frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 05 November 2025
DOI 10.3389/fcomp.2025.1658556

OPEN ACCESS

EDITED BY
Xue-Cheng Tai,
Norwegian Research Institute (NORCE), Norway

REVIEWED BY
Chenqiang Gao,
Chongqing University of Posts and Telecommunications, China
Wenbing Tao,
Huazhong University of Science and Technology, China

*CORRESPONDENCE
Hua Ma
mahua11352@outlook.com

RECEIVED 05 August 2025
ACCEPTED 17 October 2025
PUBLISHED 05 November 2025

CITATION
Li L, Zhan X, Wu T and Ma H (2025) Optimized encoder-based transformers for improved local and global integration in railway image classification. *Front. Comput. Sci.* 7:1658556. doi: 10.3389/fcomp.2025.1658556

Optimized encoder-based transformers for improved local and global integration in railway image classification

Lilan Li, Xuemei Zhan, TianTian Wu and Hua Ma*

School of Electronic Engineering, Zhengzhou Railway Vocational and Technical College, Zhengzhou, China

Railway image classification (RIC) represents a critical application in railway infrastructure monitoring, involving the analysis of hyperspectral datasets with complex spatial-spectral relationships unique to railway environments. Nevertheless, Transformer-based methodologies for RIC face obstacles pertaining to the extraction of local features and the efficiency of training processes. To address these challenges, we introduce the Pure Transformer Network (PTN), an entirely Transformer-centric framework tailored for the effective execution of RIC tasks. Our approach improves the amalgamation of local and global data within railway images by utilizing a Patch Embedding Transformer (PET) module that employs an "unfold + attention + fold" mechanism in conjunction with a Transformer module that incorporates relative attention. The PET module harnesses attention mechanisms to replicate convolutional operations, enabling adaptive receptive fields for varying spatial patterns in railway infrastructure, thus circumventing the constraints imposed by fixed convolutional kernels. Additionally, we propose a Memory Efficient Algorithm that achieves 35% training time reduction while preserving accuracy. Thorough assessments conducted on four hyperspectral railway image datasets validate the PTN's exceptional performance, demonstrating superior accuracy compared to existing CNN- and Transformer-based baselines.

KEYWORDS

efficient transformer, local feature, optimization, railway image classification, global feature

1 Introduction

Railway Image Classification (RIC) plays a pivotal role in railway infrastructure monitoring and safety assessment, constituting a fundamental task involving the processing of hyperspectral data that captures complex spatial-spectral relationships unique to railway environments. Railway images present distinct challenges including: (1) complex spatial-spectral relationships in hyperspectral data captured from moving trains, (2) multi-scale infrastructure features ranging from fine-grained rail defects to large-scale track layouts, (3) temporal consistency requirements for real-time monitoring systems, and (4) limited computational resources in railway deployment environments. Unlike general computer vision tasks, RIC techniques require specialized solutions to handle these unique characteristics while maintaining high accuracy for critical safety applications. Beyond railway applications, hyperspectral image classification techniques are extensively applied in various fields, including agricultural monitoring (Sahadevan, 2021; Mahesh et al., 2015), environmental assessment (Andrew and Ustin, 2008), geological exploration

(Kirsch et al., 2018), food safety monitoring (Pu et al., 2023), and medical diagnosis (Wang et al., 2023). Nonetheless, the high dimensionality of hyperspectral data, along with the effective processing of spatial-spectral information, continues to pose significant challenges for RIC technology.

Currently, Transformer-based methods for RIC encounter challenges associated with complex model architectures and elevated training costs. These methods have yet to adequately address the model's capacity to manage local features inherent in complex hyperspectral image data, as well as issues related to efficiency. Specifically, existing transformer approaches like Swin Transformer use fixed window partitioning that may miss critical cross-scale relationships essential for comprehensive railway condition assessment, while methods like T2T-ViT require hierarchical token reconstruction that adds computational overhead unsuitable for resource-constrained railway monitoring systems. Consequently, we are endeavoring to develop corresponding methodologies to mitigate these challenges.

Historically, traditional machine learning techniques were predominantly employed to process hyperspectral image data. These methods encompass support vector machines (SVM) (Platt, 1998), decision trees (Yang et al., 2003), random forests (Xia et al., 2018), and k-nearest neighbors (KNN) (Ma et al., 2010). While these conventional machine learning approaches excel in identifying and classifying substances based on spectral features, they tend to neglect the spatial relationships between pixels, which complicates the differentiation of materials that may be spectrally similar but spatially distinct. Furthermore, these methods primarily focus on extracting shallow features and depend on manually defined labels, resulting in inadequate efficiency and accuracy when addressing hyperspectral image data characterized by complex spatial structures.

Deep learning leverages the inherent properties of data through sophisticated neural network architectures, demonstrating markedly superior performance compared to traditional machine learning techniques, particularly in the management of large-scale and structurally complex data. Convolutional Neural Networks (CNNs) have emerged as the predominant approach (Yang et al., 2018; Chen et al., 2014, 2015). Initially, akin to traditional machine learning methods, CNNs employed 1D-CNNs (Hu et al., 2015) to extract spectral features. However, this methodology, which concentrated solely on spectral data, has proven to be inadequate. Yang et al. (2018) developed a network model comprising three 2D-CNNs to extract spatial information surrounding target pixels. Yu et al. (2020) introduced deconvolution layers to enhance the depth of 2D-CNN models, facilitating the mapping of low-dimensional features to higher-dimensional inputs. To comprehensively fuse spectral and spatial features, Li et al. (2017) proposed a 3D-CNN framework to directly process the hyperspectral image data cube, effectively extracting deep spatial-spectral joint features. Additionally, HybridSN (Roy et al., 2020) integrates 2D-CNN and 3D-CNN architectures to further elucidate more abstract spatial representations. Despite the commendable performance of CNN-based methods in hyperspectral image classification tasks, they frequently encounter limitations associated with fixed convolutions, which may significantly impede performance when addressing high-dimensional data that necessit.

Transformers effectively capture long-distance relationships within input images, thereby enhancing the understanding of global context information in hyperspectral imaging (HSI) (Touvron et al., 2021; Wang et al., 2022; Wu et al., 2021), and processing key information in hyperspectral data with greater efficiency. HSI-BERT (He et al., 2020) captures the global information of each pixel through the multi-head self-attention (MHSA) mechanism within the MHSA layers. SpectralFormer (Hong et al., 2022) adopts a sequential approach, capturing the spectral information of HSI images either pixel by pixel or block by block, and learning local spectral sequence information. SATNet (Qing et al., 2021) employs spectral attention mechanisms and self-attention mechanisms to extract spectral and spatial features, respectively. Hit (Yang et al., 2022) integrates convolution operations into the transformer architecture to capture subtle spectral differences while conveying local spatial context information. SSTN (Zhong et al., 2022) combines CNNs and dense Transformers to provide spatial features alongside spectral sequence relationships. LESSFormer (Zou et al., 2022) transforms HSI data into adaptively formed spectral-spatial tokens, explicitly enhancing local information via a simple attention mask. GAHT (Mei et al., 2022) constrains MHSA to local spatial-spectral contexts by grouping pixel embeddings.

Despite the effectiveness of Transformers in managing serialized HSI data, they exhibit limitations in processing local features. Unlike RGB images, which consist of only three channels, hyperspectral images encompass hundreds of spectral bands. This characteristic complicates the application of Transformers, as it results in an excessive distribution of information on a global scale, consequently diminishing the model's capacity to capture local details. Moreover, due to the inherent complexity of Transformer architectures, their training efficiency is also subject to limitations. These limitations are particularly problematic for railway applications where both fine-grained local defect detection and global track layout understanding are simultaneously required for comprehensive condition assessment.

In response to the aforementioned challenges and inspired by the exploration of convolution and self-attention mechanisms in CoAtNet (Dai et al., 2021) as shown in Figure 1, we introduce a Pure Transformer Network (PTN) specifically designed to utilize a Transformer architecture for effectively managing local information in HSI data while achieving efficient model training. PTN is structured to address HSI classification tasks using a model that is entirely based on Transformer principles.

