Universe Scientific Publishing

DMTCANet: Dual-Branch Multiscale CNN and Token Cross Attention Fusion for Hyperspectral and LiDAR Data Classification

Yanfen Sun, Bian Bawangdui*

School of Information Science Technology, Tibet University, Lhasa 850000, China

Abstract: Advancements in remote sensing (RS) technology have highlighted the potential of jointly classifying Hyperspectral Images (HSI) and Light Detection and Ranging (LiDAR) data, leveraging the rich spectral information of HSI and the precise 3D structural details of LiDAR. While this combination improves classification accuracy, it presents challenges due to differences in data dimensions and semantic levels. Existing deep learning approaches often struggle to effectively extract features and capture interactions between these heterogeneous sources, and traditional CNNs suffer from limited receptive fields and detail loss in complex multi-scale scenarios. To address these issues, we propose DMTCANet, a novel joint classification network that combines a dual-branch multi-scale CNN with token cross-attention (TCA) fusion. The network incorporates a multi-scale hybrid convolution module to process HSI and LiDAR data, expanding the receptive field and capturing local and global information. A TCA fusion encoder further enhances deep interactions between the two data modalities, overcoming the limitations of insufficient feature integration. Experimental results on Trento, Houston2013, and MUUFL datasets demonstrate the effectiveness of DMTCANet, outperforming existing methods.

Keywords: Multiscale CNN; Token Cross Attention (TCA) Fusion; Remote Sensing (RS) Joint Classification; Tokenization; Hyperspectral Images (HSI); Light Detection and Ranging (LiDAR)

1. Introduction

With the development of remote sensing technology, hyperspectral images (HSI) have important applications in the classification of land features due to their rich spectral information. However, the use of HSI alone is susceptible to the limitations of high-dimensional redundancy, noise interference, and insufficient spatial information, which affect the classification accuracy. LiDAR data, on the other hand, compensates for the spatial structure deficiencies of HSI by providing accurate three-dimensional spatial information, but its spectral feature expression is weak [1]. To overcome their respective limitations, joint classification of HSI and LiDAR can make full use of their complementary advantages and significantly improve the classification results. However, existing methods still have deficiencies in mining multi-modal deep interactive information. Inspired by related research, this paper proposes a joint classification network DMTCANet based on a dual-branch multi-scale CNN and a cross-tag attention (TCA) fusion to comprehensively mine the multi-source features of HSI and LiDAR and their deep associations, thereby improving the classification accuracy [2]. Its main contributions include: 1) the DMTCANet framework is designed to improve classification performance through multi-scale feature extraction and deep interaction fusion; 2) a multi-scale hybrid convolution module is proposed to expand the receptive field and capture global and local information using multi-size convolution kernels; 3) a TCA fusion encoder is introduced to enhance the deep correlation between HSI and LiDAR and significantly improve the problem of insufficient information interaction.

2. Proposed method

This section provides a detailed explanation of the classification framework of DMTCANet, as shown in Fig.1, aiming to demonstrate its efficiency and practicality.

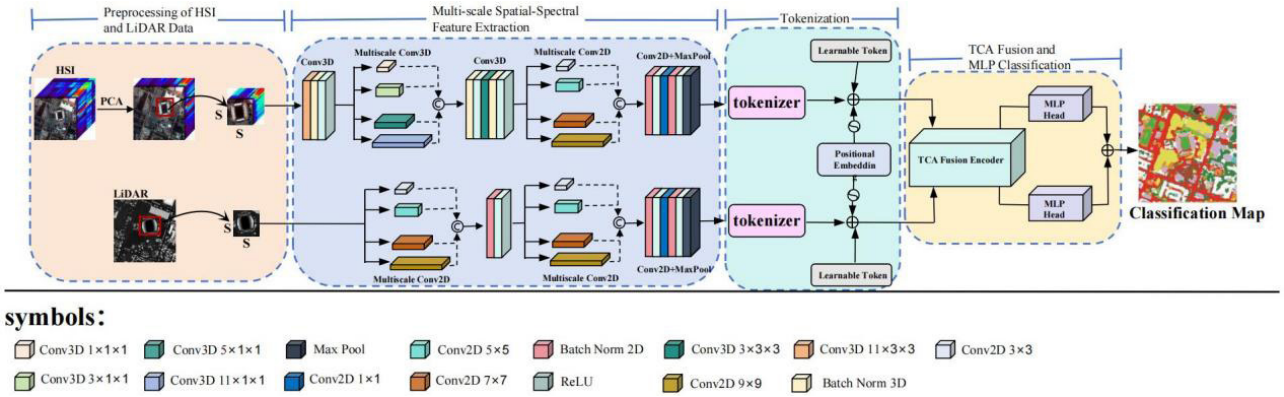

Fig.1: The model structure of our proposed DMTCANet.

