



A global-local feature adaptive fusion network for image scene classification

Guangrui Lv¹ · Lili Dong¹ · Wenwen Zhang¹ · Wenhai Xu¹

Received: 8 September 2021 / Revised: 29 June 2022 / Accepted: 19 April 2023 /
Published online: 10 June 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Convolutional neural networks (CNN) have been widely used in image scene classification and have achieved remarkable progress. However, because the extracted deep features can neither focus on the local semantics of the image, nor capture the spatial morphological variation of the image, it is not appropriate to directly use CNN to generate the distinguishable feature representations. To relieve this limitation, a global-local feature adaptive fusion (GLFAF) network is proposed. The GLFAF framework extracts multi-scale and multi-level features by using a designed CNN. Then, to leverage the complementary advantages of the multi-scale and multi-level features, we design a global feature aggregate module to discover global attention features and further learn the multiple deep dependencies of spatial scale variations among these global features. Meanwhile, a local feature aggregate module is designed to aggregate the multi-scale and multi-level features. Specially, multi-level features at the same scale are fused based on channel attention, and then spatial fused features at different scales are aggregated based on channel dependence. Moreover, spatial contextual attention is designed to refine spatial features across scales and different fisher vector layers are designed to learn semantic aggregation among spatial features. Subsequently, two different feature adaptive fusion modules are introduced to explore the complementary associations of global and local aggregate features, which can obtain comprehensive and differentiated image scene presentation. Finally, a large number of experiments on real scene datasets coming from three different fields show that the proposed GLFAF approach can more accurately realize scene classification than other state-of-the-art models.

Keywords Scene classification · Multiple deep dependencies · Spatial contextual attention · Semantic aggregation · Feature adaptive fusion

1 Introduction

Scene classification has been extensively studied in the field of computer vision due to the increasing demand of scene-centric technologies. Scene classification can help people

✉ Lili Dong
donglili@dlnu.edu.cn

¹ College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

understand the content of images, which has brought great convenience to people's lives in many applications such as smart cities, autonomous driving, video surveillance, and remote sensing detection. However, there are complex intra-class differences and inter-class similarities of objects in actual scenes, which increases the difficulty of information integration and logical reasoning in each scene. Therefore, how to extract semantic cues for image feature representation is still a challenging research topic in the field of scene classification.

To classify a scene image, the image is first characterized by a feature encoder and then can be classified by a classifier [11, 43, 65]. Generally, there are inconsistencies and differences between the information extracted from visual data and people's comprehension of the same data in a given situation, which leads to the semantic gap between feature representation and high-level understanding [48]. A lot of work has been conducted to improve the representation ability of images. Among them, constructing scene feature representations with stronger descriptive capabilities is the most critical step to bridge the semantic gap between high-level scene understanding and low-level visual attributes [67]. Early, the traditional methods mainly focus on the research of hand-crafted visual features. The descriptors such as scale invariant feature transform (SIFT) [29], speed up robust features (SURF) [3], oriented brief (ORB) [37] are mainly used to extract visual feature points from the image, and then these features are input into the classifier for classification. Subsequently, researchers have further abstracted local visual features using bag of visual words(BOVW) [69], vector of locally aggregated descriptors (VLAD) [23], or latent dirichlet allocation (LDA) [8] for image classification. However, the abovementioned features are easy to understand and implement, but are highly dependent on the prior knowledge of the designer. Therefore, these features have low semantic level and limited representation ability, making it difficult to effectively describe the high-level semantic information of complex images.

Recently, there are many deep feature-based methods trying to apply excellent CNNs to build an effective feature representation for image scene classification [17, 44, 50]. Since the CNN is a notably hierarchical network structure, there are two types of features in the CNN: convolutional features and fully connected(FC) features. Among them, convolutional features that contain different spatial structure information of image, while FC features that contain abstract semantic information. In general, it's believed that the shallow features of CNN are closer to the texture information of the image, while the middle-level and high-level features are more inclined to the semantic information of the image [6], that is, higher-level features are more discriminative in semantics. Moreover, recent works demonstrate that aggregating the intermediate features (convolutional features and FC features) [30, 42, 51] and integrating diverse features [21, 41, 71] can significantly improve the classification performance of scene image. However, these related studies usually aggregate different features directly and do not explore their multiple inter-dependencies among different features. Although different feature fusions can improve the discriminative ability of feature representations, most of which focus on global information exploration while ignoring some key local knowledge.

The local information representation is of great importance to image scene classification. Some attention mechanisms [5–7, 26, 31, 57] have been introduced into CNN architecture to consider the relative importance of different channels and spatial regions in the feature maps for improving the performance of scene classification. Different from CNNs that only explores the relationships between neighboring pixels, transformer models can directly capture the long-range correlation of the local information by the use of a self-attention mechanism [15, 28], which provides a new approach for image scene classification. In addition, multi-modal/view data objects have attracted substantial research attention because of

their heterogeneous and complementary feature spaces [53, 55, 75]. Some researchers represent images as multi-view data in different feature spaces and exploit the complementary information in multi-view data for computer vision [16, 31–33, 62, 63, 66, 68, 70]. For example, Wu et al. [62, 63] and Feng et al. [16] explored associations in multi-source data for image recognition. Xiong et al. [66], Ma et al. [32] and Xu et al. [68] proposed multi-modal network structures including global and local networks from different perspectives to extract visual features for image scene classification. While Lv et al. [31], Ni et al. [33] and Zeng et al. [70] extracted multi-modal features from a global and local perspective based on the high-level feature map from a same backbone network, and then further aggregated multi-modal features for image scene classification. However, these mentioned methods do not focus on the diversity association of multi-view features in different network layers and different spatial scales.

Based on the above discussions and aforementioned limitations, an end-to-end global-local feature adaptive fusion (GLFAF) network is proposed for image scene classification. Specially, we learn the global dependence of multi-scale and multi-level features from the CNN structure to discover globally shared texture information and semantic information. Meanwhile, a cross-scale progressive aggregation strategy is designed to enhance local semantics and generate the local aggregate features. Subsequently, the global and local features are merged to learn the multi-modal collaborative feature of the image. However, it is worth considering how to explore the fusion advantages between global aggregate features and local aggregate features. In general, it is not appropriate to merge these two features directly. Therefore, the adaptive feature fusion strategy is designed to learn the relationship between the multi-modal features and thus fuse their features rationally. All in all, the global and local aggregate features are extracted from multi-scale and multi-level spatial features generated by the same CNN. Meanwhile, the global and local aggregate features are assigned with optimal weights to get the fused features in an adaptive manner. Specially, the fusion strategy is to assign adaptive weights to confidence scores from the global aggregate features and the local aggregate features, respectively. The optimal weights of the features are trained by the overall model and is optimized in an end to end fashion.