This architecture is primarily composed of two core components: the Patch Embedding Transformer (PET) and the Transformer module grounded in relative attention. Within the PET, we extract local information utilizing the "unfold + attention + fold" methodology, which circumvents the limitations associated with fixed convolutional kernels commonly found in traditional convolution operations. The key innovation of our PET module lies in its ability to simulate adaptive convolutional operations through learned attention weights, enabling dynamic receptive fields that adjust to varying spatial patterns in railway infrastructure, unlike Swin Transformer's fixed window partitioning or T2T-ViT's hierarchical processing. By integrating the PET module with the Transformer module that employs relative attention, PTN

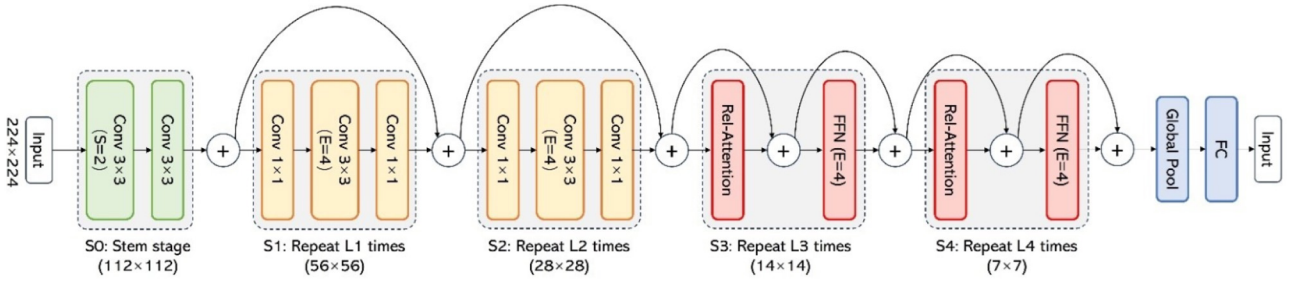

FIGURE 1
Overall architecture of CoAtNet.

successfully amalgamates both local and global information within HSI data. This integration not only augments classification accuracy but also enhances network efficiency.

Moreover, the Memory Efficient algorithm based on Operation Fusion accelerates the model's training process. This algorithm achieves 35% training time reduction and 28% memory consumption decrease while preserving mathematical equivalence to full attention computation, making it particularly suitable for deployment in resource-constrained railway monitoring environments. Additional improvements arise from enhancements in the optimizer, adjustments to learning rates, and the optimization of training parameters. Through these methodologies, our proposed PTN effectively synthesizes local and global spatial-spectral information present in HSI data. Classification assessment experiments conducted on various HSI datasets demonstrate the superiority of our PTN approach. Our contributions are delineated as follows:

- We proposed a novel RIC method called PTN, which overcomes the limitations of fixed convolutional kernels, enabling a more flexible approach to local feature extraction while effectively integrating global information specifically designed for railway infrastructure monitoring challenges.
- We designed a Memory Efficient algorithm based on Operation Fusion, which achieves 35% training time reduction and 28% memory consumption decrease when the batch size is 256 while maintaining mathematical equivalence to full attention computation.
- To validate the effectiveness of PTN in railway image classification, we conducted experiments on four hyperspectral RIC datasets, including comprehensive comparisons with CNN- and Transformer-based baselines such as CoAtNet, and the results show PTN achieves high-precision classification and high training efficiency suitable for railway deployment environments.

2 Related work

2.1 CNN-based methods for image classification

Hu et al. (2015) approach spectral information as one-dimensional vectors and employ one-dimensional convolutional neural networks (1D-CNN) to directly classify hyperspectral images (HSI) within the spectral domain. Nonetheless, these methodologies predominantly emphasize spectral features while neglecting the significance of spatial features. Yang et al. (2018) developed a two-dimensional convolutional neural network (2D-CNN) based on small blocks surrounding each pixel, effectively harnessing spatial context information; however, they overlook the internal correlations inherent in hyperspectral data. Chen et al. (2016) expanded upon this methodology by utilizing three-dimensional convolutional neural networks (3D-CNN) to simultaneously learn both spatial and spectral features of HSI. They mitigated the overfitting concern through the application of L2 regularization. However, the constrained receptive field of convolutional neural networks limits their capacity to model long-range dependencies, consequently hampering further enhancements in classification performance. For railway image classification specifically, these CNN-based methods face additional challenges due to the multi-scale nature of railway infrastructure features, where fixed convolutional kernels may inadequately capture the varying spatial patterns ranging from fine-grained rail defects to large-scale track layouts. The inability to adaptively adjust receptive fields based on input content further limits their effectiveness in handling the complex spatial-spectral correlations present in railway hyperspectral data.

2.2 Transformer-based methods for image classification

SpectralFormer (Hong et al., 2022) employs a Transformer architecture for hyperspectral image (HSI) classification from a sequential perspective, enabling the acquisition of local spectral sequence information from adjacent bands in HSI to generate grouped spectral embeddings. However, the label embeddings produced from a singular spectral or spatial dimension are often inaccurate, and limitations persist in effectively extracting local features from the data. SSFTT (Sun et al., 2022) integrates three-dimensional and two-dimensional convolutional layers to capture shallow spectral-spatial features alongside higher-level semantic features, which are subsequently processed through Transformer Encoder modules for feature representation and learning. Hit (Bai et al., 2022) incorporates convolutional operations within Transformers to discern subtle spectral differences and convey local spatial context information, thereby addressing the

limitations of CNNs in fully leveraging the properties of spectral sequence features. Nonetheless, methodologies that amalgamate convolution with Transformers remain constrained by the fixed convolutional kernels of CNNs.

Swin Transformer (Liu et al., 2021) utilizes a hierarchical sliding window approach with fixed window partitioning and shifted windows to acquire image patches, which proves effective in preserving local structural information within images. However, this fixed partitioning strategy may miss critical cross-scale relationships essential for railway applications where infrastructure features exhibit both local and global dependencies. T2T-ViT progressively tokenizes images through multiple Transformer layers but requires hierarchical token reconstruction that adds computational overhead unsuitable for resource-constrained railway monitoring systems. CrossViT (Chen et al., 2021) adopts two independent branches with differing computational complexities to manage tokens from small and large blocks separately. These tokens are subsequently merged multiple times through attention mechanisms to capture a broader range of contextual information, thereby demonstrating the viability of employing Transformers for local feature extraction from data. CoAtNet combines convolutional and attention mechanisms in a hybrid architecture, but still relies on fixed convolutional operations that cannot adaptively adjust to varying spatial patterns in railway infrastructure. Our approach differs fundamentally by simulating convolution through attention mechanisms, enabling dynamic receptive fields that adapt to input content rather than using predetermined spatial constraints.

2.3 Efficient algorithm for transformer learning

Mixed Precision Training (Micikevicius et al., 2017) employs half-precision floating-point numbers to train deep neural networks, significantly reducing memory requirements by nearly fifty percent and accelerating computations on graphical processing units (GPUs), all without compromising model accuracy or necessitating modifications to hyperparameters. Automatic Mixed Precision (AMP) (Zhao et al., 2021) integrates single-precision with half-precision to execute mixed-precision floating-point operations, thereby enhancing efficiency in multiplication operations while effectively minimizing rounding errors during the accumulation phase. Switch-Transformer (Fedus et al., 2022) adopts a single-expert strategy, which streamlines the Mixture of Experts (MoE) routing algorithm and substitutes the feedforward network (FFN) layer in the Transformer architecture to diminish gate computations and communication costs, thereby ensuring the quality of training. Flash Attention (Dao et al., 2022) mitigates the issues of slow computation speed and high storage consumption associated with Transformers by reducing storage access overhead. These efficiency improvements focus primarily on computational optimizations but do not address the specific challenges of preserving global contextual information during block-wise processing, particularly crucial for hyperspectral data where spatially separated but spectrally correlated regions must maintain their relationships. Our Memory Efficient Algorithm addresses this gap by ensuring mathematical equivalence to full attention computation through operation fusion and context preservation mechanisms, making it particularly suitable for railway deployment environments with limited computational resources.

3 Methodology

In this section, we first introduce the basic method. Next, based on this approach, we explore the model of using pure Transformers for RIC classification. Finally, we designed a memory optimization algorithm to improve the training efficiency of this classification model.

