Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PAFC-Net: A Progressive Attention Fusion Network with Semantic Consensus for Multimodal Remote Sensing Image Classification

Jiahang Lyu
Changchun University of Technology
Yuegang Fu
fuyuegang@ccut.edu.cn
Changchun University of Technology
Xueming Xiao
Changchun University of Science and Technology
Jinshuang Wu
Changchun University of Technology
Zhiwei Wang
Jilin University
Xiang Zhou
Changchun University of Technology

Research Article

Keywords: Multimodal Fusion, Remote Sensing Image Classification, Attention Mechanism, Hyperspectral Image

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11297916/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PAFC-Net: A Progressive Attention Fusion Network with Semantic Consensus for Multimodal Remote Sensing Image Classification

Abstract

Multimodal remote sensing classification using HSI and LiDAR/SAR data seeks to convert heterogeneous sensor observations into reliable joint discriminative representations. However, structural differences between modalities can impede information exchange and produce inconsistent predictions from the fused branches. To address this problem, we propose the Progressive Attention Fusion Network with Semantic Consensus (PAFC-Net). A modality-specific asymmetric encoder first preserves the spectral–spatial characteristics of hyperspectral images and the structural information of auxiliary data within a unified token representation. Rather than relying exclusively on cross-modal interaction, PAFC-Net progressively coordinates complementary information through three functionally distinct operations. It first models bidirectional cross-modal correspondences, then uses the exchanged context to enhance global semantics, and finally refines within-modality dependencies through gated local updates. A sample-specific consensus anchor subsequently calibrates the two branches against a shared semantic reference. PAFC-Net achieves overall accuracies of 80.45%, 92.48%, 93.70%, and 92.26% on the Berlin, Augsburg, Houston 2013, and Houston 2018 datasets, respectively. Notably, its OA on the Berlin dataset exceeds the strongest competing result by 4.15%. Under consistent data splits and evaluation protocols, experiments on four multisource remote sensing datasets show that PAFC-Net attains the highest OA and Kappa on all four datasets and the highest AA on three.

Keywords: Multimodal Fusion, Remote Sensing Image Classification, Attention Mechanism, Hyperspectral Image

1 Introduction

Multimodal remote sensing image classification combines observations of the same area from different sensors. It therefore provides more complete evidence for fine-grained land-cover recognition than a single sensor (Aleissaee et al. 2023; Ghamisi et al. 2019; Li et al. 2022; Sun et al. 2023). Hyperspectral images (HSI) characterize surface materials through contiguous spectral responses. Synthetic aperture radar (SAR)

and light detection and ranging (LiDAR) data provide complementary microwave-scattering and elevation information, respectively. These sensors nevertheless produce data with substantially different structures. HSI contains high-dimensional continuous spectra, whereas SAR and LiDAR primarily describe scattering properties and spatial structure (Ghamisi et al. 2019; Li et al. 2022). Early studies used convolutional neural networks (CNNs), often with attention-enhanced parallel branches, to extract and fuse multisource features (Audebert et al. 2018; Chen et al. 2017; Dong et al. 2022). Later work introduced sensor-specific asymmetric branches (Li et al. 2023b). Shared-specific learning and hierarchical representations were also developed to coordinate common and modality-dependent information (Feng et al. 2023; Guo et al. 2024; Hong et al. 2021). Cross-modal enhancement further strengthened complementarity between heterogeneous observations (Fang et al. 2022). However, several methods tightly couple their encoders to a particular fusion operator. Such designs may not preserve modality-specific information while providing a common representation for interaction. Fig. 1 contrasts cross-modal-interaction-centered fusion with our progressive collaborative fusion paradigm. Such fusion produces fused features primarily through direct information exchange between modalities. Our approach instead transfers information through successive attention operations and calibrates both branches before prediction.

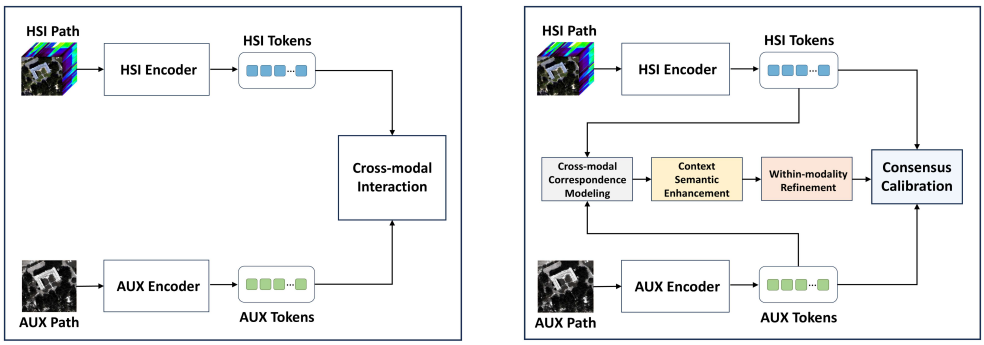


Fig. 1 Conceptual comparison between cross-modal-interaction-centered fusion and the proposed progressive collaborative fusion paradigm. (a) Cross-modal-interaction-centered fusion. (b) Proposed progressive collaborative fusion.

After modality-specific features have been extracted, the organization of cross-modal information flow becomes another determinant of fusion quality. Vision Transformers offer a general framework for modeling long-range dependencies (Liu et al. 2021). Self-attention and cross-attention can explicitly establish correspondences between modalities (Mohla et al. 2020; Wang et al. 2022; Ying et al. 2022). Subsequent studies have organized multistage information transfer through progressive fusion, cross-modal semantic enhancement, dynamic interaction, and global guidance (Han et al. 2024; Hu et al. 2025; Li et al. 2023c; Lin et al. 2025). SAR-optical Transformers and multimodal temporal attention extend this capability to cross-source spatial relations and temporal dependencies (Liu et al. 2024; Sainte Fare Garnot et al. 2022).

Recent state-space models improve the efficiency of long-range dependency modeling (Gao et al. 2025; Li et al. 2025; Pan et al. 2025a; Zhang et al. 2024b, 2026). Multistage fusion can nevertheless accumulate redundant responses and leave semantic discrepancies between its output branches (Li et al. 2023d; Pan et al. 2025b). We address these issues with the Progressive Attention Fusion Network with Semantic Consensus (PAFC-Net). Our main contributions are as follows:

- We propose PAFC-Net to reduce representational and predictive inconsistencies in heterogeneous remote sensing data. Its encoding, fusion, and calibration stages are optimized jointly for the same classification objective.
- We design MSAE to preserve sensor-specific spectral–spatial and structural information while mapping HSI and auxiliary features into a common token space for subsequent interaction.
- We develop PCAFM to progressively model cross-modal correspondences, enhance global semantics with exchanged context, and refine within-modality dependencies through dual-gated updates.
- We introduce CAMC, which models the commonality, discrepancy, and interaction between the two CLS tokens to derive a sample-specific anchor that reduces post-fusion semantic inconsistency.

2 Related Work

2.1 Modality-specific representation learning

Multisource remote sensing representation learning must preserve sensor-specific information while providing compatible representations for subsequent interaction. Early CNN-based methods commonly used parallel branches and cross-layer connections to extract local spectral–spatial features from different inputs (Audebert et al. 2018; Chen et al. 2017; Dong et al. 2022). Although this design maintains separate processing streams, similar branch structures may not fully reflect the distinct characteristics of heterogeneous sensors. Asymmetric networks address this issue by assigning different feature extractors according to the properties of each modality (Li et al. 2023b). Spectral–spatial cross-modal enhancement further improves complementarity by allowing information from one source to guide the representation of another (Fang et al. 2022). These developments shift the emphasis from parallel feature extraction to sensor-aware representation learning. A related line of research explicitly separates information shared across modalities from information unique to each sensor. Multi-view learning, feature embedding, and shared–specific decomposition organize these components within a common representation space (Feng et al. 2023; Guo et al. 2024; Hong et al. 2021; Yu et al. 2023). Such methods reduce representational discrepancies while retaining complementary cues that should not be forced into a shared subspace. Hierarchical combinations of convolution and Transformers extend this idea by capturing local structures and long-range context at different representation levels (Sun et al. 2022; Zhao et al. 2023). However, their effectiveness still depends on whether the initial encoding process preserves the information most relevant to each sensor. Representation learning has also benefited from advances in pretraining and data-efficient

learning. Masked reconstruction and multitask pretraining exploit unlabeled observations to learn transferable visual representations before task-specific optimization (Bachmann et al. 2022; He et al. 2022). Label-efficient and few-shot methods reduce the dependence on densely annotated remote sensing data (Wang et al. 2025; Zhang et al. 2025). More recently, shared–specific generative modeling and modality-aware prompts have extended multimodal models to missing or variable input modalities (Gao et al. 2026; Qu et al. 2024). In parallel, public benchmarks and data-fusion contests have established comparable evaluation protocols across different sensor combinations (Debes et al. 2014; Hong et al. 2021; Okujeni et al. 2016; Xu et al. 2019).

These studies collectively show that dimensional alignment alone cannot resolve the representational differences between heterogeneous sensors. MSAE therefore uses sensor-specific encoding paths for HSI and auxiliary data, while converting their outputs into token sequences with a consistent format. This design preserves modality-dependent information without restricting subsequent cross-modal interaction.

2.2 Cross-modal attention fusion

Attention-based fusion has moved multimodal learning beyond static concatenation and fixed feature weighting toward explicit dependency modeling. Channel, spatial, and non-local attention first established mechanisms for selecting informative responses across locations and feature dimensions (Wang et al. 2018; Woo et al. 2018). Self-attention and cross-attention subsequently enabled direct modeling of spectral–spatial dependencies and cross-source correspondences (Mohla et al. 2020; Wang et al. 2022; Ying et al. 2022). Unlike concatenation-based fusion, these mechanisms determine which information should be exchanged according to the content of each sample. Remote sensing studies have adapted this principle to the structural differences between sensor observations. Graph attention models irregular relationships, while adversarial constraints encourage compatible multimodal representations. Hybrid convolution–attention architectures and SAR–optical Transformers combine local feature extraction with long-range interaction (Cai et al. 2024; Fan et al. 2024; Li et al. 2023a; Liu et al. 2024). These methods improve cross-modal exchange but primarily focus on establishing dependencies between modalities. Progressive fusion and cross-modal semantic enhancement instead distribute information exchange across successive stages, allowing complementary evidence to be retained and refined during fusion (Han et al. 2024; Li et al. 2023c). This staged design is closely related to the motivation of PCAFM, although the organization and functions of its successive interactions remain different. Recent work has further expanded the scope and efficiency of attention-based fusion. Joint tokens and hierarchical axial attention enlarge the interaction range, whereas neighborhood contrastive learning and masked self-attention reduce interference from irrelevant responses (Liu et al. 2025; Shi et al. 2026; Yang et al. 2024; Zhu et al. 2025). Vision Transformers and CNN–Transformer hybrids connect fine spatial details with global contextual information (Srinivas et al. 2021; Yao et al. 2023; Zhang et al. 2024a). State-space models offer another route to long-range dependency modeling, with computational complexity that grows approximately linearly with sequence length (Gao et al. 2025; Pan et al. 2025a; Zhang et al. 2024b, 2026).

However, most fusion operators emphasize a particular relation type or representation scale. Few establish a continuous transition from cross-modal correspondence to global semantic enhancement and then to within-modality refinement. PCAFM organizes these interactions according to their semantic roles. It first establishes bidirectional cross-modal relations, uses them to strengthen global semantics, and then refines local dependencies within each modality. The dual-gated update regulates the contribution of local information to the semantic and CLS tokens.