2.1 HSI and LiDAR Data Preprocessing

Given raw HSI data $\boldsymbol{X}_H \in \mathbb{R}^{B\times C\times D}$ and corresponding LiDAR data $\boldsymbol{X}_L \in \mathbb{R}^{B\times C}$, where B×C is the spatial dimension and D is the number of spectral bands, PCA is applied to reduce the spectral dimensionality of the HSI data from D to b, resulting in $\boldsymbol{X}_{H_{pca}} \in \mathbb{R}^{B\times C\times b}$. Then, 3D patches of size S×S are extracted from each pixel of the reduced HSI data, forming small patch cubes $X_{H_P} \in \mathbb{R}^{S\times S\times b}$, while 2D patches of size S×S are extracted from the LiDAR data, forming small patches $X_{L_P} \in \mathbb{R}^{S\times S}$.

2.2 Multi-scale Spectral-Spatial Feature Extraction

For HSI data, first, the preprocessed cube X_{HP} is passed through the spectral-spatial feature encoder (SSFE) to extract joint spectral and spatial features, denoted as F_{ssf}. Next, these features are further processed by the spatial feature encoder (SFE) to extract spatial features, denoted as F_{sf}[4]. Finally, the maximum pooling operation (MP) is employed for generate the final spectral-spatial features M_{Hssf}. For LiDAR data, first, the preprocessed cube X_{LP} is passed through two layers of SFE to extract spatial features, denoted as F_{sf}. Then, MP is utilized to extract the elevation features from LiDAR, denoted as M_{Lsf}. The features M_{Hssf} and M_{Lsf} can be expressed as:

$$M_{H_{ssf}} = MP\left(F_{sf}\left(F_{ssf}\left(X_{H_P}\right)\right)\right) \quad (1)$$

$$M_{L_{sf}} = MP\left(F_{sf}\left(F_{sf}\left(X_{L_P}\right)\right)\right) \quad (2)$$

Additionally, the module integrates HSI and LiDAR data through a weighted fusion formula to form multimodal features, denoted as M. In addition to enhancing the model's classification capability for complex scenes, this approach also effectively lowers computational cost by utilizing spectral channel grouping and batch normalization (BN). The multimodal feature M can be determined as:

$$M = \omega \cdot M_{H_{ssf}} + (1-\omega)M_{L_{sf}} \quad (3)$$

Where represents the weighting coefficient, which can be modified.

2.3 Tokenization

To efficiently embed the features from HSI and LiDAR into the TCA network, tokenization is performed. The feature maps of HSI and LiDAR are flattened into vectors: $X_{HSI} \in \mathbb{R}^{m_H\times n_H\times v_H}$ and $X_{LiDAR} \in \mathbb{R}^{m_L\times n_L\times v_L}$, where m_H, n_H, v_H and m_L, n_L, v_L represent the height, width, and channels of the HSI and LiDAR feature maps, respectively [5]. The feature tokens are $T_{HSI} \in \mathbb{R}^{g_H\times v_H}$ and $T_{LiDAR} \in \mathbb{R}^{g_L\times v_L}$, where g_H and g_L are the number of tokens. By multiplying the input features with the learnable weights W_a and W_b, tokenization is performed to extract the key features. For the flattened features X_{HSI} and X_{LiDAR}, and T_{HSI} can be given by:

$$T_{HSI} = \underbrace{Softmax(X_{HSI}W_a)^T X_{HSI}}_{A_{HSI}} \quad (4)$$

$$T_{LiDAR} = \underbrace{Softmax(X_{LiDAR}W_b)^T X_{LiDAR}}_{A_{LiDAR}} \quad (5)$$

Where X_{HSI} W_a and X_{LiDAR} W_b represent 1×1 dot product operations. After this step, the obtained features are transposed, and the transposed feature groups are denoted as A_{HSI} and A_{LiDAR}. Finally, A_{HSI} and A_{LiDAR} are multiplied with X_{HSI} and X_{LiDAR} respectively to obtain the final feature tokens T_{HSI} and T_{LiDAR}.

