Electronic Research Archive

ERA, 32(7): 4218–4236.
DOI: 10.3934/era.2024190
Received: 05 March 2024
Revised: 06 June 2024
Accepted: 25 June 2024
Published: 03 July 2024

http://www.aimspress.com/journal/ERA

Research article

Enhanced spectral attention and adaptive spatial learning guided network for hyperspectral and LiDAR classification

Bingsheng Li[1,*], Na Li[1], Jianmin Ren[1], Xupeng Guo[2], Chao Liu[2], Hao Wang[2] and Qingwu Li[3]

[1] State Grid Qinghai Electric Power Co., LTD., Xining 810000, China
[2] State Grid Haidong Power Supply Company, Haidong 630200, China
[3] College of Information Science and Engineering, Hohai University, Changzhou 213200, China

* **Correspondence:** Email: libingsheng_123@163.com.

Abstract: Although the data fusion of hyperspectral images (HSI) and light detection and ranging (LiDAR) has provided significant gains for land-cover classification, it also brings technical obstacles (i.e., it is difficult to capture discriminative local and global spatial-spectral from redundant data and build interactions between heterogeneous data). In this paper, a classification network named enhanced spectral attention and adaptive spatial learning guided network (ESASNet) is proposed for the joint use of HSI and LiDAR. Specifically, first, by combining a convolutional neural network (CNN) with the transformer, adaptive spatial learning (ASL) and enhanced spectral learning (ESL) are proposed to learn the spectral-spatial features from the HSI data and the elevation features from the LiDAR data in the local and global receptive field. Second, considering the characteristics of HSI with a continuous, narrowband spectrum, ESL is designed by adding enhanced local self-attention to enhance the mining of the spectral correlations across the adjacent spectrum. Finally, a feature fusion module is proposed to ensure an efficient information exchange between HSI and LiDAR during spectral features and spatial feature fusion. Experimental evaluations on the HSI-LiDAR dataset clearly illustrate that ESASNet performs better in feature extraction than the state-of-the-art methods. The code is available at https://github.com/AirsterMode/ESASNet.

Keywords: hyperspectral image; spectral attention; spatial learning guided; feature fusion

1. Introduction

Remote sensing technologies enable us to gather information about the earth's surface, which may be difficult or impossible to obtain through other means, including rich information about land cover, vegetation, bodies of water, and so on. However, a single remote sensing data source is often weak to process the edges of complex regions. Hyperspectral images (HSIs) have rich spectral information, from visible light to near-infrared. However, it is difficult to recognize objects with the same spectrum within HSIs. Conversely, the digital surface model (DSM) based on a laser radar can provide accurate height and shape information, though it cannot separate different materials with the same height. HSIs and DSM based on light detection and ranging (LiDAR) can complement each other's information. In many areas, the combination of HSI and LiDAR data has been used successfully, such as forest monitoring [1], aboveground biomass estimation [2], fuel type mapping [3], and land-cover classification. The combination of HSI and LiDAR data has provided a positive influence on the classification accuracy, as well as increased the difficulty of feature extraction and provided the challenge of feature fusion between heterogeneous data.

Most works focused on improving the effectiveness of feature extraction. Machine learning technology was first applied in the HSI and LiDAR data fusion-based classification. Different mapping methods, such as multi-core learning, was applied to achieve different classification tasks [4]. However, the basic feature stacking approach is inefficient, as it results in the Hughes phenomenon. As part of the solution to the problem of a low efficiency for the feature stacking method, there have been many multi-modal data fusion-based classification methods, which are often grouped into three categories: pixel-level fusion, feature-level fusion, and decision-level fusion [5–9]. However, the above methods are highly dependent on the quality of the extracted features, which limits their applicability in complex scenes.

CNN are widely used in deep learning due to their excellent ability to extract local features. Early research focused on dual-stream networks, which were designed into two network branches to learn data of two different modalities, such as a two-branch CNN [10]. The HSI and LiDAR networks were trained separately and the features extracted by the two networks were connected; then, they were fused through the fully connected layer to obtain the classification result. In [11], a novel framework for the fusion of hyperspectral images and LiDAR-derived elevation data was proposed on the basis of CNN and composite kernels. However, as the number of CNN network layers increased, the classification accuracy began to decline. To alleviate the problem of a decrease in accuracy, Ge et al. [12] developed a fusion network on the basis of a deep residual network. The network extracts spectral and spatial information from HSI and LiDAR data, respectively, by applying extinction contour, depth residual block, and the local binary method. To further improve the classification accuracy, Zhao et al. took the detailed spectral signatures of the HSI and the elevation information of the LiDAR into account and proposed a deep CNN architecture which used a hierarchical random walk layer [13]. The local feature extraction capability based on CNN significantly improved the classification performance of the HSI and LiDAR. However, the CNN lacked perception of long-range information, which limited the classification performance.

Transformers have gradually attracted attention due to their ability to model long sequence correlations. Dosovitskiy et al. discovered the long-range correlation between image patches, and therefore applied a Transformer from natural language processing to image processing [14]. Some works have also applied it to remote sensing classification tasks. Yu et al. [15] implemented a capsule vision transformer (ViT), which effectively integrated cross contextual semantic features by utilizing

the interaction of long-range global features at different contextual scales. Roy et al. [16] proposed a multimodal fusion transformer for the joint classification of HSI and LiDAR. This method used LiDAR data as a learnable token to perform feature learning together with HSI tokens. This operation could not fully integrate the valid information from these two types of data. Feng et al. [17] proposed the Spectral Spatial Elevation Fusion Transformer, which utilized a spatial information recognition module and a sliding group spectrum embedding module to integrate the features of hyperspectral and LiDAR to a certain extent. Although the Transformer exceled at aggregating information embedded in spectral features, it was difficult to define local semantic aspects and did not fully utilize local spatial information.

To fully utilize the information in HSI and LiDAR data and improve the classification accuracy, some works attempted to combine CNN and the Transformer for the hyperspectral LiDAR classification. Zhao et al. [18] proposed a dual branch method that consisted of a layered CNN and a Transformer network to fuse multi-source heterogeneous information and to improve the joint classification performance. However, the above work ignores the domain gap between LiDAR and HSI. There is still no tailored combined strategy of a CNN and a Transformer for LiDAR and HSI data separately.

For the challenge of feature fusion between heterogeneous data, some works aim to enhance the information exchange between HSI and Lidar branches. Wang et al. proposed the modal attention (MA) module to integrate the features of hyperspectral and LiDAR and established a feature interaction between different modal data [19]. In [20], Wang et al. proposed an adaptive mutual learning fusion network, which fused features more effectively by adaptively balancing the weights of HSI and LiDAR features. Mohla et al. [21] proposed to use the cross-attention (CA) module to emphasize the spatial information of HSI by simultaneously harnessing the LiDAR-derived attention map and adopted the self-attention mechanism for a deep feature fusion. However, the above network did not fully exchange information between the spectral and spatial information of the HSI data and the elevation information of LiDAR.

To address the above challenges, in this work, we put forward a network, called the Enhanced Spectral Attention and Adaptive Spatial Learning Guided Network (ESASNet), for a more accurate data joint classification of HSI and LiDAR. We tailor the feature extraction modules for HSI and LiDAR based on a combined CNN with the Transformer. Specifically, ESL is designed to learn subtle differences between the spectral dimensions and better utilize the HSI spectral features in a local and global receptive field. The local feature extraction capabilities of a CNN are utilized to ensure the reliability of the shallow spatial features and the spectral features. HSI has a continuous, narrowband spectrum, which means the adjacent spectrum with a similar wavelength has more correlations. ESL uses the enhanced local self-attention (ELSA) module [22] to increase the channel capacity and achieve an efficient non-local feature extraction by combining attention maps and static matrices through Ghosts. LiDAR data is considered to provide accurate elevation information to complement the spectral details of HSI. ASL is proposed for LiDAR data to guarantee the global information of LiDAR while promoting the preliminary spatial feature fusion of HSI and LiDAR. For the depth feature fusion, a feature fusion (FF) module is designed. This gives more weight to the global information and further improvest the information interaction of the depth features. Cross-attention is used to exchange the global information on the spatial features and the spectral features. In short, our contributions are summarized as follows:

1) We designed a CNN-Transformer-based feature extraction structure, which provides a great assistance in extracting spectral information from the hyperspectral data and spatial information from LiDAR.