In general, the major contributions of our paper are as follows:

1. An end-to-end global-local multi-modal feature adaptive fusion network is proposed for image scene classification. It can simultaneously simulate the scale and rotation changes of images from both global and local perspectives, and can enhance the global and local semantic representation of images. Consequently, the advantages of adaptive fusion of global aggregate features and local aggregate features can be fully utilized.
2. A global feature aggregate module with multiple spatial dependence is proposed to leverage the complementary advantages of the multiple different spatial feature maps from a global perspective. It first learns spatial global attention features from multi-scale and multi-level spatial feature maps. Then, bidirectional recurrent dependencies and long-range contextual dependencies of these global attention features are learned sequentially to discover their deep correlations.
3. A local feature aggregate module with cross-scale local semantic enhancement is proposed to leverage the complementary advantages of the multiple different spatial feature maps from a local perspective. It first learns channel attention dependencies between multi-level features in the same scale space. The spatial features of different scales are then aggregated across layers to generate rich local features. Next, spatial contextual attention is designed to refine spatial region features and different fisher vector layers are introduced to learn compact associations of spatial local features.

- Two different adaptive feature fusion strategies are investigated to take full advantage of respective features. Different from the commonly used feature fusion methods, the feature adaptive fusion strategy is designed to learn the relationship between the global aggregate features and the local aggregate features with the optimal combination.

2 Related works

Our proposed image scene classification network is highly related to the following aspects.

2.1 Image scene classification methods

Early studies on the traditional features were generally low-level features developed based on the prior knowledge about images, such as lights, colors, textures, and shapes. These low-level features include not only global feature descriptors such as generalized search trees (GIST) [34] and census transform histogram (CENTRIST) [61], but also local feature descriptors such as SIFT [29], histogram of oriented gradient (HOG) [12], local binary pattern (LBP) [39] and SURF [3]. These methods are easy to implement, but their main disadvantage is that a single feature is not enough to represent high-level semantic information, resulting in low classification accuracy. These features are usually combined to integrate multiple kinds of information to yield improved performance [14]. However, these methods based on hand-crafted features are not suitable for classifying scenes with complex backgrounds due to their limited representational ability. Then, the mid-level methods that rely on unsupervised feature learning are developed to bridge the semantic gap between low-level features and semantic information. Specially, these raw pixel values or low-level local features (e.g., SIFT [29]) are used to construct a middle-level representation through BoVW [69], probabilistic latent semantic analysis (PLSA) [18], and sparse coding [40], etc. Compared to methods with hand-crafted features, mid-level based methods can obtain image representations that are more suitable for scene classification [78]. Since unsupervised mid-level methods do not use label information, the feature extraction ability is limited, which is not conducive to further improving the classification performance.

With the development of CNN, it has achieved impressive performance in image scene classification. Due to their powerful feature extraction capabilities and end-to-end learning mechanism, the performance of these methods is much better than methods based on low-level features or middle-level features. Some researchers have focused on designing the structure of the neural network for image scene classification. They designed different methods to change the network layers to add various information. For example, Cheng et al. [10] introduced a new rotation invariant layer on the basis of the existing CNN architecture for improving the performance of scene classification. Li et al. [26] proposed a multi-vector VLAD method based on CNN features for scene classification. Lu et al. [30] aggregated spatial feature maps of different scales by fusing the multi-layer features in the neural network to improve the image scene classification. Shi et al. [42] proposed a lightweight CNN architecture based on branch feature fusion (LCNN-BFF) for scene classification, which uses a lightweight branch fusion strategy to improve the computational efficiency. In order to better retain spatial information, Zhang et al. [73] replaced the original FC layer architecture in CNN with the capsule network architecture. Wang et al. [58] first utilized enhanced feature pyramid network to extract multi-scale and multi-level features, and then designed a deep semantic embedding module to learn the complementary advantages of these features.

Besides, a two-branch deep feature fusion module is introduced to aggregate the features at different levels for image scene classification.

In addition, some researchers have focused on integrating different features for image scene classification. They focused on fusing different neural networks or different features. In general, fusing different information together can add scale and rotation invariant information. For example, Zhang et al. [71] proposed a gradient boosting random convolutional network framework for scene classification, which can effectively combine multiple deep neural networks. Combining these basic neural networks can fuse scale and rotation invariant information into the deep model. Shen et al. [41] proposed a group-attention-fusion strategy to merge two different CNNs to generate refined multi-scale features for scene classification. Huang et al. [21] utilized a pre-trained CNN as a feature extractor to obtain three different features, including multi-layer convolutional features, FC features and LBP-based FC features. Then, these features are fused to fully exploit the discriminative power of the pre-trained CNN for scene classification.

Although the feature representations generated by the abovementioned CNN methods can improve the performance of scene classification tasks compared to traditional methods, they focus on the integration of global features and ignore the extraction of some key local features. However, our method can not only focus on learning global features of images, but also mine and aggregate more informative local features in an adaptive manner.

2.2 Attention mechanism

To further improve the discriminative ability of CNN features, the attention mechanism has been proposed, which aims to focus on some key component to obtain the detailed information during CNN feature learning. For example, Hu et al. [19] introduced a squeeze-and-excitation (SE) block to construct the interdependence among feature channels and adaptively readjusted the channel feature response, thereby significantly improving the performance of existing CNNs. Furthermore, Woo et al. [60] proposed a simple and effective convolutional block attention module (CBAM) that can sequentially infer attention maps along channel and spatial dimensions, and then apply the attention maps to the input feature map for adaptive feature refinement. Wang et al. [57] proposed a novel end-to-end attention recurrent convolutional network (ARCNet) to process and fuse the series of attention representation with a long short-term memory (LSTM)-based sequential processor to select a series of attention regions for scene classification. Bi et al. proposed an attention pooling-based dense connected CNN (APDC-Net) [5] and a residual attention based densely connected CNN (RADC-Net) [6] for scene classification, respectively, which have the potential to strengthen local semantic information and to preserve discriminative CNN features. Similarly, they also proposed a multiple instance densely connected CNN (MIDC-Net) [7] for scene classification, which can strengthen the local semantic representation and multi-instance feature learning. Sun et al. [51] designed gated bidirectional network based on SE block [19] to aggregate multi-layer convolutional features for image scene classification. Obviously, these methods based on regular attention aim to obtain discriminative spatial local information and key channel feature information from image scenes.

Inspired by the great success of transformer in natural language processing [52], researchers have proposed to apply transformer to solve computer vision tasks [15, 28, 59, 74]. Transformer can capture long-range contextual information through multi-head attention, which can be regarded as an enhanced version of the attention mechanism. Recently, Dosovitskiy et al. [15] proposed a fully-transformer model named vision transformer (ViT) for image classification. ViT first divides the image into fixed-size patches, and then learns

effective visual features by mining the relationships among image patches. Despite the success of ViT, the transformer architecture with full attention mechanism [52] is computationally inefficient. To improve efficiency, Liu et al. [28] proposed swin transformer for various image recognition tasks. This method designs a window-based multi-head attention mechanism, which divide the image into multiple windows and only interact inside the windows. Unlike the ViT model, which can only generate outputs of low-resolution features, Wang et al. [59] proposed the pyramid vision transformer (PVT) model for dense vision prediction tasks. Specifically, they design a progressive reduction pyramid and a spatial reduction attention layer to obtain multi-scale and higher-resolution outputs with limited resources. Similarly, Zhang et al. [74] proposed a transformer-based encoder-decoder segmentation architecture, which not only introduced a pyramidal transformer structure to encode multi-scale feature representations, but also designed a transformer parsing module to perform dual decoding of multi-scale features. Clearly, these transformer-based methods aim to obtain long-range contextual associations of local features in images.