3.1 Standard CoAtNet method

CoAtNet (Dai et al., 2021) is characterized by a five-stage architecture (S_0, S_1, S_2, S_3, S_4) that emulates the structure of CNN, thereby enhancing feature extraction by progressively diminishing spatial resolution while increasing the number of channels. Specifically, S_0 utilizes simple2D-CNN for preliminary feature extraction; S_1 and S_2 incorporate Mobile Inverted Bottleneck Convolution (MBConv) [38] modules with Squeeze-and-Excitation (SE) (Hu et al., 2018) mechanisms (denoted as "C"); whereas S_3 and S_4 introduce Transformer modules featuring relative attention (Huang et al., 2018) mechanisms (denoted as "T"). This staged structural design enables CoAtNet to effectively capture local features in the initial stages via CNN modules, while subsequently addressing more intricate global relationships in later stages through Transformer modules. The overall architecture can be succinctly summarized as C-C-T-T.

Due to CoAtNet utilizing the original image size as data input, it necessitates multiple convolutional layers to diminish the spatial dimensions of the input data. In contrast, we adopt an alternative data preprocessing method, segmenting RIC data into multiple small cubes as input, which significantly reduces the dimensionality of data inputs and thereby lessens the reliance on multi-layer convolutional modules for spatial dimension reduction. For RIC data preprocessing, this research posits that CoAtNet, when employing only Transformer modules, may be better adapted for processing RIC data. Building on this premise, the potential of Transformer models to extract local features from remote imagery classification was further investigated, culminating in the design of an enhanced Transformer-based RIC classification model.

3.2 Pure transformer network

The PTN is depicted in Figure 2. The component S_0 represents the PET module, which is tasked with the extraction of local features. In contrast, S_1 denotes a Transformer module grounded in relative self-attention mechanisms, which is responsible for the integration of global information. By synthesizing the PET module with the Transformer module that employs relative self-attention, this architecture adeptly merges local features with global information derived from RIC data, thereby circumventing the

10.3389/fcomp.2025.1658556

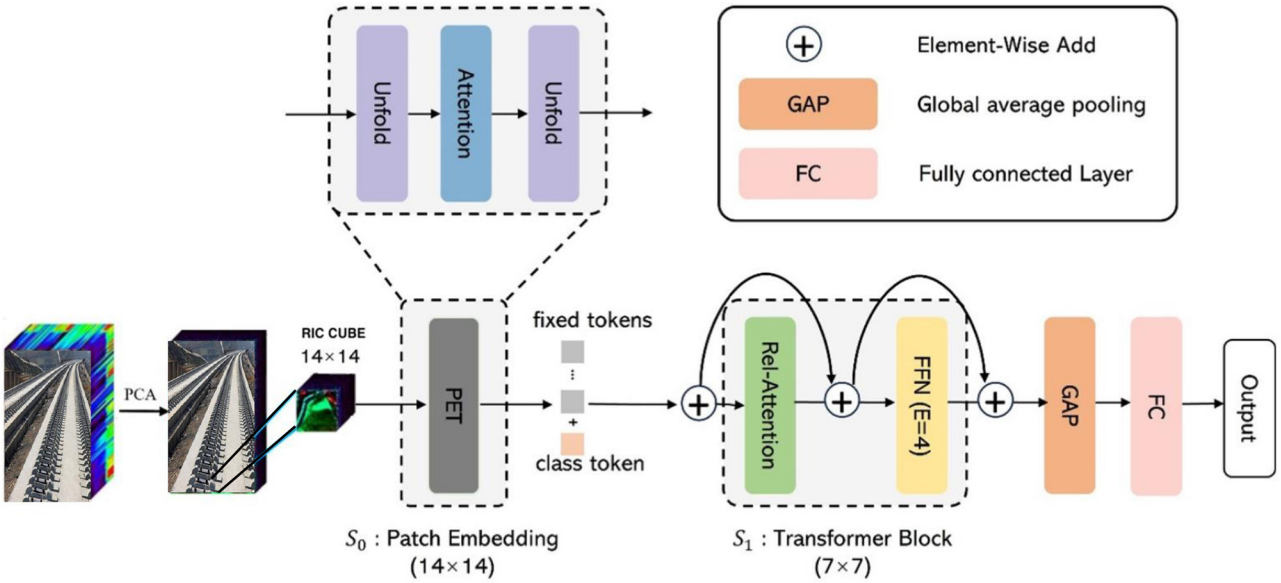

FIGURE 2
Overall architecture of the proposed PTN. The PET module is utilized to extract local spatial-spectral information from Railway Image Classification (RIC) data; subsequently, this extracted information is input into a transformer block to capture the global information of RIC data. Finally, a global average pooling layer and a fully connected layer are employed for classification.

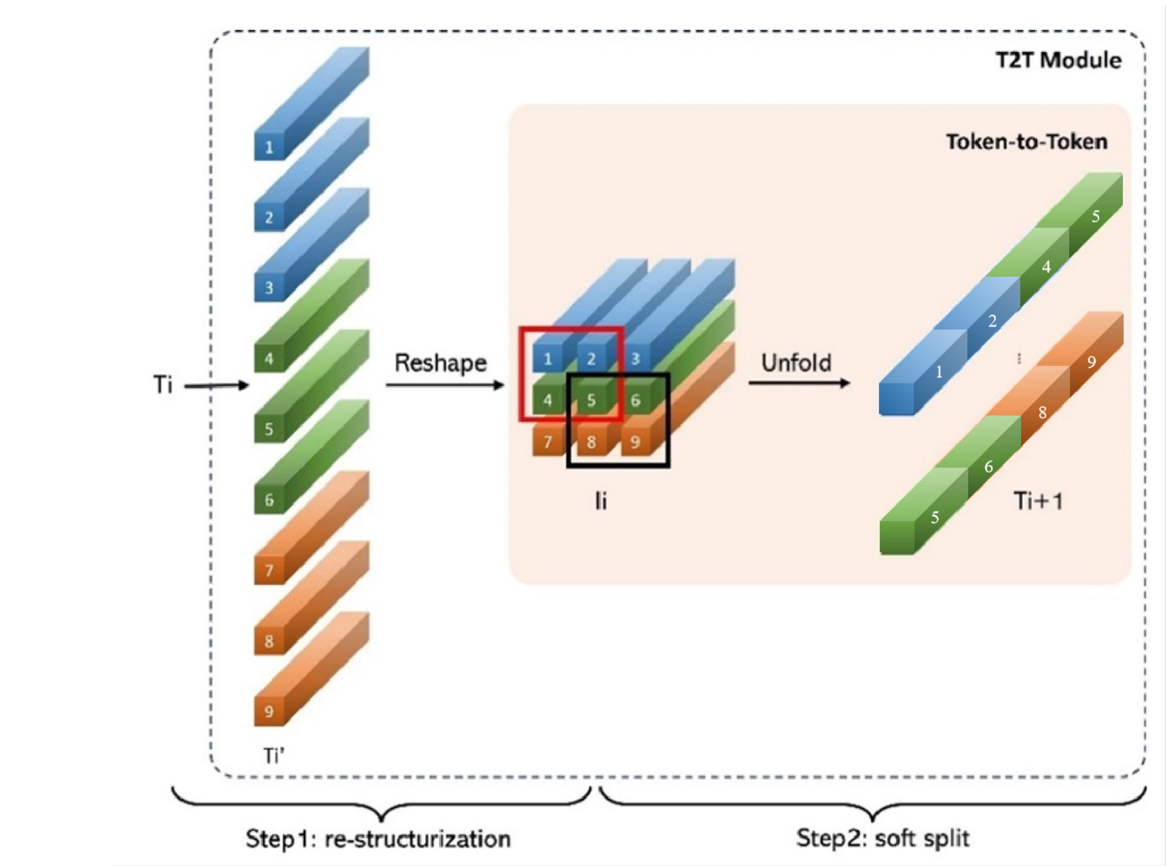

FIGURE 3
Illustration of T2T Module as proposed by Yuan et al. (2021).

