	394
						Tennis court	75	353
						Running track	120	540

Table 2. OA, AA, and kappa results of the 4 different models

Model	Dataset	OA	AA	Kappa
CMFAEN	MUUFL	0.928	0.7945	0.9037
	Trento	0.9978	0.9962	0.9965
	Houston13	0.9409	0.947	0.9339
GMV	MUUFL	0.9046	0.7507	0.8738
	Trento	0.9894	0.983	0.9858
	Houston13	0.9689	0.9708	0.9664
GMG	MUUFL	0.8943	0.7495	0.8602
	Trento	0.9933	0.9901	0.991
	Houston13	0.9772	0.9755	0.9699
GMVG	MUUFL	0.9233	0.925	0.8972
	Trento	0.9912	0.9863	0.9882
	Houston13	0.975	0.9761	0.972

For MUUFL, the confusion matrices (**Figure 4**) reveal clear robustness gaps among the four methods. Although all approaches perform well on dominant classes such as Trees and Building, the single-stream baselines show substantial off-diagonal leakage for spectrally similar vegetation/material categories and for thin urban objects. ViT-only (OA = 90.46%, AA = 75.07%, κ = 87.38) exhibits notable confusion between Mostly grass and Mixed ground. ViT-only performs poorly on boundary-sensitive minority classes, including Water (43.29%), Sidewalk (44.72%), and Yellow curb (41.34%). GRU-only (OA = 89.43%, AA = 74.95%, κ = 86.02) reduces some confusions in a few categories, but remains weak on fine-grained urban features, with particularly low accuracies for Sidewalk (31.23%) and Yellow curb (8.38), consistent with heavy misclassification concentrated in those rows of the confusion matrix.

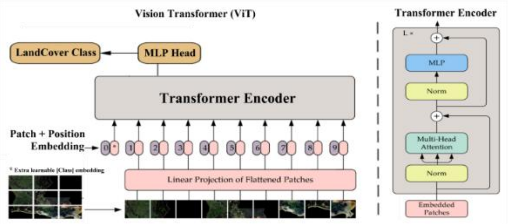


Figure 3. Detailed structure of ViT (Vision Transformer).

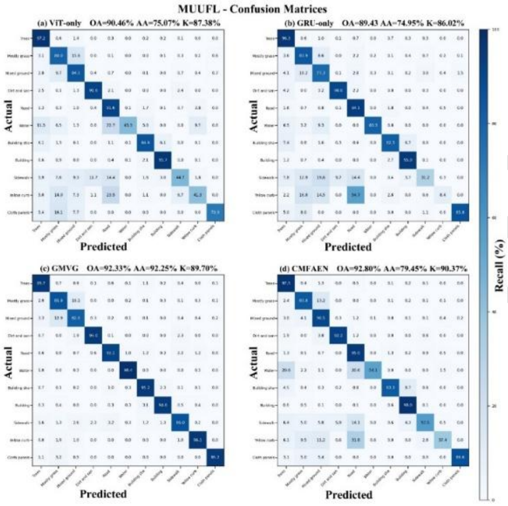


Figure 4. Confusion Matrix on MUUFL dataset of (a) ViT-only, (b) GRU-only, (c) GMVG, and (d) CMFAEN (Zhang *et al.* 2024).

In contrast, GMVG produces a much cleaner diagonal and substantially reduced off-diagonal mass, achieving OA = 92.33%, AA = 92.25%, and κ = 89.70, as also reflected in **Table 1**. The key improvement is not merely higher OA, but the dramatic increase in class-wise balance, indicating that GMVG preserves decision boundaries for categories that are typically difficult under HSI/LiDAR fusion due to mixed pixels, spectral ambiguity, and limited samples. This is strongly supported by **Table 3**, where GMVG sharply improves minority and thin classes such as Water (98.45%), Sidewalk (85.99%), and Yellow curb (96.12), while remaining competitive on dominant categories (e.g., Road and Building). Although CMFAEN achieves slightly higher OA (92.80%) and κ (90.37), its AA is much lower (79.45), indicating that its gains are less uniform across classes; this imbalance is evident in weak minority-class performance such as Water (54.11%), Sidewalk (57.48%), and Yellow curb (37.43), consistent with a more diffuse off-diagonal structure. Overall, MUUFL results indicate that MVG's main advantage is robust, class-balanced improvements, rather than accuracy gains concentrated on the easiest or largest categories. The result classification maps are shown in **Figure 5**.