Our method designs various strategies based on regular attention [19, 60] to learn channel attention in cross-layer fusion features at the same scale and spatial attention in multi-scale fusion features, and also design different soft attention to learn associative aggregation of spatial local features. Furthermore, inspired by the transformer's ability to learn long-range contextual dependencies, but different from these methods [15, 28, 59, 74] learn long-range dependencies of local region features in images, we utilize the self-attention mechanism to learn long-range association dependencies among multi-scale and multi-level spatial aggregation features.

2.3 Multi-modal learning

Recently, multi-modal or multi-view data has surged as a major stream of big data, where each modal/view encodes individual property of data objects. In general, different modalities are complementary to each other. Therefore, the research of fusing multi-modal features for comprehensive representation of data objects has received extensive attention, such as social recommendation [75], spectral clustering [55], and image retrieval [53], etc. Moreover, deep multi-view learning methods are widely employed in the field of image recognition. Ding et al. [13] first employed multiple CNNs to extract features from face images, and then concatenated all these features. Next, an auto-encoder is used to compress the merged features. Wu et al. [62, 63] proposed different deep metric learning methods for multi-spectral face recognition. Feng et al. [16] presented a cross-modality graph reasoning method for RGB-Infrared person re-identification. Their method can globally model the inter-dependency between modalities and context, and to keep semantic identity consistency between global and local representation. Xiong et al. [66] designed a global modality-specific feature learning module and a local modality-consistent feature learning module, both of which were combined to learn the specificity and consistency of multi-modal features simultaneously for scene classification. Ma et al. [32] obtained intermediate feature maps from pre-trained CNNs and built a global and local integrated model by introducing a visual attention mechanism. They also designed attention consistency model to eliminate the negative impact of attention inconsistency problems on the classification. Xu et al. [68] designed a global-local bi-branch structure to explore discriminative features of raw images and key regions, and adopted a decision-level fusion strategy to improve the image scene classification. Lv et al. [31] proposed a local-global feature fusion network for scene classification, which first utilized ResNet50 to extract the high-level feature map of the image, and then learned the local and global features of the high-level feature map from the channel

dimension and the spatial dimension, respectively. Ni et al. [33] proposed a compact global-local convolutional network based on multi-feature fusion learning, which can take full advantage of local feature distribution learning and the global cross-correlations of multi-feature statistics. Zeng et al. [70] proposed a multi-branch scene classification structure that simultaneously extracts features of global context and local small objects, which can learn robust abstract feature representations of scene images by integrating the features of both. Sun et al. [49] proposed a comprehensive scene classification representation method, which fuses three deep features of target semantic information, global appearance information and contextual appearance information. Sitaula et al. [46] introduced hybrid deep features to represent images by aggregating four types of features that include scene-based and object-based features at the whole image and part image levels.

To sum up, the abovementioned multi-modal learning methods are used for image classification by fusing deep features of multi-source data [13, 16, 62, 63] or different deep features of the same data [31–33, 46, 49, 66, 68, 70]. Unlike their methods that only use the high-level features of the network for multi-modal learning, and mostly adopt the strategy of direct feature fusion. Our method first uses multi-scale and multi-level spatial features in the same network to generate multi-modal features from different perspectives, and then designs different attention mechanisms to fuse these multi-modal features.

3 Method description

This section mainly introduces the network structure of GLFAF, as shown in Fig. 1. A deep convolutional network is designed as the backbone network for generating multi-scale and multi-level spatial features. Based on these spatial features, we first learn the global attention features, and then mine bidirectional recurrent dependencies and arbitrary pairwise dependencies of scale variations in spatial sequence features to obtain global aggregate features. Meanwhile, channel attention is used to fuse the multi-level spatial features at the same scale, and cross-scale channel dependencies is used to progressively generate the

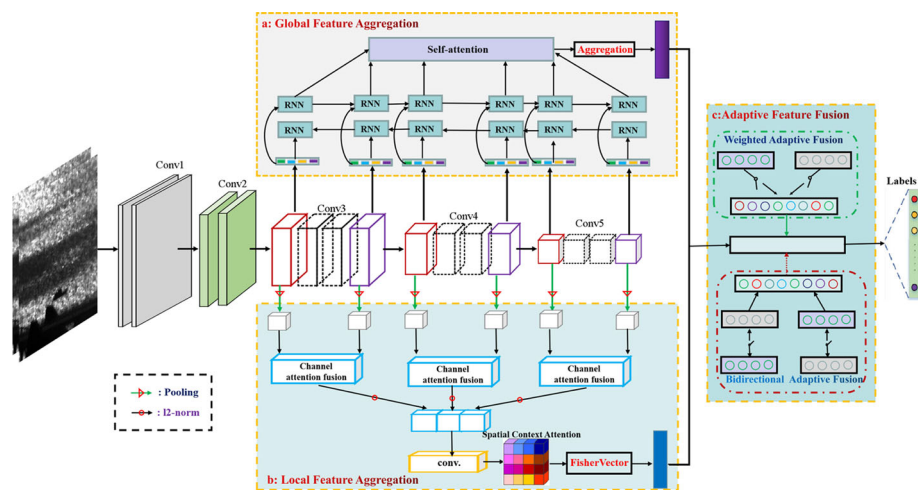


Fig. 1 The flowchart of the global-local feature adaptive fusion network for image scene classification

multi-scale aggregated spatial features. Then, spatial contextual attention is introduced to enhance spatial local features. Moreover, the learnable FisherVector layer is further designed to obtain the local aggregate features. Finally, two different adaptive fusion mechanisms are designed to merge global and local aggregate features. Next, the details of the network structure will be elaborated.

3.1 Backbone network

The backbone network of our proposed GLFAF model is consists of the five convolutional blocks, as shown Conv1-Conv5 of Fig. 1. Since each convolutional block contains several convolution layers and a maximum pooling layer, multi-level feature maps with different spatial resolutions can be generated in turn. In order to make the learned feature maps contain both high-level semantic information and sufficient spatial information, more network layers containing features of different resolutions are used to learn more spatial feature maps. It is obviously different from the previous studies of only using the features from last convolutional layer or pool layer to produce the spatial feature map. Specifically, three different convolutional blocks (Conv3-Conv5) are selected, each of which consists of four convolutional layers, can be used to generate multi-scale and multi-level spatial features. Given the input image I , the feature map X_l with different layers at each spatial resolution is obtained, and its size is $H \times W \times C$. X_l is a 3D tensor with height H , width W , and channel C , l indicates any scale space. In short, the image is input into the designed backbone network, and the multi-resolution features of different layers can be learned. These multi-scale and multi-level spatial features from different network layers can represent different visual meanings such as lines, textures, and objects. Then, we learn the relationship among the multi-scale and multi-level spatial features from the perspective of the global view and the local view to explore the discriminative feature representation.

3.2 Global feature aggregation

Based on the Conv3-Conv5 of the backbone network in Section 3.1, the three convolutional blocks can extract global information in multi-scale and multi-level feature spaces. Specifically, the global features are learned by global average pooling (GAP) of the first and last layers of the convolutional feature maps in each convolutional block. The feature values of different channels are learned by GAP, which can represent the global contextual information of the image, as follows:

$$z_l^c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_l^c(i, j) \quad (1)$$

where $X_l^c \in R^{H \times W}$ is the spatial feature map of the c -th channel extracted from a certain layer of the backbone network in GLFAF, and its width is W and height is H . $Z_l = \{z_l^c | c = 1, \dots, C\} \in R^C$ is the global feature representation of X_l .