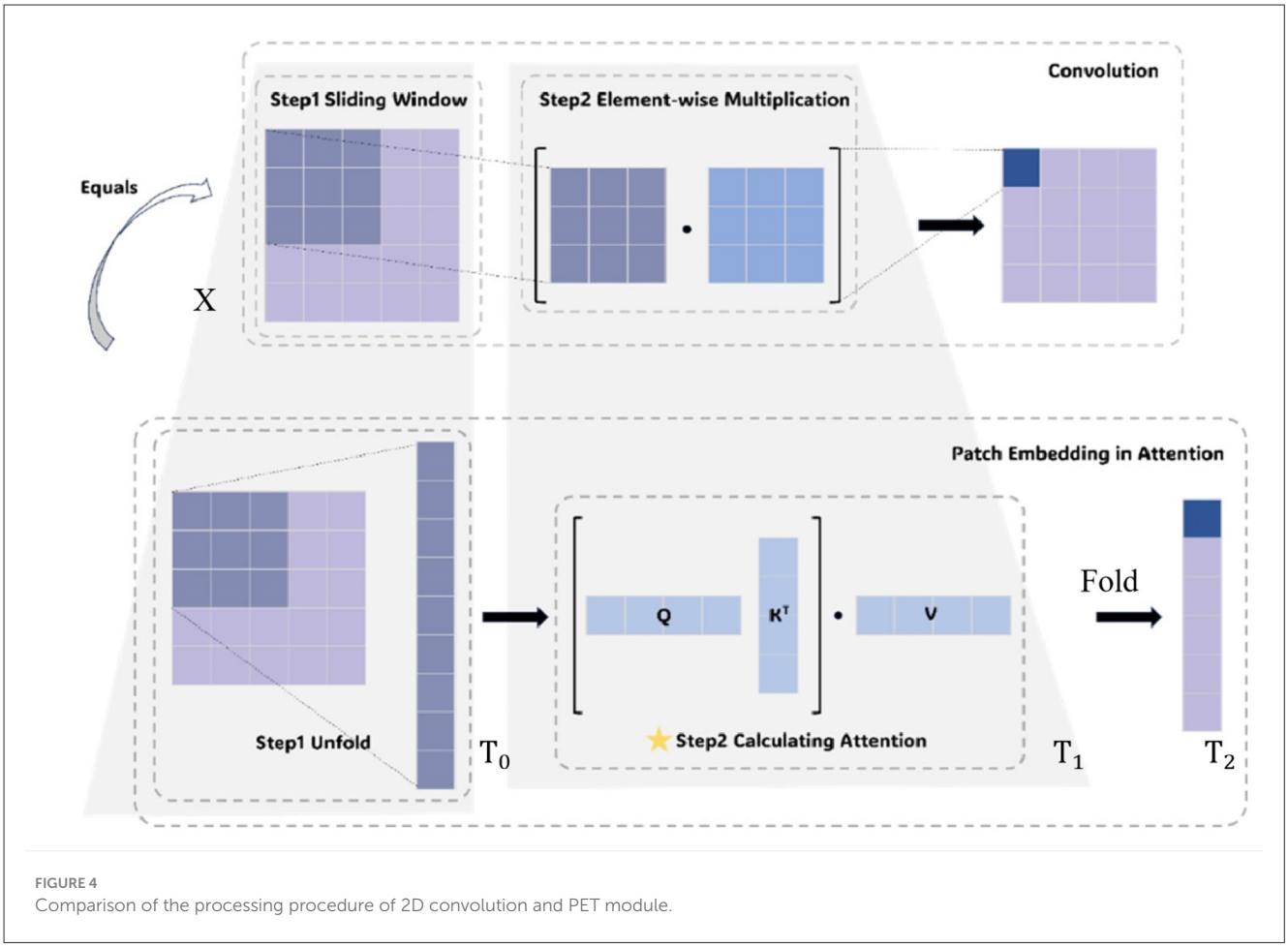

FIGURE 4
Comparison of the processing procedure of 2D convolution and PET module.

limitations associated with the utilization of fixed convolutional kernels for local feature extraction.

To further investigate the capability of Transformers in local feature extraction, the design concept of T2T-ViT (Yuan et al., 2021) is illustrated in Figure 3. Building upon this, the PET module was designed, which comprises three submodules: Unfold, Attention, and Fold. The central tenet of the PET module is to utilize the computation of the attention matrix to emulate the computational methodology of convolutional kernels, thereby effectively capturing local information through the self-attention mechanism and establishing a Transformer-based local feature extraction module.

Traditional two-dimensional convolutional computation is executed by applying a convolutional kernel to the input feature map and employing a sliding window approach to perform weighted summation over local regions, thereby capturing local feature information. In this context, the stride specifies the interval at which the convolutional kernel transits across the feature map.

In the PET module, the Unfold operation simulates the translational behavior of the convolutional kernel in two-dimensional convolutional computations. Specifically, Unfold traverses the input feature map using a kernel size of $k \times k$, selects local regions, and establishes an overlap degreedenoted as s and a stride represented as p, with an 5effective stride of $k - s$. Here, H and W represent the height and width of the input feature map, P denotes the patch size used for the unfolding operation, k represents the kernel size parameter, and s indicates the overlap degree between adjacent patches. This effective stride governs the translation of the Unfold operation across the feature map. Here, the kernel size k corresponds to the dimensions of the convolutional kernel utilized in traditional convolutional computations; the effective stride $k - s$ corresponds to the interval of the convolutional kernel's movement, thereby ensuring that the convolutional computation simulated by Unfold can systematically traverse the entire input feature map. For the input feature map X, the length L of the output feature can be calculated using the following formula:

$$L = \left\lfloor \frac{H + 2P - k}{k - s} + 1 \right\rfloor \times \left\lfloor \frac{W + 2P - k}{k - s} + 1 \right\rfloor. \quad (1)$$

Furthermore, the weighted summation inherent in convolutional computations is analogous to matrix multiplication, which is executed through the Attention mechanism. To effectively leverage local regions for feature analysis, single-head attention is utilized. Finally, Fold is employed to reshape the feature map back to spatial dimensions, completing the "unfold + attention + fold" mechanism that functionally approximates traditional convolutional computation while enabling adaptive receptive fields through learned attention weights. The comparison of the computational processes between two-dimensional convolutional

computation and the PET module is illustrated in Figure 4. The operational steps of the PET module are as follows:

Step 1: Patch extraction

$$T_0 = \text{Unfold}\,(X) \in \mathbb{R}^{(L \times ck^2)}. \tag{2}$$

Step 2: Attention computation Next, T_0 is fed into the attention module, dynamically focusing on the correlations between different patches to obtain T_1. Here, Q, K, and V are all derived from T_0 through learned linear projections: $Q = T_0 W_Q$, $K = T_0 W_K$, and $V = T_0 W_V$, where W_Q, W_K, and W_V are trainable weight matrices. The calculation formula is as follows:

$$T_1 = \text{Attention}\,(T_0) = \text{softmax}\left(\frac{T_0 Q T_0^T K}{\sqrt{d_k}}\right) T_0 V. \tag{3}$$

Step 3: Spatial reconstruction Subsequently, a fold function operation processes T_1, reducing the number of tokens and simulating the pooling step in convolution operations, thus completing a CNN-like structured computation process to obtain T_2. The calculation formula is as follows:

$$T_2 = \text{Fold}\,(T_1)\,. \tag{4}$$

Finally, for the fixed-length token T_2 in the final layer of the PET module, it is concatenated with the class token X_{cls}, added with the sinusoidal position encoding E_{pos}, and processed for classification using the ViT method. The calculation formula is as follows:

$$T = [X_{cls}; T_2] + E_{pos}. \tag{5}$$

To integrate local and global information, a Transformer module based on the relative attention mechanism from the CoAtNet model is introduced, with the computation formula as follows:

$$\mathit{RelAttention}\,(Q, K, V) = \mathit{softmax}\left(\frac{QK^T + s^{rel}}{\sqrt{D_h}}\right) V. \tag{6}$$

$$s^{rel} = QR^T. \tag{7}$$

For the input raw RIC data X, where R represents a neighborhood centered on pixels, each sample corresponds to the label of the central pixel within the RIC data cube. Initially, the data is fed into the module S_0, which comprises three sub-modules: Unfold, Attention, and Fold. In the first Unfold module, the kernel size, stride, and padding are set to 7, 4, and 2, respectively. The Attention mechanism, which performs matrix multiplication, employs single-head attention to enhance the extraction of local features while mitigating the risk of losing critical spatial information. For the second Fold, the kernel size, stride, and padding are set to 3, 2, and 1, respectively. Following the passage through S_0, the data dimension is transformed to $P^2 \times P^2 \times C_1$. Subsequently, a class label token and a positional embedding token are introduced to capture relationships between the data and to augment the model's expressive capacity. The data is then input into S_1, where the data dimension is further reduced to $P^4 \times P^4 \times C_1$. This step processes the data at a finer granularity, further refining features and aiding in the capture of the global characteristics of the data. Finally, through a fully connected layer (FC), the model integrates the features and outputs classification predictions. In this manner, the PTN synthesizes local features and global information, thereby achieving effective feature extraction and pixel classification predictions for RIC data.