For Trento, all four models achieve near-saturated performance, as shown in **Figure 6** and **Table 4**, with strong diagonal dominance and minimal off-diagonal leakage, consistent with the overall metrics in **Table 3** (all OA values ≥ 98.94%). ViT-only already provides high reliability (OA = 98.94%, AA = 98.30%, κ = 98.58), with perfect recognition of Ground and Wood (100% each), while residual errors are mainly localized to confusion between Roads and Buildings (Roads = 94.21%, Buildings = 96.19%). GRU-only further improves consistency (OA = 99.33%, AA = 99.01%, κ = 99.10), notably boosting Buildings (99.24%) and Roads (96.63%), indicating better separability when spectral cues are similar but contextual patterns help disambiguation. GMVG remains highly competitive (OA = 99.12%, AA = 98.63%, κ = 98.82), showing stable performance across vegetation types such as Apple trees (99.28%) and Vineyard (99.66%), but still trails GRU-only and CMFAEN on the most confusable urban surface (Roads, 95.37%)

Table 3. Classification accuracy by class of MUUFL dataset

Class	ViT-only	GRU-only	MVG	CMFAEN
Trees	97.2	96.25	95.74	97.51
Mostly grass	79.98	83.9	85.87	83.76
Mixed ground	84.49	77.29	82.03	90.51
Dirt and sand	91.57	89.99	94.99	92.24
Road	91.6	94.07	92.14	95.01
Water	43.29	80.3	98.45	54.11
Building shadow	84.95	82.25	95.25	83.26
Building	95.7	94.96	94.6	98.01
Sidewalk	44.72	31.23	85.99	57.48
Yellow curb	41.34	8.38	96.12	37.43
Cloth panels	70.88	85.82	95.24	86.59
OA (%)	90.46	89.43	92.33	92.8
AA (%)	75.07	74.95	92.25	79.45
Kappa (%)	87.38	86.02	89.7	90.37