Then, after the GAP, two FC layers are constructed, and $W_1 \in R^{F/r \times F}$, $W_2 \in R^{F \times F/r}$ and $b_1 \in R^F$, $b_2 \in R^{F/r}$ are the corresponding weight matrices and bias vectors, respectively. In order to capture the channel-wise dependencies, a weight factor o_c for the c -th channel can be learned via a *sigmoid* layer using an attention mechanism after training the two FC layers.

$$o_l^c = \sigma(W_2 c \delta(W_1 c z_c + b_1 c) + b_2 c) \quad (2)$$

where W_{1c} , W_{2c} and b_{1c} , b_{2c} represent weight matrices and bias vectors of two FC layers corresponding to the c -th channel, respectively. $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ is the sigmoid function. The equation (2) can learn to assign different weights to different channels. The higher weight, the more important of that channel.

Subsequently, the weight factor acts on the global contextual feature map, then global attention feature g_l^c for the c -th channel is obtained via the GAP layer:

$$g_l^c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W (o_l^c \cdot X_l^c(i, j)) \tag{3}$$

where \cdot denotes the channel-wise multiplication operation. Finally, the features of global attention can be described as follows :

$$G_l = (g_l^1, g_l^2, \dots, g_l^{C-1}, g_l^C) \tag{4}$$

As shown in a of Fig. 1, the above operations of GAP and attention are simultaneously applied to the feature maps generated by Conv3_1, Conv3_4, Conv4_1, Conv4_4, Conv5_1 and Conv5_4 of the backbone network for learning the global attention features Gaf3_1, Gaf3_4, Gaf4_1, Gaf4_4, Gaf5_1, and Gaf5_4 at different spatial resolutions and at different levels in the same spatial resolution. Based on equation (4), the above series of features can be formally expressed as $G_{3-1}, G_{3-4}, G_{4-1}, G_{4-4}, G_{5-1}, G_{5-4}$.

In order to learn the global correlation among spatial features at different scales, a bidirectional recurrent neural network (RNN) [2] approach is exploited to mine the dependence of cross-scale spatial sequence features. It can simultaneously learn global dependencies across scales in two directions which are from low spatial resolution to higher spatial resolution and from high spatial resolution to lower spatial resolution.

Specifically, when dealing with a spatial sequence of length L , we uncover the globally associative shared features of the spatial sequence from the varying spatial scales in different directions, i.e., the global aggregate features h_l^{\leftrightarrow} in the network layer of $l \in [1, \dots, L]$ can be expressed as:

$$h_l^{\rightarrow} = \varphi_h(W_{hh}^{\rightarrow} h_{l-1}^{\rightarrow} + W_{ih}^{\rightarrow} G_l^{\rightarrow} + b_h^{\rightarrow}) \tag{5}$$

$$h_l^{\leftarrow} = \varphi_h(W_{hh}^{\leftarrow} h_{l+1}^{\leftarrow} + W_{ih}^{\leftarrow} G_l^{\leftarrow} + b_h^{\leftarrow}) \tag{6}$$

$$h_l^{\leftrightarrow} = h_l^{\rightarrow} + h_l^{\leftarrow} \tag{7}$$

the symbols \rightarrow and \leftarrow represent the spatial sequence from low spatial resolution to higher spatial resolution and the spatial sequence from high spatial resolution to lower spatial resolution, respectively. G_l represents the spatial features of the l -th layer, and h_l represents the hidden layer unit. W_{ih} and W_{hh} denote the shared conversion matrices from the input state to the hidden layer state, and from the previous hidden layer to the current hidden layer, respectively. b_h is the bias term, and φ_h is the nonlinear ReLU activation function.

In a bidirectional recurrent model, the spatial sequence information needs to be recursively acquired. Although bidirectional RNNs can mine sequence information of multi-scale and multi-level features in two different directions, they ignore the semantic dependencies between different layers in the same scale space and between any two scale spaces. Usually, such dependencies are essential to understand the arbitrary invariance among global features. Since the self-attention(SA) mechanism [52] in transformer has the advantage of performing global computation on the input sequence and summarizing the information for an update. Therefore, the SA mechanism is exploited to further learn the arbitrary dependence among sequence features from multi-scale and multi-level spatial bidirectional

dependencies. That is, SA can not only mine the context information in spatial sequence features, but also capture the dependence of scale changes in spatial sequences by calculating the correlation between each pairs of global features in parallel. Specifically, SA uses the dot product operation to calculate the pair-wise relations of global features in spatial sequences, and the calculation formula is as follows:

$$SA(h_i^{\leftrightarrow}) = \text{softmax}\left(\frac{(h_i^{\leftrightarrow} W_q)(h_i^{\leftrightarrow} W_k)^T}{\sqrt{d_k}}\right)(h_i^{\leftrightarrow} W_v) \quad (8)$$

where $W_q \in R^{L \times d_q}$, $W_k \in R^{L \times d_k}$, and $W_v \in R^{L \times d_v}$ are three learnable projection matrices, which are used to project h_i^{\leftrightarrow} into the spaces of query Q , key K , and value V , respectively. QK^T uses the dot product operation to calculate the pair-wise similarity from multi-scale and multi-level global spatial sequence features. d_k and d_v are two hyperparameters. $\sqrt{d_k}$ is used to scale the attention based on the dot product and prevent the result of the dot product of Q and K from being too large. $\text{softmax}(\cdot)$ denotes a normalization function, which make the calculated value become a probability distribution with the sum of weights being 1.

The relationship among spatial features can be captured regardless of the scale difference between global features because the self-attention mechanism calculates the similarity between any two global features in the spatial sequence. Therefore, the output $SA(\cdot)$ of the self-attention learns not only the contextual relationships of the sequences, but also the long-term dependencies of the spatial sequences.

Then, the self-attention features in the spatial sequence are aggregated to generate a multi-scale dependent feature representation h_{SA} . Subsequently, a FC layer is used to further learn the association among the features, thus generating the final global aggregate features y_{SA} , as follows:

$$y_{SA} = \text{Max}(0, W_{SA}h_{SA} + b_{SA}) \quad (9)$$

where W_{SA} and b_{SA} denote the weight matrix and the bias vector, respectively, and $\text{Max}(0, \cdot)$ implements the ReLU activation function.

3.3 Local feature aggregation

Section 3.2 obtains global aggregate information by learning the overall dependence of multi-scale and multi-level spatial features. Different from Section 3.2, this section progressively enhance and aggregate local feature information based on these same spatial features to obtain discriminative feature representation, as shown in b of Fig. 1. Generally, the extraction of visual local features (e.g., SIFT [29]) can learn the scale invariance of images well. Therefore, inspired by SIFT features, CNN is used to extract rich multi-scale and multi-level local features and learn the semantic context of local features.

3.3.1 Cross-scale spatial local feature learning

Generally, convolution features of different levels contain spatial structure information of different image scenes. Shallow convolutional layers have small receptive fields, which can capture the appearance information of each detailed scene unit. However, deep convolutional layers have large receptive fields, which can harvest the spatial structure information among different scene units. It obviously provides a basis for the aggregation of different intermediate convolutional layers. However, some existing methods [56, 65] usually use traditional feature coding (e.g., VLAD [23]) methods to encode the local features generated by each convolutional layer separately. However, such traditional feature coding methods cannot explore the complementarity among convolutional features from different levels.

