



Article

Convolution-Transformer Adaptive Fusion Network for Hyperspectral Image Classification

Jiaju Li ¹ , Hanfa Xing ^{2,3}, Zurui Ao ², Hefeng Wang ^{1,4} , Wenkai Liu ² and Anbing Zhang ^{1,4,*}

¹ School of Mining and Geomatics Engineering, Hebei University of Engineering, Handan 056038, China
² Beidou Research Institute, Faculty of Engineering, South China Normal University, Foshan 528225, China
³ School of Geography, South China Normal University, Guangzhou 510631, China
⁴ Key Laboratory of Natural Resources and Spatial Information, Handan 056038, China
 * Correspondence: zhanganbing@hebeu.edu.cn; Tel.: +86-189-3102-9866

Abstract: Hyperspectral image (HSI) classification is an important but challenging topic in the field of remote sensing and earth observation. By coupling the advantages of convolutional neural network (CNN) and Transformer model, the CNN–Transformer hybrid model can extract local and global features simultaneously and has achieved outstanding performance in HSI classification. However, most of the existing CNN–Transformer hybrid models use artificially specified hybrid strategies, which have poor generalization ability and are difficult to meet the requirements of recognizing fine-grained objects in HSI of complex scenes. To overcome this problem, we proposed a convolution–Transformer adaptive fusion network (CTAFNet) for pixel-wise HSI classification. A local–global fusion feature extraction unit, called the convolution–Transformer adaptive fusion kernel, was designed and integrated into the CTAFNet. The kernel captures the local high-frequency features using a convolution module and extracts the global and sequential low-frequency information using a Transformer module. We developed an adaptive feature fusion strategy to fuse the local high-frequency and global low-frequency features to obtain a robust and discriminative representation of the HSI data. An encoder–decoder structure was adopted in the CTAFNet to improve the flow of fused local–global information between different stages, thus ensuring the generalization ability of the model. Experimental results conducted on three large-scale and challenging HSI datasets demonstrate that the proposed network is superior to nine state-of-the-art approaches. We highlighted the effectiveness of adaptive CNN–Transformer hybrid strategy in HSI classification.

Keywords: deep learning; hyperspectral image classification; convolutional neural networks; transformer; hybrid strategy; feature fusion



Citation: Li, J.; Xing, H.; Ao, Z.; Wang, H.; Liu, W.; Zhang, A. Convolution-Transformer Adaptive Fusion Network for Hyperspectral Image Classification. *Appl. Sci.* **2023**, *13*, 492. <https://doi.org/10.3390/app13010492>

Academic Editor: Jianbo Gao

Received: 10 November 2022

Revised: 25 December 2022

Accepted: 26 December 2022

Published: 30 December 2022



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hyperspectral remote sensing integrates imaging and spectral technology [1] to obtain images with rich spectral and spatial information. Hyperspectral image (HSI) has been widely used in many fields, including agriculture, forestry, mining, and marine research [2–5]. For most of these applications, HSI classification is an important basic step, which aims to assign a semantic label to each pixel in the image [6]. Even though it has attracted considerable attention, it remains a challenging problem because of the large spatial variability of spectral signatures and the limited available training samples versus the high dimensionality of hyperspectral data [7].

Extracting discriminative and robust spatial–spectral features is the key to a successful HSI classification [7]. In the early research, spatial–spectral features were designed manually according to the prior knowledge on the land cover types. Many feature descriptors in the field of computer vision were adopted for HSI feature extraction. For example, principal component analysis (PCA) [8], independent component analysis (ICA) [9] and linear discriminant analysis (LDA) [10] were used to extract spectral features, and scale-invariant

feature transform (SIFT) [11], local binary patterns (LBP) [12], and extended morphological profile (EMP) [13] were used to extract spatial features. On the basis of artificially designed features, traditional machine learning methods, such as support vector machine (SVM) [14], random forest (RF) [15], and extreme learning machine (ELM) [16] were used to classify HSI pixels into different types. The performance of these methods largely depends on the quality of the manually designed features. Generalization and robustness of these methods are generally poor because the features are designed for specific tasks, and the feature extraction requires complex processes of parameter tuning.

Different from traditional methods that rely on artificially designed features, deep learning-based methods can automatically learn multi-level nonlinear features, which are conducive to analyzing the inherent characteristics of HSI [7]. In recent years, as a representative deep learning model, the convolutional neural network (CNN) has made milestone progress in HSI classification. Relevant scholars have proposed a variety of CNN-based classification models, which can be divided into three categories according to the types of extracted features: (1) spectral CNN, which uses 1D CNN [17] to extract the spectral features of HSI; (2) spatial CNN, which uses two-dimensional CNN [18] to extract the spatial features from HSI after dimension reduction; (3) spectral-spatial CNN, which uses three-dimensional convolution [19] or dual branch network [20] to simultaneously extract spectral-spatial joint features of HSI. Although these CNN models have yielded satisfactory results in specific applications, they generally face the following two challenges: (1) due to the limited receptive field, they can hardly capture low-frequency signals, which provide global information (e.g., global shapes and structures) [21]; (2) the quality of the extracted high-frequency signals (e.g., local edges and texture) needs to be improved [22]. Several studies have attempted to improve the CNN models by directly extending the receptive field of the convolution kernel, including the use of dilated convolutions [23] and the construction of multi-scale feature pyramids [24]. Recent studies also introduced the attentional mechanism to enhance the useful components in the features while suppressing the useless ones, such as the spectral attention [25], spatial attention [26], and spectral-spatial attention [27]. Nevertheless, the above-mentioned methods do not fully overcome the limitations of CNNs, as they depend strictly on convolution operations, which are incapable of modeling long-term dependencies [28].

In recent years, the emergence of the Transformer has provided a new means for HSI classification. Transformer is a new neural network architecture consisting of a multi-head self-attention (MSA) module and a feed-forward neural network [29]. By introducing the MSA module, Transformer can effectively capture long-term dependencies. Qing et al. [30] proposed a self-attention Transformer network (SATNet) for HSI classification. SATNet employs Transformer encoders to extract image features and uses a multilevel residual structure to connect multiple encoder blocks to solve the vanishing gradient and over-fitting problems. Sun et al. [31] proposed an encoder-decoder network that fuses local-global spatial attention and spectral attention (FSSANet) for HSI classification. FSSANet introduces spectral attention into the Swin Transformer encoder [32] to encode the rich spectral-spatial information of HSI and therefore improves the classification accuracy. Although the application of Transformer as a backbone network for feature extraction has a good performance in HSI classification, it still faces two problems. First, Transformer needs to convert images to low-dimensional patch embeddings, which destroys the internal structure of the images and increases the requirement of quantity of training data to learn the unique properties of the images [33]. Secondly, Transformer cannot learn the correlation between different pixels within a patch, and it is difficult to capture local high-frequency information [34].

After reviewing the CNN-based and Transformer-based models, it was found that they complement each other, and combining use of them offers opportunities to enhance the modeling of both local high-frequency information and global long-term dependencies [35]. Sun et al. [36] proposed a spectral-spatial feature tokenization Transformer (SSFTT), which uses convolution layers to extract shallow spectral-spatial features and the Transformer

encoder to capture deep semantic information. The extracted HSI spectral–spatial semantic features from shallow to deep were fused to improve classification accuracy. Song et al. [37] proposed a two branch HSI classification framework based on three-dimensional CNN and bottleneck spatial–spectral Transformer (B2ST). In this framework, both branches use a combination of shallow CNN and deep Transformer. One branch is used to extract spatial local–global joint features, and the other is focused on extracting spectral local–global joint features. The fused spectral–spatial features can express the local global semantic information, thus achieving outstanding classification performance. Although the existing CNN–Transformer hybrid models have made progress in HSI classification, they have at least two disadvantages. First, existing methods integrate convolution and Transformer through artificially specified strategies, which are empirical and might lead to poor generalization ability. Second, most of these models adopt the image-wise classification network based on “encoder–label” structure, which make predictions through the features of the last stage and cannot make full use of the information obtained from the other stages. As a result, the “encoder–label” structure might yield incorrect results, e.g., omission of fine-grained objects and confusion between similar objects in HSI of complex scenes [21,38]. It is urgent to develop a new method that can adaptively integrate the features extracted by convolution and Transformer and effectively capture the information obtained from multiple stages to overcome the limitations of existing methods.

Studies have noted that the deep learning methods normally suffer from the data-hungry problem [39]. This problem is particularly acute in HSI classification due to a lack of high-quality benchmark dataset. Most of the previous studies have used small-scale datasets, such as Indian Pines, Salinas, and Pavia University [17,20,23,27,30,40], which comprise only hundreds of rows by hundreds of columns of pixels. It is often the case that deep learning methods yield nearly perfect classification results on these datasets, possibly because the overlapping pixels between training data and test data will lead to information leakage [40,41]. The new data partition method solves the problem of information leakage to some extent [40–42]. However, the partitioned training data usually contains only thousands of pixels, which is difficult to support the training of deep learning models (usually including millions of parameters). At the same time, a small number of test samples are insufficient to comprehensively evaluate the performance of the model. Recently, a series of large-scale and challenging datasets have been developed [43,44]. It is of interest to further test existing deep learning models using these benchmark datasets to better understand their comprehensive performances.

The main objectives of this study are to: (1) develop a convolution–Transformer adaptive fusion network (CTAFNet) using an adaptive hybrid strategy to fuse high-frequency and low-frequency signals so as to extract more robust and discriminating local–global fusion representation and improve the performance of HSI classification; (2) to evaluate the performance of several widely used deep learning models (i.e., FSSANet [31], SS3FCN [42], UNet [45], etc.) using large-scale and challenging benchmark datasets to provide a fair and comprehensive comparison between the models.

2. Datasets and Methods

2.1. Datasets

2.1.1. Data Descriptions

This paper uses three large-scale and challenging HSI datasets as benchmarks: the AeroRIT scene [43], The Data Fusion Contest 2018 (DFC2018) dataset and Xiongan New Area Matiwang Village (hereinafter referred to as Xiongan) dataset [44].