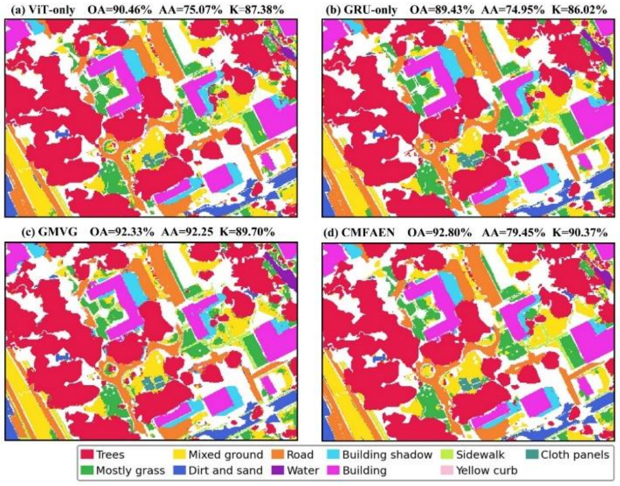


Figure 5. Classification Map of MUUFL dataset

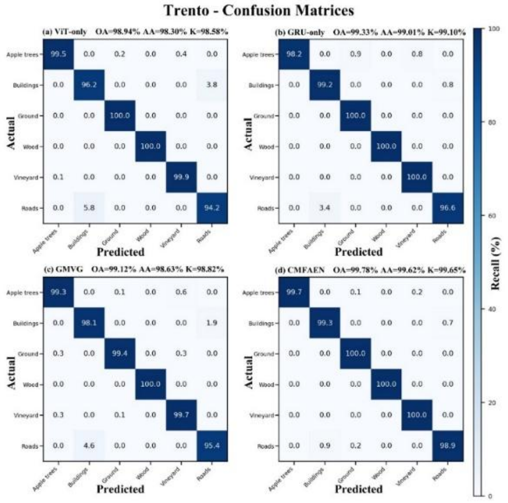


Figure 6. Confusion Matrix on Trento dataset of (a) ViT-only, (b) GRU-only, (c) GMVG, and (d) CMFAEN

Table 4. Classification accuracy by class of the Trento

Class	ViT-only	GRU-only	MVG	CMFAEN
Apple trees	99.46	98.22	99.28	99.67
Buildings	96.19	99.24	98.07	99.35
Ground	100	100	99.39	100
Wood	100	100	100	100
Vineyard	99.91	99.96	99.66	99.99
Roads	94.21	96.63	95.37	98.91
OA (%)	98.94	99.33	99.12	99.78
AA (%)	98.3	99.01	98.63	99.62
Kappa (%)	98.58	99.1	98.82	99.65

Table 5. Classification accuracy by class of the Houston 2013 dataset

Class	ViT-only	GRU-only	MVG	CMFAEN
Healthy grass	98.32	98.5	98.14	82.15
Stressed grass	97.8	96.38	98.32	96.9
Synthetic grass	100	99.31	99.83	99.21
Tree	99.11	98.67	99.38	95.45
Soil	99.82	100	100	99.91
Water	100	99.2	100	100
Residential	97.82	95.82	98.78	97.01
Commercial	94.84	94.48	95.91	95.35
Road	90.72	93.55	92.31	92.07
Highway	99.19	99.1	99.37	95.75
Railway	98.74	99.55	98.74	99.43
Parking lot 1	90.48	97.57	91.91	81.36
Parking lot 2	89.34	99.75	93.15	87.37
Tennis court	100	100	99.72	99.19
Running track	100	100	100	100
OA (%)	96.89	97.72	97.5	94.09
AA (%)	97.08	97.55	97.61	94.7
Kappa (%)	96.64	96.99	97.2	93.39

The strongest overall performance is achieved by CMFAEN (OA = 99.78%, AA = 99.62%, κ = 99.65), driven largely by improved handling of the built-environment confusion, especially Roads and Buildings. Collectively, Trento results suggest that the dataset's class separability is high, and performance differences are primarily dictated by subtle Roads–Buildings boundary ambiguity. The result classification maps are shown in **Figure 7**.

For Houston, the confusion matrices again show strong discrimination across most categories, as shown in **Figure 8** and **Table 5**, but with meaningful differences in robustness driven by specific hard classes. ViT-only achieves high performance (OA = 96.89%, AA = 97.08%, κ

= 96.64), with near-perfect accuracies for Synthetic grass (100%), Soil (99.82%), Water (100%), Tennis court (100%), and Running track (100%). However, its errors cluster in spectrally similar urban surfaces, particularly Road (90.72%) and the two parking-lot categories (Parking lot 1 = 90.48%, Parking lot 2 = 89.34). GRU-only improves global accuracy (OA = 97.72%, AA = 97.55%, κ = 96.99) and substantially strengthens the most confusing categories—most notably Parking lot 1 (97.57%), Parking lot 2 (99.75%), and Road (93.55)—suggesting that sequential/contextual modeling helps separate surfaces with subtle material and structural differences. GMVG delivers the most balanced overall outcome, achieving OA = 97.50%, the highest AA = 97.61%, and the best κ = 97.20%, matching **Table 4**, indicating strong agreement beyond chance and consistently reliable class-wise behavior. Its stability is reflected in improved residential and commercial discrimination (Residential = 98.78%, Commercial = 95.91) while preserving near-perfect performance on structurally distinct categories. In contrast, CMFAEN is clearly less robust on the Houston 2013 dataset (OA = 94.09%, AA = 94.70%, κ = 93.39), with pronounced drops for Healthy grass (82.15%) and Parking lot 1 (81.36), aligning with stronger off-diagonal leakage in the confusion matrix. The result classification maps are shown in **Figure 9**.