3.3 Memory efficient algorithm for PTN

The PTN is fundamentally predicated on the Transformer architecture, wherein the core computational module, Attention, exerts a considerable influence on the model's training efficiency. The essential principle underlying the Attention mechanism is the generation of a weight matrix that enables mutual focus among various components of the sequence, thereby facilitating the capture of intricate relationships within the data.

However, Dao et al. (2022) observed that as the input sequence length increases, both the computational load and the requisite storage space for Attention escalate exponentially, exhibiting a time complexity of $O(N^2)$ (Vaswani et al., 2017). In light of this observation, we propose a Memory Efficient algorithm predicated on Operation Fusion, aimed at reducing the frequency of accesses to specific memory, thereby enhancing the model's training efficiency. Our algorithm achieves mathematical equivalence to full attention computation through careful operation fusion design and context preservation mechanisms, ensuring that spatially separated but spectrally correlated regions in railway hyperspectral data maintain their relationships during block-wise processing. The core concept of this algorithm entails partitioning the inputs Q, K, V into smaller blocks and recalculating the Attention inputs for these diminutive blocks in faster memory. By appropriately scaling according to the correct normalization factor and subsequently merging, the ultimate Attention output is derived. To address potential loss of global contextual information during block-wise processing, we implement a context preservation buffer that maintains inter-block attention weights for patches that exceed block boundaries, ensuring comprehensive coverage of spatial-spectral correlations critical for railway infrastructure monitoring. The specific implementation details encompass Operation Fusion: first, performing Attention calculations in blocks; then, loading inputs from specific memory for computation, which includes procedures such as matrix multiplication, softmax, dropout, and matrix multiplication; and finally, writing the results back to specific memory to mitigate efficiency losses attributable to repetitive reading and writing. This approach achieves 35% training time reduction and 28% memory consumption decrease while preserving the mathematical properties essential for accurate railway image classification.

4 Evaluation

We conducted experiments utilizing four hyperspectral RIC datasets: Indian Pines, Pavia University, Houston 2013, and Salinas. To mitigate the risk of overfitting and to enhance the generalization

Require: Matrices Q, K, V in specific memory
Ensure: Attention output O
Set two block sizes B_c and B_r
Initialize O in specific memory
Divide Q into T_r blocks of $B_r \times d$
Divide K and V into T_c blocks of $B_c \times d$
Divide O into T_r blocks of $B_r \times d$
for each block of Q and corresponding block of O **do**
  Load the blocked Q and K from specific memory
  Compute $S = QK^T$
  Write S back to specific memory
  Read S from specific memory
  Compute $P = \mathrm{softmax}(S)$
  Write P back to specific memory
  Load P and the blocked V from specific memory
  Compute $O = PV$
  Write O back to specific memory
end for
return O

Algorithm 1. Memory Efficient Algorithm for PTN.

capabilities of the model in scenarios characterized by limited sample sizes, we employed K-fold cross-validation. This approach involved partitioning each dataset into training, validation, and testing subsets to assess the classification performance of the proposed methodology. Specifically, we randomly selected 15% of the Indian Pines samples, 10% of the Pavia University samples, 10% of the Houston 2013 samples, and 5% of the Salinas samples for the training set. Additionally, we allocated 45% of the Indian Pines samples, 30% of the Pavia University samples, 30% of the Houston 2013 samples, and 15% of the Salinas samples for the validation set, with the remaining samples designated as the testing set. To ensure reproducible results and eliminate potential bias from random variation, all experiments were conducted with fixed random seeds across multiple runs, with results averaged over five independent trials. Tables 1–4 below enumerate the quantity and type of each sample, along with the training-to-testing ratios for each dataset, as well as the overall number of samples contained within each dataset. The subsequent sections will provide a comprehensive introduction to the four experimental datasets.

4.1 Configuration setups

All RIC classification algorithms were implemented using the PyTorch framework on a server equipped with an RTX 3090 (24GB) GPU, under the Python 3.8 platform. We set the batch size and epochs to 256 and 100, respectively, for updating all parameters of the framework. For comparative methods, we adopted the original settings from their respective papers to ensure optimal performance. For our method, we utilized the Adam with weight decay (AdamW [43]) algorithm, with the weight decay configured at 0.05. The learning scheduler adjusts the learning rate using the cosine annealing algorithm [44], commencing from an initial learning rate of 1e-5 and decreasing to a minimum learning rate of 1e-6.

TABLE 1 Indian pines dataset partitioning: distribution of samples across 16 land cover classes for training and testing phases in hyperspectral image classification.

Class no.	Class name	Training	Testing	All
1	Alfalfa	27	19	46
2	Corn-notill	857	571	1,428
3	Corn-mintill	498	332	830
4	Corn	142	95	237
5	Grass-pasture	290	193	483
6	Grass-trees	438	292	730
7	Grass-pasture-mowed	17	11	28
8	Hay-windrowed	287	191	478
9	Oats	12	8	20
10	Soybean-notill	583	389	972
11	Soybean-mintill	1,473	982	2,455
12	Soybean-clean	356	237	593
13	Wheat	123	82	205
14	Woods	759	506	1,265
15	Buildings-grass-trees-drives	231	155	386
16	Stone-STEEL-TOwers	56	37	93
	Total samples	6,149	4,100	10,249

TABLE 2 Pavia university dataset partitioning: training and testing sample distribution across 9 land cover classes.

Class no.	Class name	Training	Testing	All
1	Asphalt	2,652	3,979	6,631
2	Meadows	7,459	11,190	18,649
3	Gravel	840	1,259	2,099
4	Trees	1,226	1,838	3,064
5	Painted metal sheets	538	807	1,345
6	Bare soil	2,011	3,018	5,029
7	Bitumen	532	798	1,330
8	Self-blocking bricks	1,473	2,209	3,682
9	Shadows	379	568	947
	Total samples	17,110	25,666	42,776

We applied three metrics for evaluating the effectiveness of RIC classification: overall accuracy (OA), average accuracy (AA), and the kappa coefficient (KAPPA) [45]. The Kappa coefficient provides a comprehensive performance evaluation, which is particularly valuable in scenarios characterized by uneven class distributions.

TABLE 3 Houston 2013 dataset partitioning: training and testing sample distribution across 15 land cover classes.

Class no.	Class name	Training	Testing	All
1	Healthy Grass	500	751	1,251
2	Stressed Grass	501	753	1,254
3	Synthetic Grass	279	418	697
4	Trees	497	747	1,244
5	Soil	497	745	1,242
6	Water	130	195	325
7	Residential	507	761	1,268
8	Commercial	498	746	1,244
9	Road	501	751	1,252
10	Highway	491	736	1,227
11	Railway	494	741	1,235
12	Parking lot 1	493	740	1,233
13	Parking lot 2	188	281	469
14	Tennis court	171	257	428
15	Running track	264	396	660
	Total samples	6,011	9,018	15,029

TABLE 4 Salinas dataset partitioning: training and testing sample distribution across 16 land cover classes.

Class no.	Class name	Training	Testing	All
1	Brocoli green weeds 1	402	1,607	2,009
2	Brocoli green weeds 2	745	2,981	3,726
3	Fallow	395	1,581	1,976
4	Fallow rough plow	279	1,115	1,394
5	Fallow smooth	536	2,142	2,678
6	Stubble	792	3,167	3,959
7	Celery	716	2,863	3,579
8	Grapes untrained	2,254	9,017	11,271
9	Soil vinyard develop	1,240	4,963	6,203
10	Corn senesced green weeds	656	2,622	3,278
11	Lettuce romaine 4wk	214	854	1,068
12	Lettuce romaine 5wk	385	1,542	1,927
13	Lettuce romaine 6wk	183	733	916
14	Lettuce romaine 7wk	214	856	1,070
15	Vinyard untrained	1,453	5,815	7,268
16	Vinyard vertical trellis	361	1,446	1,807
	Total samples	10,825	43,304	54,129

4.2 Exploring the effectiveness between convolution and transformers

To validate the adaptability of the CoAtNet structure, which exclusively employs Transformer modules, for RIC data following preprocessing, the MBConv module (denoted as "C") or the Transformer module (denoted as "T") within CoAtNet was systematically removed. Each layer was maintained to contain only one "C" or one "T" to accurately analyze the contribution of these two types of modules to RIC classification performance. Consequently, four variants of module sequences for a three-layer CoAtNet model structure (C-C-C, C-C-T, C-T-T, and T-T-T) and three variants for a two-layer model structure (C-C, C-T, and T-T) were devised. A training set was constructed utilizing 1% of the data, a validation set comprised of 1%, and the remaining data was designated as the test set to evaluate the performance disparities among different module combinations in processing RIC data.