2) We designed an ESL module which fully utilizes a spectral correlation to efficiently extract more spectral information.

3) We designed a more effective FF module based on the cross attention. This module can fully promote the interaction between the spectral and spatial information, thus enabling classification networks to better balance the information of the spectral and spatial branches.

4) We conducted sufficient experiments to validate the adaptability of our network in different scenarios and the effectiveness of various network modules.

2. Materials and methods

Two kinds of remote sensing sources, namely HSI and LiDAR, are combined to classify pixel-based images.

Specifically, a hyperspectral image $\mathbf{X}_{hsi} \in \mathbb{R}^{M\times N\times D}$ and the associated LiDAR image $\mathbf{X}_{lidar} \in \mathbb{R}^{M\times N\times 1}$ covering the same area on the surface of the earth are given as follows, where M and N represent the width and height of these two images, respectively, and D stands for the number of original HSI channels.

We need to perform a Principal components analysis operation on $\mathbf{X}_{hsi}$ to reduce the computational redundancy. The HSI channel will be compressed into d as $\mathbf{X}_{hsi} \in \mathbb{R}^{M\times N\times d}$.

For the training set, its patched image can be expressed as $\mathbf{X}=\{\mathbf{X}^i|\mathbf{X}^i_{hsi},\mathbf{X}^i_{lid}\}^T_{i=1}$, where T is the total number of samples used to train the model.

$\mathbf{X}^i_{hsi} \in \mathbb{R}^{m\times n\times d}$ and $\mathbf{X}^i_{lid} \in \mathbb{R}^{m\times n\times 1}$ are the HSI and LiDAR training patches, respectively. $\mathbf{Y}=\{y^i\}^T_{i=1}$ is the ground truth label, where $\mathbf{y}^i \in \{1,2,...,C\}$, and C represent the classes of the ground truth.

The ESASNet contains two main parts: a spectral-spatial feature learning (SSFL) module and a FF module. The detailed architecture of the proposed ESASNet network is shown within Figures 1 and 2. The pseudocode is provided within Algorithm 1.

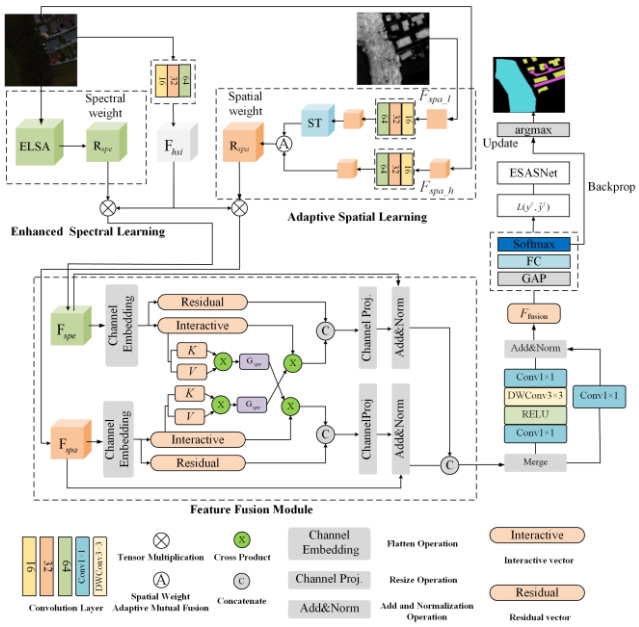

Figure 1. Network structure diagram.

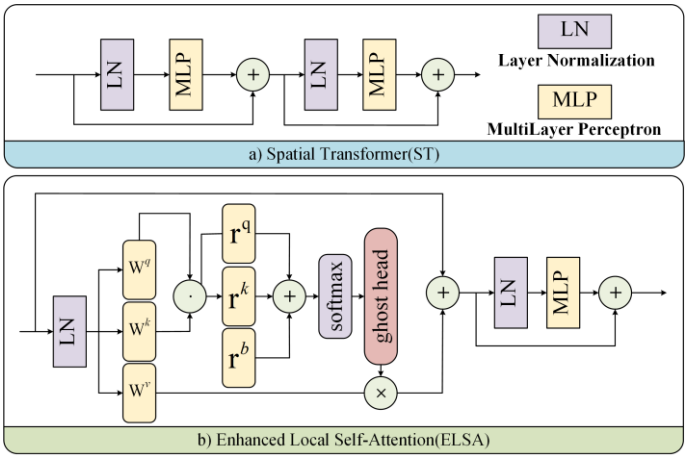

Figure 2. a) The content of spatial transformer. b) The content of enhanced local self-attention.

Algorithm 1 Pseudocode of the proposed ESASNet in Pytorch-Like style

Require:
1: X_HSI, X_LiDAR: Dimension reduced hyperspectral image and corresponding LiDAR data;
2: Network Hyper-parameters: Adam optimizer, Learning rate = 0.01, Batch size = 256, Epochs = 100;
Ensure: Ensembles of classifier;
3: # X_hsi, X_lid: Preprocess original inputa data to m × n patch data; Y: GroundTruth label
4: for (X_hsi, X_lid, Y) in data_loader: # load a mini-batch data
5: # Spectral-spatial feature learning (SSFL) Module
6: # F_hsi (HSI-based features): m × n × 64; F_spectral, F_spatial: m × n × 64
7: F_hsi = Feature_HSI (X_hsi)
8: R_spectral, R_spatial = Spectral_Weight (X_hsi), Spatial_Weight (X_hsi, X_lid, α1, α2)
9: F_spectral, F_spatial = R_spectral × F_hsi, R_spatial × F_hsi
10: # Feature fusion (FF) module
11: res1, res2, inter1, inter2 = information_exchange (F_spectral, F_spatial)
12: v1,v2 = cross_attn (inter1, inter2)
13: out1 = norm (F_spectral + cat (res1, v1))
14: out2 = norm (F_spatial + cat (res2, v2))
15: feature_fusion = channel_emb (out1, out2)
16: # Multimodal data classification (MDC) module
17: loss = Margin_Loss (nn.softmax (nn.Linear (nn.AdaptiveAvgPool (feature_fusion))), Y)
18: # Update trainable parameters
19: Return classification map;

2.1. Spectral spatial feature learning module

1) Hyperspectral feature extractor: The hyperspectral feature extractor consists of 16, 32, and 64 filters to extract the HSI's primary features. The primary feature $\mathbf{F}_{hsi}$ is a patch of size $m \times n \times 64$. Then, the extractor multiplies $\mathbf{F}_{hsi}$ by the spectral feature weight $\Re_{spe}$, and the spatial feature weight $\Re_{spa}$ (obtained from the following two methods) ensures the stability of the spectral and spatial features.

2) ESL module: ELSA is a self-attention mechanism, which overcomes the weakness of the self-attention mechanism in local finer-level feature learning and has an acceptable cost. In terms of the spectral information extraction, using ELSA to highlight the spectral channel of every pixel will further enhance the extraction of the spectral features. There are two parts within the module; one is a Hadamard attention, which makes up for the lack of LSA in local fine feature learning while reducing the memory usage; and the other is the Ghost head, which improves the channel capacity, instead of the attention. Specifically, we perform a convolution operation with 64 filters on the input HSI patch $\mathbf{X}_{hsi-p} \in \mathbb{R}^{d\times p\times p}$. It is made into the size of $(64, p, p)$; then, through the enhanced self-local attention block, the spectral-level feature weight $\Re_{spe}$ is as follows:

$$\Re_{spe} = Norm(MLP(\mathbf{X}_{hsi-p} + ELSA(Norm(\mathbf{X}_{hsi-p})))), \tag{1}$$

where $Norm(\cdot)$ refers to the general regularization method, $MLP(\cdot)$ is the multi-layer perceptron, and $ELSA(\cdot)$ is the ELSA method. These methods do not change the size of the input data. The final spectral feature to be obtained is $\mathbf{F}_{spe} \in \mathbb{R}^{64\times p\times p}$.