2.3 Multimodal feature calibration and joint optimization

Cross-modal interaction can strengthen complementary information, but it may also propagate redundant or inconsistent responses into the two output branches. Existing approaches address this problem through global guidance, attention calibration, multiscale split–reorganization, dynamic interaction, or self-distillation (Hu et al. 2025; Li et al. 2023d, 2025; Lin et al. 2025; Pan et al. 2025b). Although these strategies operate at different levels, they share the objective of suppressing unreliable information after or during multimodal interaction. Feature-level methods refine the fused representation directly, whereas prediction-level constraints encourage agreement between modality branches.

However, branch agreement does not necessarily provide an explicit semantic reference for feature calibration. Most existing approaches regulate interactions or predictions without constructing a sample-specific representation jointly supported by both modalities. CAMC addresses this distinction by deriving a consensus anchor from the common, discrepant, and interactive relations between the two CLS tokens. The anchor provides a shared semantic reference for calibrating both fused branches while retaining their modality-specific information.

3 Methodology

We propose the Progressive Attention Fusion Network with Semantic Consensus (PAFC-Net) for joint classification of HSI and auxiliary remote sensing data (SAR/LiDAR). PAFC-Net progressively coordinates heterogeneous sensor observations to improve joint representation learning for multimodal remote sensing classification. It integrates modality-specific encoding, progressive collaborative attention, and consensus-anchored calibration within a unified framework. Fig. 2 presents the overall architecture of PAFC-Net. The network comprises MSAE, PCAFM, CAMC, and a dual-branch classification head. In PAFC-Net, principal component analysis first reduces the original HSI to 30 spectral components. We then extract local patches of size $1 \times 30 \times 11 \times 11$ and $C_{\mathrm{A}} \times 11 \times 11$ around each target pixel from the reduced HSI and the coregistered auxiliary data, respectively. MSAE applies two complementary encoding paths to accommodate the spectral and spatial properties of HSI and auxiliary data. Following learnable feature tokenization (Sun et al. 2022), both encoded features undergo the same semantic tokenization procedure. A learnable CLS token and positional embedding are then added to each stream. This produces dimensionally aligned HSI and AUX token sequences. PCAFM progressively exchanges information through bidirectional cross-modal interaction, global semantic enhancement, and

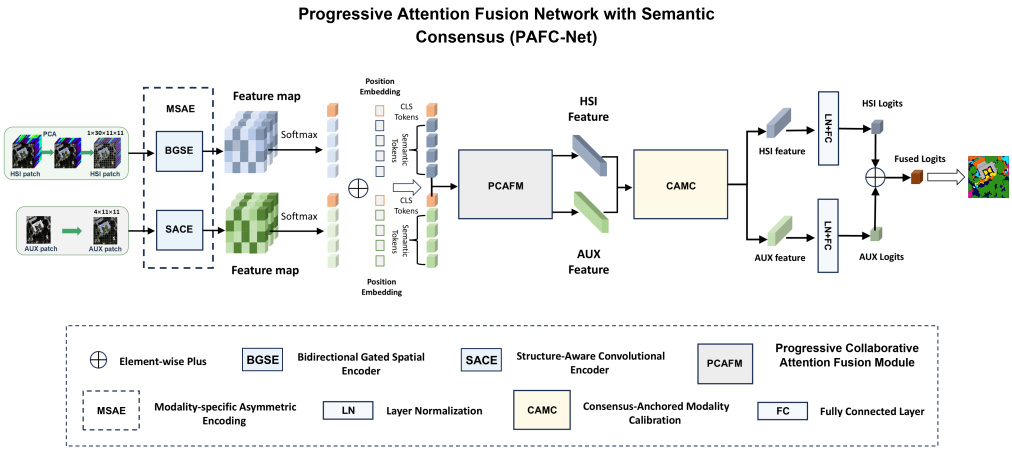


Fig. 2 Overall architecture and end-to-end information flow of the proposed PAFC-Net.

within-modality local context modeling. CAMC subsequently encodes the common, discrepant, and interactive relations between the two CLS tokens. These relations define an adaptive consensus anchor. Anchor-conditioned attention generates a calibration response for each modality, which is injected into the corresponding fused feature through a gated residual connection. The calibrated CLS token from each sequence is passed to an independent linear classifier. Element-wise addition of the two branch logits produces the final prediction.

3.1 Modality-specific asymmetric encoding

HSI and AUX data exhibit distinct representational characteristics because of their different sensing mechanisms. Accordingly, MSAE employs a spectral–spatial encoding path for HSI and a structure-aware convolutional path for AUX. The resulting features are then mapped into a common feature space for subsequent interaction while preserving their modality-specific characteristics.

As shown in Fig. 3, MSAE uses separate HSI and AUX branches to extract modality-specific information in a common representation space. Let the HSI input patch be $X_h \in \mathbb{R}^{1\times 30\times 11\times 11}$. A three-dimensional convolution with a $3\times 3\times 3$ kernel jointly models adjacent bands and local spatial neighborhoods. It produces an initial spectral–spatial representation. The channel and spectral dimensions are then rearranged. A two-dimensional convolution with a 3×3 kernel further aggregates spatial information to obtain the shallow feature map $F_h^0 \in \mathbb{R}^{C_h\times H_h\times W_h}$. We flatten F_h^0 along its spatial dimensions to form the sequence S_h. The sequence is processed in forward and reverse order by depthwise one-dimensional convolutions with kernel size 3 and padding 1. This bidirectional design reduces the directional bias of a single scan. In each direction, Tanh generates candidate responses, while Sigmoid estimates

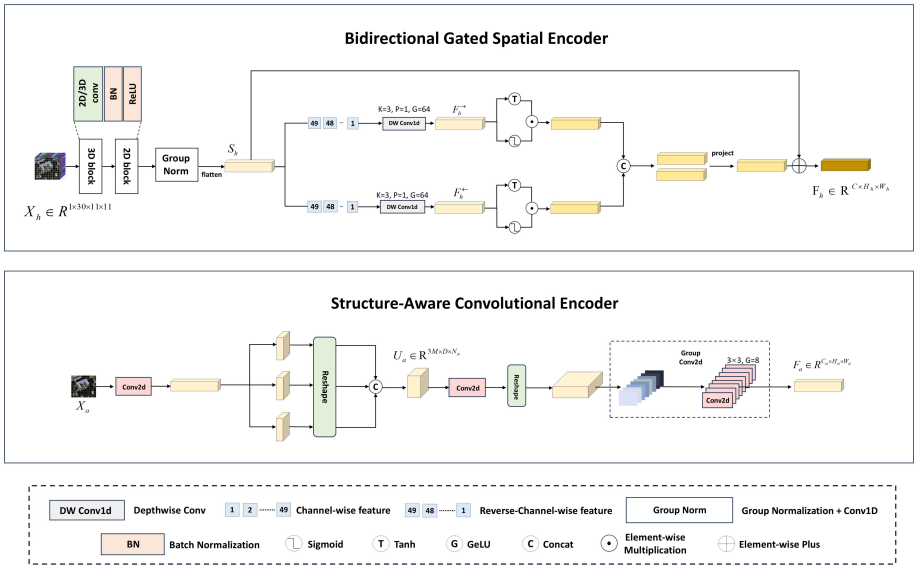


Fig. 3 Detailed architecture of modality-specific asymmetric encoding in MSAE.

selection weights. The weighted responses are concatenated along the feature dimension. A gated residual connection then filters their contributions and yields the HSI representation F_h with both local and bidirectional context:

$$F_h = \mathcal{P}_h\big(U_f^h \parallel U_b^h\big) + S_h. \tag{1}$$

Here, U_f^h and U_b^h denote the forward and backward contextual responses, respectively. Each response is produced by depthwise convolution, Tanh candidate generation, and Sigmoid gating. The backward response is restored to the original spatial order before fusion.

The auxiliary modality primarily describes spatial morphology and structure. Its branch first applies a two-dimensional convolution with a 3×3 kernel for local feature mapping. Group reorganization then forms distinct feature subspaces. A grouped two-dimensional convolution with a 3×3 kernel extracts local features from each subspace. Aggregation and channel projection produce the AUX representation as follows:

$$F_a = \mathcal{C}_a(\mathrm{GConv}_{2\mathrm{D}}(\mathcal{G}(\mathrm{Conv}_{2\mathrm{D}}(X_a)))) \in \mathbb{R}^{C_a \times H_a \times W_a}. \tag{2}$$

In Eq. (2), X_a is the input AUX patch, $\mathrm{Conv}_{2\mathrm{D}}$ denotes the initial two-dimensional convolution, $\mathcal{G}$ represents feature reorganization into predefined groups, and $\mathrm{GConv}_{2\mathrm{D}}$ denotes the grouped convolution. $\mathcal{C}_a$ denotes feature aggregation and channel projection. The resulting AUX feature is F_a, whose channel number, height, and width are C_a, H_a, and W_a, respectively.

This design preserves the spatial structure of the auxiliary modality while matching the HSI feature dimension. We then tokenize F_h and F_a into a fixed number of semantic tokens and add positional embeddings. The operation is defined as:

$$T_m = \left[c_m \,;\, \mathrm{Softmax}\left(Z_m W_m^{\top}\right) V_m\right] + P_m, \qquad m \in \{h, a\}. \tag{3}$$

Here, m indexes the modality, with h and a denoting HSI and AUX, respectively. Z_m is the flattened feature sequence from encoder m, W_m denotes the learnable projection matrix for token weights, and P_m is the positional embedding. The semicolon indicates concatenation along the token dimension. Normalized assignment weights aggregate the projected value features into a fixed number of semantic tokens. Following learnable feature tokenization (Sun et al. 2022), a learnable CLS token is concatenated with each token sequence before positional embeddings are added. This gives the unified HSI and AUX sequences (T_h, T_a), which form the two inputs to PCAFM.

3.2 Progressive collaborative attention fusion module

After MSAE encoding and tokenization, HSI and AUX are represented by dimensionally aligned token sequences (T_h, T_a). Dimensional alignment alone, however, does not ensure sufficient interaction. The modalities still differ in semantic distribution and local structure. Relying solely on cross-modal interaction does not explicitly address how the exchanged information should enhance global semantics and refine dependencies within each modality. PCAFM therefore performs progressive fusion from cross-modal to within-modality relations through three attention stages. A dual-gated unit then controls the final information injection.

As shown in Fig. 4, PCAFM comprises bidirectional cross-modal attention (BCMA), semantic enhancement attention (SEA), and local multi-head self-attention (LMSA). BCMA first establishes bidirectional dependencies between the HSI and AUX tokens. SEA uses the exchanged global relations to enhance the original content of each modality. LMSA then refines local token dependencies within each modality. Finally, the dual-gated unit (DGU) updates the local and CLS tokens through separate paths. The overall mapping of PCAFM is:

$$(T_h^{\mathrm{ref}}, T_a^{\mathrm{ref}}) = \mathcal{G}(\mathrm{LMSA}(\mathrm{SEA}(\mathrm{BCMA}(T_h, T_a)))), \tag{4}$$

where $\mathcal{G}$ denotes the dual-gated residual update performed by DGU. T_h^{ref} and T_a^{ref} are the refined HSI and AUX feature sequences produced by PCAFM, respectively.

PCAFM begins with BCMA, whose structure is shown in Fig. 5(a). BCMA establishes fine-grained correspondences between HSI and AUX. In the AUX-guided HSI branch, the HSI sequence is projected as the query, while the AUX sequence is projected as the key and value. The two modalities exchange these roles in the HSI-guided AUX branch. Both directions use the same attention formulation:

$$T'_m = \mathrm{MHA}(Q_m, K_n, V_n) W_m^O + T_m, \qquad m, n \in \{h, a\},\ m \neq n. \tag{5}$$

Here, m denotes the target modality to be updated and n is the source modality providing complementary information. d is the dimension of each attention head, $\|$ denotes multi-head concatenation, and W_m^O is the output projection. The residual term T_m preserves the original target representation. T'_m is the cross-modally enhanced sequence produced by BCMA.