In order to aggregate the intermediate convolutional layers to generate richer local features, a convolutional feature fusion module is designed to convert the features of intermediate convolutional layers into an aggregated convolutional representation. It consists of three parts: 1) multi-level spatial feature fusion in the same scale; 2) multi-scale spatial feature fusion; 3) attention-enhanced spatial feature learning. Unlike traditional feature coding methods that encode each layer of features separately, our proposed convolutional feature coding module simultaneously takes all intermediate convolutional features as input to generate a convolutional representation.

The main idea of the feature encoding module is to learn the convolutional representation using nonlinear convolution. First, a channel attention fusion method is designed to merge the multi-level convolutional features in the same scale space. Then the pooling operation is used to unify the sizes of the convolutional features in different scale spaces. Next, the concatenation operation is used to merge different convolutional features, and the convolution of 1×1 with ReLU is used to explore the complementarity among all convolutional features on the channel because the convolutional operation based on ReLU is a simple and efficient operation to increase the nonlinear interaction across the channel features. Finally, the spatial contextual attention is designed to generate discriminative spatial feature map. Next, the convolutional feature fusion module will be described.

As shown in b of Fig. 1, the lowest-level convolutional feature map Conv_L.1 and the highest-level convolutional feature map Conv_L.4 in each scale space are combined in the dimension of the channel to obtain the multi-level aggregated spatial feature map Conv_L.1-4, denoted as U . Since the above operations all merge multi-level features in the channel dimension, we further explore the feature dependence among channels to better aggregate the multi-level features in the same spatial resolution. In recent years, channel attention modules such as SE block [19] and ACNet [20] have been designed to make the network focus on more informative channels. More precisely, the proposed channel attention fusion module (as shown in Fig. 2) is inspired by channel attention mechanism in CBAM [60]. Specially, the maximum pooling f_{max} and average pooling f_{avg} are applied to spatial feature map U to generate one-dimensional global channel representations v^m and v^a . Subsequently, v^m and v^a are learned through multi-layer perceptron (MLP) to analyze their channel dependencies. Therefore, the different deep channel-dependent representations are fused by summing and activated with sigmoid to obtain the channel attention representation $Z^c(U)$. The final channel attention map U_c is generated by performing element-wise multiplication between $Z^c(U)$ and U . The expression of channel attention is as follows:

$$Z_1^c(U) = F(U, W) = f_{c2}(\delta(f_{c1}(f_{max}(U), W_1), W_2)) \tag{10}$$

$$Z_2^c(U) = F(U, W) = f_{c2}(\delta(f_{c1}(f_{avg}(U), W_1), W_2)) \tag{11}$$

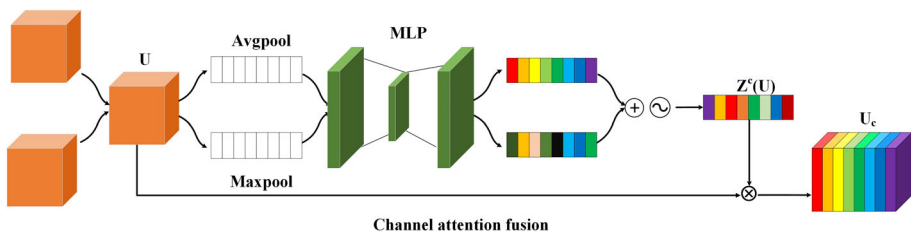


Fig. 2 Channel attention fusion

$$Z^c(U) = \sigma(Z_1^c(U) + Z_2^c(U)) \quad (12)$$

$$U_c = Z^c(U) \otimes U \quad (13)$$

where f_c represents the MLP operation, $W_1 \in R^{C/r \times C}$ and $W_2 \in R^{C \times C/r}$ denote the learning weights of the two FC layers in the MLP, $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ is the sigmoid function.

The abovementioned multi-level channel attention fusion operation is applied to spatial maps of different scales to construct three-scale multi-level aggregated spatial feature maps, namely Conv3_1_4, Conv4_1_4 and Conv5_1_4. The maximum pooling operation in the backbone network in GLFAF will sequentially generate convolutional feature maps of different sizes, which can help increase the receptive field. Obviously, the sizes of the three different spatial convolution feature maps, Conv3_1_4, Conv4_1_4, and Conv5_1_4, are different.

In order to aggregate convolution features of different sizes, the width and height of all intermediate convolution features need to be unified to the same size. Specifically, average pooling operations with different strides are used to unify the width and height of the three different multi-level aggregated spatial feature maps, and keep the number of spatial channels unchanged. In addition, the l_2 -normalization is employed to normalize the multi-level aggregated convolutional features across the channel. Since the magnitude of values in different convolutional feature is completely different, the l_2 -normalization can effectively avoid numerical problems. Then, the formula of the l_2 -normalization across the channel is expressed as follows:

$$\bar{r}_{h,w,c} = \frac{r_{h,w,c}}{\sqrt{\sum_{i=1}^C r_{h,w,i}^2 + \varepsilon}} \quad (14)$$

where $r \in R^{H,W,C}$ denotes the convolutional feature, $\bar{r} \in R^{H,W,C}$ denotes the normalized convolutional feature, H denotes the height of r and \bar{r} , W denotes the width of r and \bar{r} , C denotes the number of channels of r and \bar{r} , and ε denotes used to avoid the divisor, 0. After the l_2 -normalization, the proposed GLFAF can converge more steadily.

After a series of pooling and normalization operations, three convolutional feature maps of uniform size are generated. To avoid introducing weight parameters, a concatenation operation is used to merge different convolutional features. That is, the three uniform convolutional features are directly stacked in the channel dimension. Finally, the convolutional operation of 1×1 with ReLU is used to further explore the complementarity among channels in the convolutional feature. It not only increases the nonlinear interaction of cascading features among channels [50], but also keeps the width and height of the features unchanged. Therefore, a multi-scale deep aggregated spatial feature map $U_m \in R^{N \times N \times D}$ is generated. In summary, multi-level spatial feature in the same scale and multi-scale spatial feature are both progressively fused from the channel dimension of the convolutional feature map to obtain rich features. However, there is no reinforcement learning about the differences in different positions of spatial features in the spatial (width and height) dimensions. Therefore, the spatial attention mechanism should be further designed to highlight meaningful spatial feature units.

Spatial attention is used to make the network focus on more informative regions. The CBAM [60] utilizes a 7×7 convolution to learn the spatial attention mask after concatenating the max-pooled and average-pooled spatial features. In order to better capture the spatial context-aware information, we modified the original spatial attention sub-module in

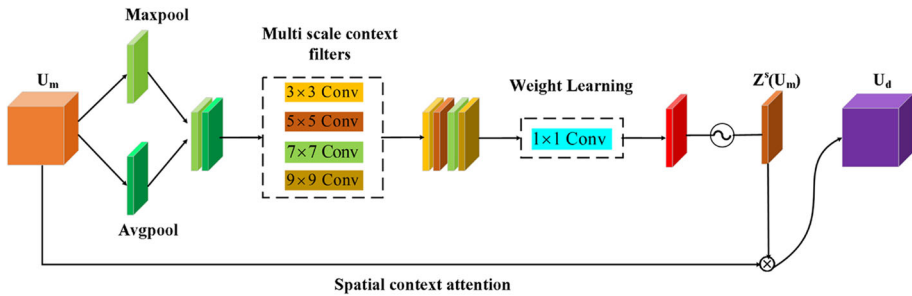


Fig. 3 Spatial attention fusion

CBAM. That is, the convolutions with different receptive fields are used to generate intermediate feature masks, rather than using a 7×7 convolution only. Then, these intermediate masks are concatenated and a 1×1 convolution is used to learn weights.