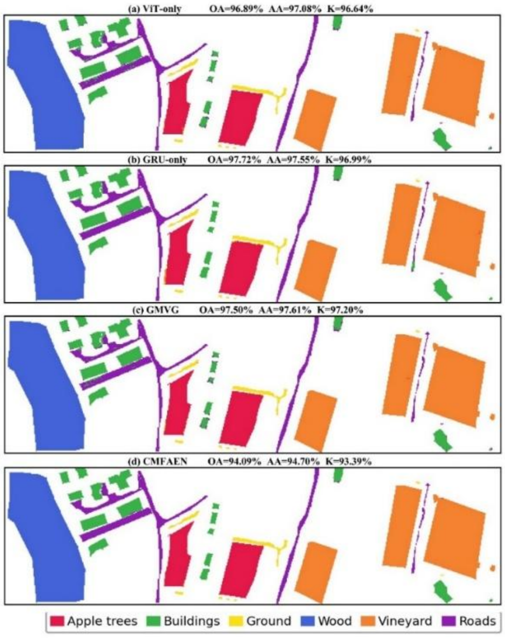


Figure 7. Classification map of the Trento dataset

Taken together, MUUFL is the most challenging benchmark, Trento is the easiest, and Houston occupies an intermediate regime where a few confusing urban classes dominate remaining errors. Within this landscape, GMVG's primary strength is its ability to improve class balance and reduce systematic confusions, particularly on the more difficult datasets (MUUFL and Houston), as evidenced by its markedly higher AA in MUUFL (92.25%) and the best κ in Houston (97.20%). This is especially important for practical LULC mapping, where thin or minority classes (e.g., curbs, sidewalks, water) are often the most operationally valuable and where OA alone can mask substantial weaknesses in rare-category performance.

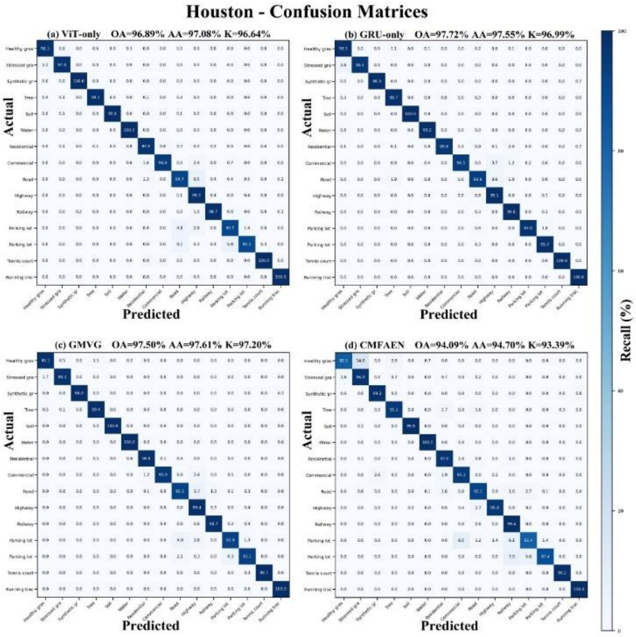


Figure 8. Confusion Matrix on Houston 2013 dataset of (a) ViT-only, (b) GRU-only, (c) GMVG, and (d) CMFAEN