2.4 TCA Fusion Encoder

In RS data, effective feature fusion is crucial for successfully constructing multimodal feature representations. This paper introduces a TCA fusion encoder to process HSI and LiDAR feature information. The TCA module utilizes the CLS token from one modality as a bridge to facilitate the exchange of information with tokens from the other modality, projecting the exchanged information back into the original feature tokens.

Taking HSI feature tokens as an example, its CLS token is merged with LiDAR feature tokens to align dimensions. The specific process is as follows:

$$t'_{cls\_HSI} = f_{HSI}\left(t_{cls_{HSI}}\right) \tag{6}$$

$$T_{tca\_HSI} = \left[t'_{cls_{HSI}} \cup \left(T_{LiDAR} - t_{cls_{LiDAR}}\right)\right] \tag{7}$$

Where $t_{cls\ HSI}$ has a dimension of $1\times1\times Z_H$ and t'_{clsHSI} matches the dimension of the LiDAR CLS token, $1\times z_L$. The new feature token T_{tcaHSI} replaces the LiDAR CLS token with the transformed HSI CLS token.

Subsequently, the TCA module interacts using t'_{clsHSI} as the query item. The mathematical formulation of this process is shown below.

$$\mathrm{Q} = t'_{cls\_HSI}U_q,\ K = T_{tca\_HSI}U_k,\ V = T_{tca_{HSI}}U_v \tag{8}$$

$$\mathrm{A} = \mathrm{Softmax}\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d/m}}}\right),\ \ \mathrm{TCA}(T_{tca\_HSI}) = AV \tag{9}$$

Where $U_q, U_k, U_v \in \mathbb{R}^{d\times(d/m)}$ are learnable weight matrices, d denotes the new embedding dimension size, and m is the number of attention heads. This method reduces complexity compared to traditional attention mechanisms.

Finally, combining layer normalization (LN) and residual connections, the output of the TCA module T_{outHSI} can be defined as:

$$Y'_{cls\_HSI} = t'_{cls\_HSI} + MH_{A\left(LN\left(T_{tca_{HSI}}\right)\right)} \tag{10}$$

$$Y_{cls\_HSI} = g_{HSI}\left(Y'_{cls_{HSI}}\right) \tag{11}$$

$$T_{out\_HSI} = \left[Y_{cls_{HSI}} \cup \left(T_{HSI} - t_{cls_{HSI}}\right)\right] \tag{12}$$

Similar to HSI, LiDAR features also undergo the same processing flow, with the output being T_{out_LiDAR}, where the learnable CLS token is defined as Y_{cls_LiDAR}.

2.5 MLP Layer Classification

After the TCA module, classification tokens and are passed through a multilayer perceptron (MLP) with two linear layers and GELU activations. The final layer applies Softmax to compute the classification probabilities, where the output dimension matches the number of classes. The two resulting probability vectors are summed, and the class with the highest probability is selected as the classification result for the pixel.

3. Experiments And Analysis

3.1 Dataset Description

To evaluate the proposed model's effectiveness and underlying rationale, experiments were carried out on three publicly available datasets, with a thorough discussion and analysis of the results. Table 1 presents detailed information about the three datasets.

Table I: A detailed description of the three datasets: Trento, Houston2013, and MUUFL.

Description	Sensor	Spectral Bands	Total Training Samples	Total Testing Samples	Image Size (pixels)	Spatial Size (m)
Trento	HSI(AISA Eagle)	63	819	29395	600166	1
	LiDAR(Optech ALTM 3100EA)	1				
Houston2013	HSI (NCALM)	144	2832	12197	3491905	2.5
	LiDAR (NCALM)	1				
MUUFL	HSI (AISA Eagle)	72	1650	52037	325220	1
	LiDAR (Optech ALTM 3100EA)	1				

3.2 Parameter Analysis

(1) Patch Size: The patch size determines the local image region considered by the model, affecting its ability to perceive texture, shape, and detailed information. Smaller patch sizes capture finer local features but may miss larger patterns, while larger patches capture broader context but may lose fine details. To explore its impact, we tested patch sizes of 7, 9, 11, 13, and 15. As shown in Fig. 2(a), the optimal patch size is 11 for all three datasets.