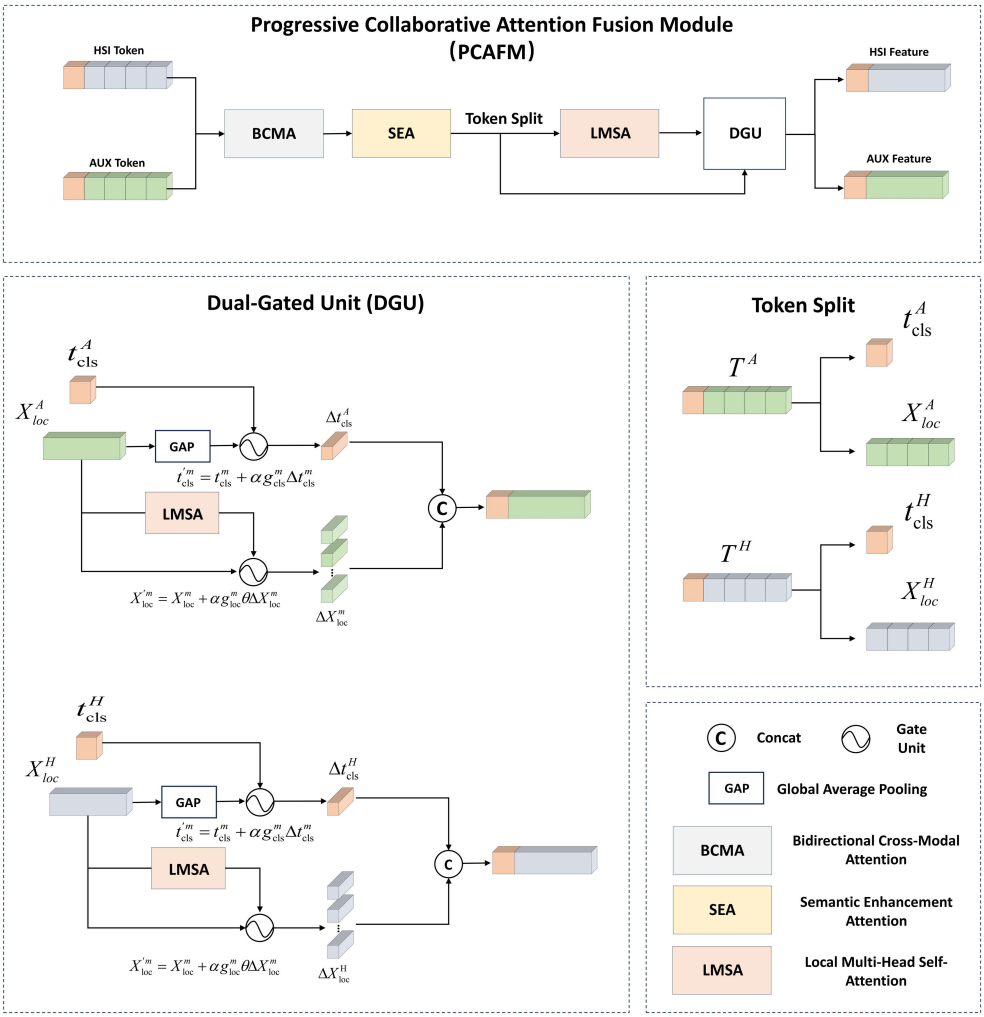


Fig. 4 Progressive cross-modal interaction and dual-gated feature refinement in PCAFM.

The enhanced features are passed to SEA, shown in Fig. 5(b). SEA uses the learned relation structure to recalibrate the original modality content:

$$T_m^{\text{sem}} = \text{Softmax}\left(\frac{Q_m K_m^\top}{\sqrt{d}}\right) V_m, \qquad m \in \{h, a\}. \tag{6}$$

In Eq. (6), Q_m and K_m are projected from the cross-modally updated sequence T'_m, whereas V_m is projected from the original sequence T_m. Thus, SEA uses the updated sequence to determine semantic affinities while aggregating values that retain the original modality content.

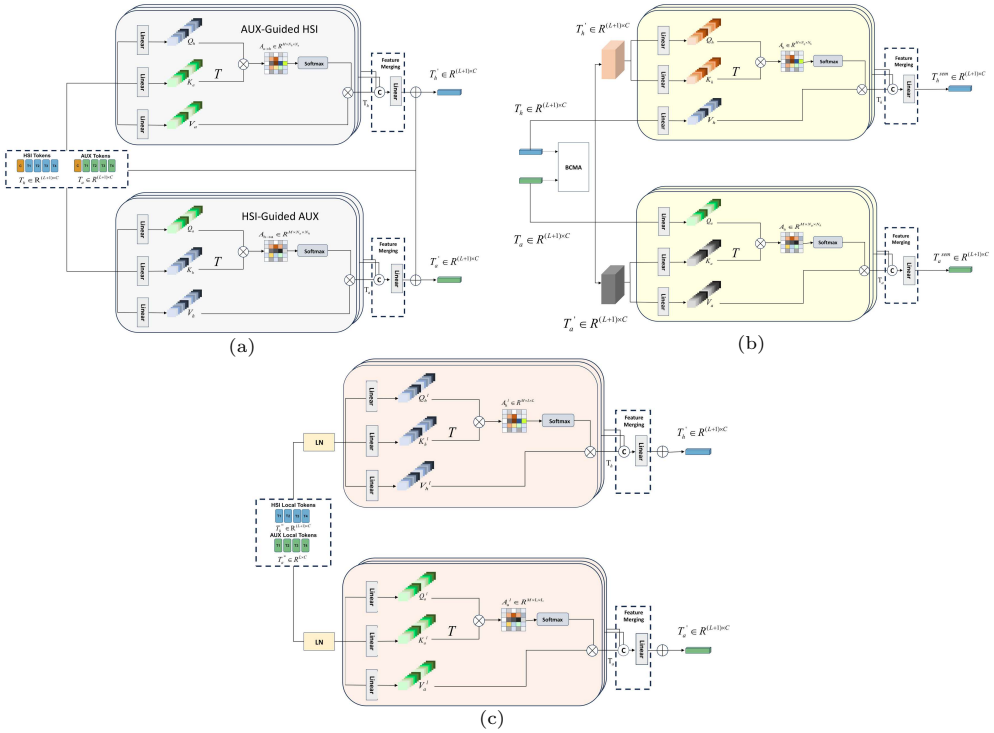


Fig. 5 Architectures of the attention operations in PCAFM. (a) Bidirectional cross-modal attention. (b) Semantic enhancement attention. (c) Local multi-head self-attention.

After semantic enhancement, LMSA models dependencies among local tokens within each modality. As shown in Fig. 5(c), it separates the CLS token from the local tokens and applies layer normalization and multi-head linear projections. Its local attention increment is:

$$\Delta L_m = \mathrm{Concat}_{j=1}^{M}\left[\mathrm{Softmax}\left(\frac{Q_{m,j}K_{m,j}^{\top}}{\sqrt{d}}\right)V_{m,j}\right]W_m^L, \qquad m \in \{h, a\}. \tag{7}$$

Here, L_m denotes the local-token sequence of modality m, M is the number of attention heads, j is the head index, and W_m^L is the output projection after multi-head concatenation. The DGU then performs separate local and CLS updates:

$$L_m^{\mathrm{ref}} = L_m + \alpha g_m^l \Delta L_m, \tag{8}$$

$$c_m^{\mathrm{ref}} = c_m + \alpha g_m^c \,\mathrm{Mean}(\Delta L_m), \tag{9}$$

$$T_m^{\mathrm{ref}} = [c_m^{\mathrm{ref}}; L_m^{\mathrm{ref}}]. \tag{10}$$

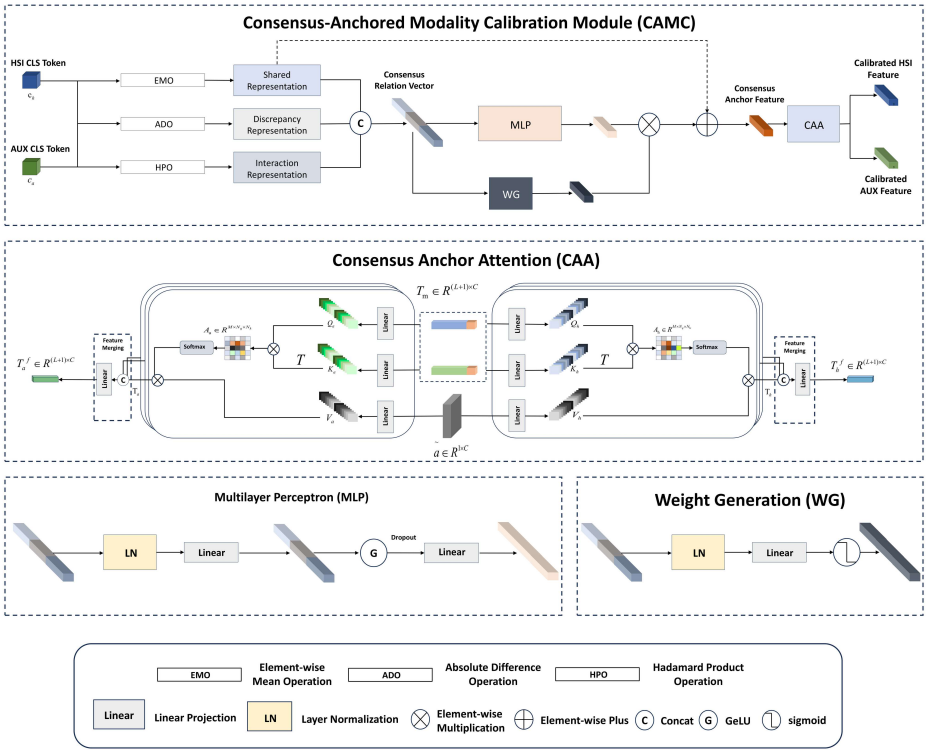


Fig. 6 Consensus-anchor construction and anchor-guided dual-branch feature calibration in CAMC.

The gates g_m^l and g_m^c are generated separately for local and CLS updates, and α is a learnable residual scale initialized to 0.05. Across these stages, PCAFM first establishes bidirectional correspondences between HSI and AUX. It then uses cross-modal context to enhance global semantics and models local dependencies within each modality. DGU controls how the local updates affect the local and CLS tokens.

3.3 Consensus-anchored modality calibration module

PCAFM progressively transfers cross-modal information through token correlations. However, its outputs remain in two separate modality branches and may retain redundant or locally inconsistent discriminative responses. Independent classifiers alone may therefore produce divergent decisions. CAMC explicitly encodes common, discrepant, and interactive relations between the two CLS tokens. It uses these relations to construct a sample-adaptive consensus anchor that calibrates both branches against the same semantic reference.

As shown in Fig. 6, CAMC first computes the element-wise mean, absolute difference, and element-wise multiplication of the two CLS tokens:

$$s = \frac{c_h + c_a}{2}, \qquad d = |c_h - c_a|, \qquad p = c_h \odot c_a, \qquad r = [s; d; p]. \tag{11}$$

Here, s, d, and p denote the shared, discrepancy, and interaction representations, respectively. The consensus relation vector r determines a candidate anchor increment and its weight. The sample-adaptive consensus anchor is:

$$a = s + \sigma(\mathrm{WG}(r)) \odot \mathrm{MLP}(r). \tag{12}$$

CAMC applies consensus anchor attention (CAA) for conditional calibration instead of directly adding the anchor to both modalities. The anchor is broadcast to the length of the current modality sequence. Each PCAFM output generates the queries and keys, whereas the broadcast anchor generates the values. For modality $m \in \{h, a\}$, the CAA response is:

$$A_m = \mathrm{Softmax}\left(\frac{Q_m K_m^\top}{\sqrt{d}}\right) V_m. \tag{13}$$

Here, Q_m and K_m are linear projections of the modality sequence and describe semantic relations among its tokens. V_m is projected from the broadcast consensus anchor a and supplies a common semantic reference to each token. To preserve the modality-specific information retained by PCAFM, CAMC does not inject the consensus response without control. The CAA response and the PCAFM output jointly estimate a modality-dependent injection weight. A residual connection then adds the weighted response to each branch and produces the calibrated sequences. Their calibrated CLS tokens are passed to independent linear classifiers, and element-wise addition of the two branch logits yields the final prediction.