The spatial attention map can be considered as the weighted sum of feature masks. An illustration of our spatial context attention sub-module is shown in Fig. 3. The $S \in R^{N \times N \times 2}$ is generated by concatenating squeezed feature masks $f_{\max}^c(U_m)$ and $f_{\text{avg}}^c(U_m)$. Here, f_{\max}^c and f_{avg}^c represent max-pooling and average-pooling along the channel dimensions, respectively. To exploit the spatial contextual information, four different scales of context filters (3×3 , 5×5 , 7×7 , and 9×9) are used. The feature mask is produced by concatenating these channel masks generated by the 3×3 , 5×5 , 7×7 , and 9×9 context filters. Then, a 1×1 convolution is used to learn and accumulate weights. The spatial contextual attention can be computed as:

$$S = f_{\max}^c(U_m) \oplus f_{\text{avg}}^c(U_m) \tag{15}$$

$$Z^s(U_m) = \delta(f^{1 \times 1}(f^{3 \times 3}(s); f^{5 \times 5}(s); f^{7 \times 7}(s); f^{9 \times 9}(s))) \tag{16}$$

$$U_d = Z^s(U_m) \otimes U_m \tag{17}$$

where $f^{n \times n}$ denotes a $n \times n$ convolution. Consequently, the modulated spatial feature map U_d is obtained by element-wise multiplication between the multi-scale aggregated spatial feature map U_m and the final spatial attention mask $Z^s(U_m)$.

3.3.2 Fisher vector network layer

The discriminative spatial feature map U_d is aggregated into an visual feature representation with local invariance, which is specifically implemented by the designed fisher layer. The traditional fisher vector(FV) [35] is hard-assigned and non-differentiable, so it cannot be learned end-to-end through deep networks. Therefore, we modify the traditional non-differentiable learning method in FV [35] to an end-to-end learning method, and call it the fisher layer, as shown in Figs. 4 and 5.

In fisher’s aggregation learning, $U_d \in R^{N \times N \times D}$ can be decomposed into $N \times N$ number of D -dimensional local region features $U_d = \{u_{i,j} | u_{i,j} \in R^D, i, j = 1, 2, \dots, N\}$, where each $u_{i,j}$ represents a local feature at a specific location of the input image. Then, FisherNet and FisherNext are proposed by using two different designs of spatial region $u_{i,j}$ as the smallest unit and grouped $u_{i,j}$ as the smallest unit.

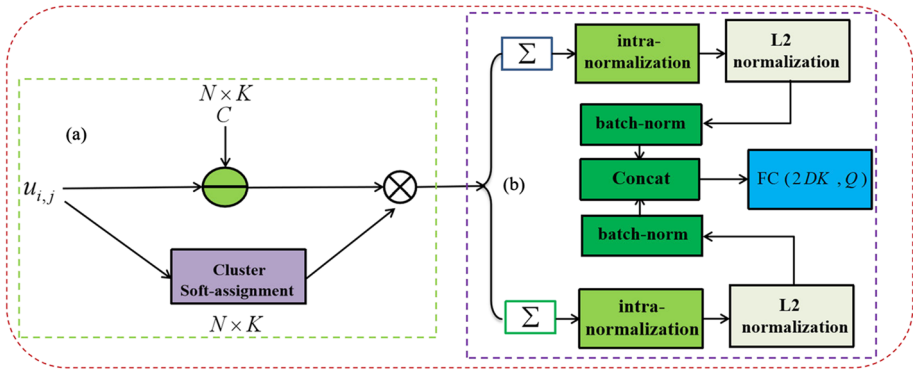


Fig. 4 Schematic diagram of FisherNet layers

FisherNet For a series of inputable spatial local features $u_{i,j} \in R^D$, a soft-assigned learnable probability function is as follows:

$$a_k(u_{i,j}) = \frac{e^{W_k^T u_{i,j} + b_k}}{\sum_{k'} e^{W_{k'}^T u_{i,j} + b_{k'}}} \tag{18}$$

where W_k and b_k are learnable parameters. In other words, the soft assignment of the local features $u_{i,j}$ to the semantic centers c_k uses the interval range of 0 to 1 to measure the correlation weight of the local feature $u_{i,j}$ to the semantic center c_k . In the traditional hard allocation method, if the local feature $u_{i,j}$ and the semantic centers c_k are closely related, then $a_k(u_{i,j})$ is equal to 1; otherwise, it is equal to 0.

In order to be able to integrate into the deep network end-to-end, $a_k(u_{i,j})$ will define the soft assignment between the local region features $u_{i,j}$ and the learnable K semantic centers $\{c_1, c_2, \dots, c_K | c_k \in R^D\}$. As shown in (a) of Fig. 4, the differentiable fisher vector can be

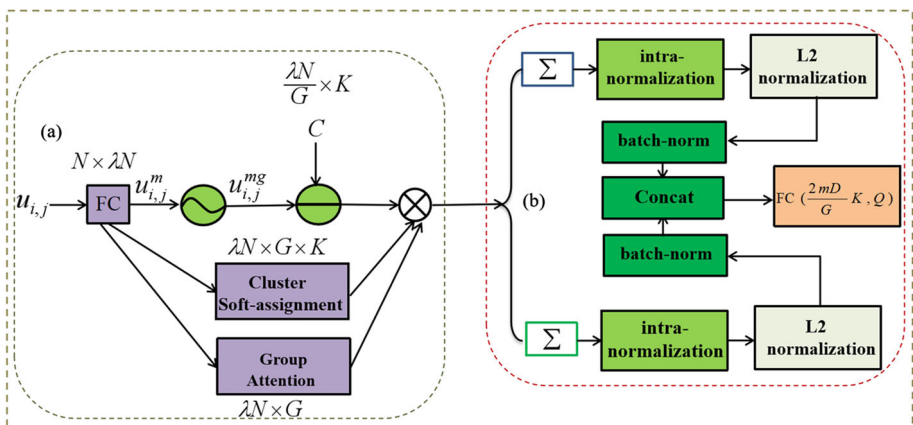


Fig. 5 Schematic diagram of FisherNext layers

expressed as follows:

$$V_1(:, q, k) = \sum_{i,j}^{N,N} a_k(u_{i,j}) \left(\frac{u_{i,j} - c_k(q)}{\delta_k(q)} \right) \quad (19)$$

$$V_2(:, q, k) = \sum_{i,j}^{N,N} a_k(u_{i,j}) \left(\left(\frac{u_{i,j} - c_k(q)}{\delta_k(q)} \right)^2 - 1 \right) \quad (20)$$

where $V_1(:, q, k)$ and $V_2(:, q, k)$ capture the first-order and second-order FisherNet aggregation representation, respectively. $c_k, k \in [1, K]$ is the learnable semantic center, and $\delta_k, k \in [1, K]$ is the diagonal covariance of the semantic center. To define $\delta_k, k \in [1, K]$ as positive, a gaussian noise with unit mean and small variance is first used to initialize their values randomly, and then square their values during the training process to keep them positive.