(2) Learning Rate: The learning rate determines the size of parameter adjustments, influencing both the speed and stability of the convergence process. A lower learning rate allows finer exploration but increases training time, while a higher rate accelerates training but risks missing the optimal solution. We tested learning rates of 0.0001, 0.0003, 0.0005, 0.001, and 0.005. As shown in Fig. 2(b), a learning rate of 0.001 gives the best performance across all datasets.

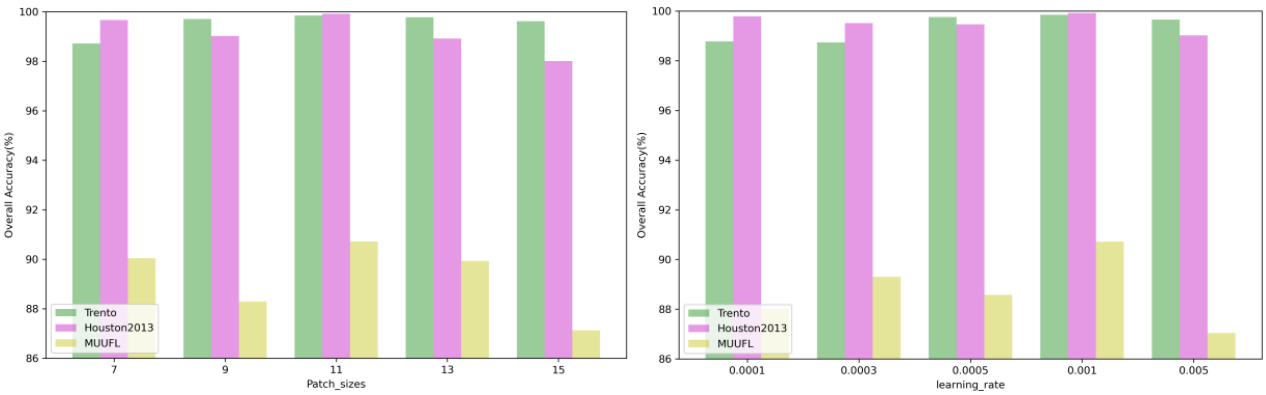

Fig.2: The effect of various parameters on the three datasets: (a) patch size, (b) learning rate.

3.3 Classification Results and Analysis

To demonstrate the effectiveness and robustness of the proposed network architecture, experiments were conducted to compare it with several well-known methods, such as CCRNet [8], CoupledCNN [9], MHST [7], EndNet [10], HCT [6], and FusAtNet [3]. The implementation of these methods followed the parameter configurations specified in their original papers.

Table II-III and Fig. 3-4, present the classification results and maps for various models on the Houston2013, and MUUFL datasets. For Houston2013, DMTCANet achieved 99.92%, significantly outperforming CCRNet (95.28%) and MHST (96.19%), with clearer maps and better classification in complex regions. On MUUFL, DMTCANet obtained an OA of 90.72%, surpassing MHST (88.71%) and CCRNet (83.12%), showing precise object boundaries and minimal confusion. These results highlight the superior performance and classification quality of DMTCANet across all datasets.

Table II: The classification accuracy (%) of different methods on the Houston2013 dataset.

No.	CCRNet	CoupledCNN	MHST	EndNet	HCT	FusAtNet	DMTCANet
1	93.92	83.06	98.05	83.09	82.86	83.10	99.81
2	94.46	82.79	98.22	79.45	83.23	96.05	100
3	99.82	94.51	99.26	99.40	93.06	100	99.80
4	99.52	93.24	99.59	91.86	95.64	93.09	98.76
5	99.69	99.75	99.35	99.88	98.69	99.43	100
6	99.44	98.83	99.67	96.87	95.54	100	100
7	96.57	91.56	95.59	84.42	92.79	93.53	98.97
8	92.39	81.19	90.77	76.26	79.74	92.12	97.81
9	90.66	87.12	89.20	72.14	83.51	83.63	99.33
10	92.61	63.22	96.77	54.83	60.41	64.09	100
11	95.27	91.56	94.32	85.53	93.54	90.13	99.53
12	91.28	90.43	93.82	95.48	86.89	91.93	99.61
13	96.81	90.09	97.33	71.93	92.28	88.42	97.19
14	100	98.56	100	100	98.81	100	100
15	99.75	97.66	100	99.74	99.78	99.15	100
OA(%)	95.28	87.76	96.19	83.90	88.03	89.98	99.92
AA(%)	96.15	89.65	96.80	85.86	89.68	94.65	99.94
k100	94.87	86.76	95.88	82.57	87.03	89.13	99.91

TableIII: The classification accuracy (%) of different methods on the MUUFL dataset.