Algorithm 1 summarizes the end-to-end training procedure.

Algorithm 1 Training procedure of PAFC-Net

Require: HSI patches X_h, AUX patches X_a, and labels y
Ensure: Fused logits z and training loss $\mathcal{L}$

$(F_h, F_a) \leftarrow \mathrm{MSAE}(X_h, X_a)$
$T_m \leftarrow \mathrm{Tokenize}(F_m) + P_m$ for $m \in \{h, a\}$
$U_h \leftarrow \mathrm{BCMA}(T_h, T_a)$; $U_a \leftarrow \mathrm{BCMA}(T_a, T_h)$
$V_m \leftarrow \mathrm{SEA}(U_m, T_m)$; $\Delta L_m \leftarrow \mathrm{LMSA}(V_m)$
$T_m^{\mathrm{ref}} \leftarrow \mathrm{DGU}(U_m, V_m, \Delta L_m)$
Compute s, d, p, and r from the two CLS tokens
$a \leftarrow s + \sigma(\mathrm{WG}(r)) \odot \mathrm{MLP}(r)$
$A_m \leftarrow \mathrm{CAA}(T_m^{\mathrm{ref}}, a)$ for $m \in \{h, a\}$
$T_m^{\mathrm{c}} \leftarrow \mathrm{CAMC}(T_m^{\mathrm{ref}}, A_m)$
$z_m \leftarrow \mathcal{C}_m(T_m^{\mathrm{c}}[:, 0])$; $z \leftarrow z_h + z_a$
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{focal}} + \lambda_b \mathcal{L}_{\mathrm{branch}} + \lambda_m \mathcal{L}_{\mathrm{mutual}}$
Update Θ using Adam and return z and $\mathcal{L}$

3.4 Balanced multi-branch loss

PAFC-Net is optimized with a fused focal loss, branch-wise cross-entropy supervision, and bidirectional mutual consistency. The focal term is applied to the fused logits and

reduces the gradient contribution from abundant easy samples. Cross-entropy terms maintain the independent discriminative capacity of the HSI and AUX branches. A symmetric temperature-scaled KL term encourages their predictive distributions to remain compatible during collaborative fusion. The overall objective is:

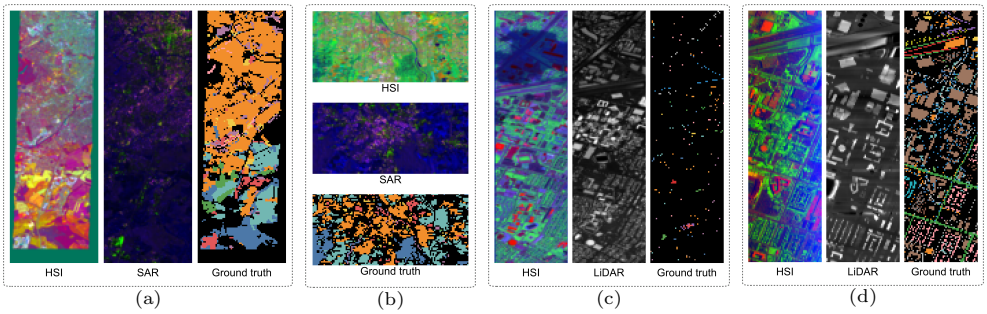
$$\mathcal{L} = \mathcal{L}_{\text{focal}}(z, y) + \lambda_b \left[\mathcal{L}_{\text{CE}}(z_h, y) + \mathcal{L}_{\text{CE}}(z_a, y)\right] + \lambda_m \mathcal{L}_{\text{mutual}}. \tag{14}$$

Here, $\mathcal{L}_{\text{focal}}$ is applied to the fused logits. The two $\mathcal{L}_{\text{CE}}$ terms supervise the HSI and AUX branch logits, respectively. $\mathcal{L}_{\text{mutual}}$ is the symmetric temperature-scaled KL divergence between their predictive distributions. λ_b and λ_m weight the branch-supervision and mutual-consistency terms.

4 Experimental Results

4.1 Datasets and implementation details

We evaluated PAFC-Net on four public multimodal remote sensing datasets: Berlin, Augsburg, Houston 2013, and Houston 2018. The Berlin and Augsburg benchmarks were compiled by Hong et al. (2021). The Berlin HSI was simulated from HyMap data as a 30-m EnMAP product (Okujeni et al. 2016). We used HSI and SAR data for both benchmarks. Houston 2013 and Houston 2018 originate from the 2013 and 2018 IEEE GRSS Data Fusion Contests, respectively. Both datasets combine HSI with LiDAR data (Debes et al. 2014; Xu et al. 2019). The four benchmarks differ in scene extent, class distribution, and auxiliary modality. They therefore assess cross-modal classification under varied conditions. Fig. 7 shows their modality composition and land-cover distribution. Tables 1–4 list the classes and the numbers of training and test samples. The Berlin dataset contains coregistered HSI and SAR images



(a) (b) (c) (d)

Fig. 7 Spatial distributions and modality compositions of the four multimodal remote sensing datasets. (a) Berlin. (b) Augsburg. (c) Houston 2013. (d) Houston 2018.

with a spatial size of 1723 × 476 pixels. The HSI contains 244 spectral bands, and the SAR image provides four polarimetric feature channels. Eight land-cover classes are included: Forest, Residential Area, Industrial Area, Low Plants, Soil, Allotment, Commercial Area, and Water. The predefined split contains 2,820 labeled training samples and 461,851 labeled test samples. Berlin covers an extensive area with a

Table 1 Numbers of training and test samples in the Berlin benchmark dataset.

No.	Class	Color	Training	Test
1	Forest		443	54,511
2	Residential Area		423	268,219
3	Industrial Area		499	19,067
4	Low Plants		376	58,906
5	Soil		331	17,095
6	Allotment		280	13,025
7	Commercial Area		298	24,526
8	Water		170	6,502
		Total	**2,820**	**461,851**

Table 2 Numbers of training and test samples in the Augsburg benchmark dataset.

No.	Class	Color	Training	Test
1	Forest		146	13,361
2	Residential Area		264	30,065
3	Industrial Area		21	3,830
4	Low Plants		248	26,609
5	Allotment		52	523
6	Commercial Area		7	1,638
7	Water		23	1,507
		Total	**761**	**77,533**

Table 3 Numbers of training and test samples in the Houston 2013 benchmark dataset.

No.	Class	Color	Training	Test	No.	Class	Color	Training	Test
1	Healthy Grass		198	1,053	9	Road		193	1,059
2	Stressed Grass		190	1,064	10	Highway		191	1,036
3	Synthetic Grass		192	505	11	Railway		181	1,054
4	Trees		188	1,056	12	Parking Lot 1		192	1,041
5	Soil		186	1,056	13	Parking Lot 2		184	285
6	Water		182	143	14	Tennis Court		181	247
7	Residential		196	1,072	15	Running Track		187	473
8	Commercial		191	1,053					
							Total	**2,832**	**12,197**

Table 4 Numbers of training and test samples in the Houston 2018 benchmark dataset.

No.	Class	Color	Training	Test	No.	Class	Color	Training	Test
1	Healthy Grass		1,000	38,196	11	Sidewalks		1,000	135,035
2	Stressed Grass		1,000	129,008	12	Crosswalks		1,000	5,059
3	Artificial Turf		1,000	1,736	13	Major Thoroughfares		1,000	184,438
4	Evergreen Trees		1,000	53,322	14	Highways		1,000	38,438
5	Deciduous Trees		1,000	19,172	15	Railways		1,000	26,748
6	Bare Earth		1,000	17,064	16	Paved Parking Lots		1,000	44,932
7	Water		500	564	17	Unpaved Parking Lots		250	337
8	Residential Buildings		1,000	157,995	18	Cars		1,000	25,289
9	Non-residential Buildings		1,000	893,769	19	Trains		1,000	20,479
10	Roads		1,000	182,283	20	Stadium Seats		1,000	26,296
							Total	**18,750**	**2,000,160**

complex spatial distribution and evaluates joint spectral–radar representation learning under severe class imbalance.

The Augsburg dataset also combines coregistered HSI and SAR data and has a spatial size of 332×485 pixels. Its HSI contains 180 spectral bands, while its SAR data comprise four feature channels. The scene includes seven classes. We used 761 labeled samples for training and 77,533 for testing. Augsburg provides fewer training samples and a more imbalanced class distribution than Berlin, which tests whether spectral and structural information can be fused under limited supervision.

Houston 2013 consists of coregistered HSI and LiDAR data with a spatial size of 349×1905 pixels. The HSI contains 144 spectral bands, and LiDAR provides one elevation channel. We followed the official split of 2,832 labeled training samples and 12,197 test samples. Houston 2018 spans 1202×4768 pixels, contains 50 HSI spectral bands and one LiDAR elevation channel, and is divided into 20 classes. We used the predefined split of 18,750 labeled training samples and 2,000,160 test samples. To reduce spectral redundancy and computational cost, principal component analysis compressed each HSI to 30 components. The HSI and AUX channels were then standardized separately using z-score normalization. For every labeled pixel, we extracted spatially corresponding patches from the two coregistered modalities. For Berlin, Houston 2013, and Houston 2018, the input patch size was 11×11. Augsburg used 9×9 because its training set is smaller and more imbalanced. Both encoders mapped their inputs to a 64-dimensional feature space, and the number of semantic tokens was fixed at $L = 4$. The loss weights were $\lambda_b = 0.3$ and $\lambda_m = 0.05$ for all datasets. Focal loss used $\alpha = 0.25$ and $\gamma = 2$, while the temperature of the mutual-consistency term was set to 2.0. We used Adam with an initial learning rate of 1×10^{-3}. Training and test batch sizes were 128 and 512, respectively. All experiments were implemented in PyTorch with CUDA 11.8 and run on an NVIDIA GeForce RTX 2080 Ti GPU. We fixed the random seed at 2026.

No geometric or spectral data augmentation was applied. The encoders and attention modules used dropout 0.1. Adam used $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and zero weight decay. The learning rate remained fixed at 10^{-3} without a scheduler. All models were trained from scratch for 40 epochs on Berlin, Augsburg, and Houston 2013, and for 100 epochs on Houston 2018.

4.2 Comparison with state-of-the-art methods

4.2.1 Quantitative comparison

We used overall accuracy (OA), average accuracy (AA), and the Kappa coefficient as the main evaluation metrics. OA measures the proportion of correctly classified test samples. AA averages per-class accuracies and is informative under class imbalance. Kappa measures agreement between predictions and reference labels after accounting for chance. Tables 5–8 also report per-class accuracy. Bold values indicate the best result for each class or metric, and underlined values indicate the second-best result.

We compared PAFC-Net with nine representative methods: TBCNN (Xu et al. 2018), FusAtNet (Mohla et al. 2020), ExViT (Yao et al. 2023), HCT (Zhao et al. 2023), MACN (Li et al. 2023d), HLMamba (Liao et al. 2024), MSFMamba (Gao et al. 2025), S^2CrossMamba (Zhang et al. 2024b), and S^2ENet (Fang et al. 2022).