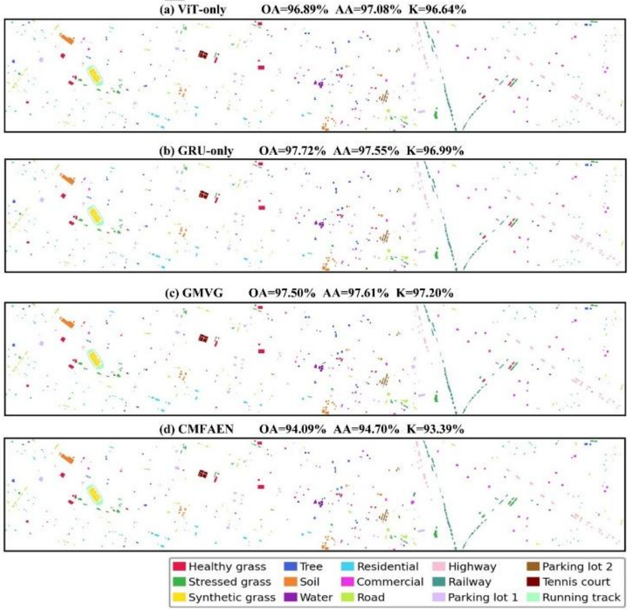


Figure 9. Classification map of the Houston 2013 dataset

4. Discussion

The experimental results on MUUFL, Houston, and Trento demonstrate that dataset characteristics strongly mediate the relative strengths of competing fusion architectures. Trento exhibits near-saturated performance for all methods, indicating high intrinsic class separability, whereas MUUFL is markedly more challenging due to the presence of spectrally similar materials and spatially thin urban classes. Houston occupies an intermediate regime: most categories are well separated, but residual errors concentrate in visually and spectrally similar impervious-surface classes (e.g., roads and parking lots). These patterns are consistently reflected in both the confusion matrices and the OA/AA/κ summaries.

On the Trento dataset, all four methods exhibit strong diagonal dominance and minimal off-diagonal leakage, indicating that most classes are inherently well separated. Even the single-stream baselines perform extremely well: ViT-only achieves OA = 98.94%, AA = 98.30%, and κ = 98.58, with perfect recognition of Ground and Wood (100% each), while its remaining errors are largely confined to subtle confusion between Roads and Buildings (Roads = 94.21%, Buildings = 96.19%). GRU-only further improves these built-environment categories (OA = 99.33%, AA = 99.01%, κ = 99.10), boosting Buildings to 99.24% and Roads to 96.63%, suggesting that sequential/contextual modeling provides additional separability when spectral cues are similar but spatial context differs. In this near-saturation regime, GMVG remains competitive (OA = 99.12%, AA = 98.63%, κ = 98.82) and performs strongly on vegetation discrimination (Apple trees = 99.28%, Vineyard = 99.66%), though it trails GRU-only and CMFAEN on the most confusable class (Roads = 95.37%). CMFAEN yields the best overall Trento results (OA = 99.78%, AA = 99.62%, κ = 99.65), primarily by reducing the small remaining Roads–Buildings ambiguity (Roads = 98.91%, Buildings = 99.35%). These findings indicate that when the dataset is relatively homogeneous, differences among architectures are driven by a narrow set of residual confusions rather than broad representational limitations.

In contrast, the MUUFL dataset reveals substantial robustness gaps and highlights the limitations of single-stream modeling. While ViT-only and GRU-only perform adequately on dominant categories (e.g., Trees and Building), both struggle severely on minority and spatially thin classes. ViT-only (OA = 90.46%, AA = 75.07%, κ = 87.38) exhibits pronounced confusion between vegetation and mixed-material classes (Mostly grass vs. Mixed ground) and performs poorly on Water (43.29%), Sidewalk (44.72%), and Yellow curb (41.34). GRU-only (OA = 89.43%, AA = 74.95%, κ = 86.02) reduces some confusions in a few categories but remains weak on fine-grained urban features, with particularly low performance on Sidewalk (31.23%) and Yellow curb (8.38), consistent with concentrated off-diagonal leakage in those rows of the confusion matrix. These results suggest that, on MUUFL, neither global attention alone (ViT-only) nor sequential dependency modeling alone (GRU-only) is sufficient to resolve the severe spectral and structural ambiguity present in thin, heterogeneous urban classes.