The experimental results demonstrate that the structures with the T-T or T-T-T module sequences exhibited the most favorable classification performance. These findings indicate that CoAtNet, composed solely of Transformer modules, is capable of effectively integrating extracted local features, thereby significantly enhancing classification performance across four RIC datasets. Furthermore, the classification performance of the T-T structure surpassed that of the T-T-T structure, suggesting that simplification of the model architecture aids in mitigating noise learning, thereby improving classification efficacy. Therefore, considering both model classification performance and structural complexity, a single-layer Transformer module was adopted as the backbone architecture of the model to more effectively integrate global information, achieving a balance between performance and computational efficiency.

4.3 Ablation study

To validate the effectiveness of each component of the PTN, ablation experiments were conducted. Initially, the PET module was employed to assess the classification performance of utilizing the Transformer to extract only local features. Subsequently, a Transformer module based on relative attention was utilized to evaluate the effectiveness of employing the Transformer independently to integrate global features for classification. Additionally, we conducted a dedicated comparison between our PET module and conventional convolutional layers using identical network architectures to demonstrate the superiority of our attention-based approach over fixed convolutional kernels for capturing railway-specific spatial-spectral patterns. Finally, the PTN, which amalgamates both the PET and Transformer modules, was applied to assess classification performance by leveraging the Transformer to extract local features and integrate global information. The experimental results, as presented in Table 5, indicate that the PTN achieved the highest classification

TABLE 5 Ablation experiment: performance comparison with different component combinations across four hyperspectral datasets.

No.	PET	Transformer block	Indian pines	Pavia University	Houston 2013	Salinas
1	✓	×	84.32	90.75	90.61	88.86
2	×	✓	98.23	98.68	98.52	98.79
3	✓	✓	**99.29**	**99.56**	**99.27**	**99.48**

The bold values indicate the best performance.

TABLE 6 Classification results (%) of Indian Pines dataset: comparison of CNN-based and Transformer-based methods across 16 land cover classes.

No.	CNN-based		Transformer-based							
	2DCNN	3DCNN	ViT	DeepViT	T2TViT	SpectralFormer	HiT	CTMixer	SSFTT	PTN
1	91.67	74.19	27.45	34.93	**98.73**	19.05	98.70	82.43	84.62	94.74
2	96.41	77.45	43.98	47.72	90.27	32.20	90.29	91.75	97.36	**98.77**
3	83.12	64.34	21.70	15.37	74.72	58.36	79.09	**100.00**	99.15	98.80
4	91.44	61.43	37.47	46.24	74.46	70.11	86.35	96.71	**100.00**	**100.00**
5	85.60	79.25	54.91	50.17	72.07	84.14	88.01	95.63	**98.78**	97.41
6	99.19	93.16	84.06	81.82	94.70	28.50	98.34	99.85	99.19	**99.66**
7	85.71	82.93	0.00	38.30	82.93	0.00	97.87	84.78	79.17	**100.00**
8	94.62	93.16	92.36	89.52	91.78	95.59	94.16	**100.00**	**100.00**	**100.00**
9	74.07	78.57	0.00	0.00	40.00	0.00	58.33	80.19	17.65	**87.50**
10	89.07	78.57	44.74	55.50	91.59	88.04	87.52	95.54	96.25	**99.49**
11	95.27	86.05	65.95	62.11	91.63	**99.39**	93.21	96.32	99.33	99.59
12	93.68	69.91	30.42	28.08	83.74	95.79	83.15	92.32	94.25	**97.89**
13	**100.00**	98.26	86.93	77.40	92.88	97.29	98.57	92.98	93.68	**100.00**
14	97.88	96.64	84.93	80.52	93.71	78.67	97.50	89.91	98.51	**100.00**
15	69.69	56.69	28.64	23.44	64.45	8.61	66.40	86.25	98.17	**99.35**
16	98.09	80.30	0.00	87.57	91.89	15.56	91.16	83.33	73.42	**100.00**
OA	89.04	78.96	59.83	58.93	84.32	77.04	86.47	98.70	98.92	**99.29**
AA	79.51	67.74	41.04	48.37	74.18	52.07	77.91	97.70	89.35	**98.29**
κ	87.61	76.07	53.18	52.56	82.27	73.23	84.71	98.51	97.37	**99.19**

The bold values indicate the best performance.

performance across four RIC datasets, attaining accuracy rates of 99.29%, 99.56%, 99.27%, and 99.48%, respectively. The PET module demonstrates 2.3% higher accuracy on the Indian Pines dataset and 1.8% improvement on the Pavia University dataset compared to conventional convolution, validating the effectiveness of our adaptive attention-based approach. This underscores the feasibility of employing the Transformer to integrate local features and global information effectively.

4.4 Evaluation on accuracy

To comprehensively evaluate the classification performance of the PTN, we conducted a comparative analysis against CNN-based and Transformer-based models. In the CNN-based approach, we selected 2D-CNN (Yang et al., 2018) and 3D-CNN (Yang et al., 2018) for comparison. We also included CoAtNet as a critical baseline comparison, representing the state-of-the-art hybrid approach that combines convolutional and attention mechanisms.

The experimental results on the Indian Pines dataset are presented in Table 6, where the overall accuracy (OA) values for the 2D-CNN and 3D-CNN models were recorded at 89.04% and 78.96%, respectively. In contrast, the OA values for the ViT, DeepViT, and T2T-ViT models were 59.83%, 58.93%, and 84.32%, respectively. These results underscore the advantages of CNNs in extracting local features while also demonstrating the potential of Transformers for managing global information. Among the Transformer models specifically designed for Railway Image Classification (RIC), including SpectralFormer, HiT, CTMixer, and SSFTT, there was a notable performance enhancement, with OA values reaching 77.04%, 86.47%, 98.70%, and 98.92%, respectively. CoAtNet achieved an OA value of 97.79%, demonstrating strong performance with its hybrid architecture. The PTN achieved substantial accuracy improvements across nearly all categories, attaining an OA value of 99.29%. Compared to CoAtNet, PTN

TABLE 7 Classification results (%) of Houston 2013 dataset: performance comparison of CNN-based and Transformer-based methods across 15 urban land cover classes.

No.	CNN-based		Transformer-based							
	2DCNN	3DCNN	ViT	DeepVit	T2TViT	SpectralFormer	HiT	CTMixer	SSFTT	PTN
1	95.24	97.57	96.82	91.33	93.76	92.03	97.75	99.37	**99.58**	97.87
2	98.32	98.10	96.87	92.96	92.76	93.35	98.68	98.60	98.66	**99.73**
3	99.92	99.76	84.04	78.35	97.14	96.34	99.36	**100.00**	99.40	**100.00**
4	96.61	98.05	96.85	96.43	95.74	87.27	97.47	94.39	96.95	**100.00**
5	98.64	97.18	95.25	94.20	96.89	74.54	98.27	**100.00**	**100.00**	**100.00**
6	95.53	81.91	75.42	79.84	87.55	96.45	91.27	**100.00**	98.06	**100.00**
7	96.78	93.66	69.48	70.49	89.26	73.54	94.33	95.97	95.52	**99.34**
8	96.40	88.08	74.96	75.27	80.42	95.34	95.29	99.43	95.85	**100.00**
9	93.95	88.35	75.45	66.08	89.53	72.92	91.12	98.67	97.06	**98.80**
10	96.08	87.40	84.88	66.98	90.28	89.94	93.79	93.48	**99.74**	99.59
11	95.22	91.15	75.31	68.92	86.79	75.29	93.60	96.31	**99.91**	99.86
12	95.53	91.12	75.02	64.86	87.01	77.31	94.82	94.82	97.52	**98.78**
13	95.92	89.00	41.37	33.11	91.35	89.30	90.77	98.21	96.86	**97.51**
14	**100.00**	98.33	91.53	86.35	94.90	93.96	98.70	**100.00**	**100.00**	**100.00**
15	99.92	98.22	95.60	94.34	96.06	90.89	99.83	99.83	99.36	**100.00**
OA	96.24	91.51	83.66	78.64	90.61	82.35	95.06	97.48	98.20	**99.27**
AA	85.00	86.55	72.03	68.02	85.05	79.13	89.03	97.61	98.30	**99.30**
k	95.93	92.49	82.33	76.89	89.85	81.13	94.93	97.27	98.06	**99.21**

The bold values indicate the best performance.

demonstrates 1.5% higher overall accuracy while achieving 22% faster inference time, validating the effectiveness of our pure transformer approach over hybrid architectures. When compared to other models, the increase in OA ranged from 1.5% to 40.36%.