We can multiply the spectral level feature $\Re_{spe}$ by the primary feature $\mathbf{F}_{hsi}$ as follows:

$$\mathbf{F}_{spe} \in \Re_{spe} \otimes \mathbf{F}_{hsi}. \tag{2}$$

3) ASL module: We use HSI and LiDAR data to extract the spatial information. A two-layer convolution (with 32 and 64 filters, respectively) is used to process the HSI patch $\mathbf{X}_{hsi-p} \in \mathbb{R}^{d\times p\times p}$ and the LiDAR patch data $\mathbf{X}_{lidar-p} \in \mathbb{R}^{1\times p\times p}$, and ensures that they end up with the same number of channels ($\mathbf{X}_{hsi-conv} \in \mathbb{R}^{64\times d\times d}$, $\mathbf{X}_{lidar-conv} \in \mathbb{R}^{64\times d\times d}$), where $p\times p$ is the size of the patch.

Taking the benefits of lidar imagery for spatial information representation into account, we use a Transformer to enhance the performance of LiDAR in the spatial information. However, we still set up an adaptive spatial feature fusion layer to ensure the spatial performance of HSI, as shown below:

$$\Re_{spa} = \alpha_1 \mathbf{X}_{hsi-conv} + \alpha_2 \text{Transformer}(\mathbf{X}_{lidar-conv}), \tag{3}$$

where α_1 and α_2 represent various HSI and spatial weights of the enhanced LiDAR images, respectively. This makes it possible to adjust the weight ratio more flexibly to increase precision, and to adapt the loss function to the whole learning process. $\text{Transformer}(\cdot)$ is a two-layer MLP composition vision Transformer method. Consequently, to obtain the final spatial features $\mathbf{F}_{spa} \in \mathbb{R}^{64\times p\times p}$, we multiply the spatial-level feature weight $\mathbf{F}_{spa}$ by the main feature $\mathbf{F}_{hsi}$ as follows:

$$\mathbf{F}_{spa} \in \Re_{spa} \otimes \mathbf{F}_{hsi}. \tag{4}$$

2.2. Feature fusion module

The FF module is composed of an information exchange stage and a fusion stage. This module finally fuses the input the spectral-spatial features ($\mathbf{F}_{\text{spe}} \in \mathbb{R}^{64\times p\times p}$ and $\mathbf{F}_{spa} \in \mathbb{R}^{64\times p\times p}$) into a final fusion weight.

2.2.1. Information exchange stage

At this stage, the previously obtained spatial and spectral features exchange their information by means of a symmetrical cross-attention structure. Since the input spectral features and the spatial features have the same size, the spectral path is used as a representative for the introduction.

We flatten the input features of size $\mathbb{R}^{64\times p\times p}$ to $\mathbb{R}^{64\times N}$, where $N = p\times p$. Then, the linear embeddings are used to generate two $\mathbb{R}^{C_i\times N}$ vectors of the same size, which represent the residual vector $\mathbf{X}_{res}$ and the interaction vector $\mathbf{X}_{inter}$.

We use a cross-attention mechanism, which is applied on the interaction vectors of the spatial features and the spectral features, to achieve a sufficient information exchange between the two features. Specifically, the interaction vector will be embedded into the key K and value V of each head, both of which are of size $\mathbb{R}^{C_{head}\times N}$. The output is the multiplication of this interaction vector with the context vector of another modality path as a cross-attention result:

$$\mathbf{G}_{spe} = \mathbf{K}_{spe}^{T}\mathbf{V}_{spe}, \tag{5}$$

$$\mathbf{G}_{spa} = \mathbf{K}_{spa}^{T}\mathbf{V}_{spa}, \tag{6}$$

$$\mathbf{U}_{spe} = \mathbf{X}_{spe}^{inter}\mathrm{Softmax}(\mathbf{G}_{spa}), \tag{7}$$

$$\mathbf{U}_{spa} = \mathbf{X}_{spa}^{inter}\mathrm{Softmax}(\mathbf{G}_{spe}). \tag{8}$$

Among them, $\mathbf{G}$ stands for the global context vector and $\mathbf{U}$ stands for the results generated by the two branches. Then, the obtained result vector $\mathbf{U}$ and the residual vector $\mathbf{X}_{res}$ are concatenated. Finally, we take the second linear embedding and change the size of the features to $\mathbb{R}^{64\times p\times p}$.

2.2.2. Fusion stage

In the second stage of the FFM, where a precise fusion is required, we use a simple channel embedding to merge the features of the two paths, which is achieved by a 1 × 1 convolutional layer. In addition, a deep convolution layer DWConv 3 × 3 is implemented to realize the skip connection structure. In this way, the pooled features of size $\mathbb{R}^{128\times p\times p}$ are fused into the final output of size $\mathbb{R}^{64\times p\times p}$, namely the final fused features O_i.

2.3. *Multi-modal data classification module*

Our task goal is the same as AM^{3}Net [20]; therefore, we refer to the MDC module in it. For the *i*-th training sample, we take the final fused features $\mathbf{O}_i$ to the multimodal data classification module (MDC module) to classify the input pixels. After global average pooling (GAP) and fully connected (FC) layers, we convert $\mathbf{O}_i$ to a vector of size (1, C), where C is the number of pixel categories. Finally, we use the Softmax function to convert the obtained result into a probability distribution

$\hat{\mathbf{y}}^{i} = Softmax(FC(GAP(\mathbf{O}_i))) \in \mathbb{R}^{1\times C}$. In addition, we use MarginLoss to evaluate the difference between the prediction result of the model $\hat{\mathbf{y}}^{i}$ and the GroundTruth $\mathbf{y}^{i}$, then continuously modify the weight of the network.

After the training of ESASNet is completed, the test pixels $\mathbf{X}^{test}$ and their corresponding prediction results $\hat{\mathbf{y}}^{test} \in \mathbb{R}^{1\times C}$ can be determined according to the maximum probability as follows:

$$Class(\mathbf{X}^{test}) = \arg\max(\hat{\mathbf{y}}^{test}), \tag{9}$$

where $Class(\mathbf{X}^{test}) \in [1,...,C]$. Therefore, we perform the same processing on all pixels; then, we can obtain the whole classification map.

3. Results and discussion

3.1. Datasets

Experiments on the Houston2013, Trento, and Augsburg datasets are carried out in this section. The Houston2013 dataset consists of an HSI and a LiDAR image, with the HSI data was acquired by an AISA Eagle sensor. This dataset has a total of 15 different classes [23]. Figure 3(a) is the pseudo-color image of Houston. Figure 3(c),(e) are the 2D LiDAR-derived DSM image and the groundtruth, respectively. The Trento dataset is also an HSI-LiDAR pair dataset, which contains 6 classes [24]. Figure 3(b) is the pseudo-color image of Trento, (d) is the 2D LiDAR-derived DSM image of Trento, and (f) is the groundtruth. Figure 3(g),(h) are the class names of the Houston and Trento datasets, respectively. The Augsburg dataset contains 7 classes [25]. Figure 4 is the detail visualization of Augsburg. Specially, Figure 4(a) is a pseudo-color image for the HSI, (b) is the 2D LiDAR-derived DSM image, (c) is the groundtruth, and (d) are class names.

As shown by the exact number of the training and test samples in Table 1, the training and test data sets for the Houston, Trent, and Augsburg data sets do not overlap.