The AeroRIT scene is a HSI of the Rochester Institute of Technology’s university campus captured by the Headwall Micro E sensor. The sensor captures a total of 372 spectral bands. We use HSI of a total of 51 bands obtained by sampling every tenth band from 400 nm to 900 nm. This dataset has a ground sampling distance (GSD) of 0.4 m/px, resulting in a 1973×3975 px image. This dataset is marked with five types of ground objects, as shown in Figure 1. This dataset has problems, such as small target recognition, effects of

glint, and shadows. Although there are few types of ground objects, accurate classification is still challenging.

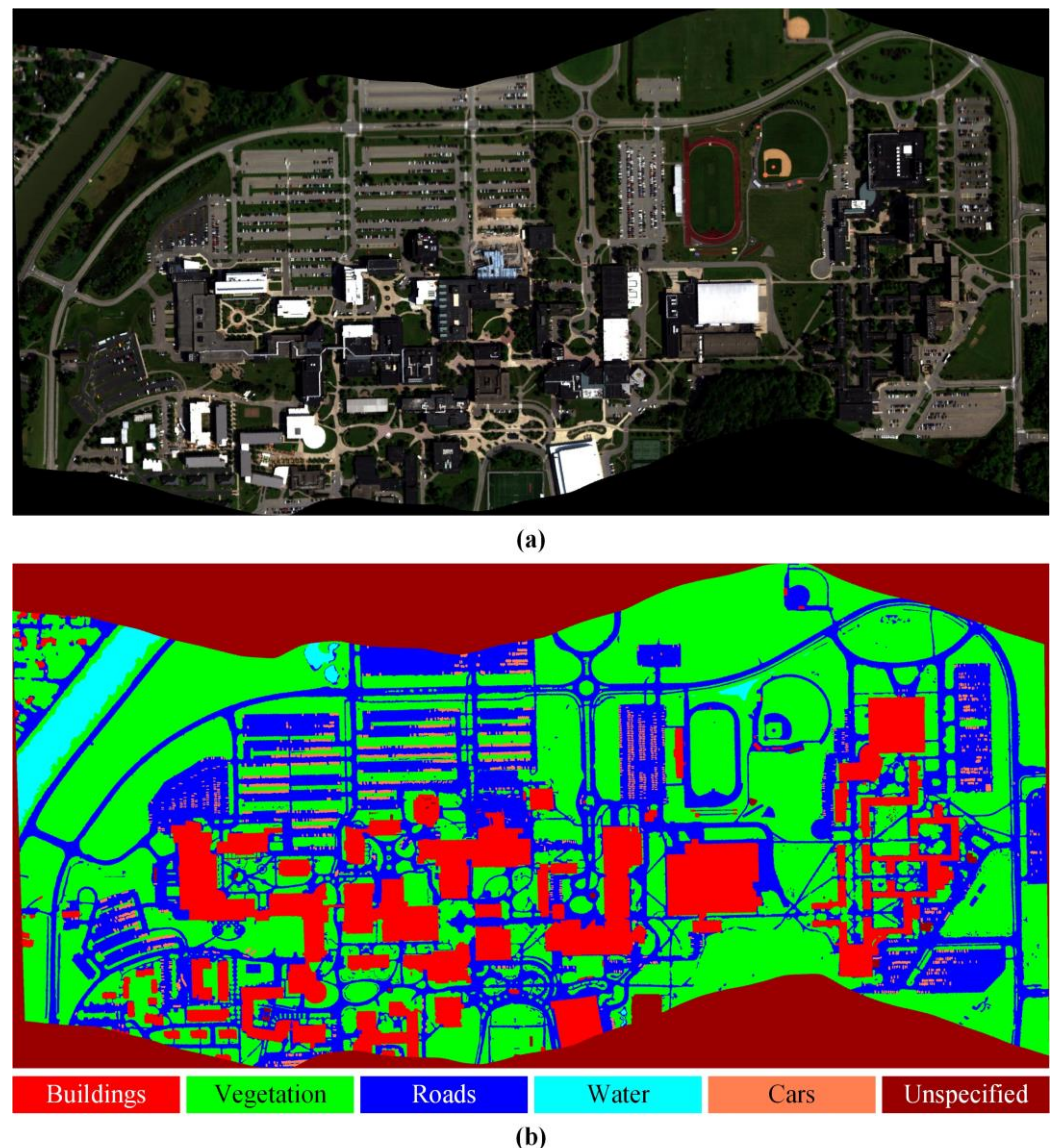


Figure 1. The AeroRIT scene. (a) RGB image; (b) Ground-truth classification map.

The DFC2018 hyperspectral data were acquired over Central Houston, Texas, USA, using an ITRES-CASI 1500 airborne sensor. It covers a 380–1050 nm spectral range over 48 contiguous bands at 1 m GSD, resulting in a 1202×4768 px image. This dataset has a total of 20 types of labels and contains many fine-grained objects. The number of artificial turn, water, and unpaved parking lots classes is too small to be partitioned into the training set and the test set at the same time. The method adopted in this paper cannot evaluate the performance of these three classes, therefore, these three classes are merged into the unspecified class. After processing the labels, there are 17 types of ground objects left, as shown in Figure 2.

The Xiongan dataset is a HSI of Matiwan Village in Xiongan New Area of China, which is acquired using the visible and near-infrared imaging spectrometer. The spectral range is 400~1000 nm with 256 bands, and the spatial resolution is 0.5 m, resulting in a 1580×3750 px image. This dataset has a number of fine-grained objects, which are mainly croplands. For the same reason as DFC2018 dataset, we merged the acacia and sparse

forests into the unspecified class. After merging, there are 18 types of ground objects, as shown in Figure 3.

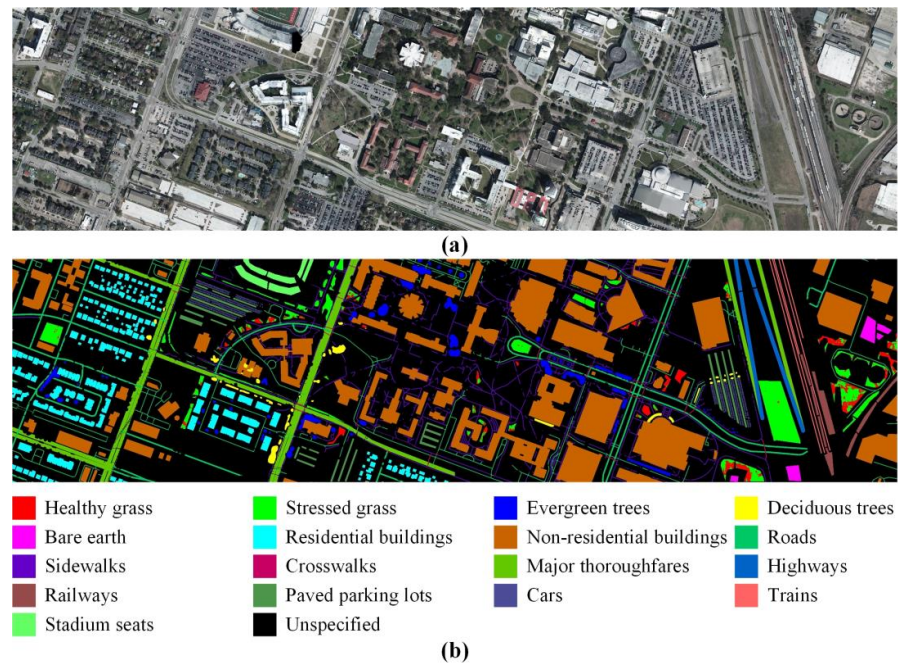


Figure 2. The DFC2018 dataset. (a) RGB image; (b) ground-truth classification map.



Figure 3. The Xiongan dataset. (a) three-band false color composite; (b) ground-truth map.

2.1.2. Data Partition Method

The AeroRIT dataset is partitioned to training, validation, and test sets according to the method in the paper [43], with a patch size of 64×64 . The number of samples of each type is shown in Table 1. The DFC2018 dataset and the Xiongan dataset use the same data partition method as shown in Figure 4. We first crop the image into a number of 64×64 non-overlapping patches and then randomly partition these patches into the training set and test set. The number of samples of the DFC2018 dataset and the Xiongan dataset are shown in Tables 2 and 3, respectively.

Table 1. Sample size of each class in each set after partitioning the AeroRIT dataset.

Class	Train	Val	Test
Buildings	423,605	141,424	352,788
Vegetation	1,277,105	349,211	1,551,317
Roads	843,770	319,228	781,508
Water	112,946	0	5718
Cars	70,313	19,537	42,243
Total	2,727,739	829,400	2,733,574

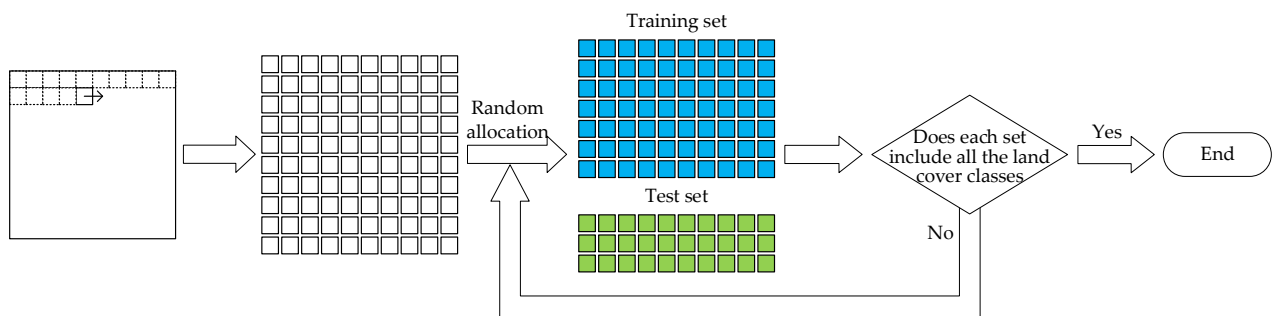


Figure 4. Dataset partition method.