No.	CCRNet	CoupledCNN	MHST	EndNet	HCT	FusAtNet	DMTCANet
1	84.67	87.15	91.81	86.15	90.87	90.75	91.99
2	84.19	86.78	85.70	82.67	86.09	74.20	86.43
3	67.13	75.34	72.78	77.18	73.63	64.45	86.02
4	96.12	96.17	89.81	92.30	96.24	87.49	96.54
5	83.78	89.76	88.29	91.86	82.88	87.22	88.62
6	99.12	98.85	100	98.73	100	100	100
7	86.45	90.92	93.05	88.96	91.98	92.54	94.58
8	95.23	96.88	95.52	92.99	89.11	93.06	94.09
9	67.14	68.98	82.19	81.94	75.06	71.77	83.89
10	83.78	97.18	92.68	93.94	78.79	82.11	90.91
11	98.57	99.11	99.04	99.16	91.60	97.61	99.16
OA(%)	83.12	88.43	88.71	86.55	86.94	85.45	90.72
AA(%)	86.08	90.34	90.08	89.63	86.93	85.56	92.02
k100	77.95	85.07	85.32	82.53	82.95	81.15	87.81

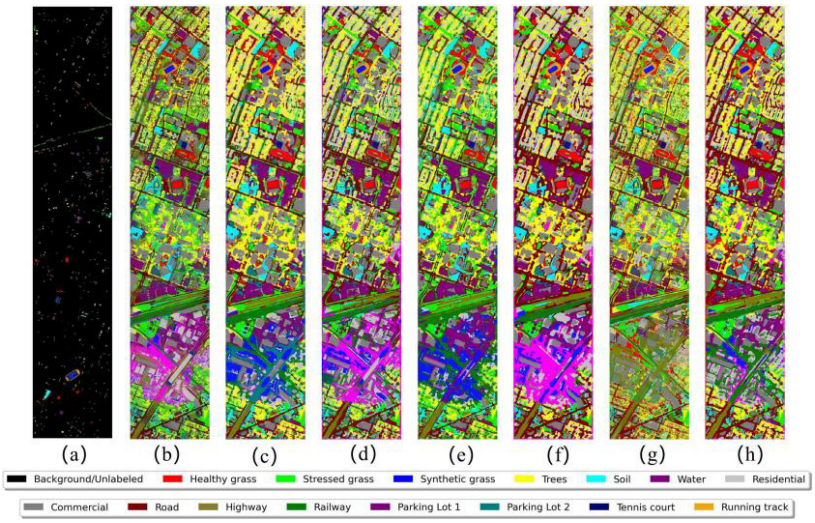

Fig.3: Classification maps of various methods for the Houston2013 dataset. (a) Ground truth, (b) CCRNet, (c) CoupledCNN, (d) MHST, (e) EndNet, (f) HCT, (g) FusAtNet, (h) DMTCANet.