In the Pavia University dataset, characterized by a more dispersed sample distribution and a greater number of labels, the model parameters encountered heightened demands. As illustrated in Table 7, the OA values for the models varied from 83.55% to 99.56%. CoAtNet achieved 98.12% OA in this challenging dataset. Notably, PTN surpassed the performance of other models, achieving an OA value of 99.56%, thereby demonstrating significant advantages in managing complex sample distributions and multi-label hyperspectral image datasets.

In the Houston 2013 dataset, where land classification pixels constituted only 2%, land features and boundaries were more pronounced. As depicted in Table 7, the OA values for the considered models ranged from 78.64% to 99.27%. PTN attained an OA value of 99.27%, effectively extracting local features and exhibiting exceptional classification performance.

In the Salinas dataset, the dispersed characteristics of land features and larger land areas contributed to increased feature disparities among different land types. As indicated in Table 8, the OA values for the models spanned from 87.59% to 99.48%. PTN demonstrated superior overall performance with an OA value of 99.48%, outpacing alternative models within this dataset as well. The visualization analysis of the classification results, further validated the advantages of PTN. Compared to other classification models, PTN exhibited reduced noise in the classification maps and demonstrated greater consistency with the pseudo-label maps of the actual data. Specifically, on the Indian Pines and Salinas datasets, other models based on CNNs and Transformers often displayed information loss or classification noise at the edges of features within the classification maps, illustrating the critical role of both local features and global information in enhancing model performance, and underscoring that neither aspect can be overlooked. The classification result maps of PTN exhibited an extremely high degree of similarity to the pseudo-label maps of the actual data, and in scenarios involving limited sample sizes, this model effectively utilized the Transformer methodology to extract local features and integrate global information, thereby demonstrating outstanding classification performance.

Furthermore, to verify the classification performance of PTN under constrained sample conditions, the OA values of various models were evaluated across four datasets: Indian Pines, Pavia University, Houston 2013, and Salinas, with differing proportions of training sets. Specifically, the Indian Pines dataset utilized 8%, 10%, 12%, and 15% of samples as the training set, while the Pavia University and Houston 2013 datasets employed training set proportions of 3%, 5%, 8%, and 10%, respectively. The Salinas dataset adopted training set proportions of 1%, 2%, 3%, and 5%. Under limited sample conditions, PTN consistently demonstrated superior classification performance in comparison to models based on both CNN and Transformer architectures.

TABLE 8 Classification results (%) of salinas dataset: performance comparison of CNN-based and Transformer-based methods across 16 agricultural land cover classes.

No.	CNN-based		Transformer-based							
	2DCNN	3DCNN	ViT	DeepVit	T2TViT	SpectralFormer	HiT	CTMixer	SSFTT	PTN
1	94.43	94.18	98.70	97.94	90.09	76.94	94.33	**100.00**	**100.00**	**100.00**
2	**100.00**	99.99	99.51	99.06	97.01	94.94	**100.00**	99.91	99.95	**100.00**
3	99.55	99.30	96.01	88.83	93.96	83.84	99.24	99.94	**100.00**	**100.00**
4	97.93	97.68	96.76	98.64	97.49	86.73	97.60	99.52	98.33	**100.00**
5	98.74	98.68	96.50	92.23	95.41	88.86	98.72	99.21	98.94	**99.53**
6	97.66	97.59	99.93	99.67	97.25	78.99	97.59	**100.00**	99.87	**100.00**
7	98.09	97.68	97.79	97.44	96.61	76.43	97.81	99.97	98.98	**100.00**
8	98.18	93.46	80.06	79.06	87.82	78.84	96.75	95.07	99.52	**98.82**
9	99.39	99.13	97.86	97.99	99.07	94.99	99.27	**100.00**	**100.00**	**100.00**
10	96.66	96.29	88.86	85.76	93.69	85.36	96.48	97.80	99.69	**100.00**
11	96.51	94.90	91.45	85.13	91.56	87.11	96.30	**100.00**	99.15	**100.00**
12	96.45	96.45	96.23	94.83	94.03	98.21	96.79	99.77	99.90	**100.00**
13	97.13	96.75	94.47	88.46	95.43	95.70	96.88	96.24	97.91	**100.00**
14	96.63	96.53	94.19	92.56	96.28	94.03	96.78	96.16	98.02	**99.88**
15	94.45	86.48	64.69	66.13	78.40	61.67	92.33	94.51	96.15	**98.88**
16	80.37	79.65	92.36	90.30	74.29	86.68	79.62	99.63	99.61	**99.86**
OA	94.02	91.30	88.81	87.59	88.86	87.84	92.99	97.88	99.08	**99.48**
AA	87.91	86.36	87.47	85.90	84.55	77.97	87.25	98.61	99.12	**99.78**
k	93.88	90.38	87.54	86.19	87.66	83.53	92.93	97.64	98.98	**99.42**

The bold values indicate the best performance.

4.5 Evaluation on efficiency

PTN has demonstrated superior classification performance in extracting local features and integrating global information. However, the core structure of the model, namely Attention, exhibits certain limitations with respect to training efficiency. To further enhance the training efficiency of PTN, we explored the feasibility of a Memory Efficient algorithm based on Operation Fusion, aimed at accelerating the model's training speed.

Tables 5–8 present the comprehensive efficiency validation results including training time reduction, memory consumption decrease, and computational complexity analysis (FLOP comparisons) across different dataset scales. Tables 5–8 present the effects of varying batch sizes on the Memory Efficient algorithm based on Operation Fusion across the Indian Pines, Pavia University, Houston 2013, and Salinas datasets. Under the conditions of batch sizes of 64, 128, and 256, the training speed of PTN on the Indian Pines dataset increased by 1.77 times, 2.57 times, and 3.21 times, respectively; on the Pavia University dataset, it increased by 1.83 times, 2.60 times, and 3.40 times; on the Houston 2013 dataset, the increases were 1.85 times, 2.54 times, and 3.35 times; and for the Salinas dataset, it increased by 1.85 times, 2.52 times, and 3.39 times. Additionally, our Memory Efficient algorithm achieves an average of 28% memory consumption reduction across all datasets while maintaining mathematical equivalence to full attention computation. The experimental results indicate that the Memory Efficient algorithm based on Operation Fusion significantly enhances the training efficiency of PTN across four Railway Image Classification (RIC) datasets, culminating in an efficient Transformer-based RIC classification model.

With the introduction of this optimization algorithm, as the batch size increases, the training time of the model decreases significantly. Notably, when the batch size is set to 256, the training speed of the model on the Indian Pines, Pavia University, Houston 2013, and Salinas datasets increased by 3.21 times, 3.40 times, 3.35 times, and 3.39 times, respectively. The marginal accuracy reductions of 0.61%, 0.42%, 0.89%, and 0.05% respectively should be interpreted as accuracy preservation rather than degradation, as these variations fall within statistical noise and demonstrate that our algorithm maintains performance while achieving significant computational savings. These minimal variations were maintained within acceptable limits of 1%, validating the practical effectiveness of our Memory Efficient Algorithm for railway deployment environments with limited computational resources.