By comparison, GMVG provides markedly more balanced performance on MUUFL, achieving OA = 92.33%, AA = 92.25%, and κ = 89.70, with large gains in difficult categories (Water = 98.45%, Sidewalk = 85.99%, Yellow curb = 96.12%). Notably, although CMFAEN attains a slightly higher OA (92.80%) and κ (90.37), its AA (79.45%) remains much lower, indicating that its accuracy is less uniformly distributed and still limited on minority/high-confusion categories (Water = 54.11%, Sidewalk = 57.48%, Yellow curb = 37.43). Thus, MUUFL emphasizes that class-balanced reliability (AA) is the most informative indicator of robustness, and GMVG's primary advantage is its ability to reduce systematic confusion rather than only improve majority-class performance.

For Houston, all methods achieve strong discrimination overall, but key differences emerge in challenging urban surfaces and spectrally similar categories. ViT-only already yields high performance (OA = 96.89%, AA = 97.08%, κ = 96.64), with near-perfect results for classes such as Synthetic grass, Soil, Water, Tennis court, and Running track, but it shows lower accuracy for Road (90.72%) and both parking-lot categories (Parking lot 1 = 90.48%, Parking lot 2 = 89.34). GRU-only improves global metrics (OA = 97.72%, AA = 97.55%, κ = 96.99) and significantly strengthens the most confusing classes, especially Parking lot 1 (97.57%) and Parking lot 2 (99.75), suggesting that sequential/contextual modeling helps separate impervious surfaces with subtle material differences. GMVG achieves the most balanced overall behavior in the Houston dataset, reaching OA = 97.50% with the highest AA (97.61%) and best κ (97.20%), indicating consistent agreement beyond chance across classes. While GMVG does not outperform GRU-only on both parking-lot classes, it improves stability across the broader label space (e.g., Residential = 98.78%, Commercial = 95.91%, Highway = 99.37), reflecting fewer systematic biases toward easy categories. In contrast, CMFAEN shows reduced robustness on Houston (OA = 94.09%, AA = 94.70%, κ = 93.39), with pronounced drops for Healthy grass (82.15%) and Parking lot 1 (81.36), consistent with stronger off-diagonal leakage in its confusion matrix.

Architecturally, these outcomes align with the complementary strengths of the compared models. ViT-only benefits from global self-attention and is effective when classes are well separated or when large-scale contextual cues dominate, but it can underperform on thin or minority classes when spectral ambiguity is high. GRU-only provides efficient sequential dependency modeling that can improve separability for structured surfaces (e.g., roads and parking lots in Houston), yet it lacks explicit mechanisms for rich spatial–topological reasoning. GMVG integrates these advantages by combining ViT-based global context, GRU-based sequential refinement, and GNN message passing for spatial–topological interactions, while leveraging Mamba state-space modeling to capture long-range dependencies with near-linear complexity. This synergy is most evident on difficult benchmarks, where GMVG substantially improves AA by reducing systematic confusions in minority and high-overlap categories. Nevertheless, the small performance trade-offs observed in near-saturated settings (e.g., slightly lower Roads accuracy on Trento) suggest that additional refinement—such as uncertainty-guided loss weighting, adaptive class balancing, or targeted augmentation for the most confusable classes—could further improve stability.

Finally, the results underscore that model advances alone cannot fully overcome data constraints. Performance degrades most strongly in heterogeneous urban scenes with class imbalance, shadows, and high spectral overlap (as in MUUFL), highlighting the need for robust preprocessing, domain adaptation, and augmentation. Moreover, the limited availability of diverse HSI–LiDAR benchmarks restricts the evaluation of transferability. Future work would benefit from larger, geographically and thematically diverse datasets, particularly spanning complex urban–vegetation and wetland–built interfaces, to more rigorously test cross-scene generalization.