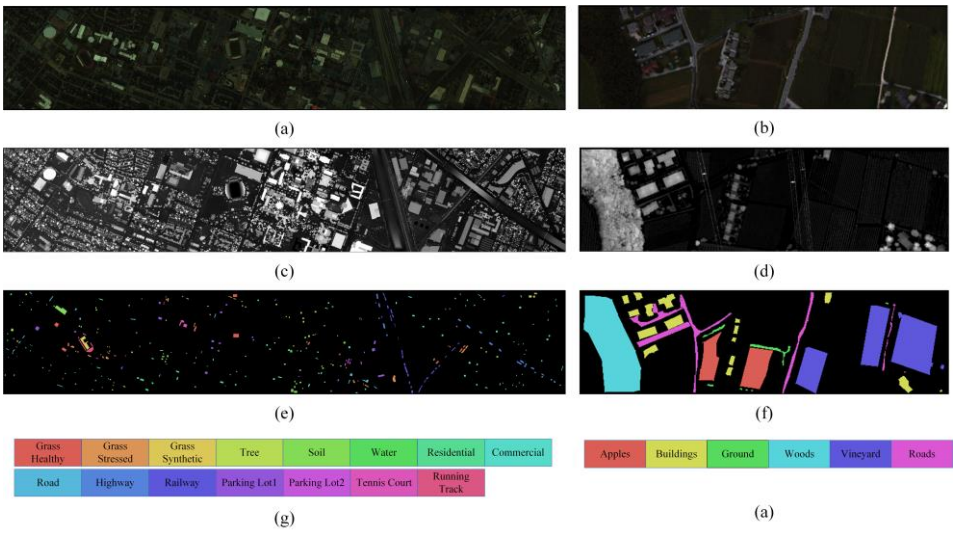

Figure 3. Visualization of the Houston and Trento datasets.

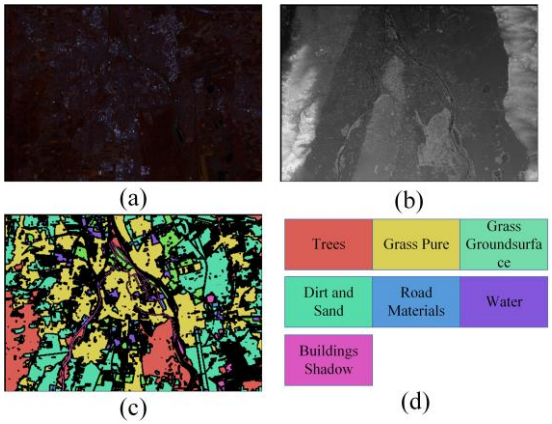

Figure 4. Visualization of the Augsburg dataset.

Table 1. Train and test samples in each class of Houston2013, Trento, and Augsburg datasets.

Houston2013							
No.	Training	Test	Totle	No.	Training	Test	Totle
1	198	1053	1251	9	193	1059	1252
2	190	1064	1254	10	191	1036	1227
3	192	505	697	11	181	1054	1235
4	188	1056	1244	12	192	1041	1233
5	186	1056	1242	13	184	285	469
6	182	143	325	14	181	247	428
7	196	1072	1268	15	187	473	660
8	191	1053	1244	Totle	2832	12197	15029
Trento							
No.	Training	Test	Totle	No.	Training	Test	Totle
1	129	3905	4034	5	184	10317	10501
2	125	2778	2903	6	122	3052	3174
3	105	374	479	Totle	819	29395	30214
4	154	8969	9123				
Augsburg							
No.	Training	Test	Totle	No.	Training	Test	Totle
1	146	13361	13507	5	52	523	575
2	264	30065	30329	6	7	1638	1645
3	21	3830	3851	7	23	1507	1530
4	248	26609	26857	Totle	761	77533	78294

3.2. Quantitative metrics

The evaluation of classification accuracy in this paper includes overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa). These are the three common target indicators that we have specified to objectively and effectively assess the classification effect of the HSI-LiDAR fusion images. In particular, the percentage of correctly classified pixels throughout all tests is determined by

the OA value, and the average of all the class accuracies is shown by the AA value. A multivariate statistical technique that accounts for the classification process's uncertainty elements is kappa statistics, which is related to the classification accuracy.

3.3. Ablation study

The ablation experiment will arrange and combine the three proposed modules (ESL, ASL, and FF) to demonstrate their effectiveness.

3.3.1. Impact of ESL

As shown in Table 2, when we used ESL to learn the spectral features of the HSI data, we achieved a better performance on the three datasets. The experimental results demonstrate that the ESL module extracts information more efficiently in the hyperspectral branch. However, this accuracy improvement is limited, mainly due to an insufficient information exchange during the fusion of the original hyperspectral features and the LiDAR features, which results in a loss of a lot of information.

Table 2. Ablation analysis of the designed ESL module, ASL module and FF module in terms of OA (%), AA (%), and KAPPA (× 100) on Houston2013, Trento and Augsburg datasets.

Module			Trento			Houston			Augsburg		
ESL	ASL	FF	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
×	×	×	98.12 (0.56)	97.29 (0.67)	97.47 (0.76)	95.55 (1.95)	96.30 (1.52)	95.16 (2.12)	77.15 (0.89)	36.70 (0.39)	65.79 (1.32)
√	×	×	98.56 (0.39)	97.78 (0.45)	98.06 (0.52)	95.62 (3.27)	96.40 (0.24)	95.23 (3.56)	77.40 (1.36)	36.82 (0.63)	66.11 (2.05)
×	√	×	98.22 (0.25)	97.43 (0.63)	97.60 (0.34)	96.73 (0.76)	97.24 (0.66)	96.44 (0.83)	79.82 (2.24)	43.34 (5.83)	70.44 (2.70)
×	×	√	99.04 (0.12)	98.59 (0.12)	98.70 (0.16)	98.05 (0.22)	98.50 (0.16)	97.88 (0.24)	80.90 (2.76)	67.28 (1.21)	73.23 (3.27)
√	√	×	98.85 (0.18)	98.63 (0.17)	98.46 (0.24)	96.98 (1.66)	97.54 (1.23)	96.71 (1.81)	83.03 (1.45)	39.52 (2.98)	74.58 (1.95)
√	×	√	99.06 (0.18)	98.53 (0.15)	98.73 (0.25)	98.52 (0.19)	98.54 (0.18)	98.21 (0.22)	83.58 (2.45)	68.93 (0.91)	76.69 (3.00)
×	√	√	99.03 (0.10)	98.56 (0.16)	98.69 (0.14)	98.40 (0.20)	98.68 (0.17)	98.26 (0.22)	84.02 (0.99)	64.99 (0.85)	77.05 (1.31)
√	√	√	**99.26 (0.09)**	**98.91 (0.13)**	**99.01 (0.12)**	**98.69 (0.18)**	**98.82 (0.17)**	**98.57 (0.19)**	**85.11 (1.04)**	**68.99 (0.89)**	**78.79 (1.34)**

3.3.2. Impact of ASL

As shown in Table 2, we only use the effects of the ASL module. For the three datasets, the OA

increased by 0.10, 1.18, and 2.67% respectively, the AA increased by 0.14, 0.94, and 6.64% respectively, and Kappa increased by 0.13, 1.28, and 4.65% respectively. This proves our point of using a Transformer to preserve the global information of LiDAR. However, in the network with ESL and ASL, we noticed that although the OA improved, the AA of the Augsburg dataset decreased. The reason could be that ESL pays more attention to local finer spectral details, while ASL pays more attention to the global information. The FF module of the original network did not fully complete the information exchange during the fusion process, which led to the classification results being more biased towards certain classes. Overall, the ASL module was more effective.

3.3.3. Impact of FF

As shown in Table 2, we only used the FF module in the network. As the most important part of the entire network, the FF module significantly improved the accuracy of all three datasets. The improvement of the three metrics (OA, AA, and Kappa) proved the effectiveness of the FF module. It is worth noting that training the network without the FF module will result in the classification being focused on specific categories. This is particularly significant on the Augsburg dataset (with a low AA metric). After adding the FF module, it uses cross attention to exchange the global information between the spectral and spatial features. The FF module enables an effective interaction between the hyperspectral information and the LiDAR information, which avoids a reliance on a single branch to solve the classification problems. All networks with added FF modules performed better in AA on the Augsburg dataset. The classification result is no longer focused on certain classes.