As shown in Table 5, PAFC-Net achieved an OA of 80.45%, an AA of 66.16%, and a Kappa of 69.32% on Berlin. All three aggregate metrics exceeded those of the compared methods. Relative to the strongest competing result for each metric, the gains were 4.15, 0.62, and 5.72 percentage points, respectively. PAFC-Net obtained the highest accuracy for Low Plants at 89.64% and reached 87.36% for Residential Area. Commercial Area and other underrepresented classes remained difficult. Even so, the

Table 5 Classification performance (%) of different methods on the Berlin dataset.

Class	TBCNN	FusAtNet	ExViT	HCT	MACN	HLMamba	MSFMamba	S^2CrossMamba	S^2ENet	PAFC-Net (Ours)
Forest	77.88	46.45	77.72	80.91	69.08	78.17	81.49	81.08	**82.82**	80.01
Residential Area	78.87	84.06	80.70	**88.83**	84.11	82.75	80.31	81.89	77.69	87.36
Industrial Area	54.71	43.02	60.11	39.86	45.13	51.43	57.42	51.23	**71.18**	52.76
Low Plants	70.21	65.86	47.41	60.16	81.26	58.55	78.40	64.41	68.74	**89.64**
Soil	87.69	**90.73**	70.47	80.08	77.95	65.53	87.97	69.47	78.47	79.95
Allotment	71.12	62.44	70.69	44.38	67.11	**72.74**	61.60	59.47	67.56	65.59
Commercial Area	20.36	21.20	19.10	18.96	16.92	23.72	27.13	**41.09**	30.90	20.85
Water	63.52	51.40	51.43	43.86	47.22	**71.58**	49.03	53.14	42.79	53.11
OA	73.43	71.45	70.91	76.30	75.57	73.62	75.75	74.64	73.65	**80.45**
AA	65.54	58.15	59.70	57.13	61.10	63.06	65.42	62.72	65.02	**66.16**
Kappa	60.47	55.68	56.35	62.58	62.00	60.16	63.60	61.59	61.24	**69.32**

Table 6 Classification performance (%) of different methods on the Augsburg dataset.

Class	TBCNN	FusAtNet	ExViT	HCT	MACN	HLMamba	MSFMamba	S^2CrossMamba	S^2ENet	PAFC-Net (Ours)
Forest	95.25	93.40	95.16	97.72	96.67	93.26	91.67	94.33	**98.55**	96.03
Residential Area	97.08	96.07	96.87	98.33	98.36	98.95	99.13	98.34	98.20	**99.19**
Industrial Area	31.36	44.28	46.34	53.37	54.13	59.09	20.18	46.34	44.20	**72.77**
Low Plants	82.09	82.83	91.66	96.23	**98.20**	92.32	95.12	88.55	96.79	94.69
Allotment	46.65	60.61	52.58	**72.85**	40.92	58.70	48.57	44.93	68.26	61.38
Commercial Area	1.16	**8.91**	2.50	5.31	3.91	3.17	1.78	0.00	6.96	0.31
Water	49.64	0.00	16.06	**52.55**	14.13	49.44	36.68	12.87	52.42	49.17
OA	85.08	84.56	88.43	92.26	91.81	90.47	88.96	87.62	92.09	**92.48**
AA	57.60	55.16	57.31	**68.05**	58.04	64.99	56.16	55.05	66.48	67.65
Kappa	78.48	77.71	83.25	88.80	88.04	86.21	83.68	81.97	88.53	**89.11**

Table 7 Classification performance (%) of different methods on the Houston 2013 dataset.

Class	TBCNN	ExViT	HCT	MACN	HLMamba	MSFMamba	S^2CrossMamba	S^2ENet	PAFC-Net (Ours)
Healthy Grass	81.29	81.01	82.91	82.53	81.10	81.86	79.87	82.43	**85.47**
Stressed Grass	83.27	81.39	95.30	99.72	86.28	99.62	77.26	82.14	**100.00**
Synthetic Grass	98.02	96.83	98.22	96.63	97.62	97.82	96.24	96.83	**99.80**
Trees	87.88	91.48	92.14	97.44	86.46	**99.43**	86.65	88.45	97.35
Soil	99.91	**100.00**	99.53	**100.00**	**100.00**	99.71	99.91	**100.00**	**100.00**
Water	95.80	79.02	**100.00**	98.60	95.80	95.80	95.10	97.20	91.61
Residential	82.65	**90.49**	89.74	77.15	83.96	74.53	83.68	76.59	88.81
Commercial	86.23	**97.25**	94.21	79.87	95.73	93.65	82.15	82.05	92.21
Road	91.22	90.37	83.38	95.18	**98.58**	86.21	84.80	82.81	83.29
Highway	**99.81**	71.04	93.82	74.90	86.68	61.10	94.79	95.66	91.12
Railway	94.88	**98.20**	89.75	94.59	93.93	95.97	96.30	96.68	93.17
Parking Lot 1	98.75	92.03	98.94	97.98	98.75	98.37	97.60	98.75	**99.71**
Parking Lot 2	87.37	88.77	90.18	90.18	85.26	90.88	79.65	87.37	**91.23**
Tennis Court	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Running Track	**100.00**	**100.00**	**100.00**	**100.00**	99.79	**100.00**	**100.00**	**100.00**	99.37
OA	91.40	90.15	92.74	90.92	91.83	90.18	89.16	89.60	**93.70**
AA	92.47	90.52	93.87	92.32	92.66	91.66	90.26	91.13	**94.21**
Kappa	90.66	89.32	92.12	90.14	91.14	89.34	88.23	88.71	**93.16**

model improved overall accuracy and agreement while retaining competitive class-balanced accuracy. These results indicate a more stable joint representation when HSI and SAR provide markedly different information. Table 6 reports the results on Augsburg. PAFC-Net achieved the highest OA and Kappa, reaching 92.48% and 89.11%, respectively. These values exceeded those of the second-ranked HCT by 0.22 and 0.31 percentage points. Its AA was 67.65%, which ranked second and was 0.40 percentage points below HCT. PAFC-Net attained the highest accuracies for Residential Area and Industrial Area at 99.19% and 72.77%. The gain for Industrial Area over the next-best method was 13.68 percentage points. This result indicates improved discrimination between urban classes with similar spectra but different spatial structures.

On Houston 2013, PAFC-Net achieved the best OA, AA, and Kappa values of 93.70%, 94.21%, and 93.16%, respectively, as shown in Table 7. The corresponding gains over the next-best results were 0.96, 0.34, and 1.04 percentage points. PAFC-Net also obtained the highest accuracies for Healthy Grass, Stressed Grass, Synthetic Grass, Parking Lot 1, and Parking Lot 2. The scene contains interwoven vegetation,

Table 8 Classification performance (%) of different methods on the Houston 2018 dataset.

Class	TBCNN	ExViT	HCT	MACN	HLMamba	MSFMamba	S^2CrossMamba	S^2ENet	PAFC-Net (Ours)
Healthy Grass	93.69	96.74	93.99	95.49	95.98	94.39	93.41	**98.16**	95.35
Stressed Grass	87.91	87.12	**93.30**	90.88	91.38	90.79	88.67	87.69	92.39
Artificial Turf	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Evergreen Trees	97.72	97.39	99.25	98.96	**99.45**	99.18	98.51	99.19	98.98
Deciduous Trees	96.56	96.76	98.79	98.98	99.16	99.23	98.10	**99.58**	99.21
Bare Earth	**99.99**	99.66	99.98	99.95	**99.99**	99.98	99.88	99.93	**99.99**
Water	99.65	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Residential Buildings	91.28	88.86	97.11	95.73	96.84	96.31	96.16	**97.30**	96.41
Non-residential Buildings	87.78	87.28	92.79	93.63	94.81	92.10	92.86	94.69	**94.86**
Roads	58.99	39.59	75.02	78.04	74.28	75.79	71.37	75.90	**79.89**
Sidewalks	63.47	67.56	78.08	77.82	79.47	79.14	76.22	79.83	**81.31**
Crosswalks	80.53	84.72	96.84	98.22	98.70	96.40	95.24	**99.27**	97.81
Major Thoroughfares	69.87	63.87	84.92	78.81	**87.39**	82.51	82.15	87.23	85.37
Highways	94.68	94.03	97.77	97.64	98.21	94.92	96.98	**98.38**	98.08
Railways	99.21	99.57	99.79	99.81	99.63	99.80	99.69	99.62	**99.90**
Paved Parking Lots	96.19	96.70	97.44	97.43	99.15	97.71	97.47	**99.77**	98.13
Unpaved Parking Lots	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Cars	95.56	94.57	98.52	98.22	97.45	98.54	97.78	97.71	**98.89**
Trains	99.38	98.21	99.97	99.97	**99.99**	99.75	99.83	**99.99**	99.89
Stadium Seats	99.74	99.90	99.99	99.99	99.93	**100.00**	**100.00**	**100.00**	**100.00**
OA	83.56	81.09	90.69	90.52	91.79	90.04	89.57	91.75	**92.26**
AA	90.61	89.63	95.18	94.98	95.59	94.83	94.21	95.71	**95.82**
Kappa	79.16	76.09	88.05	87.81	89.41	87.24	86.60	89.37	**90.01**

roads, and artificial surfaces. The results show that PAFC-Net combines HSI spectra with LiDAR structure to improve separation among these land-cover classes.

Houston 2018 is larger and contains a more complex class composition. PAFC-Net achieved the best OA, AA, and Kappa values of 92.26%, 95.82%, and 90.01%, respectively, in Table 8. These results exceeded the next-best values by 0.47, 0.11, and 0.60 percentage points. The method also ranked first for Bare Earth, Non-residential Buildings, Roads, Sidewalks, Railways, and Cars. Its advantages on roads and buildings suggest effective integration of spectral responses and spatial geometry in large urban areas.

PAFC-Net achieved the highest OA and Kappa on all four datasets. Its AA was also the highest on Berlin, Houston 2013, and Houston 2018, and ranked second on Augsburg. These datasets cover both HSI–SAR and HSI–LiDAR combinations and vary in scale, class count, and sample distribution. The results indicate that the gains are not confined to one sensor combination or a single scene. Instead, the network forms reliable joint representations across varied multimodal classification settings.