5. Conclusion

This study evaluated ViT-only, GRU-only, CMFAEN, and the proposed GMVG framework for HSI–LiDAR land-cover classification across MUUFL, Houston, and Trento. The results demonstrate that GMVG is particularly effective in improving class balance and reducing systematic confusions on challenging benchmarks. On MUUFL, GMVG achieves a substantial increase in class-wise reliability (AA = 92.25%) and markedly improves difficult minority/thin categories such as Water, Sidewalk, and Yellow curb. In Houston, GMVG attains the strongest overall agreement beyond chance (κ = 97.20) and the highest AA (97.61), reflecting consistent performance across vegetation, built, and transportation classes. In Trento, where all models approach saturation, CMFAEN yields the best overall metrics (OA = 99.78%, AA = 99.62%), while GMVG remains competitive and demonstrates strong vegetation discrimination, with residual differences largely attributable to subtle Roads–Buildings boundary ambiguity.

Beyond accuracy improvements, these findings support the broader conclusion that combining graph-based spatial reasoning, global-context attention, and efficient long-range state-space modeling provides a practical pathway for mitigating spectral redundancy and local ambiguity in multi-modal remote sensing classification. However, limitations remain—especially for minority and highly overlapped classes under severe urban heterogeneity—suggesting future gains from class-aware learning strategies (e.g., adaptive reweighting, uncertainty-guided loss design) and stronger domain generalization.

Finally, while hyperspectral imagery offers unmatched spectral detail for fine-grained material discrimination, its limited temporal coverage constrains long-term monitoring applications. A promising direction is to integrate HSI–LiDAR fusion with dense multispectral time series (e.g., Landsat or Sentinel) to combine fine spectral–structural mapping with regular temporal sampling. Extending GMVG toward spatiotemporal fusion, incorporating uncertainty estimation and (where appropriate) physics-informed constraints, would help move from static land-cover mapping toward dynamic, process-aware monitoring frameworks that better support long-term environmental assessment and decision-making.

Statements and Declarations

The evaluation datasets that support the findings of this study are openly available at https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. The newly constructed LA dataset is available upon request. Please contact the corresponding author for a copy of the dataset.

Disclosure Statement

The authors report there are no competing interests to declare.

References

Alzubaidi, L., *et al.*, Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data, 2021. 8(1): p. 53.

Anzum, H., M.N.S. Sammo, and S. Akhter. Leveraging Data Efficient Image Transformer (DeIT) for Road Crack Detection and Classification. in 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS). 2024. IEEE.

Butt, M.H.F., *et al.* Mitigating Hughes Phenomenon: Improving Hyperspectral Imaging Classification Through Active Learning for Generalization Enhancement. in 2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). 2023. IEEE.

Chen, K., *et al.*, Rsmamba: Remote sensing image classification with state space model. IEEE Geoscience and Remote Sensing Letters, 2024.

Chung, J., *et al.*, Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.

DiPietro, R. and G.D. Hager, Deep learning: RNNs and LSTM, in Handbook of medical image computing and computer assisted intervention. 2020, Elsevier. p. 503-519.

Dosovitskiy, A., *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. in International Conference on Learning Representations. 2020.

Du, B., *et al.*, Mapping wetland plant communities using unmanned aerial vehicle hyperspectral imagery by comparing object/pixel-based classifications combining multiple machine-learning algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021. 14: p. 8249-8258.

Emmert-Streib, F., *et al.*, An Introductory Review of Deep Learning for Prediction Models with Big Data. Front Artif Intell, 2020. 3: p. 4.

Han, K., *et al.*, Vision gnn: An image is worth graph of nodes. Advances in Neural Information Processing Systems, 2022. 35: p. 8291-8303.