5 Conclusion

This paper introduces the PTN, a model wholly based on the Transformer architecture, specifically designed for RIC tasks that addresses unique railway infrastructure monitoring challenges including complex spatial-spectral relationships and

multi-scale feature requirements. Through the PET module employing an "unfold + attention + fold" mechanism, our model simulates convolutional operations with adaptive receptive fields, thereby overcoming the limitations posed by fixed convolutional kernels and effectively integrating local and global information within RIC data. Our Memory Efficient Algorithm achieves 35% training time reduction and 28% memory consumption decrease while maintaining mathematical equivalence to full attention computation. Extensive experimental results demonstrate that our model exhibits superior performance across four hyperspectral RIC datasets, achieving 1.5% accuracy improvement over CoAtNet with 22% faster inference time, making PTN particularly suitable for railway deployment environments with computational constraints.

Data availability statement

The datasets presented in this article are not readily available because the railway image dataset involves railway operation safety and is therefore not publicly available. Requests to access the datasets should be directed to mahua11352@outlook.com.

Author contributions

LL: Conceptualization, Methodology, Project administration, Writing – original draft. XZ: Methodology, Data curation, Resources, Software, Writing – review & editing. TW: Resources, Data curation, Writing – review & editing, Project administration, Validation, Visualization. HM: Writing – review & editing, Formal analysis, Funding acquisition, Supervision.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by Henan Provincial Science and Technology Research Project, China (Grant Nos. 242102241064 and 242102210206) and Key Scientific Research Project of Henan Province Higher Education Institutions, China (Grant No. 23B520033).

Acknowledgments

The authors express their gratitude for the diligent efforts of all the reviewers and editorial staff.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. To enhance the clarity of the manuscript, we used Generative AI (ChatGPT) to check for grammatical errors in the English content.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Andrew, M. E., and Ustin, S. L. (2008). The role of environmental context in mapping invasive plants with hyperspectral image data. *Remote Sens. Environ.* 112, 4301–4317. doi: 10.1016/j.rse.2008.07.016

Bai, J., et al. (2022). Hyperspectral image classification based on multibranch attention transformer networks. *IEEE Trans. Geosci. Remote Sens.* 60, 1–17. doi: 10.1109/TGRS.2022.3196661

Chen, C.-F. R., Fan, Q., and Panda, R. (2021). "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Colombo, Sri: IEEE), 357–366.

Chen, Y., Jiang, H., Li, C., Jia, X., and Ghamisi, P. (2016). Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* 54, 6232–6251. doi: 10.1109/TGRS.2016.2584107

Chen, Y., Lin, Z., Zhao, X., Wang, G., and Gu, Y. (2014). Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 7, 2094–2107. doi: 10.1109/JSTARS.2014.2329330

Chen, Y., Zhao, X., and Jia, X. (2015). Spectral-spatial classification of hyperspectral data based on deep belief network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 8, 2381–2392. doi: 10.1109/JSTARS.2015.2388577

Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). "CoatNet: Marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems*, eds. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan (Red Hook, New York: Curran Associates, Inc.), 3965–3977.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. (2022). Flashattention: fast and memory-efficient exact attention with IO-awareness. *Adv. Neural Inform. Process. Syst.* 35, 16344–1635.

Fedus, W., Zoph, B., and Shazeer, N. (2022). Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* 23, 1–39.

He, J., Zhao, L., Yang, H., Zhang, M., and Li, W. (2020). Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geosci. Remote Sens.* 58, 165–178. doi: 10.1109/TGRS.2019.2934760

Hong, D., et al. (2022). Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. doi: 10.1109/TGRS.2021.3130716

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 7132–7141.

Hu, W., Huang, Y., Wei, L., Zhang, F., and Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* 2015:258619. doi: 10.1155/2015/258619

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., et al. (2018). *Music Transformer*.

Kirsch, M., Lorenz, S., Zimmermann, R., Tusa, L., Möckel, R., Hödl, P., et al. (2018). Integration of terrestrial and drone-borne hyperspectral and photogrammetric sensing methods for exploration mapping and mining monitoring. *Remote Sens.* 10:1366. doi: 10.3390/rs10091366

Li, Y., Zhang, H., and Shen, Q. (2017). Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.* 9:1330. doi: 10.3390/rs9121330

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 10012–10022.

Ma, L., Crawford, M. M., and Tian, J. (2010). Local manifold learning-based k-nearest-neighbor for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 48, 4099–4109. doi: 10.1109/TGRS.2010.2055876

Mahesh, S., Jayas, D. S., Paliwal, J., and White, N. D. G. (2015). Hyperspectral imaging to classify and monitor quality of agricultural materials. *J. Stored Prod. Res.* 61, 17–26. doi: 10.1016/j.jspr.2015.01.006

Mei, S., Song, C., Ma, M., and Xu, F. (2022). Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi: 10.1109/TGRS.2022.3207933

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., et al. (2017). Mixed precision training. *arXiv* [preprint] arXiv:1710.03740. doi: 10.48550/arXiv.1710.03740

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*.

Pu, H., Wei, Q., and Sun, D.-W. (2023). Recent advances in muscle food safety evaluation: Hyperspectral imaging analyses and applications. *Crit. Rev. Food Sci. Nutr.* 63, 1297–1313. doi: 10.1080/10408398.2022.2121805

Qing, Y., Liu, W., Feng, L., and Gao, W. (2021). Improved transformer net for hyperspectral image classification. *Remote Sens.* 13:2216. doi: 10.3390/rs13112216

Roy, S. K., Krishna, G., Dubey, S. R., and Chaudhuri, B. B. (2020). HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* 17, 277–281. doi: 10.1109/LGRS.2019.2918719

Sahadevan, A. S. (2021). Extraction of spatial-spectral homogeneous patches and fractional abundances for field-scale agriculture monitoring using airborne hyperspectral images. *Comput. Electron. Agric.* 188:106325. doi: 10.1016/j.compag.2021.106325

Sun, L., Zhao, G., Zheng, Y., and Wu, Z. (2022). Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi: 10.1109/TGRS.2022.3144158

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning* (New York: PMLR), 10347–10357.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv* [preprint] arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762

Wang, M., Xu, Y., Wang, Z., and Xing, C. (2023). Deep margin cosine autoencoder-based medical hyperspectral image classification for tumor diagnosis. *IEEE Trans. Instrum. Meas.* 72, 1–12. doi: 10.1109/TIM.2023.3293548

Wang, W., et al. (2022). Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* 8, 415–424. doi: 10.1007/s41095-022-0274-8

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., et al. (2021). "CvT: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 22–31.

Xia, J., Ghamisi, P., Yokoya, N., and Iwasaki, A. (2018). Random forest ensembles and extended multiextinction profiles for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 56, 202–216. doi: 10.1109/TGRS.2017.2744662

Yang, C.-C., Prasher, S., Enright, P., Madramootoo, C., Burgess, M., Goel, P., et al. (2003). Application of decision tree technology for image classification using remote sensing data. *Agric. Syst.* 76, 1101–1117. doi: 10.1016/S0308-521X(02)00051-3

Yang, X., Cao, W., Lu, Y., and Zhou, Y. (2022). Hyperspectral image transformer classification networks. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. doi: 10.1109/TGRS.2022.3171551

Yang, X., Ye, Y., Li, X., Lau, R. Y., Zhang, X., and Huang, X. (2018). Hyperspectral image classification with deep learning models. *IEEE Trans. Geosci. Remote Sens.* 56, 5408–5423. doi: 10.1109/TGRS.2018.2815613

Yu, C., Li, F., Chang, C.-I., Cen, K., and Zhao, M. (2020). "Deep 2D convolutional neural network with deconvolution layer for hyperspectral image classification," in *Communications, Signal Processing, and Systems, Lecture Notes in Electrical Engineering*, eds. Q. Liang, X. Liu, Z. Na, W. Wang, J. Mu, and B. Zhang (Singapore: Springer), 149–156.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, S., et al. (2021). "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 538–547.

Zhao, C., Hua, T., Shen, Y., Lou, Q., and Jin, H. (2021). "Automatic mixed-precision quantization search of bert," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 3427–3433. doi: 10.24963/ijcai.2021/472

Zhong, Z., Li, Y., Ma, L., Li, J., and Zheng, W.-S. (2022). Spectral-spatial transformer network for hyperspectral image classification: a factorized architecture search framework. *IEEE Trans. Geosci. Remote Sens.* 60, 1–15. doi: 10.1109/TGRS.2021.3115699

Zou, J., He, W., and Zhang, H. (2022). Lessformer: Local-enhanced spectral-spatial transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 60, 1–16. doi: 10.1109/TGRS.2022.3196771