3.4. Comparative experiments

For the classification of HSI and LiDAR data, we performed comparative experiments on the following joint methods. The evaluation methods included a visual contrast and a quantitative analysis of the following:

TB-CNN [26]: A two-tunnel CNN network that extracts hyperspectral and LiDAR features, respectively.

EndNet [27]: A network that follows the deep encoder-decoder network architecture. This network uses the reconstruction strategy to merge the characteristics.

MDL-Late [28]: A concatenation-based fusion framework, whose feature fusion process is in the late stage.

MDL-Cross [28]: The goal of the framework is to understand the compact feature representations across modalities by interactively updating the parameters.

S2Enet [29]: A spatial-spectral enhancement module used by the framework to facilitate a cross-modal information exchange. Specifically, the spectral enhancement module improves the spectral representation of the LiDAR data using hyperspectral features, and the spatial enhancement module improves the spatial presentation of the hyperspectral data using LiDAR features.

AM^{3}Net [20]: An adaptive network using multi-scale fusion and mutual learning strategies. It uses a combination operator to process the HSI data and uses a spatial feature extraction module to extract spatial features of HSI and LiDAR and adaptively fuse them.

S2EFT [17]: A classification method of the HSI and LiDAR data based on a spectral-spatial-elevation fusion Transformer (S2EFT) framework. Moreover, the Transformer framework is introduced into the task of a multi-source RS image classification.

GAMF [30]: A novel, graph-attention based, multimodal fusion network (GAMF). It employs three major components, including an HIS-LiDAR feature extractor, a graph-attention based fusion

module, and a classification module.

Table 3. The classification result of Houston2013 dataset.

No.	Classes	TB-CNN	EndNet	MDL-Late	MDL-Cross	S2ENet	AM^{3}Net	S2EFT	GAMF	Ours
1	Grass-healthy	83.07 (5.80)	83.09 (3.21)	83.55 (0.54)	81.35 (2.24)	80.02 (15.82)	94.94 (1.46)	82.43 (6.23)	92.34 (2.52)	99.24 (0.60)
2	Grass-stressed	96.13 (3.88)	95.25 (4.81)	99.08 (1.76)	98.79 (1.56)	99.87 (0.28)	93.87 (1.89)	94.92 (1.41)	90.38 (2.67)	99.56 (0.44)
3	Grass-synthetic	98.38 (1.48)	99.83 (0.16)	99.29 (0.71)	92.06 (4.42)	99.91 (0.26)	97.57 (1.12)	99.40 (0.35)	99.56 (0.32)	99.61 (0.21)
4	Tree	99.18 (0.77)	97.09 (1.56)	99.35 (0.61)	98.14 (2.13)	98.42 (3.22)	94.02 (1.55)	97.34 (1.72)	99.83 (0.12)	95.64 (1.10)
5	Soil	99.14 (1.74)	98.25 (1.26)	99.74 (0.21)	96.95 (6.58)	99.99 (0.03)	98.87 (0.86)	98.57 (0.96)	97.43 (1.35)	99.97 (0.06)
6	Water	94.93 (3.13)	94.83 (4.12)	97.34 (1.95)	98.88 (1.68)	97.52 (3.31)	98.25 (0.47)	94.40 (2.79)	84.76 (5.64)	97.30 (1.13)
7	Residential	82.81 (4.09)	79.87 (4.75)	90.79 (2.74)	92.38 (4.60)	88.31 (6.17)	92.87 (1.52)	91.79 (5.28)	96.59 (1.75)	99.08 (0.61)
8	Commercial	84.92 (9.78)	81.32 (6.44)	86.14 (9.90)	92.62 (4.37)	90.62 (7.93)	95.92 (1.64)	78.44 (11.78)	80.12 (7.54)	98.51 (0.51)
9	Road	85.51 (6.50)	71.32 (15.59)	88.10 (3.43)	91.73 (5.87)	85.30 (6.99)	93.59 (1.55)	78.75 (13.23)	86.01 (4.73)	96.44 (0.73)
10	Highway	58.86 (14.04)	70.35 (11.86)	82.71 (9.97)	64.95 (13.47)	84.74 (7.87)	99.63 (0.62)	55.69 (21.67)	76.97 (10.21)	99.98 (0.06)
11	Railway	93.76 (2.70)	94.74 (3.43)	92.53 (3.29)	95.50 (3.39)	96.46 (2.17)	99.35 (0.67)	80.07 (7.43)	90.82 (3.46)	99.95 (0.09)
12	Parking Lot-1	82.93 (4.52)	75.48 (13.47)	83.43 (3.83)	64.97 (13.38)	81.38 (7.54)	89.26 (21.96)	73.48 (9.80)	84.87 (2.89)	97.26 (1.32)
13	Parking Lot-2	92.91 (2.24)	77.47 (7.32)	92.60 (0.66)	92.56 (2.81)	89.14 (3.54)	98.95 (1.99)	62.45 (17.26)	94.29 (1.93)	100.00 (0.00)
14	Tennis court	97.13 (2.24)	99.62 (0.77)	97.21 (2.24)	96.92 (2.68)	99.05 (1.74)	100.00 (0.00)	97.57 (1.12)	96.39 (1.48)	100.00 (0.00)
15	Running track	98.54 (1.95)	98.45 (1.04)	99.83 (0.32)	95.11 (7.63)	99.69 (0.62)	97.37 (1.52)	97.67 (1.26)	98.59 (0.83)	99.78 (0.39)
OA		88.08 (1.80)	86.12 (2.14)	91.55 (1.26)	88.71 (1.90)	91.50 (2.15)	95.55 (1.95)	84.41 (1.78)	90.66 (0.97)	**98.69 (0.18)**
AA		89.88 (1.46)	87.80 (1.82)	92.78 (0.96)	90.19 (1.68)	92.69 (1.75)	96.60 (1.52)	85.54 (2.27)	91.26 (1.84)	**98.82 (0.17)**
Kappa		87.07 (1.96)	84.95 (2.32)	90.84 (1.37)	87.75 (2.07)	90.78 (2.34)	95.16 (2.12)	83.10 (2.69)	89.91 (2.07)	**98.57 (0.19)**

We conducted extensive experiments, with the percentage of different methods being averaged after 20 repeated experiments. The numbers in parentheses in the table represent the standard deviation of the repeated experiments.

As shown in Figure 5 and Table 3, the Houston dataset has a large scale and a sparse distribution of labeled samples, leading to an increased similarity between classes. ESASNet resists this similarity problem to a certain extent. Compared with the other methods, ESASNet achieved the most consistent results with the GroundTruth, and the OA, AA, and Kappa improved by 2.96, 2.47, and 3.22%, respectively, compared with the baseline network.