Han, Y., *et al.* Vision hgnn: An image is more than a graph of nodes. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Hussain, M., *et al.*, A Comprehensive Review On Deep Learning-Based Data Fusion. IEEE Access, 2024.

Ignatious, H.A., A novel data fusion fr amework to enhance contextual awareness of the autonomous vehicles for accurate decision making. 2024.

Jaelani, L.M., *et al.*, Mapping Mangrove Species Distribution and Density Using Sentinel-2 Satellite Imagery and Spectral Analysis. Journal of Human, Earth, and Future, 2025. 6(1): p. 1-11.

Kadhim, N. and N.M. Salih, Assessment of urban changes at the residential neighbourhood level based on satellite imageries. Civil Engineering Journal, 2025. 11(1): p. 58-72.

Kattenborn, T., *et al.*, Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS journal of photogrammetry and remote sensing, 2021. 173: p. 24-49.

Kumar, A., *et al.*, Correlation filter based single object tracking: A review. Information Fusion, 2024: p. 102562.

Li, L., *et al.*, KNN-GNN: A powerful graph neural network enhanced by aggregating K-nearest neighbors in common subspace. Expert Systems with Applications, 2024. 253: p. 124217.

Liu, X., *et al.* EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Macarringue, L.S., É.L. Bolfe, and P.R.M. Pereira, Developments in land use and land cover classification techniques in remote sensing: A review. Journal of Geographic Information System, 2022. 14(1): p. 1-28.

Madroñero Palacios, S.M. and D.A. Muñoz Guerrero, Projections of Land-Cover Change in a Tropical High-Andean Lake. Civil Engineering Journal, 2025. 11(9): p. 3840-3856.

Munir, M., *et al.* GreedyViG: Dynamic Axial Graph Construction for Efficient Vision GNNs. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

Qureshi, I., *et al.*, Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. Information Fusion, 2023. 90: p. 316-352.

Sherstinsky, A., Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020. 404: p. 132306.

Sun, L., *et al.*, Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 2022. 60: p. 1-14.

Touvron, H., *et al.* Training data-efficient image transformers & distillation through attention. in International conference on machine learning. 2021. PMLR.

Wang, P.-Y., *et al.*, Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE access, 2021. 9: p. 55244-55259.

Wang, Q., *et al.*, DBCT-Net: A dual branch hybrid CNN-transformer network for remote sensing image fusion. Expert Systems with Applications, 2023. 233: p. 120829.

Xie, Z., *et al.*, Multilayer global spectral–spatial attention network for wetland hyperspectral image classification. IEEE transactions on geoscience and remote sensing, 2021. 60: p. 1-13.

Xue, Z., *et al.*, Deep hierarchical vision transformer for hyperspectral and LiDAR data classification. IEEE Transactions on Image Processing, 2022. 31: p. 3095-3110.

Yang, J., A. Rusak, and A. Belozubov. Enhancing Brain Tumor Classification Using Data-Efficient Image Transformer. in 2024 International Russian Smart Industry Conference (SmartIndustryCon). 2024. IEEE.

Yao, T., *et al.*, Cross-modality interaction reasoning for enhancing vision-language pre-training in image-text retrieval. Applied Intelligence, 2024. 54(23): p. 12230-12245.

Zeng, Z., *et al.* Joint Classification of Hyperspectral and Lidar Data Using Cross-Modal Hierarchical Frequency Fusion Network. in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. IEEE.

Zhang, X., Y.n. Zhou, and J. Luo, Deep learning for processing and analysis of remote sensing big data: A technical review. Big Earth Data, 2022. 6(4): p. 527-560.

Zhang, Y., *et al.*, A cross-modal feature aggregation and enhancement network for hyperspectral and LiDAR joint classification. Expert Systems with Applications, 2024. 258: p. 125145.

Zhao, X. and M. Zhang, Review of Filter-Based Image Denoising Methods. International Journal of Computer Science and Information Technology, 2024. 2(2): p. 36-42.