Table 2. Number of samples in train and test sets for the DFC2018 datasets.

Class	Train	Test
Healthy grass	24,546	14,650
Stressed grass	95,176	34,832
Evergreen trees	42,446	11,876
Deciduous trees	13,973	6199
Bare earth	11,844	6220
Residential buildings	119,456	39,539
Non-residential buildings	616,125	278,644
Roads	129,820	53,463
Sidewalks	94,431	41,604
Crosswalks	3875	2184
Major thoroughfares	134,167	51,271
Highways	29,175	10,263
Railways	18,308	9440
Paved parking lots	31,833	14,099
Cars	19,875	6414
Trains	14,524	6955
Stadium seats	17,074	10,222
Total	1,416,648	597,875

Table 3. Number of samples in train and test sets for the Xiongan datasets.

Class	Train	Test
Acer negundo	140,953	84,694
Willow	119,010	61,756
Elm	12,523	2830
Paddy	342,682	109,462
Sophora japonica	305,611	169,980
Fraxinus chinensis	123,224	46,118
Goldenrain tree	19,126	4178
Waters	122,790	42,857
Bare ground	27,526	10,883
Stubble	134,439	59,391
Corn	43,796	15,369
Pyrus	764,745	261,768
Soybean	6682	469
Poplar	62,489	28,583
Vegetable field	20,822	8326
Grass	315,130	106,660
Peach	43,220	22,294
Building	18,452	11,164
Total	2623,220	1,046,782

2.2. Method

We propose a convolution–Transformer adaptive fusion network (CTAFNet) for pixel-wise HSI classification. CTAFNet uses a novel local–global fusion feature extraction unit, called the convolution–Transformer adaptive fusion kernel, to capture both the local high-frequency features and the sequential low-frequency information. An adaptively feature fusion strategy was designed to obtain a more robust and discriminative representation of the HSI data. Moreover, CTAFNet adopts an encoder–decoder structure to improve the flow of fused local–global information between different stages, thus ensuring the generalization ability of the model.

2.2.1. CTAFNet Architecture Overview

The overall architecture of the proposed CTAFNet for HSI classification is presented in Figure 5a. CTAFNet has a CNN–Transformer hybrid architecture, which mainly composed of encoder and decoder. The encoder follows the hierarchical pyramid architecture equipped with a CTAFK in each stage, which is used to capture local–global feature representations at different levels. The decoder uses bilinear interpolation for up-sampling to recover the spatial resolution of the feature map and concatenates the feature map from the previous and the current stage of encoders. The encoder–decoder framework enables CTAFNet to effectively utilize the local and global information extracted at each stage to improve HSI classification accuracy. The output feature size for each stage is shown in the Table 4.

2.2.2. Convolution–Transformer Adaptive Fusion Kernel

The artificially specified CNN–Transformer hybrid strategy has poor generalization ability and is difficult to satisfy the recognition requirements of fine-grained objects in complex scenes. To overcome this challenge, a feature extraction unit called the convolution–Transformer adaptive fusion kernel (CTAFK) was designed, as shown in Figure 5b.

CTAFK captures the local high-frequency features using a convolution block (Conv block) and extracts the global and sequential low-frequency information using a Transformer block (Trans block). Afterwards, the local high-frequency and global low-frequency features are adaptively weighted and fused to provide a more generalized and discriminative representation of the HSI data. The details of Conv block, Trans block, and CTAFK workflow are shown, as follows.

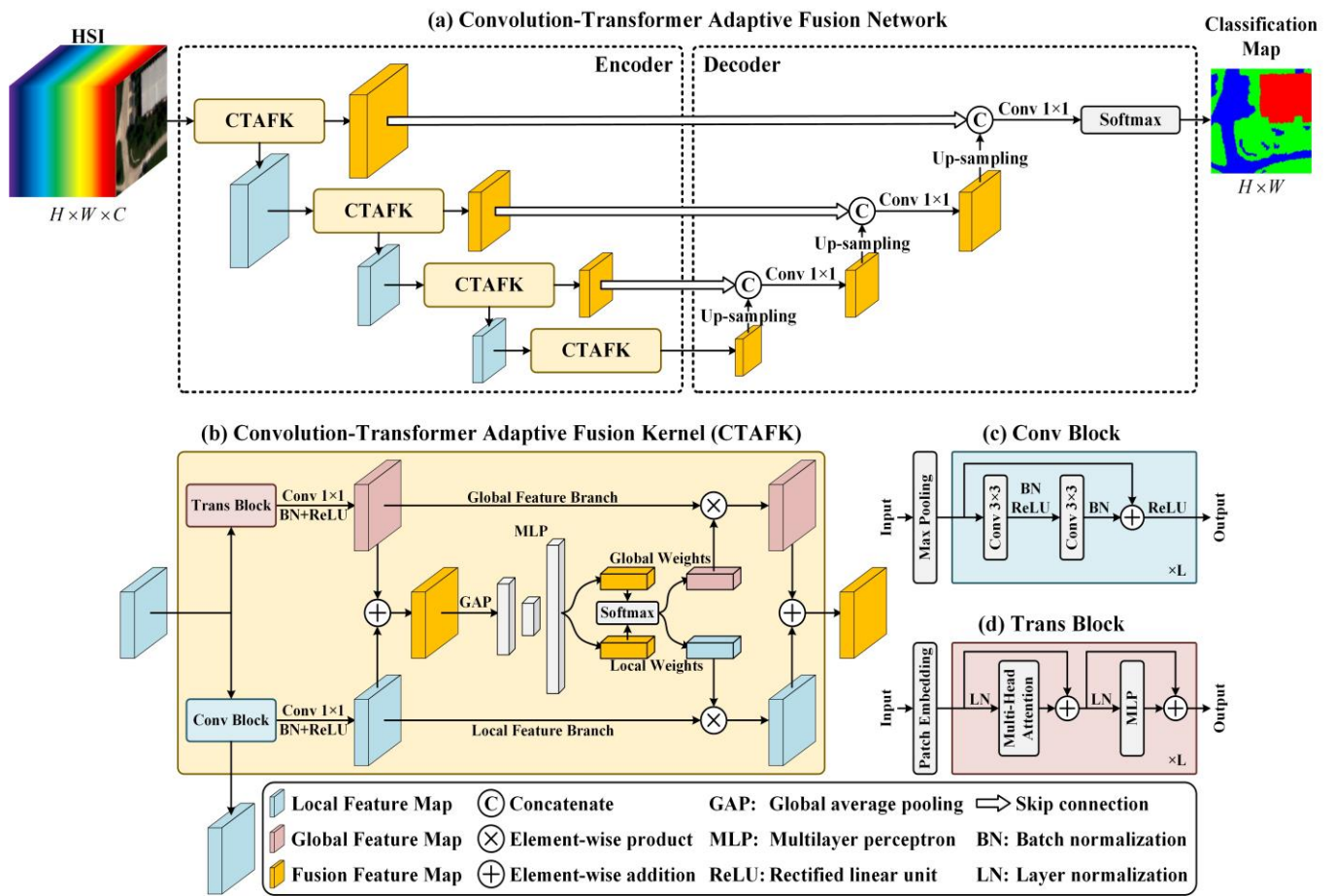


Figure 5. The framework of CTAFFNet for HSI classification.

Table 4. Output feature size for each stage.

Stage	Encoder ($H \times W \times C$)	Decoder ($H \times W \times C$)
1	$64 \times 64 \times 96$	$16 \times 16 \times 96$
2	$32 \times 32 \times 96$	$32 \times 32 \times 96$
3	$16 \times 16 \times 96$	$64 \times 64 \times K^1$
4	$8 \times 8 \times 96$	

¹ K represents the number of classes.

1. Conv Block

Conv block is the basic module for extracting local features in CTAFK, which is composed of max pooling and residual block [46], as shown in Figure 5c. Among them, max pooling is used to halve the resolution of the input feature map, thus capturing multi-scale information. Note that we did not use the max pooling layer in the first stage of the encoder to ensure that the output image size is the same as the input image size. Residual block consists of a Conv 3×3 , of a batch normalization (BN) [47], of a rectified linear unit (ReLU) [48] can efficiently encode spatial local information, and solve the degradation problem of deep network through shortcut connections, thus reducing the difficulty of model optimization. The number L of residual blocks for each stage is set to {1, 1, 3, 1}. The calculation formula of residual block is as follows:

$$y = \sigma(\mathcal{F}(x, \{\text{Conv}_i\}) + x) \quad (1)$$

$$\mathcal{F} = \mathcal{B}(\text{Conv}_2(\sigma(\mathcal{B}(\text{Conv}_1 x)))) \quad (2)$$

where x and y are input and output vectors, the function \mathcal{F} represents the residual mapping to be learned, σ is the ReLU activation function, and \mathcal{B} denotes BN and the biases are omitted for simplifying notations.

2. Trans Block

Trans block is the basic module for extracting global features in CTAFK, which is composed of patch embedding and Transformer encoder blocks [29], as shown in Figure 5d. As with the Conv block, the number L of the Transformer encoder for each of the four stages is also set to $\{1, 1, 3, 1\}$.