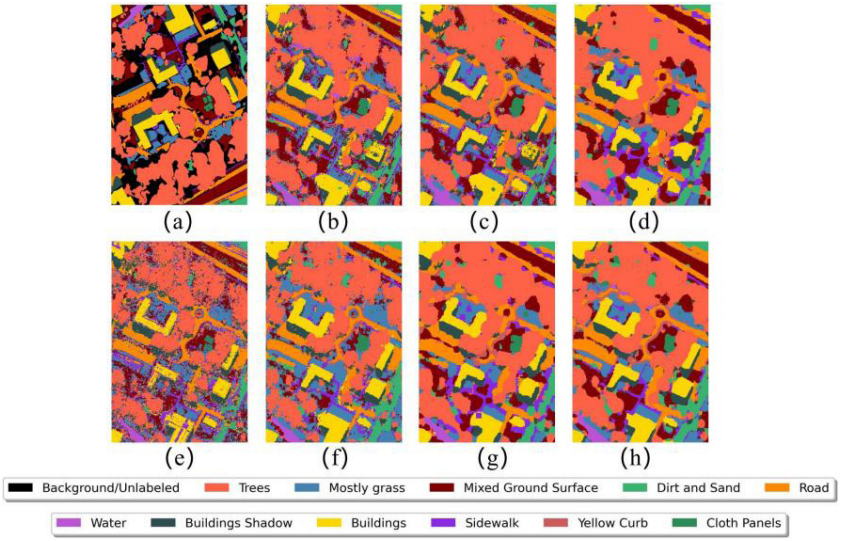

Fig.4: Classification maps of various methods for the MUUFL dataset. (a) Ground truth, (b) CCRNet, (c) CoupledCNN, (d) MHST, (e) EndNet, (f) HCT, (g) FusAtNet, (h) DMTCANet.

3.4 Ablation Study

Table IV shows the ablation study results on the Houston2013 dataset. The framework comprises four key components: the Conv3D module for HSI, the Conv2D module for HSI and LiDAR, the Tokenizers for both modalities, and the TCA fusion encoder.

The results indicate that removing the Conv2D module caused the lowest classification accuracy, highlighting its importance. Removing the Conv3D module led to a slight accuracy increase, showing the importance of multi-scale feature extraction. When both Conv2D and Conv3D were removed and replaced with ViT's Patch Embedding(PE), the accuracy reached 95.21%. Removing the TCA fusion encoder resulted in an OA of 98.57%, slightly lower than the optimal, but still underscoring the significance of the TCA encoder in enhancing performance. These findings validate the framework's effectiveness and the importance of each component.

Table IV: Ablation study of the proposed model components on the Houston2013 dataset.

Cases		1	2	3	4	5	6	7
Component	Conv3D	-	√	-	√	√	√	√
	Conv2D	√	-	-	√	√	√	√
	Tokenization	√	√	PE	PE	-	√	√
	TCA	√	√	√	√	-	-	√
Indicators	OA(%)	92.45	88.23	95.21	96.23	94.33	98.57	99.92
	AA(%)	89.07	65.78	93.89	94.78	93.45	98.43	99.94
	k100	91.26	74.61	93.43	95.02	92.41	99.21	99.91

4. Conclusion

In this work, we introduced an innovative framework called DMTCANet, which integrates a dual-branch multi-scale CNN with a TCA fusion mechanism. The multi-scale CNN extracts features at various scales, enhancing the receptive field to capture information at both local and global scales. The TCA mechanism improves the correlation between HSI and LiDAR data, addressing the challenge of limited information exchange. The experimental outcomes indicated that DMTCANet achieved substantial performance enhancements on three publicly available datasets. Compared to existing approaches, DMTCANet demonstrated superior performance, thus validating its effectiveness and advantages in joint classification tasks.

References:

[1]B. Li, Q.-W. Wang, J.-H. Liang, E.-Z. Zhu, and R.-Q. Zhou, "Squconvnet: Deep sequencer convolutional network for hyperspectral

image classification," Remote Sensing, vol. 15, no. 4, p. 983, 2023.

[2]D. Song, Y. Tang, B. Wang, J. Zhang, and C. Yang, "Two-branch generative adversarial network with multiscale connections for hyperspectral image classification," IEEE Access, vol. 11, pp. 7336–7347, 2022.

[3]S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 92–93.

[4]W. Wang, C. Li, P. Ren, X. Lu, J. Wang, G. Ren, and B. Liu, "Dualbranch feature fusion network based cross-modal enhanced cnn and transformer for hyperspectral and lidar classification," IEEE Geoscience and Remote Sensing Letters, 2024.

[5]L. Sun, X. Wang, Y. Zheng, Z. Wu, and L. Fu, "Multiscale 3-d–2-d mixed cnn and lightweight attention-free transformer for hyperspectral and lidar classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–16, 2024.

[6]G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and lidar data using a hierarchical cnn and transformer," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–16, 2022.

[7]K. Ni, D. Wang, Z. Zheng, and P. Wang, "Mhst: Multiscale head selection transformer for hyperspectral and lidar classification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.

[8]X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–10, 2021.

[9]R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and lidar data using coupled cnns," IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 7, pp. 4939–4950, 2020.

[10]D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder–decoder networks for classification of hyperspectral and lidar data," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2020.