4.2.2 Qualitative comparison

Figs. 8–11 compare classification maps for Berlin, Augsburg, Houston 2013, and Houston 2018, respectively. Red boxes identify representative regions, and each panel includes an enlarged view for examining class confusion and boundary recovery. Several comparison methods recovered the dominant land-cover patterns but produced scattered errors, fragmented regions, and displaced boundaries. The maps from PAFC-Net were more spatially continuous and preserved regional structures more faithfully. In the enlarged Berlin region in Fig. 8, ExViT, HCT, and S^2ENet produced responses from incorrect classes, creating visible discontinuities in otherwise coherent land-cover areas. PAFC-Net reduced local interference among Allotment, Soil, and Commercial Area, and more accurately preserved the boundary between Forest and Low Plants. The enlarged Augsburg region in Fig. 9 contains narrow objects and boundaries between adjacent classes. Several methods confused classes in the lower part of the region. PAFC-Net produced a more continuous main region and retained narrow

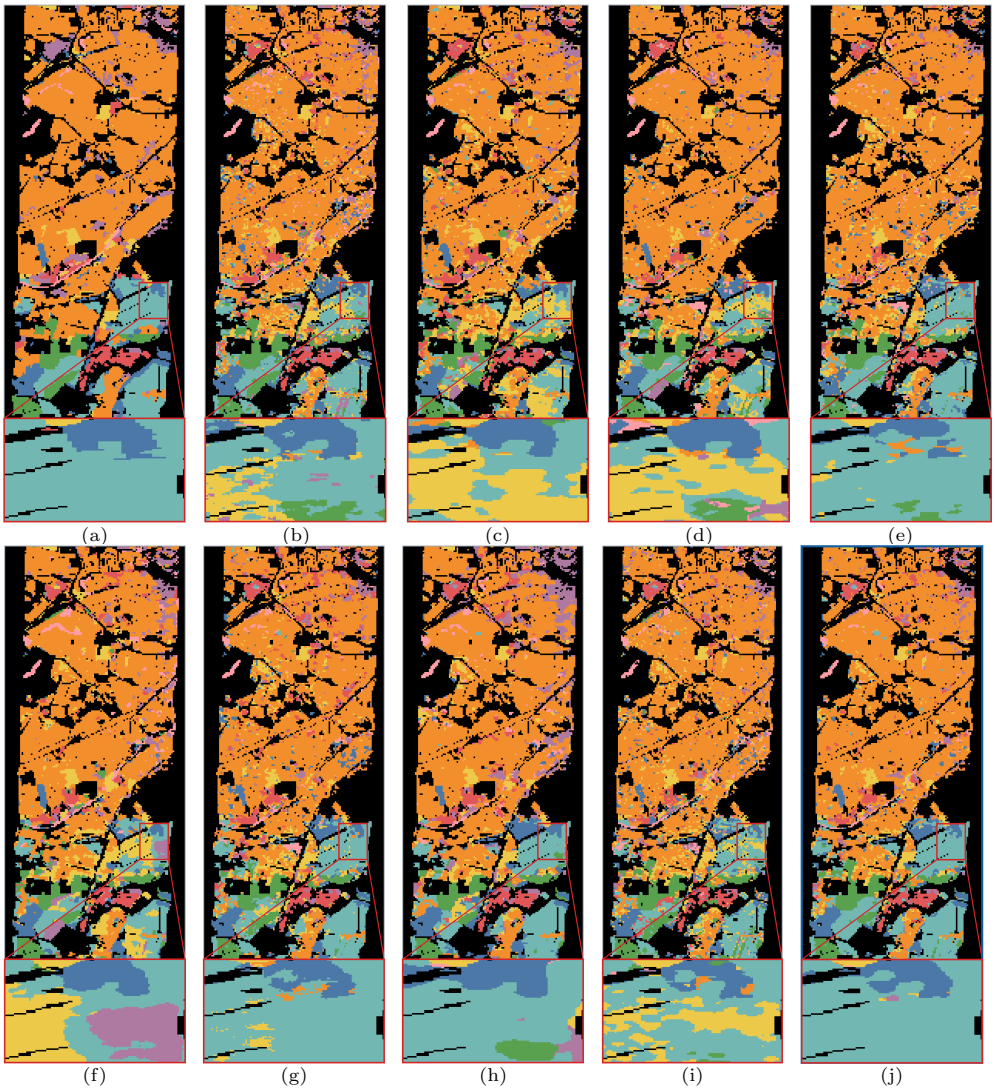


Fig. 8 Classification maps and enlarged local regions on the Berlin dataset. (a) Ground truth. (b) TBCNN. (c) ExViT. (d) HCT. (e) MACN. (f) HLMamba. (g) MSFMamba. (h) S^2CrossMamba. (i) S^2ENet. (j) PAFC-Net.

structures and their boundaries more faithfully, thereby reducing fragmented local predictions. Houston 2013 contains densely distributed urban objects with similar shapes. As shown in Fig. 10, several comparison methods distorted object geometry in the enlarged region. PAFC-Net better recovered the central outline and the surrounding strip-like structures, leading to predictions that were more consistent with the reference map. In the enlarged Houston 2018 region in Fig. 11, roads, vegetation strips, and adjacent buildings form complex local arrangements. PAFC-Net better preserved these spatial relationships and reduced scattered errors within linear structures. Across

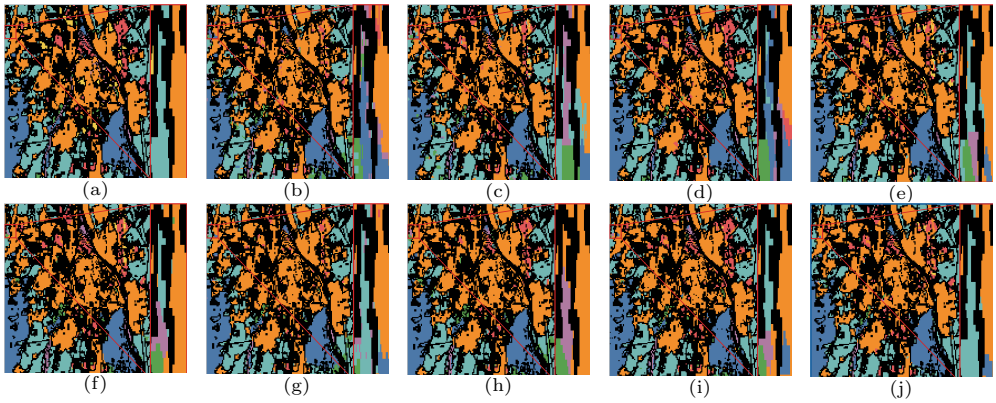


Fig. 9 Classification maps and enlarged local regions on the Augsburg dataset. (a) Ground truth. (b) TBCNN. (c) ExViT. (d) HCT. (e) MACN. (f) HLMamba. (g) MSFMamba. (h) S^2CrossMamba. (i) S^2ENet. (j) PAFC-Net.

Table 9 Ablation study of the proposed components on four datasets (OA, %).

MSAE	PCAFM	CAMC	BML	Berlin	Augsburg	Houston 2013	Houston 2018
✓	×	×	×	76.80	90.56	91.77	89.91
×	✓	×	×	77.80	91.37	91.32	89.59
✓	✓	×	×	78.71	91.83	91.95	90.46
✓	✓	✓	×	79.33	91.87	92.63	90.96
✓	✓	✓	✓	**80.45**	**92.48**	**93.70**	**92.26**

Figs. 8–11, the qualitative results are consistent with the quantitative comparisons, indicating that the proposed fusion process improves spatial continuity and boundary recovery across different sensor combinations.

4.3 Ablation studies

4.3.1 Effectiveness of the model components

We conducted component, modality-input, and parameter-sensitivity experiments to evaluate the contributions of PAFC-Net and its main design choices. We further performed an internal ablation of PCAFM to isolate the contributions of its constituent stages and their execution order. We constructed five configurations by progressively adding MSAE, PCAFM, CAMC, and the balanced multi-branch loss (BML). Table 9 reports the OA of each configuration on the four datasets.

Table 9 evaluates the four proposed design elements through controlled combinations across the four datasets. Combining MSAE and PCAFM consistently outperformed the stronger single-component configuration, yielding OA gains of 0.91, 0.46, 0.18, and 0.55 percentage points on Berlin, Augsburg, Houston 2013, and Houston 2018, respectively. The inclusion of BML produced additional gains ranging from 0.61 to 1.30 percentage points. The complete configuration consequently achieved OA values of 80.45%, 92.48%, 93.70%, and 92.26% on the four datasets.

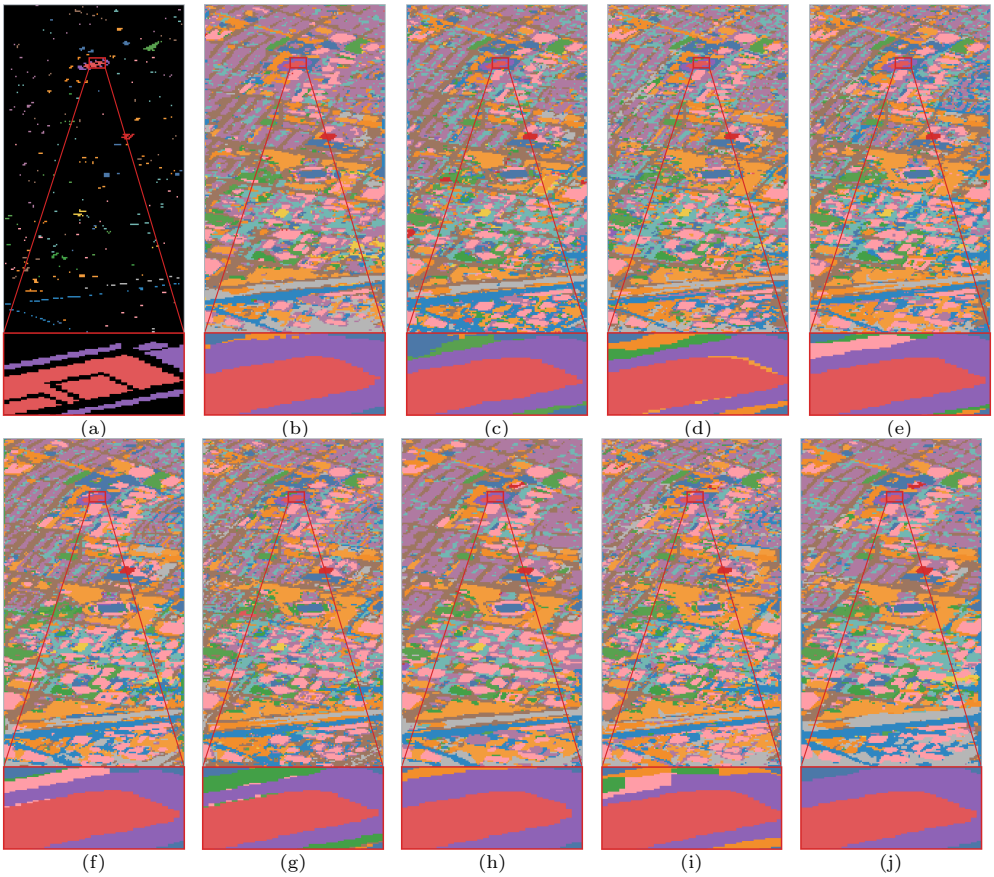


Fig. 10 Classification maps and enlarged local regions on the Houston 2013 dataset. (a) Ground truth. (b) TBCNN. (c) ExViT. (d) HCT. (e) MACN. (f) HLMamba. (g) MSFMamba. (h) S^2CrossMamba. (i) S^2ENet. (j) PAFC-Net.

Having established the contributions of the main components, we next examined how information is progressively processed within PCAFM. Table 10 evaluates the roles of BCMA, SEA, and the local refinement path comprising LMSA and DGU, together with the effect of changing their execution order. The complete PCAFM achieved an OA of 80.45% on Berlin. Removing BCMA, SEA, or the LMSA–DGU path reduced OA to 79.32%, 76.68%, and 79.45%, respectively. Placing local refinement before SEA also decreased OA to 79.14%, supporting the adopted progression from cross-modal correspondence modeling to semantic enhancement and local refinement.

As components were added progressively, Fig. 12 shows tighter within-class clusters and clearer boundaries between classes. The complete PAFC-Net reduced feature overlap relative to the basic dual-branch configuration and produced a more discriminative embedding. This pattern agrees with the quantitative improvements in Table 9 and further supports the complementary roles of the proposed components.