Patch embedding can convert two-dimensional images into one-dimensional token sequences with the standard Transformer as input, as shown in Figure 6a, and the calculation formula is as follows:

$$z_0 = [x_p^1 \mathbf{E}; x_p^2 \mathbf{E}; \dots x_p^N \mathbf{E}], \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D} \quad (3)$$

where $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ is a sequence of flattened two-dimensional patches. $N = HW/P^2$ is the resulting number of patches, \mathbf{E} is a trainable linear projection, which can map the patches to D dimension, $P \times P$ is the resolution of image patches, and the P is set to $\{1, 2, 2, 2\}$ for each stage. The Transformer encoder consists of multi-headed self-attention (MSA), multilayer perceptron (MLP) block, layer normalization (LN) [49], and shortcut connections. This process is expressed by the following equation:

$$z'_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1}, \quad \ell = 1 \dots n \quad (4)$$

$$z_\ell = \text{MLP}(\text{LN}(z'_\ell)) + z'_\ell, \quad \ell = 1 \dots n \quad (5)$$

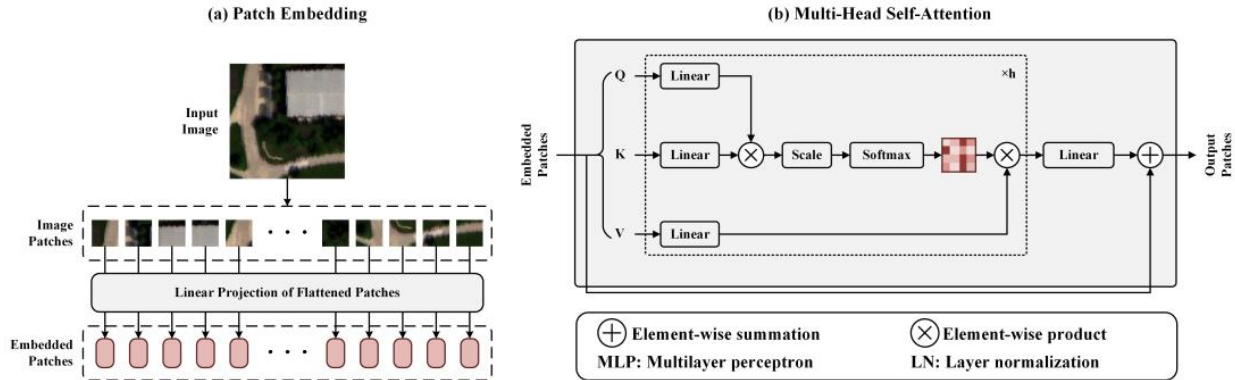


Figure 6. Patch embedding and multi-head self-attention mechanism.

As shown in Figure 6b, MSA mainly captures the correlation between input patches through the attention function. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query (Q), keys (K), values (V), and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. This process is expressed by this equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

MSA involves multiple groups of the weight matrix in mapping Q , K , and V , using the same operation process to calculate the attention value. MSA allows the model to jointly

attend to information from different representation subspaces at different positions. This process is expressed by this equation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (7)$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$, $W_i^V \in \mathbb{R}^{d \times d_v}$, and $W_i^O \in \mathbb{R}^{hd_v \times d}$ are parameter matrices.

MLP block is used to enhance the non-linear expression ability of Transformer and consists of two full connection layers. Between the two full connection layers, we use Gaussian error linear units (GELU) [50] as the activation functions.

3. CTAFK workflow

Overall, CTAFK deals with the input feature map $X \in \mathbb{R}^{H \times W \times C}$ via three steps, i.e., feature extraction and integration, adaptive weight calculation, and weighted fusion.

Feature extraction and integration: first, Conv block and Trans block are used to extract the local and global features of the given feature map X , respectively. The two transformations can be expressed as $\mathcal{F}_{\text{Conv}} : X \rightarrow \tilde{U} \in \mathbb{R}^{H \times W \times C}$ and $\mathcal{F}_{\text{Trans}} : X \rightarrow \hat{U} \in \mathbb{R}^{H \times W \times C}$. Then, a convolution layer composed of Conv 1×1 , BN and ReLU activation function is used to unify the number of channels of the two branches. Finally, integrate information from local feature branch and global feature branch via an element-wise addition:

$$U = \tilde{U} + \hat{U} \quad (8)$$

Adaptive weight calculation: the weights of local and global features are calculated through a squeeze-and-excitation module [51]. First, a statistic $s \in \mathbb{R}^C$ is generated by shrinking U through spatial dimensions $H \times W$:

$$s_c = \mathcal{F}_{\text{GAP}}(U_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W U_c(i, j) \quad (9)$$

where s_c denotes the c -th element of s . Then, a MLP is used to describe the dependency of the two branches as $g \in \mathbb{R}^{2C}$, so as to guide the adaptive fusion. The used MLP consists of two full connection layers, its calculation formula is as follows:

$$g = \mathcal{F}_{\text{MLP}}(s) = W_2(W_1s + b_1) + b_2 \quad (10)$$

where W_i and b_i ($i = 1, 2$) denote the weights and biases, respectively. Finally, g is divided into two vectors, i.e., $\tilde{g} \in \mathbb{R}^C$ and $\hat{g} \in \mathbb{R}^C$, and the weights are calculated by:

$$\tilde{w}_c = \frac{\tilde{g}_c}{\tilde{g}_c + \hat{g}_c} \quad (11)$$

$$\hat{w}_c = \frac{\hat{g}_c}{\tilde{g}_c + \hat{g}_c} \quad (12)$$

Weighted fusion: the fused feature map $V \in \mathbb{R}^{H \times W \times C}$ is obtained by weighting \tilde{U} and \hat{U} :

$$V_c = \tilde{w}_c \cdot \tilde{U}_c + \hat{w}_c \cdot \hat{U}_c \quad (13)$$

where $\tilde{w}_c + \hat{w}_c = 1$, V_c denotes the c -th channel of V , and $V_c \in \mathbb{R}^{H \times W}$.

In addition, the output of CTAFK was input to the CTAFK in the next stage to transfer the inductive bias of convolution and the learned location information.

2.2.3. Comparison Methods

Nine representative deep learning models were selected for comparison. Among them, four models are the most popular solutions for semantic segmentation tasks in

the computer vision domain, including UNet [45], Deeplab v3+ [52], SegFormer [53], and Swin-UNet [54]. The remaining five are the most advanced models in the current HSI classification tasks, including UNet-m-se(prelu)-gan [44] (hereinafter referred to as UNet-m), SS3FCN [42], ENLFCN [26], SSDGL [55], and FSSANet [31]. Among them, UNet and SS3FCN are methods based on two-dimensional and three-dimensional convolution respectively. Deeplab v3+ belongs to the method of directly expanding the receptive field of the convolution kernel, UNet-m, ENLFCN, and SSDGL belong to the methods that introduce the attention mechanism, and SegFormer, Swin-UNet, and FSSANet are methods based on Transformer. Please refer to the corresponding paper for details and parameter settings of the model.

3. Experiments and Analysis

3.1. Implementation Details and Metrics

During the training period, all models used the he-normal [56] method to initialize the weight parameters and the AdamW [57] optimizer to optimize the weight parameters. The initial learning rate is set to 0.001, and the weight decay is set to 0.0001. A weighted cross entropy loss function (Equation (14)) was used to handle the unspecific class in the dataset. Note that we used the same loss function and optimizer for all models to ensure a fair comparison. All models were implemented under the Pytorch 1.10 deep learning open-source framework using a NVIDIA GeForce RTX 3070 GPU with 8 GB memory.

$$\text{Loss} = - \sum_{k=1}^K \frac{t_0}{t_k} \times y_k \log(p_k) \quad (14)$$

where K is the number of classes, y and p are the real and the predicted classes, respectively, t_0 is the number of unspecified sample, and t_k is the sample size of a single class.

In this paper, the pixel accuracy, the intersection over union (IoU) of each class, and IoU averaged over all classes (mIoU) are selected as the evaluation metrics for the quantitative evaluation.

3.2. Experiment Result

3.2.1. Experiment Results on AeroRIT Dataset

Figure 7 shows the visually comparison of different methods on the AeroRIT test set. Overall, the classification map obtained through CTAFNet is highly consistent with the ground truth, and there are few misclassification phenomena. The classification map generated by CTAFNet is the smoothest, while ENLFCN and SS3FCN have more salt and pepper noise. As can be seen in the enlarged images, the CTAFNet correctly extracted the racetrack and effectively preserved shapes. The car boundaries obtained from CTAFNet and SS3FCN are clearer than that from the other methods, whereas the object edges generated by CTAFNet are the closest to the ground truth.

Table 5 shows the experimental results of different methods in the AeroRIT test set. The experimental results demonstrate that the proposed CTAFNet method achieves the best performance, with 95.07% pixel accuracy and 81.41% mIoU. Specifically, CTAFNet proposed in this paper is 0.36% higher with regards to pixel accuracy and 1.69% mIoU with regard to the second-best model SS3FCN, while SegFormer has the worst performance, with 88.51% pixel accuracy and 57.32% mIoU. CTAFNet achieved the best score in four classes of Buildings, Roads, Water and Cars, which is 1.10%, 0.71%, 1.53%, and 1.20% higher than the second place, respectively.

3.2.2. Experiment Results on DFC2018 Dataset

Figure 8 shows the visually comparison of different methods on the DFC2018 dataset. One can observe that CTAFNet obtained more accuracy classification map than the other methods. Moreover, the classification map generated by CTAFNet is the cleanest, with little

noise in buildings and roads, while the classification map generated by other methods has noticeable noises.