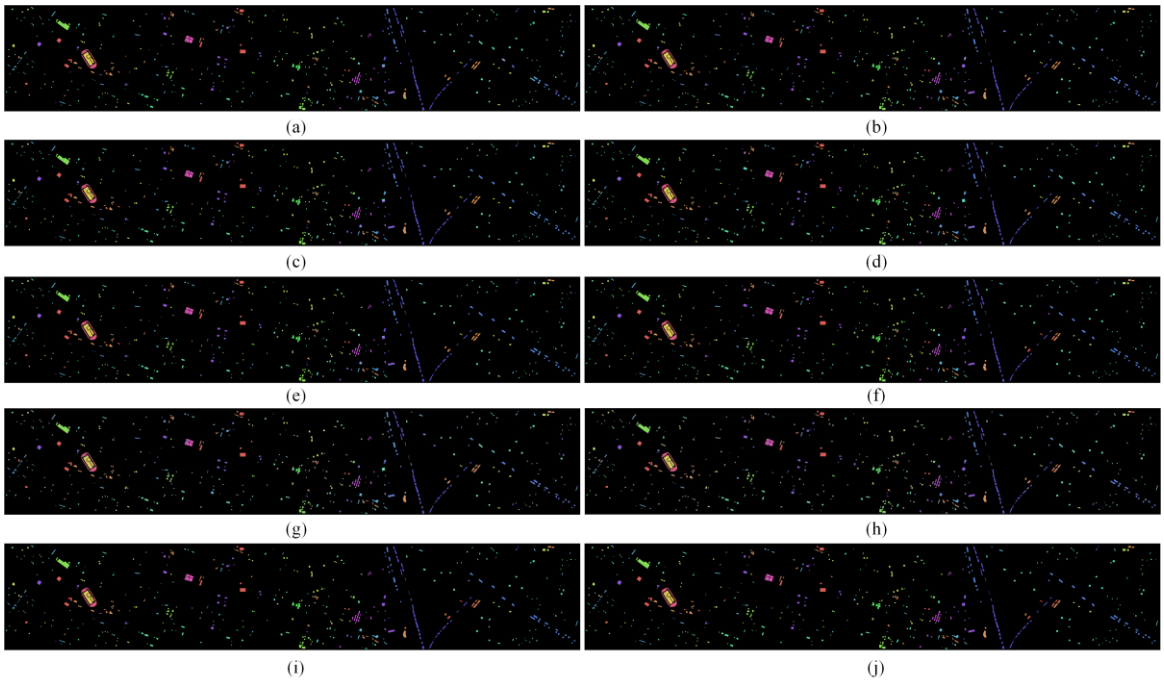

Figure 5. Classification maps obtained by seven methods on the Houston dataset: (a) TB-CNN (80.08%), (b) EndNet (86.12%), (c) MDL-Late (91.55%), (d) DML-Cross (88.71%), (e) S2ENet (91.50%), (f) AM^{3}Net (95.55%), (g) S2EFT (84.41%), (h) GAMF (90.66%), (i) Ours (98.69%), and (j) GroundTruth.

The evaluation criteria and classification maps comparison outcomes for the Trento dataset are presented in both Table 4 and Figure 6. The samples of different categories in this dataset are often concentrated in different regions, which is convenient for an accurate classification. The buildings and roads significantly differed in the elevation information; due to the concentration of the classification samples, the LiDAR data will improve the classification accuracy. For some networks that do not use the attention mechanism, such as TB-CNN, there are occasional error points in the central area of the classification block. ESASNet uses the cross-attention mechanism to ensure the global information of the feature; therefore, it achieved better results. The OA, AA, and Kappa increased by 1.14, 1.62, and 1.54%, respectively. The comparative experiments and classification performance on the Augsburg dataset are listed in Table 5 and Figure 7. The distribution of the labeled samples in the Augsburg dataset is highly dense, and the training samples are less than those in the Houston, Trento, and other datasets, therefore making the classification task more difficult. Because AM^{3}Net does not have an effective attention mechanism, it is easy to cause the classification to be biased towards specific categories in the case of a small number of training samples, which results in a poor performance. ESASNet uses ELSA to enhance the local finer features, uses a Transformer as attention to enhance the global information of LiDAR, and fully exchanges information during FF to obtain

obvious performance gains. From the index results, ESASNet improved by 7.96% on OA, 32.29% on AA, and 13% on Kappa compared to AM^{3}Net. As proposed, the ESASNet model yields smoother classification outcomes as observed through visual comparisons, where it exhibits a reduced intra-class noise and sharper classification boundaries.

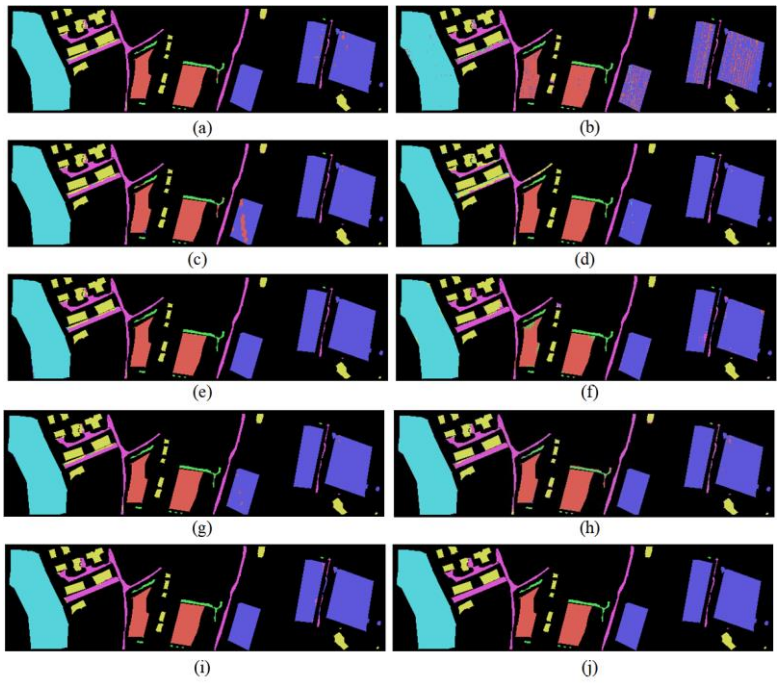

Figure 6. Classification maps obtained by seven methods on the Trento dataset: (a) TB-CNN (97.23%), (b) EndNet (86.43%), (c) MDL-Late (97.57%), (d) DML-Cross (97.65%), (e) S2ENet (93.69%), (f) AM^{3}Net (98.12%), (g) S2EFT (97.08%), (h) GAMF (97.08%), (i) Ours (99.26%), and (j) GroundTruth.

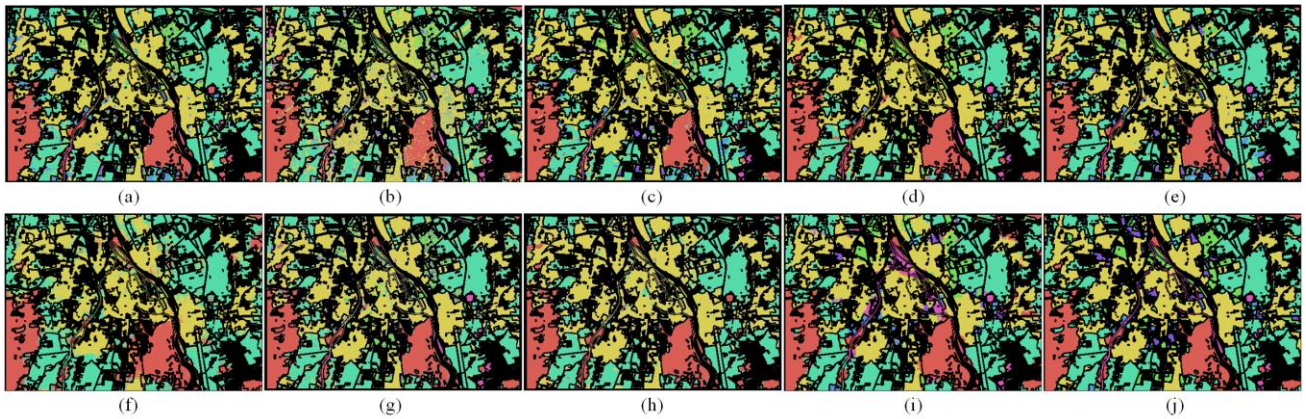

Figure 7. Classification maps obtained by seven methods on the Augsburg dataset: (a) TB-CNN (81.46%), (b) EndNet (74.56%), (c) MDL-Late (83.40%), (d) DML-Cross (85.09%), (e) S2ENet (84.45%), (f) AM^{3}Net (77.15%), (g) S2EFT (73.35%), (h) GAMF (83.83%), (i) Ours (85.11%), and (j) GroundTruth.

Table 4. The classification result of Trento dataset.