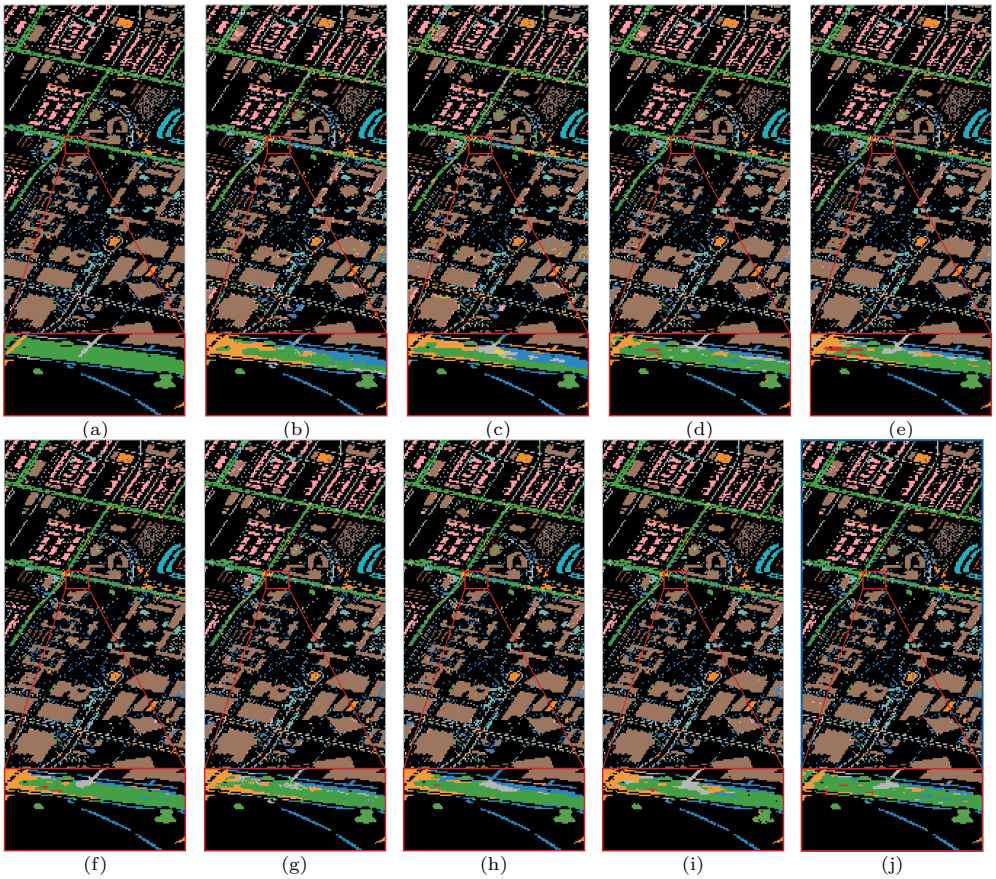


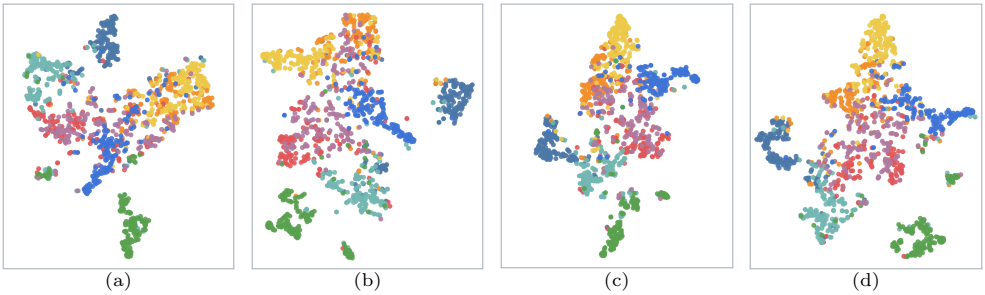
Fig. 11 Classification maps and enlarged local regions on the Houston 2018 dataset. (a) Ground truth. (b) TBCNN. (c) ExViT. (d) HCT. (e) MACN. (f) HLMamba. (g) MSFMamba. (h) S^2CrossMamba. (i) S^2ENet. (j) PAFC-Net.

Table 10 Ablation study of the internal stages and ordering of PCAFM on the Berlin dataset.

BCMA	SEA	LMSA + DGU	Stage order	OA (%)
✓	✓	✓	BCMA → SEA → LMSA	**80.45**
×	✓	✓	SEA → LMSA	79.32
✓	×	✓	BCMA → LMSA	76.68
✓	✓	×	BCMA → SEA	79.45
✓	✓	✓	BCMA → LMSA → SEA	79.14

4.3.2 Effectiveness of multimodal fusion

To determine whether the gains arose from effective complementarity, we evaluated PAFC-Net with HSI-only, AUX-only, and HSI+AUX inputs. Each single-modality



(a) (b) (c) (d)

Fig. 12 t-SNE visualization of feature distributions produced by different PAFC-Net variants on the Berlin dataset. (a) Basic dual-branch configuration. (b) MSAE configuration. (c) MSAE–PCAFM configuration. (d) Full PAFC-Net.

Table 11 Ablation analysis of different modality inputs on four datasets (%).

Input setting	Berlin			Augsburg			Houston 2013			Houston 2018		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
HSI only	75.85	62.84	62.67	91.35	63.63	87.42	89.33	90.97	88.42	87.30	93.23	83.73
AUX only	48.21	36.81	29.90	82.90	42.62	74.56	45.05	44.62	40.66	40.54	29.99	29.36
HSI + AUX	**80.45**	**66.16**	**69.32**	**92.48**	**67.65**	**89.11**	**93.70**	**94.21**	**93.16**	**92.26**	**95.82**	**90.01**

configuration retained only the relevant encoder and classifier path. The joint-input configuration used the complete network.

As shown in Table 11, HSI-only consistently outperformed AUX-only, confirming that spectral information provides the primary discriminative evidence. Adding AUX to HSI increased OA by 4.60, 1.13, 4.37, and 4.96 percentage points on Berlin, Augsburg, Houston 2013, and Houston 2018, respectively. AA and Kappa also improved in every case. These results indicate that PAFC-Net effectively incorporates complementary AUX cues while preserving the dominant discriminative contribution of HSI.

4.3.3 Parameter sensitivity analysis

In this section, we examined the effects of spatial patch size P, semantic token number L, and auxiliary-branch supervision weight λ_b on Berlin by varying one parameter at a time while holding the others fixed. Fig. 13 reports the corresponding changes in OA, AA, and Kappa. Although $P = 13$ produced the highest OA and Kappa, $P = 11$ yielded a higher AA. We therefore selected $P = 11$. Increasing the number of tokens from 2 to 4 improved performance, whereas further increases caused a slight decline. Thus, $L = 4$ provided an effective balance between semantic capacity and representation compactness. Performance improved as λ_b increased to 0.3 and then saturated; therefore, $\lambda_b = 0.3$ was used in all experiments.

4.3.4 Computational complexity analysis

Besides, we compared parameter counts, estimated FLOPs, and per-sample inference latency under the same software and hardware environment.

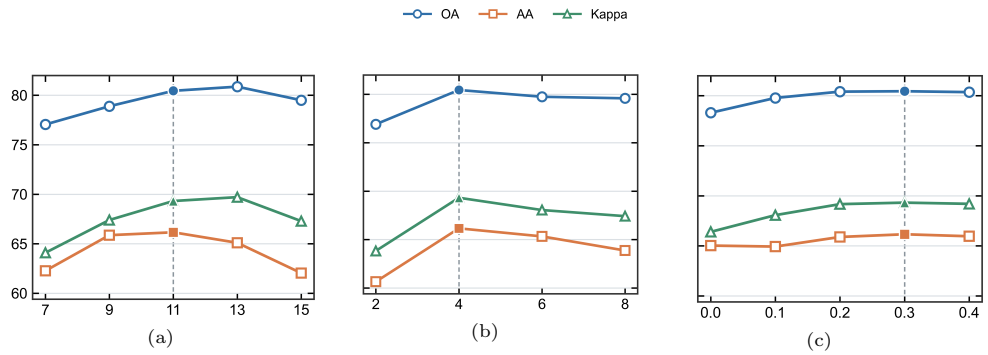


Fig. 13 Parameter sensitivity analysis of PAFC-Net on the Berlin dataset. (a) Patch size P. (b) Number of semantic tokens L. (c) Auxiliary-branch supervision weight λ_b.

Table 12 Computational complexity and inference efficiency of different methods.

Metric	TBCNN	MACN	ExViT	HLMamba	S^2CrossMamba	MSFMamba	FusAtNet	PAFC-Net
Params (M)	0.103	0.383	0.203	0.232	0.263	2.958	35.992	0.344
FLOPs (G)	6.578	20.300	115.345	110.045	36.878	105.148	6697.151	21.527
Latency (ms/sample)	0.0125	0.3719	0.1153	0.0530	0.0512	0.2401	0.3700	0.2016
OA (%, Berlin)	73.43	75.57	70.91	73.62	74.64	75.75	71.45	**80.45**

As shown in Table 12, PAFC-Net contains only 0.344 M parameters. Its estimated computational cost is 21.527 G FLOPs, lower than ExViT, HLMamba, S2CrossMamba, MSFMamba, and FusAtNet. Its per-sample inference latency is 0.2016 ms. These results indicate a favorable balance between classification accuracy and computational complexity.

5 Conclusion and Discussion

Multimodal remote sensing classification remains limited by the difficulty of turning heterogeneous sensor observations into a coherent joint representation. PAFC-Net addresses this difficulty by preserving sensor-specific evidence during encoding, coordinating complementary information through progressive attention, and calibrating the fused branches against a shared semantic reference.

The ablation results support this progression. Removing SEA reduced the OA on Berlin by 3.77 percentage points, while changing the interaction order caused a 1.31 percentage-point decrease. Modality-input experiments and feature visualizations further showed that AUX observations complemented HSI spectra with spatial and structural information. Across the four datasets, PAFC-Net achieved OA values from 80.45% to 93.70% with only 0.344 M parameters.

Future work will extend PAFC-Net with uncertainty-aware modality selection and more efficient cross-modal interaction. These developments will broaden its applicability to diverse sensor combinations and operational scenarios while retaining its compact architecture and reliable classification performance.

Declaration of Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aleissaee AA, Kumar A, Anwer RM, Khan S, Cholakkal H, Xia GS, Khan FS (2023) Transformers in Remote Sensing: A Survey. Remote Sensing 15(7):1860. https://doi.org/10.3390/rs15071860

Audebert N, Le Saux B, Lefèvre S (2018) Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140:20-32. https://doi.org/10.1016/j.isprsjprs.2017.11.011

Bachmann R, Mizrahi D, Atanov A, Zamir A (2022) MultiMAE: Multi-modal Multi-task Masked Autoencoders. In: Computer Vision – ECCV 2022, pp 348-367. https://doi.org/10.1007/978-3-031-19836-6_20

Cai J, Zhang M, Yang H, He Y, Yang Y, Shi C, Zhao X, Xun Y (2024) A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and LiDAR data. Expert Systems with Applications 249:123587. https://doi.org/10.1016/j.eswa.2024.123587

Chen B, Huang B, Xu B (2017) Multi-source remotely sensed data fusion for land cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 124:27-39. https://doi.org/10.1016/j.isprsjprs.2016.12.008

Debes C, Merentitis A, Heremans R, Hahn J, Frangiadakis N, van Kasteren T, Liao W, Bellens R, Pizurica A, Gautama S, Philips W, Prasad S, Du Q, Pacifici F (2014) Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(6):2405-2418. https://doi.org/10.1109/JSTARS.2014.2305441

Dong W, Zhang T, Qu J, Xiao S, Zhang T, Li Y (2022) Multibranch Feature Fusion Network With Self- and Cross-Guided Attention for Hyperspectral and LiDAR Classification. IEEE Transactions on Geoscience and Remote Sensing 60:1-12. https://doi.org/10.1109/TGRS.2022.3179737

Fan Y, Qian Y, Gong W, Chu Z, Qin Y, Muhetaer P (2024) Multi-level interactive fusion network based on adversarial learning for fusion classification of hyperspectral and LiDAR data. Expert Systems with Applications 257:125132. https://doi.org/10.1016/j.eswa.2024.125132

Fang S, Li K, Li Z (2022) S^2ENet: Spatial–Spectral Cross-Modal Enhancement Network for Classification of Hyperspectral and LiDAR Data. IEEE Geoscience and Remote Sensing Letters 19:1-5. https://doi.org/10.1109/LGRS.2021.3121028