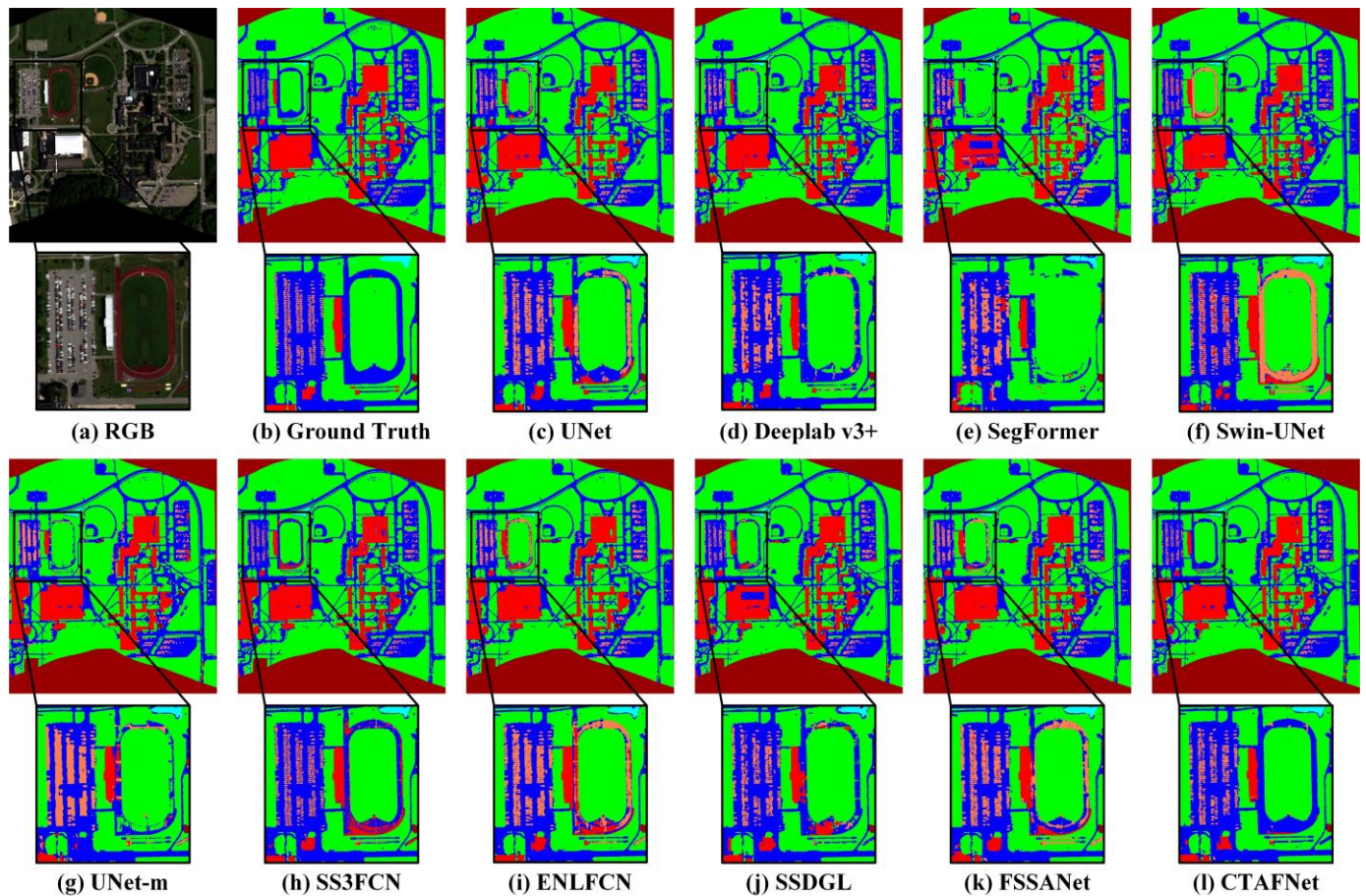


Figure 7. Visually comparison of different methods on the AeroRIT test set.

Table 5. Classification results for the AeroRIT test set.

Class	UNet	DeepLab v3+	SegFormer	Swin-UNet	UNet-m	SS3FCN	ENLFCN	SSDGL	FSSANet	CTAFNet
Buildings	85.47	85.53	63.47	82.89	82.64	83.86	85.17	81.66	84.93	86.63
Vegetation	94.87	94.99	93.09	95.94	95.45	95.71	95.25	95.72	95.8	95.72
Roads	81.17	82.09	66.95	81.07	80.04	83.93	80.77	82.01	81.62	84.64
Water	72.11	68.78	30.75	75.52	76.74	74.65	76.89	75.23	75.06	78.42
Cars	48.94	47.74	32.34	38.4	47.17	60.44	47.56	57.38	45.84	61.64
pixel acc.	93.86	94.08	88.51	93.72	93.61	94.71	93.88	94.2	94.09	95.07
mIoU	76.52	75.83	57.32	74.76	76.41	79.72	77.13	78.4	76.65	81.41

Table 6 shows the experimental results of different methods in the DFC2018 test set. The experimental results demonstrate that the proposed CTAFNet method achieves the best performance, with 92.59% pixel accuracy and 82.10% mIoU. Specifically, the pixel accuracy and mIoU of CTAFNet is 1.47% and 2.50% higher than the second-best model SS3FCN, respectively. SegFormer has the worst performance, achieving only 73.44% pixel accuracy and 58.11% mIoU. In terms of classification accuracy of each class, CTAFNet obtained the best score in 12 of the 17 classes. Among them, IoU in residential buildings, major thoroughfare, and cars was far higher than the second place (7.46%, 10.93% and 8.57% higher, respectively).

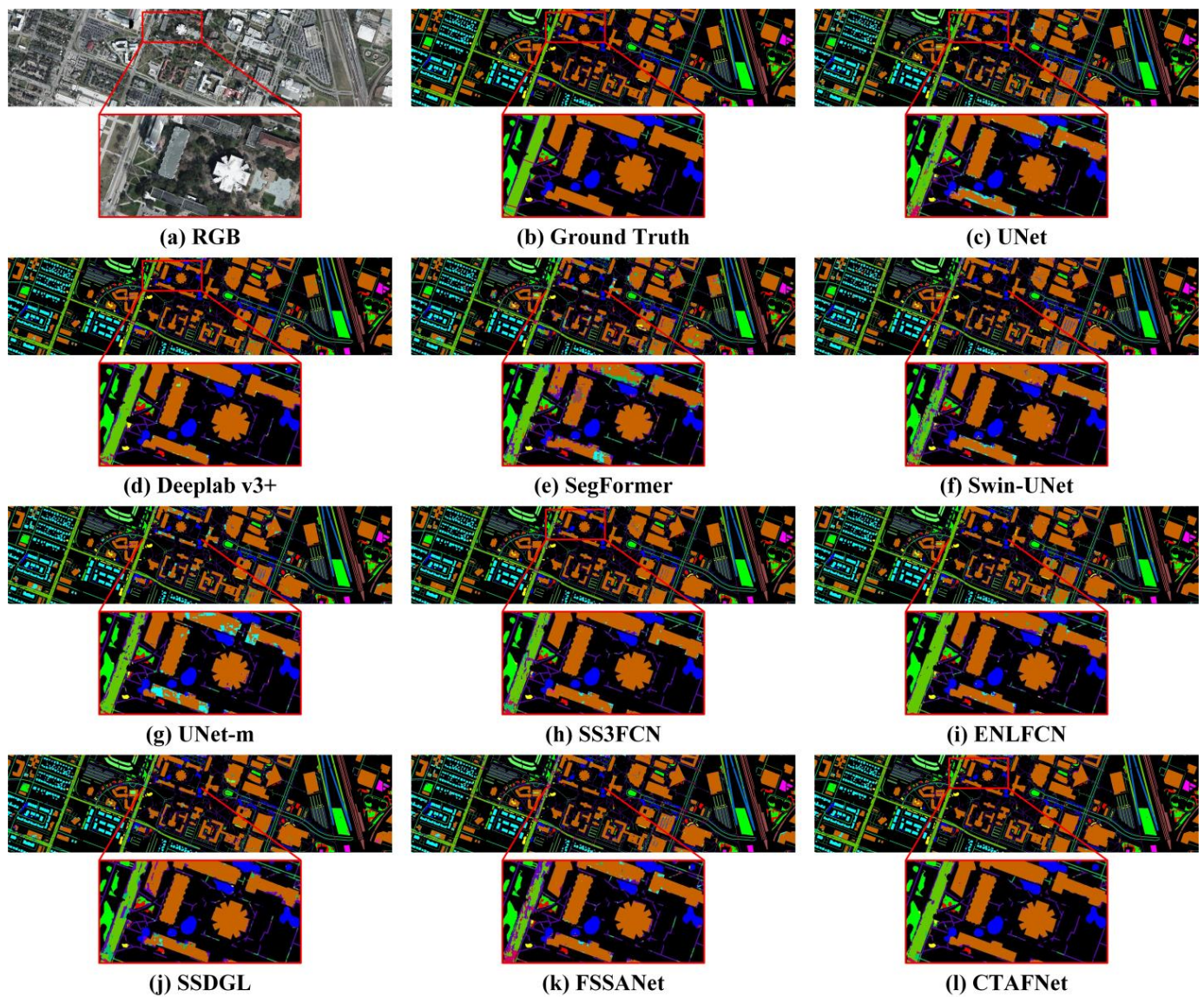


Figure 8. Visually comparison of different methods on the DFC2018 dataset.

Table 6. Classification results for the DFC2018 test set.

Class	UNet	DeepLab v3+	SegFormer	Swin-UNet	UNet-m	SS3FCN	ENLFCN	SSDGL	FSSANet	CTAFFNet
Healthy grass	78.88	74.67	64.39	87.59	82.68	87.01	80.77	84.56	85.36	77.79
Stressed grass	86.36	83.94	69.95	88.68	88.92	90.34	86.86	85.30	87.41	88.08
Evergreen trees	81.53	84.70	78.05	81.47	83.75	87.67	82.83	85.48	80.27	89.47
Deciduous trees	61.14	68.57	36.77	68.18	65.83	75.80	60.05	51.59	61.20	77.39
Bare earth	84.67	93.53	71.37	82.90	86.68	85.84	88.95	70.52	89.12	93.53
Residential buildings	66.30	83.75	64.26	70.33	70.05	81.89	79.40	78.61	72.18	91.21
Non-residential buildings	86.99	89.68	69.82	86.89	86.82	93.09	88.03	85.80	89.92	93.69
Roads	55.15	57.47	32.47	50.39	55.79	65.72	51.22	48.82	53.57	69.71
Sidewalks	50.69	55.47	34.60	47.74	56.89	64.23	53.96	43.93	53.19	62.98
Crosswalks	2.61	12.13	4.20	14.15	23.34	29.23	17.57	8.84	11.31	14.00
Major thoroughfares	38.41	72.58	48.41	47.83	69.96	72.84	66.66	61.08	53.62	83.77
Highways	76.58	80.54	75.60	38.04	86.48	76.20	74.53	82.80	53.67	90.06
Railways	79.06	90.99	59.26	94.60	92.41	96.99	92.36	98.96	93.89	97.17
Paved parking lots	67.52	88.01	69.21	81.18	92.97	92.85	92.46	84.20	91.47	95.96
Cars	69.57	79.15	54.47	54.41	69.04	63.23	84.98	81.62	65.22	93.56
Trains	96.19	94.57	90.42	72.24	96.13	94.80	95.83	98.43	88.52	99.16
Stadium seats	57.83	85.46	64.68	82.01	93.90	95.40	88.87	62.24	81.17	96.52
pixel acc.	84.84	88.66	73.44	83.56	87.56	91.12	87.24	84.72	85.97	92.74
mIoU	68.29	76.19	58.11	67.57	76.57	79.60	75.61	71.34	71.24	83.18

3.2.3. Experiment Results on Xiongan Dataset

Figure 9 shows the visually comparison of different methods on the Xiongan dataset. It can be found that the classification map obtained by CTAFNet keeps highly consistent with the ground truth, and there is little noise in the classification map. Compared with CTAFNet, the other three transformer models (SegFormer, Swin-UNet and FSSANet) produce more noises and misclassifications. As can be seen in the enlarged area, CTAFNet can completely classify the Grass (pink block in the middle) without any omissions or errors, which indicates that CTAFNet can effectively integrate the advantages of CNN and Transformer.