No.	Classes	TB-CNN	EndNet	MDL-Late	MDL-Cross	S2ENet	AM^{3}Net	S2EFT	GAMF	Ours
1	Apples	98.94 (2.89)	89.68 (11.60)	99.76 (0.15)	99.75 (0.21)	92.58 (22.04)	97.37 (2.16)	95.56 (3.62)	97.41 (0.47)	99.64 (0.27)
2	Buildings	98.46 (0.44)	96.43 (0.75)	97.41 (0.42)	96.80 (1.44)	97.67 (0.89)	97.15 (2.35)	98.45 (1.83)	99.90 (0.04)	97.53 (0.87)
3	Ground	88.55 (4.22)	96.93 (1.36)	86.84 (4.79)	92.03 (6.28)	95.95 (15.98)	97.75 (1.94)	92.78 (5.96)	84.93 (7.43)	99.96 (0.10)
4	Woods	99.82 (0.37)	98.50 (1.39)	99.69 (0.32)	99.93 (0.11)	91.20 (22.45)	99.42 (0.77)	100 (0.00)	99.93 (0.07)	99.97 (0.05)
5	Vineyard	96.68 (6.14)	70.64 (18.51)	97.92 (1.02)	99.54 (0.48)	95.82 (9.52)	99.01 (0.87)	97.87 (1.29)	99.80 (0.21)	99.67 (0.17)
6	Roads	89.65 (1.17)	89.21 (0.63)	89.33 (1.90)	83.84 (7.24)	91.21 (1.29)	93.05 (2.35)	87.05 (2.20)	85.98 (0.46)	96.70 (0.95)
OA		97.23 (2.08)	86.43 (5.65)	97.57 (0.42)	97.65 (0.68)	93.69 (9.91)	98.12 (0.56)	97.08 (1.88)	97.91 (0.18)	**99.26 (0.09)**
AA		95.35 (1.07)	90.23 (2.20)	95.16 (0.78)	95.31 (1.29)	94.07 (9.64)	97.29 (0.67)	95.29 (2.25)	94.66 (1.33)	**98.91 (0.13)**
Kappa		96.34 (2.70)	82.53 (6.86)	96.77 (0.56)	96.87 (0.91)	91.53 (13.73)	97.47 (0.76)	96.10 (1.21)	97.20 (0.24)	**99.01 (0.12)**

Table 5. The classification result of Augsburg dataset.

No.	Classes	TB-CNN	EndNet	MDL-Late	MDL-Cross	S2ENet	AM^{3}Net	S2EFT	GAMF	Ours
1	Trees	86.41 (10.46)	82.46 (11.11)	84.12 (5.12)	89.45 (4.44)	86.97 (4.67)	88.15 (1.57)	85.43 (3.52)	92.94 (2.53)	94.67 (0.74)
2	Grass pure	87.34 (20.30)	82.55 (5.29)	94.92 (2.52)	92.66 (8.77)	92.92 (6.95)	84.13 (2.48)	80.92 (4.55)	96.57 (1.54)	86.79 (2.32)
3	Grass surface	49.25 (10.98)	36.63 (10.98)	58.38 (14.51)	53.72 (24.89)	62.28 (14.37)	0.00 (0.00)	44.30 (16.89)	0.00 (0.00)	58.16 (2.18)
4	Dirt and sand	84.70 (3.54)	73.79 (6.76)	80.86 (4.09)	87.09 (4.88)	84.24 (4.58)	84.62 (1.95)	70.51 (8.72)	88.61 (2.84)	87.84 (1.60)
5	Road materials	58.16 (15.74)	24.89 (15.16)	56.93 (19.11)	54.06 (18.65)	53.13 (18.57)	0.00 (0.00)	36.90 (12.68)	0.00 (0.00)	83.35 (1.65)
6	Water	12.82 (7.26)	10.78 (4.30)	16.12 (9.35)	3.67 (4.61)	13.18 (6.63)	0.00 (0.00)	8.11 (4.53)	0.00 (0.00)	25.73 (3.09)
7	Building Shadow	27.72 (11.25)	42.24 (4.57)	37.87 (10.80)	39.45 (7.32)	41.73 (7.95)	0.00 (0.00)	22.76 (11.69)	0.00 (0.00)	46.38 (4.80)
OA		81.46 (8.53)	74.56 (2.35)	83.40 (2.07)	85.09 (2.91)	84.45 (3.41)	77.15 (0.89)	73.35 (2.73)	83.83 (0.79)	**85.11 (1.04)**
AA		58.06 (5.90)	50.47 (2.76)	61.32 (3.24)	60.01 (4.60)	62.06 (3.88)	36.70 (0.39)	49.85 (3.32)	39.73 (0.47)	**68.99 (0.89)**
Kappa		73.67 (11.99)	63.47 (3.00)	76.53 (2.62)	78.84 (3.72)	77.97 (4.89)	65.79 (1.32)	61.71 (2.82)	76.20 (1.83)	**78.79 (1.34)**

Table 6 represents the impact of different patch sizes on the network on three datasets. We use OA for comparison, with standard deviation in parentheses. The table shows an optimal value for accuracy with patch size of 25 × 25 for the Trento dataset and the Augsburg dataset, and 29 × 29 for the Houston dataset. These results also indicate that our network requires a larger patch to enhance the ability of the feature extraction.

Table 6. Impact of different patch size for the OA on three datasets.

Patch size	Trento	Houston	Augsbarg
9 × 9	94.45 (0.71)	95.63 (0.38)	65.22 (2.92)
13 × 13	97.82 (0.61)	98.19 (0.24)	71.63 (4.99)
17 × 17	98.53 (0.51)	98.15 (0.21)	80.20 (1.95)
21 × 21	98.42 (0.31)	98.30 (0.24)	81.24 (2.05)
25 × 25	**99.26 (0.09)**	98.41 (0.21)	**85.11 (1.04)**
29 × 29	98.69 (0.18)	**98.69 (0.18)**	81.94 (1.63)
33 × 33	98.04 (0.23)	98.31 (0.14)	81.43 (1.79)

3.5. Impart of training samples

Evaluating the performance of a model under different sample sizes is a key aspect of evaluating the model quality. Deep learning models typically require a small amount of training data to achieve advanced learning capabilities due to the current task typically using training samples determined by predecessors. Therefore, we used the OA obtained from 20, 40, 60, 80, and 100% of the original training samples for comparison.

For the convenience of the comparison, we only used four representative networks (S2EFT, GAMF, AM^{3}Net, and Our) for comparison on the Trento dataset. As shown in Figure 8, we used OA as a measurement metric.

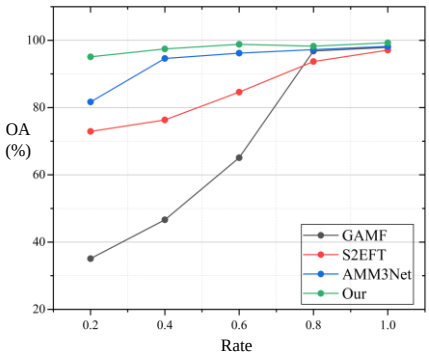

Figure 8. OA of original training samples with different rates (20, 40, 60, 80, 100%).

GAMF and S2EFT, which are excellent hyperspectral classification networks, performed well on the original sample size. However, the small number of training samples lead to a rapid decrease in their accuracy. The shallow feature extraction of GAMF and the lack of shallow feature processing in S2EFT resulted in an excessive loss of their shallow feature information. AM^{3}Net adopts dual branch shallow feature extraction to ensure an information extraction. Therefore, the accuracy is ensured under different samples.

Due to its efficient spectral spatial feature extraction module, our network minimizes the loss of the shallow feature information. Additionally, the efficient fusion module ensures the effectiveness of the information exchange. Our network enables us to maintain a good performance even in low sample situations.