Feng J, Zhang J, Zhang Y (2023) Multiview Feature Learning and Multilevel Information Fusion for Joint Classification of Hyperspectral and LiDAR Data. IEEE Transactions on Geoscience and Remote Sensing 61:1-13. https://doi.org/10.1109/TGRS.2023.3325818

Gao F, Jin X, Zhou X, Dong J, Du Q (2025) MSFMamba: Multiscale Feature Fusion State Space Model for Multisource Remote Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing 63:1-16. https://doi.org/10.1109/TGRS.2025.3535622

Gao Q, Qu J, Dong W (2026) T-APT: Text-Guided Modality-Aware Prompt Tuning for Arbitrary Multimodal Remote Sensing Data Joint Classification. Proceedings of the AAAI Conference on Artificial Intelligence 40(6):4158-4166. https://doi.org/10.1609/aaai.v40i6.42411

Ghamisi P, Gloaguen R, Atkinson PM, Benediktsson JA, Rasti B, Yokoya N, Wang Q, Hofle B, Bruzzone L, Bovolo F, Chi M, Anders K (2019) Multisource and Multitemporal Data Fusion in Remote Sensing: A Comprehensive Review of the State of the Art. IEEE Geoscience and Remote Sensing Magazine 7(1):6-39. https://doi.org/10.1109/MGRS.2018.2890023

Guo F, Meng Q, Li Z, Ren G, Wang L, Zhang J, Xin R, Hu Y (2024) Multisource Feature Embedding and Interaction Fusion Network for Coastal Wetland Classification With Hyperspectral and LiDAR Data. IEEE Transactions on Geoscience and Remote Sensing 62:1-16. https://doi.org/10.1109/TGRS.2024.3367960

Han W, Miao W, Geng J, Jiang W (2024) CMSE: Cross-Modal Semantic Enhancement Network for Classification of Hyperspectral and LiDAR Data. IEEE Transactions on Geoscience and Remote Sensing 62:1-14. https://doi.org/10.1109/TGRS.2024.3368509

He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 16000-16009. https://doi.org/10.1109/CVPR52688.2022.01553

Hong D, Hu J, Yao J, Chanussot J, Zhu XX (2021) Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. ISPRS Journal of Photogrammetry and Remote Sensing 178:68-80. https://doi.org/10.1016/j.isprsjprs.2021.05.011

Hu WS, Li W, Li HC, Huang FH, Tao R (2025) Global Clue-Guided Cross-Memory Quaternion Transformer Network for Multisource Remote Sensing Data Classification. IEEE Transactions on Neural Networks and Learning Systems 36(4):7357-7371. https://doi.org/10.1109/TNNLS.2024.3406735

Li J, Hong D, Gao L, Yao J, Zheng K, Zhang B, Chanussot J (2022) Deep learning in multimodal remote sensing data fusion: A comprehensive review. International Journal of Applied Earth Observation and Geoinformation 112:102926. https://doi.org/10.1016/j.jag.2022.102926

Li K, Wang D, Wang X, Liu G, Wu Z, Wang Q (2023) Mixing Self-Attention and Convolution: A Unified Framework for Multisource Remote Sensing Data Classification. IEEE Transactions on Geoscience and Remote Sensing 61:1-16. https://doi.org/10.1109/TGRS.2023.3310521

Li W, Gao Y, Zhang M, Tao R, Du Q (2023) Asymmetric Feature Fusion Network for Hyperspectral and SAR Image Classification. IEEE Transactions on Neural Networks and Learning Systems 34(10):8057-8070. https://doi.org/10.1109/TNNLS.2022.3149394

Li X, Zhang G, Cui H, Hou S, Chen Y, Li Z, Li H, Wang H (2023) Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images. ISPRS Journal of Photogrammetry and Remote Sensing 195:178-191. https://doi.org/10.1016/j.isprsjprs.2022.11.015

Li Z, Sui H, Luo C, Guo F (2023) Morphological Convolution and Attention Calibration Network for Hyperspectral and LiDAR Data Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16:5728-5740. https://doi.org/10.1109/JSTARS.2023.3284655

Li D, Li B, Liu Y (2025) Mamba Cross-Modal Information Fusion Self-Distillation Model for Joint Classification of LiDAR and Hyperspectral Data. IEEE Transactions on Geoscience and Remote Sensing 63:1-13. https://doi.org/10.1109/TGRS.2025.3600692

Liao D, Wang Q, Lai T, Huang H (2024) Joint Classification of Hyperspectral and LiDAR Data Based on Mamba. IEEE Transactions on Geoscience and Remote Sensing 62:1-15. https://doi.org/10.1109/TGRS.2024.3459709

Lin J, Gao F, Qi L, Dong J, Du Q, Gao X (2025) Dynamic Cross-Modal Feature Interaction Network for Hyperspectral and LiDAR Data Classification. IEEE Transactions on Geoscience and Remote Sensing 63:1-16. https://doi.org/10.1109/TGRS.2025.3550587

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

Liu R, Ling J, Zhang H (2024) SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 218:277-293. https://doi.org/10.1016/j.isprsjprs.2024.09.012

Liu S, Yu T, Zhou J, Xing G, Chen H (2025) Cross-source transformer-based neighborhood contrastive learning for joint classification of hyperspectral and LiDAR Data. Information Fusion 124:103225. https://doi.org/10.1016/j.inffus.2025.103225

Mohla S, Pande S, Banerjee B, Chaudhuri S (2020) FusAtNet: Dual Attention based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://doi.org/10.1109/CVPRW50498.2020.00054

Okujeni A, van der Linden S, Hostert P (2016) Berlin-Urban-Gradient dataset 2009 - An EnMAP Preparatory Flight Campaign (Datasets). GFZ Data Services. https://doi.org/10.5880/ENMAP.2016.002

Pan H, Zhao R, Ge H, Liu M, Zhang Q (2025) Multimodal Fusion Mamba Network for Joint Land Cover Classification Using Hyperspectral and LiDAR Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 18:17328-17345. https://doi.org/10.1109/JSTARS.2025.3585640

Pan H, Ren B, Wang L, Ge H, Shi C, Liu M, Yan H, Li X (2025) Multiscale Split-Recombination Cooperative Fusion Network for Hyperspectral and LiDAR Land Cover Classification. IEEE Transactions on Geoscience and Remote Sensing 63:1-24. https://doi.org/10.1109/TGRS.2025.3589253

Qu J, Yang Y, Dong W, Yang Y (2024) LDS2AE: Local Diffusion Shared-Specific Autoencoder for Multimodal Remote Sensing Image Classification with Arbitrary Missing Modalities. Proceedings of the AAAI Conference on Artificial Intelligence 38(13):14731-14739. https://doi.org/10.1609/aaai.v38i13.29391

Sainte Fare Garnot V, Landrieu L, Chehata N (2022) Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS Journal of Photogrammetry and Remote Sensing 187:294-305. https://doi.org/10.1016/j.isprsjprs.2022.03.012

Shi L, Li C, Zeng Z, Duan P, Rasti B, Plaza A (2026) Masked Self-Attention Fusion Network for Joint Classification of Hyperspectral and LiDAR Data. IEEE Transactions on Image Processing 35:346-360. https://doi.org/10.1109/TIP.2025.3648926

Srinivas A, Lin TY, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck Transformers for Visual Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 16519-16529. https://doi.org/10.1109/CVPR46437.2021.01625

Sun L, Zhao G, Zheng Y, Wu Z (2022) Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing 60:1-14. https://doi.org/10.1109/TGRS.2022.3144158

Sun X, Tian Y, Lu W, Wang P, Niu R, Yu H, Fu K (2023) From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Science China Information Sciences 66(4):140301. https://doi.org/10.1007/s11432-022-3588-0

Wang X, Girshick R, Gupta A, He K (2018) Non-Local Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

Wang Y, Chen X, Cao L, Huang W, Sun F, Wang Y (2022) Multimodal Token Fusion for Vision Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12186-12195. https://doi.org/10.1109/CVPR52688.2022.01187

Wang J, Song W, Chen H, Ren J, Zhao H (2025) FusDreamer: Label-Efficient Remote Sensing World Model for Multimodal Data Classification. IEEE Transactions on Geoscience and Remote Sensing 63:1-14. https://doi.org/10.1109/TGRS.2025.3554862

Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: Convolutional Block Attention Module. In: Computer Vision – ECCV 2018, pp 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

Xu X, Li W, Ran Q, Du Q, Gao L, Zhang B (2018) Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. IEEE Transactions on Geoscience and Remote Sensing 56(2):937-949. https://doi.org/10.1109/TGRS.2017.2756851

Xu Y, Du B, Zhang L, Cerra D, Pato M, Carmona E, Prasad S, Yokoya N, Hansch R, Le Saux B (2019) Advanced Multi-Sensor Optical Remote Sensing for Urban Land Use and Land Cover Classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12(6):1709-1724. https://doi.org/10.1109/JSTARS.2019.2911113

Yang B, Wang X, Xing Y, Cheng C, Jiang W, Feng Q (2024) Modality Fusion Vision Transformer for Hyperspectral and LiDAR Data Collaborative Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17:17052-17065. https://doi.org/10.1109/JSTARS.2024.3415729

Yao J, Zhang B, Li C, Hong D, Chanussot J (2023) Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework. IEEE Transactions on Geoscience and Remote Sensing 61:1-15. https://doi.org/10.1109/TGRS.2023.3284671

Ying X, Chuah MC (2022) UCTNet: Uncertainty-Aware Cross-Modal Transformer Network for Indoor RGB-D Semantic Segmentation. In: Computer Vision – ECCV 2022, pp 20-37. https://doi.org/10.1007/978-3-031-20056-4_2

Yu W, Gao L, Huang H, Shen Y, Shen G (2023) HI-D-FNet: Hyperspectral Intrinsic Image Decomposition Guided Data Fusion Network for Hyperspectral and LiDAR Classification. IEEE Transactions on Geoscience and Remote Sensing 61:1-15. https://doi.org/10.1109/TGRS.2023.3308109

Zhang G, Zhang L, Zhang Z, Deng J, Bian L, Yang C (2024) DECT: Diffusion-Enhanced CNN–Transformer for Multisource Remote Sensing Data Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17:19288-19301. https://doi.org/10.1109/JSTARS.2024.3479212

Zhang G, Zhang Z, Deng J, Bian L, Yang C (2024) S-CrossMamba: Spatial–Spectral Cross-Mamba for Multimodal Remote Sensing Image Classification. IEEE Geoscience and Remote Sensing Letters 21:1-5. https://doi.org/10.1109/LGRS.2024.3488036

Zhang J, Cao M, Yang X, Jiang K, Li Y (2025) DiffCLIP: Few-shot Language-driven Multimodal Classifier. Proceedings of the AAAI Conference on Artificial Intelligence 39(21):22443-22451. https://doi.org/10.1609/aaai.v39i21.34401

Zhang Y, Gao H, Chen Z, Zhang C, Ghamisi P, Zhang B (2026) E-Mamba: Efficient Mamba network for hyperspectral and LiDAR joint classification. Information Fusion 126:103649. https://doi.org/10.1016/j.inffus.2025.103649

Zhao G, Ye Q, Sun L, Wu Z, Pan C, Jeon B (2023) Joint Classification of Hyperspectral and LiDAR Data Using a Hierarchical CNN and Transformer. IEEE Transactions on Geoscience and Remote Sensing 61:1-16. https://doi.org/10.1109/TGRS.2022.3232498

Zhu F, Shi C, Shi K, Wang L (2025) Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Multimodal Feature Aggregation-Based Multihead Axial Attention Transformer. IEEE Transactions on Geoscience and Remote Sensing 63:1-17. https://doi.org/10.1109/TGRS.2025.3533475