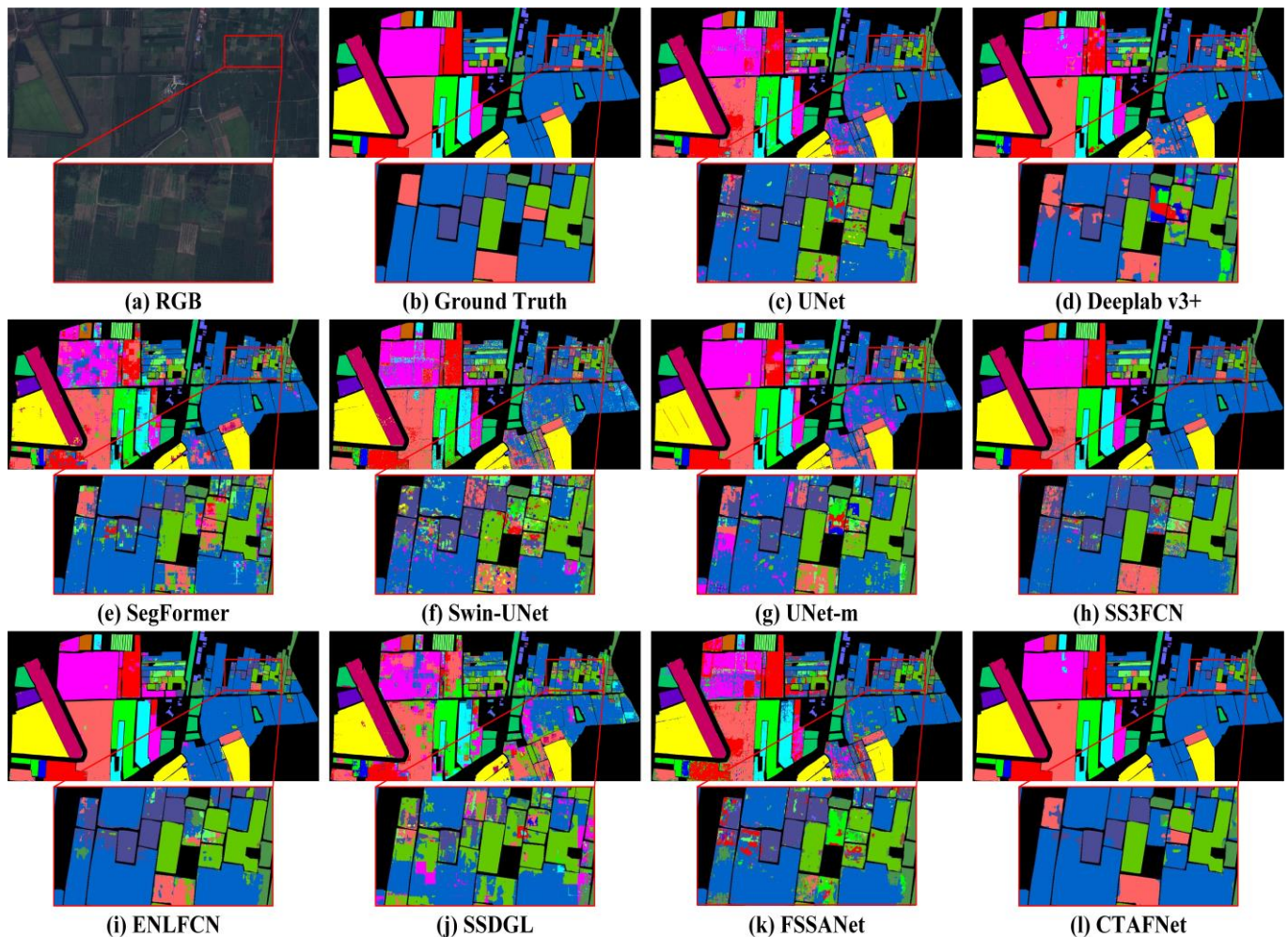


Figure 9. Visually comparison of different methods on the Xiongan dataset.

Table 7 shows the experimental results of different methods in the Xiongan test set. The experimental results demonstrate that the proposed CTAFNet achieves the best performance with 96.17% pixel accuracy and 86.84% mIoU. Specifically, the pixel accuracy and mIoU of CTAFNet is 1.80% and 3.98% higher than the second-best model ENLFCN, respectively. SegFormer has the worst performance, achieving only 66.58% pixel accuracy and 42.37% mIoU. In terms of classification accuracy of each class, CTAFNet obtained the best score in 10 of the 18 classes. Among them, its IoU scores on the 3th, 12th, 13th, 16th, and 17th categories of ground objects far exceeded values for the second-best model, which were 6.43%, 6.93%, 11.04%, 4.29% and 14.70% higher, respectively.

Table 7. Classification results for the Xiongan test set.

Class	UNet	DeepLab v3+	SegFormer	Swin-UNet	UNet-m	SS3FCN	ENLFCN	SSDGL	FSSANet	CTAFNet
Acer negundo	68.54	52.18	14.18	63.24	72.43	85.71	88.25	20.27	49.43	89.29
Willow	85.62	77.04	56.18	70.77	85.10	91.62	98.05	43.78	72.67	97.74
Elm	79.88	57.30	0.00	62.33	59.82	82.54	86.08	16.71	45.42	92.53
Paddy	94.82	97.62	89.89	93.76	97.05	98.68	98.95	95.53	95.76	99.20
Sophora japonica	73.32	78.77	35.37	70.00	70.31	91.88	90.62	47.58	52.18	91.31
Fraxinus chinensis	80.43	77.90	34.64	49.20	77.37	91.14	93.15	41.71	74.34	85.19
Goldenrain tree	99.59	93.01	7.80	87.24	74.96	97.43	99.86	50.12	97.93	99.52
Waters	96.95	94.97	89.31	88.36	95.54	94.83	94.14	87.16	91.59	95.11
Bare ground	93.60	96.37	76.12	93.02	91.82	98.88	95.36	85.76	85.60	94.39
Stubble	87.53	98.01	82.62	95.13	97.72	99.56	98.17	88.09	97.36	99.96
Corn	58.02	56.89	21.06	34.57	56.34	65.95	77.61	8.97	61.65	80.21
Pyrus	76.00	80.09	58.47	61.74	68.51	86.90	86.94	48.97	65.91	93.86
Soybean	8.52	23.32	0.00	13.01	9.57	19.71	9.85	4.91	9.50	34.36
Poplar	74.69	71.24	28.68	59.29	69.95	72.20	70.30	28.48	48.72	78.53
Vegetable field	31.12	32.62	12.56	15.03	54.24	36.28	44.92	21.25	25.99	53.72
Grass	73.27	68.16	40.42	61.54	67.21	84.69	88.61	39.88	67.36	92.90
Peach	78.26	78.22	39.09	51.83	81.30	80.58	79.71	43.56	49.74	95.99
Building	83.70	83.49	76.26	60.84	88.33	88.25	90.83	67.33	64.61	89.38
pixel acc.	87.29	87.52	66.58	79.63	86.17	93.79	94.37	64.91	79.38	96.17
mIoU	74.66	73.18	42.37	62.83	73.20	81.49	82.86	46.67	64.21	86.84

4. Discussion

4.1. Ablation Study

In order to verify the effectiveness of the proposed adaptive hybrid strategy, we conducted ablation experiments on CTAFK. Models with different hybrid strategies were used as backbone networks in CTAFNet. The experimental results on three datasets are shown in Table 8. Among them, CCTT means two layers of convolution and two layers of Transformer encoder stacking. Add represents the direct element-wise addition of local and global feature maps. Cat represents concatenate of local and global feature maps along the spectral axis. Adapt means the adaptive hybrid strategy proposed in this paper. Comparing the experimental results, it can be found that the pure CNN model has achieved much better performance than the pure Transformer model, and a possible explanation is that Transformer model requires more training data, and the data size of the current HSI open-source dataset is not enough to train a pure Transformer model with good performance [33]. When using the CNN–Transformer hybrid architecture, selecting an unsuitable hybrid strategy may cause performance degradation. For example, the hybrid model with the add strategy performs worse than pure CNN model on the three datasets, the hybrid model with the CCTT strategy performs poorly on the aerial dataset and DFC2018 dataset, and the hybrid model with the cat strategy performs worse than pure CNN model on the DFC2018 dataset. The reason for this phenomenon might be that the attention mechanism has mixed the intra-class and inter-class contexts when extracting the global information [58], and the artificially specified hybrid strategies cannot distinguish between different contexts, which leads to poor generalization ability of the model. By contrast, the proposed adaptive hybrid strategy can enhance the useful context information by adjusting the weight, thus extracting more generalized and discriminative features and overcoming the limitations of artificially specified hybrid strategies. The model using adaptive hybrid strategy achieves the best classification accuracy on three datasets, which proves the effectiveness of the proposed adaptive hybrid strategy.