4. Conclusions

This paper proposed a network called ESASNet for HSI-LiDAR data fusion and classification. ESASNet proposed a CNN-Transformer-based feature extraction structure to provide a great assistance in extracting spectral information from hyperspectral data and spatial information from LiDAR. To fully utilize the spectral correlation and extract more spectral information more efficiently, the ESL module was designed to maximize the extraction of spectral information from HSI. A more efficient FF module was proposed to assign more weight to the global information and further improve the information exchange of the deep features. Experiments on HSI-LiDAR fusion datasets, such as Houston, Trento, and AugsBurg, showed that the ESASNet network had a higher accuracy and was better than some of the most advanced methods at present.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the Technology Project of State Grid Qinghai Electric Power Co., Ltd., grant number, SGQHHD00YXJS2310532.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. J. Mäyrä, S. Keski-Saari, S. Kivinen, T. Tanhuanpää, P. Hurskainen, P. Kullberg, et al., Tree species classification from airborne hyperspectral and LiDAR data using 3D convolutional neural networks, *Remote Sens. Environ.*, **256** (2021), 112322. https://doi.org/10.1016/j.rse.2021.112322
2. C. T. de Almeida, L. S. Galvao, J. P. H. B. Ometto, A. D. Jacon, F. R. de Souza Pereira, L. Y. Sato, et al., Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms, *Remote Sens. Environ.*, **232** (2019), 111323. https://doi.org/10.1016/j.rse.2019.111323
3. M. R. Soosai, Y. C. Joshya, R. S. Kumar, I. G. Moorthy, S. Karthikumar, N. T. L. Chi, et al., Versatile image processing technique for fuel science: A review, *Sci. Total Environ.*, **780** (2021), 146469. https://doi.org/10.1016/j.scitotenv.2021.146469

4. Y. Gu, Q. Wang, X. Jia, J. A. Benediktsson, A novel MKL model of integrating LiDAR data and MSI for urban area classification, *IEEE Trans. Geosci. Remote Sens.*, **53** (2015), 5312–5326. https://doi.org/10.1109/TGRS.2015.2421051
5. J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, N. Canagarajah, Pixel-and region-based image fusion with complex wavelets, *Inf. Fusion*, **8** (2007), 119–130. https://doi.org/10.1016/j.inffus.2005.09.006
6. Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, X. Wang, Deep learning for pixel-level image fusion: Recent advances and future prospects, *Inf. Fusion*, **42** (2018), 158–173. https://doi.org/10.1016/j.inffus.2017.10.007
7. S. Li, X. Kang, L. Fang, H. Yin, Pixel-level image fusion: A survey of the state of the art, *Inf. Fusion*, **33** (2017), 100–112. https://doi.org/10.1016/j.inffus.2016.05.004
8. Y. Tong, Y. Quan, W. Feng, G. Dauphin, Y. Wang, P. Wu, et al., Multi-scale feature extraction and total variation based fusion method for HSI and lidar data classification, in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, IEEE, Brussels, Belgium, (2021), 5433–5436. https://doi.org/10.1109/IGARSS47720.2021.9554337
9. R. Luo, W. Liao, H. Zhang, Y. Pi, W. Philips, Classification of cloudy hyperspectral image and LiDAR data based on feature fusion and decision fusion, in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, Beijing, China, (2016), 2518–2521. https://doi.org/10.1109/IGARSS.2016.7729650
10. X. Xu, W. Li, Q. Ran, Q. Du, L.Gao, B. Zhang, Multisource remote sensing data classification based on convolutional neural network, *IEEE Trans. Geosci. Remote Sens.*, **56** (2017), 937–949, https://doi.org/10.1109/TGRS.2017.2756851
11. H. Li, P. Ghamisi, U. Soergel, X. X. Zhu, Hyperspectral and LiDAR fusion using deep three-stream convolutional neural networks, *Remote Sens.*, **10** (2018), 1649. https://doi.org/10.3390/rs10101649
12. C. Ge, Q. Du, W. Sun, K. Wang, J. Li, Y. Li, Deep residual network-based fusion framework for hyperspectral and LiDAR data, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **14** (2021), 2458–2472. https://doi.org/10.1109/JSTARS.2021.3054392
13. X. Zhao, R. Tao, W. Li, H. C. Li, Q. Du, W. Liao, et al., Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture, *IEEE Trans. Geosci. Remote Sens.*, **58** (2020), 7355–7370. https://doi.org/10.1109/TGRS.2020.2982064
14. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
15. Y. Yu, T. Jiang, J. Gao, H. Guan, D. Li, S. Gao, et al., CapViT: Cross-context capsule vision transformers for land cover classification with airborne multispectral LiDAR data, *Int. J. Appl. Earth Obs. Geoinf.*, **111** (2022), 102837. https://doi.org/10.1016/j.jag.2022.102837
16. S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, J. Chanussot, Multimodal fusion transformer for remote sensing image classification, *IEEE Trans. Geosci. Remote Sens.*, **61** (2020), 1–20. https://doi.org/10.1109/TGRS.2023.3286826
17. Y. Feng, J. Zhu, R Song, X. Wang, S2EFT: Spectral-spatial-elevation fusion transformer for hyperspectral image and LiDAR classification, *Knowledge-Based Syst.*, **283** (2024), 111190. https://doi.org/10.1016/j.knosys.2023.111190

18. G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, B. Jeon, Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–16. https://doi.org/10.1109/TGRS.2022.3232498
19. X. Wang, Y. Feng, R. Song, Z. Mu, C. Song, Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data, *Inf. Fusion*, **82** (2022), 1–18. https://doi.org/10.1016/j.inffus.2021.12.008
20. J. Wang, J. Li, Y. Shi, J. Lai, X. Tan, AM³Net: Adaptive mutual-learning-based multimodal data fusion network, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 5411–5426. https://doi.org/10.1109/TCSVT.2022.3148257
21. S. Mohla, S. Pande, B. Banerjee, S. Chaudhuri, 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)*, IEEE, Seattle, WA, USA, (2020), 416–425, https://doi.org/10.1109/CVPRW50498.2020.00054
22. J. Zhou, P. Wang, F. Wang, Q. Liu, H. Li, R. Jin, Elsa: Enhanced local self-attention for vision transformer, preprint, arXiv:2112.12786, https://doi.org:10.48550/arXiv.2112.12786
23. M. Khodadadzadeh, J. Li, S. Prasad, A. Plaza, Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **8** (2015), 2971–2983. https://doi.org/10.1109/JSTARS.2015.2432037
24. B. Rasti, P. Ghamisi, R. Gloaguen, Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis, *IEEE Trans. Geosci. Remote Sens.*, **55** (2017), 3997–4007. https://doi.org/10.1109/TGRS.2017.2686450
25. D. Hong, J. Hu, J. Yao, J. Chanussot, X. X. Zhu, Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model, *ISPRS J. Photogramm. Remote Sens.*, **178** (2021), 68–80. https://doi.org/10.1016/j.isprsjprs.2021.05.011
26. X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, B. Zhang, Multisource remote sensing data classification based on convolutional neural network, *IEEE Trans. Geosci. Remote Sens.*, **56** (2017), 937–949. https://doi.org/10.1109/TGRS.2017.2756851
27. D. Hong, L. Gao, R. Hang, B. Zhang, J. Chanussot, Deep encoder–decoder networks for classification of hyperspectral and LiDAR data, *IEEE Geosci. Remote Sens. Lett.*, **19** (2020), 1–5. https://doi.org/10.1109/LGRS.2020.3017414
28. D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, et al., More diverse means better: Multimodal deep learning meets remote-sensing imagery classification, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021), 4340–4354. https://doi.org/10.1109/TGRS.2020.3016820
29. S. Fang, K. Li, Z. Li, S²ENet: Spatial-spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data, *IEEE Geosci. Remote Sens. Lett.*, **19** (2021), 1–5. https://doi.org/10.1109/LGRS.2021.3121028
30. J. Cai, M. Zhang, H. Yang, Y. He, Y. Yang, C. Shi, et al. A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and LiDAR data, *Expert Syst. Appl.*, **249** (2024), 123587. https://doi.org/10.1016/j.eswa.2024.123587

AIMS Press