As shown in (b) of Fig. 4, the matrix V_1 and V_2 is L2-normalized column-wise (intra-normalization) first, and stretch it into a vector, L2-normalized again in its entirety, batch-normalized again, and finally, different orders of spatial aggregation features V_1^n and V_2^n are obtained. Subsequently, the two features are fused and a non-linear FC layer is used to explore the internal associations among the fused features to obtain the final feature representation of local invariance. Its calculation is as follows:

$$V_m^n = \text{Max}(0, W_{net}(V_1^n \oplus V_2^n) + b_{net}) \quad (21)$$

where \oplus denotes the operation of feature splicing, W_{net} and b_{net} are the weight matrix and the bias vector respectively, and $\text{Max}(0, \cdot)$ realizes the ReLU activation function.

FisherNext Unlike FisherNet, which takes the generated image local region as the smallest unit, FisherNext further divides the image local region into groups. It can also calculate the association among spatial local features, and can greatly reduce the network parameters of the fisher layer.

Specifically, as shown in (a) of Fig. 5, a FC layer is first used to expand the feature vector $u_{i,j}$ in each spatial region, as shown in the following equation:

$$u_{i,j}^m = \text{Max}(0, W_{fc}u_{i,j} + b_{fc}) \quad (22)$$

where $W_{fc} \in R^{D \times 2D}$ and $b_{fc} \in R^D$ denote the weight matrix and the bias vector, respectively, and $\text{Max}(0, \cdot)$ implements the ReLU activation function. This operation relearns the spatial region feature $u_{i,j} \in R^D$ as $u_{i,j}^m \in R^{2D}$.

Then, in order to finely discover the contribution of the features in each spatial region to the overall semantics, all spatial region features are divided into $\{G \mid g \in \{1, \dots, G\}\}$ groups, each of which contains mD/G -dimensional feature vectors. That is, the spatial region feature map $U_{N,N}^m$ with a shape of (N^2, mD) is divided into G lower dimensional feature vectors $U_{N,N}^{mg}$ with a shape of $(N^2, G, mD/G)$.

Different from the idea of soft assignment in FisherNet, we try to design a hierarchical soft assignment idea to adaptively learn higher-order related information in the semantic centers. Formally, take the given $u_{i,j}^m$ in $U_{N,N}^m$ and the corresponding $u_{i,j}^{mg}$ in $U_{N,N}^{mg}$ as inputs, K semantic centers c_k^{mg} as learnable parameters over groups. FisherNext-based adaptive aggregation means that the compact association of local features $u_{i,j}^{mg}$ in all groups from learnable semantic centers c_k^{mg} in the same grouped lower-dimensional space. The

calculation is as follows:

$$V_1^g(\cdot, q, k) = \sum_g \sum_{i,j}^{N,N} a_g^m(u_{i,j}^m) \bar{a}_{gk}^m(u_{i,j}^m) \left(\frac{u_{i,j}^{mg}(q) - c_k^{mg}(q)}{\delta_k^{mg}(q)} \right) \quad (23)$$

$$V_2^g(\cdot, q, k) = \sum_g \sum_{i,j}^{N,N} a_g^m(u_{i,j}^m) \bar{a}_{gk}^m(u_{i,j}^m) \left(\left(\frac{u_{i,j}^{mg}(q) - c_k^{mg}(q)}{\delta_k^{mg}(q)} \right)^2 - 1 \right) \quad (24)$$

where $V_1^g(\cdot, q, k)$ and $V_2^g(\cdot, q, k)$ capture the first-order and second-order FisherNext aggregation representation, respectively. $\{c_k^{mg} | k \in [1, K]\}$ is the learnable semantic center, and $\{\delta_k^{mg} | k \in [1, K]\}$ is the diagonal covariance of the semantic center. To define $\{\delta_k^{mg} | k \in [1, K]\}$ as positive, we first use a gaussian noise with unit mean and small variance to initialize their values randomly, and then square their values during the training process to keep them positive.

Among them, $a_g^m(u_{i,j}^m)$ denotes the attention function over groups G , and the calculation process is as follows:

$$a_g^m(u_{i,j}^m) = \sigma(W_g u_{i,j}^m + b_g) \quad (25)$$

where $W_g \in R^{mD \times G}$ and $b_g \in R^G$ denote the weight matrix and bias vector, respectively, while $\sigma(\cdot)$ implements the sigmoid function with an output scale of 0 to 1. Thus, we can obtain probability values to measure the importance of each subgroup g .

In addition, $\bar{a}_{gk}^m(u_{i,j}^m)$ denotes the soft assignment of grouped local feature $u_{i,j}^m$ to K different semantic center c_k^{mg} . The attention function of the learnable soft-assignment is denoted as follows:

$$\bar{a}_{gk}^m(u_{i,j}^m) = \frac{e^{BN(W_{gk} u_{i,j}^m + b_{gk})}}{\sum_{s=1}^K e^{BN(W_{gs} u_{i,j}^m + b_{gs})}} \quad (26)$$

where $W_{gk} \in R^{mD \times Gk}$ and $b_{gk} \in R^{Gk}$ denote the weight matrix and the bias vector, respectively. $BN(\cdot)$ implements the batch normalization function, and the conversion process of batch normalization is as follows:

$$BN(h; \gamma; \beta) = \beta + \gamma \frac{h - Mean[h]}{\sqrt{Var[h] + \varepsilon}} \quad (27)$$

where h is the vector passing through FC layer over a complete minibatch, γ and β are model parameters that determine the mean and standard deviation of the normalized activation, ε is a regularization hyperparameter. The statistics $Mean[h]$ and $Var[h]$ are estimated by the sample mean and sample variance of the current minibatch.

As shown in (b) of Fig. 5, the matrix V_1^m and V_2^m are L2-normalized column-wise (intra-normalization) first, and stretch it into a vector, L2-normalized again in its entirety, batch-normalized again, and finally, different orders of spatial aggregation features V_{1n}^m and V_{2n}^m are obtained. Subsequently, the two features are fused and a non-linear FC layer is used to explore the internal associations among the fused features to obtain the final feature representation V_n^m of local invariance. Its calculation is as follows:

$$V_n^m = Max(0, W_V (V_{1n}^m \oplus V_{2n}^m) + b_V) \quad (28)$$

where \oplus denotes the operation of feature splicing, W_V and b_V are the weight matrix and the bias vector respectively, and $Max(0, \cdot)$ realizes the ReLU activation function.