4.2. Parameter Sensitivity Analysis

CTAFNet contains two hyperparameters, i.e., the head numbers in MSA and the number of channels. To understand the sensitivity of CTAFNet to the hyperparameters, we tested the performance of different configurations in HSI classification. Table 9 shows the impact of head numbers in MSA on CTAFNet performance. With the increase in head numbers, the two performance indicators increase first and then decrease. This tendency is consistent across datasets. Table 10 shows the impact of different channel numbers on the performance of CTAFNet. The results show that, with the increase in channel numbers, the performance of the model also increases first and then decreases. This is possibly because

the complexity of model increases together with the head numbers in MSA and the number of channels, and overfitting problem may occur when the model becomes too complex. When the head numbers of MSA is set to 2, and the model channel number is set to 96, CTAFFNet performs best on the three datasets. Since the optimal hyperparameters are constants and no additional parameter tuning is needed, and the proposed CTAFFNet is highly applicable in different scenes.

Table 8. Backbone network ablation experiment.

Conv	Trans	Hybrid Strategy	Aerial		DFC2018		Xiongan	
			Pixel Acc.	mIoU	Pixel Acc.	mIoU	Pixel Acc.	mIoU
✓		None	94.90	76.54	92.31	81.40	87.86	76.66
	✓	None	93.12	72.70	75.59	56.34	61.27	40.61
✓	✓	CCTT	94.02	75.97	88.46	76.41	90.30	76.75
✓	✓	Add	94.25	76.15	89.45	76.46	88.62	71.32
✓	✓	Cat	94.69	77.79	91.12	80.62	92.36	79.05
✓	✓	Adapt	95.07	81.41	92.74	83.18	96.17	86.84

Table 9. Influence of the number of heads in MSA on the performance of CTAFFNet.

Number of Heads	Aerial		DFC2018		Xiongan	
	Pixel Acc.	mIoU	Pixel Acc.	mIoU	Pixel Acc.	mIoU
1	94.51	77.98	92.08	82.37	94.05	81.79
2	95.07	81.41	92.74	83.18	96.17	86.84
4	94.83	79.25	91.76	82.00	94.49	81.35
8	94.94	80.22	91.96	81.83	93.95	80.20

Table 10. Influence of number of channels on the performance of CTAFFNet.

Number of Channels	Aerial		DFC2018		Xiongan	
	Pixel Acc.	mIoU	Pixel Acc.	mIoU	Pixel Acc.	mIoU
32	94.91	78.02	90.37	78.30	89.80	74.74
64	94.53	77.95	91.81	81.25	93.76	75.97
96	95.07	81.41	92.74	83.18	96.17	86.84
128	94.84	78.78	91.75	82.21	94.19	82.10

4.3. Generalization of CTAFFNet on Small Dataset

In order to verify the generalization of CTAFFNet on small datasets, this section compares CTAFFNet with nine methods on the widely used the Salinas dataset, and the results are shown in Table 11. The results demonstrate that the proposed CTAFFNet method achieves the best performance, with 94.62% pixel accuracy and 95.70% mIoU, indicating that CTAFFNet generalizes well to small datasets. In addition, Deeplab v3+, which performed well on the three large datasets, performed poorly on the Salinas dataset. It may be because the Deeplab v3+ has a large number of parameters, and overfitting occurs when there are fewer training samples.

4.4. Limitations and Future Works

Although the proposed CTAFFNet outperforms the other models, there are still some limitations to overcome in the future. For example, the boundary of ground objects generated by this method is not as clear as that generated by SS3FCN. This is possibly because CTAFFNet uses max pooling layer to reduce the calculation of the model, resulting in the loss of boundary information. By comparison, SS3FCN does not involve any downsampling operation, which helps preserve the shapes and boundaries. Current research in the field of deep learning shows that introducing post-processing (such as conditional random

fields [59]) or improving loss functions (such as using Hausdorff distance loss [60]) can help solve this problem. It is of interest to further improve the CTAFFNet through the above-mentioned methods to better preserve the boundaries.

Table 11. Comparison of generalization performance on the Salinas dataset.

Class	UNet	DeepLab v3+	SegFormer	Swin-UNet	UNet-m	SS3FCN	ENLFCN	SSDGL	FSSANet	CTAFNet
Broccoli green weeds 1	99.14	0.00	16.63	77.59	97.06	100.00	100.00	9.49	100.00	100.00
Broccoli green weeds 2	100.00	74.01	1.33	89.49	99.87	100.00	100.00	4.20	100.00	100.00
Fallow	95.41	11.29	39.90	82.89	100.00	85.28	85.04	17.39	83.23	100.00
Fallow rough plow	86.45	73.03	72.03	98.29	95.02	98.97	91.11	89.13	97.92	98.62
Fallow smooth	95.22	95.72	94.34	94.18	99.48	99.27	98.33	99.58	98.35	99.69
Stubble	100.00	67.84	64.73	94.63	99.56	99.85	100.00	94.87	96.76	100.00
Celery	99.71	99.86	7.71	97.29	100.00	99.28	100.00	23.18	97.57	100.00
Grapes untrained	38.36	37.28	40.66	61.13	65.59	72.46	75.94	18.15	66.20	75.06
Soil vineyard develop	98.82	83.38	83.81	99.88	95.67	99.71	99.88	97.85	97.76	99.51
Corn senesced green weeds	67.75	50.51	51.65	62.56	80.10	79.92	83.03	45.93	81.58	98.66
Lettuce romaine 4 wk	87.41	24.68	38.46	69.05	92.38	86.62	85.62	52.24	84.62	100.00
Lettuce romaine 5 wk	96.45	5.80	13.77	83.41	85.33	97.24	98.60	51.62	98.24	100.00
Lettuce romaine 6 wk	93.64	0.00	4.37	80.88	93.16	89.52	96.09	87.70	77.00	100.00
Lettuce romaine 7 wk	88.19	64.58	70.08	75.21	87.26	77.41	83.72	79.46	69.49	99.13
Vineyard untrained	56.43	0.00	31.97	64.59	67.97	66.49	67.11	17.89	54.25	60.48
Vineyard vertical trellis	18.59	0.00	0.00	82.10	0.00	91.86	88.28	0.00	78.49	100.00
pixel acc.	85.70	63.44	61.24	88.72	90.99	92.95	93.55	59.84	90.44	94.62
mIoU	82.60	43.00	39.47	82.07	84.90	90.24	90.80	49.29	86.34	95.70

Although the proposed CTAFFNet achieves the best overall accuracy, the classification results are relatively poor on the categories with a small number of samples (for example, vegetation in the aerial dataset, crosswalks in the DFC2018 dataset, and *Fraxinus chinensis* in the Xiongan dataset). A possible reason is that the objective of model training is to optimize the overall accuracy, rather than the accuracy of a specific class. Therefore, the trained model will bias towards the classes with large numbers of samples and away from and the classes with small numbers of samples. Future works will be focused on addressing the problem of sample imbalance through data enhancement and dynamic weighting in loss function [61].

In this paper, we compared the HSI classification accuracy of nine state-of-the-art models on three large-scale and challenging datasets to provide an insight into their performance. In addition to these models, many deep learning-based HSI classification methods have been developed in recent years [17–20,23–25,27,28,30,35–37]. It is of interest to conduct a cross-comparison to guide users to select the most suitable methods in specific applications and assist scholars in designing advanced models. Such a comparison requires large number of datasets with challenging cases in HSI classification, e.g., images with fine-grained object, sample imbalance, inter-class variation, and cloud contamination. However, due to the high cost of HSI acquisition and labeling [62], the amount of samples in the used datasets is still small as compared with the datasets in the field of computer vision [63,64]. A recent study has developed a high-quality HSI classification benchmark dataset [65], which provides the starting point to construct a standard dataset. We plan to evaluate more HSI classification models after this dataset is released for public use so as to further understand the comprehensive performance differences between models.

5. Conclusions

In this article, a novel CTAFFNet is proposed for HSI classification. We designed a CTAFF module to capture the local high-frequency features using a convolution module and extracts the global and sequential low-frequency information using a Transformer module. Afterwards, the local high-frequency and global low-frequency features are adaptively weighted and fused to provide a more robust and discriminative representation of the HSI data. An encoder–decoder structure was adopted in the CTAFFNet to improve the flow of fused local-global information between different stages, thus ensuring the generalization ability of the model. Experimental results conducted on three large-scale challenging HSI datasets demonstrate that the proposed network is superior to nine state-of-the-art approaches. The developed adaptive feature fusion strategy can effectively overcome the

limitations of the existing hybrid strategy and improve the accuracy of HSI classification. Our research provides promising methods for HSI classification and keen insights into the comprehensive performance differences between models.

Author Contributions: Conceptualization, J.L., H.X. and H.W.; methodology, J.L.; data curation, J.L.; writing—original draft preparation, J.L. and Z.A.; writing—review and editing, J.L., Z.A. and W.L.; funding acquisition, H.X. and A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Nos. 41971406, 42071246, 42271470), the Natural Science Foundation of Hebei Province, China (No. D2021402007) and the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515011586).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The AeroRIT dataset utilized in this study are freely available at <https://github.com/aneesh3108/AeroRIT> (accessed on 25 October 2022). The DFC2018 dataset utilized in this study is freely available at https://hyperspectral.ee.uh.edu/?page_id=1075 (accessed on 25 October 2022). The Xiong'an dataset utilized in this study is freely available at <http://www.hrs-cas.com/a/share/shujuchanpin/2019/0501/1049.html> (accessed on 25 October 2022). The Salinas dataset utilized in this study is freely available at http://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 25 October 2022).

Acknowledgments: The authors thank the Hyperspectral Image Analysis Laboratory at the University of Houston, as well as the IEEE GRSS Image Analysis and Data Fusion Technical Committee for acquiring and providing the HSI data used in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Li, X.; Li, Z.; Qiu, H.; Hou, G.; Fan, P. An Overview of Hyperspectral Image Feature Extraction, Classification Methods and The Methods Based on Small Samples. *Appl. Spectrosc. Rev.* **2021**, *11*, 1–34. [CrossRef]
- Khan, M.J.; Khan, H.S.; Yousaf, A.; Khurshid, K.; Abbas, A. Modern Trends in Hyperspectral Image Analysis: A Review. *IEEE Access* **2018**, *6*, 14118–14129. [CrossRef]
- Adão, T.; Hruška, J.; Pádua, L.; Bessa, J.; Peres, E.; Morais, R.; Sousa, J.J. Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sens.* **2017**, *9*, 1110. [CrossRef]
- Krupnik, D.; Khan, S. Close-Range, Ground-Based Hyperspectral Imaging for Mining Applications at Various Scales: Review and Case Studies. *Earth-Sci. Rev.* **2019**, *198*, 102952. [CrossRef]
- Liu, B.; Liu, Z.; Men, S.; Li, Y.; Ding, Z.; He, J.; Zhao, Z. Underwater Hyperspectral Imaging Technology and Its Applications for Detecting and Mapping the Seafloor: A Review. *Sensors* **2020**, *20*, 4962. [CrossRef]
- Chen, Y.; Xing, Z.; Jia, X. Spectral-Spatial Classification of Hyperspectral Data Based on Deep Belief Network. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.* **2015**, *8*, 2381–2392. [CrossRef]
- Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
- Prasad, S.; Bruce, L.M. Limitations of Principal Components Analysis for Hyperspectral Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 625–629. [CrossRef]
- Li, W.; Prasad, S.; Fowler, J.E.; Bruce, L.M. Locality-Preserving Dimensionality Reduction and Classification for Hyperspectral Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1185–1198. [CrossRef]
- Liao, W.; Pizurica, A.; Scheunders, P.; Philips, W.; Pi, Y. Semisupervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 184–198. [CrossRef]
- Al-khafaji, S.L.; Zhou, J.; Zia, A.; Liew, A.W. Spectral-Spatial Scale Invariant Feature Transform for Hyperspectral Images. *IEEE Trans. Image Process.* **2018**, *27*, 837–850. [CrossRef] [PubMed]
- Li, W.; Chen, C.; Su, H.; Du, Q. Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3681–3693. [CrossRef]
- Gu, Y.; Liu, H. Sample-Screening MKL Method via Boosting Strategy for Hyperspectral Image Classification. *Neurocomputing* **2016**, *173*, 1630–1639. [CrossRef]
- Melgani, F.; Bruzzone, L. Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]

15. Ham, J.; Chen, Y.; Crawford, M.M.; Ghosh, J. Investigation of the Random Forest Framework for Classification of Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501. [\[CrossRef\]](#)
16. Zhou, Y.; Peng, J.; Chen, C.L.P. Extreme Learning Machine with Composite Kernels for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2351–2360. [\[CrossRef\]](#)
17. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [\[CrossRef\]](#)
18. Zhao, W.; Du, S. Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4544–4554. [\[CrossRef\]](#)
19. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [\[CrossRef\]](#)
20. Zhang, H.; Li, Y.; Zhang, Y.; Shen, Q. Spectral-Spatial Classification of Hyperspectral Imagery Using a Dual-Channel Convolutional Neural Network. *Remote Sens. Lett.* **2017**, *8*, 438–447. [\[CrossRef\]](#)
21. Li, J.S.; Xia, X.; Li, W. Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios. *arXiv* **2022**, arXiv:2207.05501v2.
22. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8681–8691. [\[CrossRef\]](#)
23. Zhao, F.; Zhang, J.; Meng, Z.; Liu, H. Densely Connected Pyramidal Dilated Convolutional Network for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 3396. [\[CrossRef\]](#)
24. Liu, D.; Han, G.; Liu, P.; Yang, H.; Sun, X.; Li, Q.; Wu, J. A Novel 2D-3D CNN with Spectral-Spatial Multi-Scale Feature Fusion for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 4621. [\[CrossRef\]](#)
25. Fang, B.; Li, Y.; Zhang, H.; Chan, J.C.-W. Hyperspectral Images Classification Based on Dense Convolutional Networks with Spectral-Wise Attention Mechanism. *Remote Sens.* **2019**, *11*, 159. [\[CrossRef\]](#)
26. Shen, Y.; Zhu, S.; Chen, C.; Du, Q.; Xiao, L.; Chen, J.; Pan, D. Efficient Deep Learning of Nonlocal Features for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6029–6043. [\[CrossRef\]](#)
27. Wang, L.; Peng, J.; Sun, W. Spatial-Spectral Squeeze-and-Excitation Residual Network for Hyperspectral Image Classification. *Remote Sens.* **2019**, *11*, 884. [\[CrossRef\]](#)
28. Xue, Z.H.; Xu, Q.; Zhang, M.X. Local Transformer with Spatial Partition Restore for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4307–4325. [\[CrossRef\]](#)
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. *arXiv* **2017**, arXiv:1706.03762.
30. Qing, Y.; Liu, W.; Feng, L.; Gao, W. Improved Transformer Net for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 2216. [\[CrossRef\]](#)
31. Sun, J.; Zhang, J.; Gao, X.; Wang, M.; Ou, D.; Wu, X.; Zhang, D. Fusing Spatial Attention with Spectral-Channel Attention Mechanism for Hyperspectral Image Classification via Encoder-Decoder Networks. *Remote Sens.* **2022**, *14*, 1968. [\[CrossRef\]](#)
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 12 October 2021; pp. 10012–10022. [\[CrossRef\]](#)
33. D’Ascoli, S.; Touvron, H.; Leavitt, M.; Morcos, A.; Biroli, G.; Sagun, L. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. *arXiv* **2021**, arXiv:2103.10697. [\[CrossRef\]](#)
34. Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in Transformer. *arXiv* **2021**, arXiv:2103.00112.
35. Yang, L.; Yang, Y.; Yang, J.; Zhao, N.; Wu, L.; Wang, L.; Wang, T. FusionNet: A Convolution-Transformer Fusion Network for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 4066. [\[CrossRef\]](#)
36. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral-Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [\[CrossRef\]](#)
37. Song, R.; Feng, Y.; Cheng, W.; Mu, Z.; Wang, X. BS2T: Bottleneck Spatial-Spectral Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5532117. [\[CrossRef\]](#)
38. Park, N.; Kim, S. How Do Vision Transformers Work? *arXiv* **2022**, arXiv:2202.06709.
39. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#)
40. Nalepa, J.; Myller, M.; Kawulok, M. Validating Hyperspectral Image Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1264–1268. [\[CrossRef\]](#)
41. Liang, J.; Zhou, J.; Qian, Y.; Wen, L.; Bai, X.; Gao, Y. On the Sampling Strategy for Evaluation of Spectral-Spatial Methods in Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 862–880. [\[CrossRef\]](#)
42. Zou, L.; Zhu, X.; Wu, C.; Liu, Y.; Qu, L. Spectral-Spatial Exploration for Hyperspectral Image Classification via the Fusion of Fully Convolutional Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 659–674. [\[CrossRef\]](#)
43. Rangnekar, A.; Mokashi, N.; Ientilucci, E.J.; Kanan, C.; Hoffman, M.J. AeroRIT: A New Scene for Hyperspectral Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8116–8124. [\[CrossRef\]](#)
44. Cen, Y.; Zhang, L.; Zhang, X.; Wang, Y.; Qi, W.; Tang, S.; Zhang, P. Aerial Hyperspectral Remote Sensing Classification Dataset of Xiongan New Area (Matiwan Village). *J. Remote Sens.* **2020**, *24*, 10–17. [\[CrossRef\]](#)

45. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Lecture Notes in Computer Science in Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [\[CrossRef\]](#)
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; pp. 770–778. [\[CrossRef\]](#)
47. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015*; pp. 448–456. [\[CrossRef\]](#)
48. Glorot, X.; Bordes, A.; Bengio, Y.S. Deep Sparse Rectifier Neural Networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 11–13 April 2011*; pp. 315–323. Available online: <http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf> (accessed on 25 October 2022).
49. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
50. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
51. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018*; pp. 7132–7141. [\[CrossRef\]](#)
52. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611v3.
53. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203v2.
54. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.
55. Zhu, Q.; Deng, W.; Zheng, Z.; Zhong, Y.; Guan, Q.; Lin, W.; Zhang, L.; Li, D. A Spectral-Spatial-Dependent Global Learning Framework for Insufficient and Imbalanced Hyperspectral Image Classification. *IEEE Trans. Cybern.* **2021**, *52*, 11709–11723. [\[CrossRef\]](#)
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015*; pp. 1026–1034. [\[CrossRef\]](#)
57. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.
58. Yu, C.; Wang, J.; Gao, C.; Yu, G.; Shen, C.; Sang, N. Context Prior for Scene Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; pp. 12413–12422. [\[CrossRef\]](#)
59. Sun, Z.; Liu, M.; Liu, P.; Li, J.; Yu, T.; Gu, X.; Yang, J.; Mi, X.; Cao, W.; Zhang, Z. SAR Image Classification Using Fully Connected Conditional Random Fields Combined with Deep Learning and Superpixel Boundary Constraint. *Remote Sens.* **2021**, *13*, 271. [\[CrossRef\]](#)
60. Karimi, D.; Salcudean, S.E. Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. *IEEE Trans. Med. Imaging* **2020**, *39*, 499–513. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Sinha, S.; Ohashi, H.; Nakamura, K. Class-Wise Difficulty-Balanced Loss for Solving Class-Imbalance. In *Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December 2020*; pp. 549–565. [\[CrossRef\]](#)
62. Wambugua, N.; Chen, Y.; Xiao, Z.; Tan, K.; Wei, M.; Liu, X.; Li, J. Hyperspectral Image Classification on Insufficient-Sample and Feature Learning Using Deep Neural Networks: A Review. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102603. [\[CrossRef\]](#)
63. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; pp. 3213–3223. [\[CrossRef\]](#)
64. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; Yuille, A.L. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014*; pp. 891–898. [\[CrossRef\]](#)
65. Xu, Y.; Gong, J.; Huang, X.; Hu, X.; Li, J.; Li, Q.; Peng, M. Luojia-HSSR: A High Spatial-Spectral Resolution Remote Sensing Dataset for Land-Cover Classification with a New 3D-HRNet. *Geo-Spat. Inf. Sci.* **2022**. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.