3.4 Adaptive feature fusion

In general, fusing features from different modal spaces can provide superior performance over a single modality. However, direct cascading of different modal features [21, 30, 41] often does not provide the best performance. In order to alleviate this limitation, two adaptive fusion networks are constructed: 1) weighted adaptive feature fusion; 2) bidirectional adaptive feature fusion. For ease of representation, the global aggregate features are denoted as h_f^g , and the local aggregate features are denoted as h_f^l . Then the two fusion models are introduced as follows:

3.4.1 Weighted adaptive feature fusion

The attention model is designed to calculate the optimal fusion weight for each feature h_f^g and h_f^l , which can adaptively weigh the importance of h_f^g and h_f^l . Figure 6 shows the schematic diagram of weighted adaptive feature fusion. Specifically, according to the content of the respective characteristics, the respective attention weights λ_g and λ_l are adaptively learned to generate the optimal solution to satisfy $\lambda_g + \lambda_l = 1$. The learned attention is subsequently applied to the input features h_f^g and h_f^l respectively, and the calculation process is as follows:

$$h_g = \delta(W_g h_f^g + b_g) \tag{29}$$

$$h_l = \delta(W_l h_f^l + b_l) \tag{30}$$

$$\lambda_g = \sigma(W_h(h_g \oplus h_l)) \tag{31}$$

$$h_{gl} = \lambda_g \odot h_g \oplus (1 - \lambda_g) \odot h_l \tag{32}$$

$$o = \delta(W_{gl} h_{gl} + b_{gl}) \tag{33}$$

where W_* and b_* are the parameters of learnable weight matrix and bias vector, respectively, and h_g and h_l are the global aggregate features and local aggregate features after single-layer perceptron transformation, respectively. $\delta(\cdot)$ implements the ReLU activation

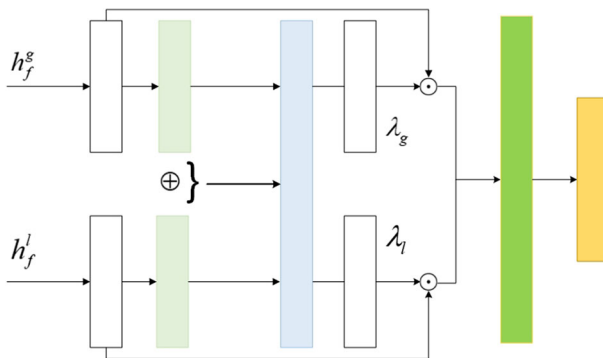


Fig. 6 Schematic diagram of weighted adaptive feature fusion

function, $\sigma(\cdot)$ is the *sigmoid* activation function of logistic regression, and \oplus is the connection operation. λ_g represents the attention weight applied to the global invariant features, and $1 - \lambda_g$ represents the attention weight applied to the local aggregate features. Therefore, h_{gl} represents the global-local features of weighted adaptive fusion. In addition, o represents the final output features of exploring the internal associations in h_{gl} .

3.4.2 Bidirectional adaptive feature fusion

Inspired by the recurrent neural network [2], a bidirectional adaptive feature fusion method from “global-local” and “local-global” is proposed to merge global aggregate features and local aggregate features. The framework of bidirectional adaptive feature fusion is shown in Fig. 7.

In the global to local direction, there are two input nodes in the model. The first input feature is the global aggregate feature h_f^g , and the second input feature is the local aggregate feature h_b^g . The reset ratio and the update ratio are calculated by Eqs.(34) and (35), respectively. The primary fusion feature h_f^p is calculated by Eq.(36). The primary fusion feature and the global aggregate feature are used as the two input features to calculate the intermediate fusion feature h_f^i by Eq.(37).

$$z_f = \sigma(W_z h_f^l + U_z h_f^g) \tag{34}$$

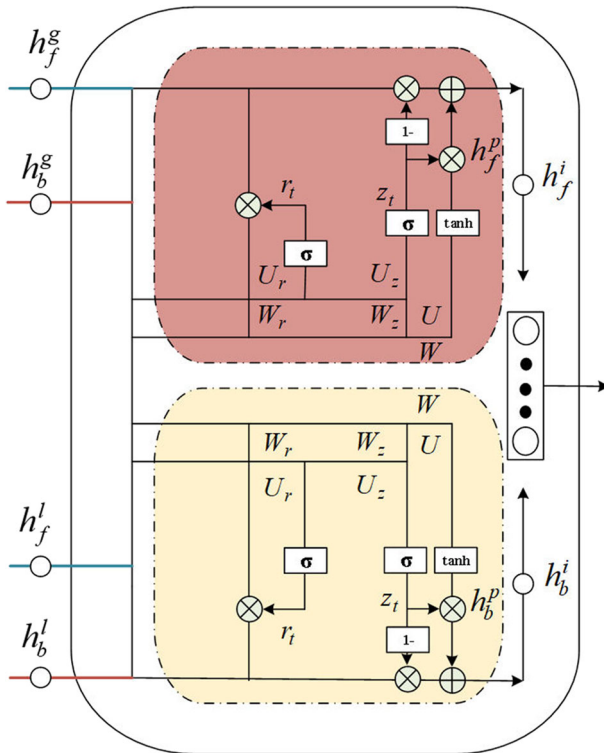


Fig. 7 Schematic diagram of bidirectional adaptive feature fusion

$$r_f = \sigma(W_r h_f^l + U_r h_f^g) \quad (35)$$

$$h_f^p = \tanh(W h_f^l + r_f \circ U h_f^g) \quad (36)$$

$$h_f^i = z_f * h_f^g + (1 - z_f) * h_f^p \quad (37)$$

In the local to global direction, the first input feature is the local aggregate feature h_b^l and the second input feature is the global aggregate feature h_b^g . The reset ratio and the update ratio are also calculated by Eqs.(38) and (39), respectively. The primary fusion feature h_b^p is calculated by Eq.(40). The primary fusion feature and the local aggregate feature are used as the two input features to calculate the intermediate fusion feature h_b^i by Eq.(41).

$$z_b = \sigma(W_z h_b^g + U_z h_b^l) \quad (38)$$

$$r_b = \sigma(W_r h_b^g + U_r h_b^l) \quad (39)$$

$$h_b^p = \tanh(W h_b^g + r_b \circ U h_b^l) \quad (40)$$

$$h_b^i = z_b * h_b^l + (1 - z_b) * h_b^p \quad (41)$$

In bidirectional feature fusion, h_b^i and h_f^i are calculated together to get the final fusion feature o . The formula is show as follows:

$$o = W_f h_f^i + W_b h_b^i + b_o \quad (42)$$

In all these formula Eqs.(34)-(42). $W_z, W_r, U_z, U_r, W, U, W_f, W_b, b_o$ are all the weight parameters that need to be learned during the training phase. $\sigma(\cdot)$ is the *sigmoid* activation function of logistic regression.

3.5 Classification

In summary, an end-to-end network structure is proposed to learn the global-local features, and to explore the adaptive fusion of global aggregate features and local aggregate features. As shown in Fig. 1, the scene image I is input to the proposed GLFAF model, and the softmax classifier is used to obtain the semantic label of the scene image I . The goal of training the GLFAF is to minimize the cross entropy loss of the classifier. The formula of cross entropy loss is as follows:

$$\begin{aligned} J(\theta; \{o^{(m)}, y^{(m)}\}_{m=1}^M) \\ = -\frac{1}{M} \sum_{m=1}^M \sum_{k=1}^K 1\{y^{(m)} = c\} \log \frac{e^{\theta_k^T o^{(m)}}}{\sum_{n=1}^K e^{\theta_n^T o^{(m)}}} \end{aligned} \quad (43)$$

where o is the adaptive fusion feature representation extracted from the scene image, m is the label corresponding to the image, θ is the weight parameter of softmax, $\{c|c \in \{1, \dots, K\}\}$ is the total number of semantic labels in the training database, and M is the number of scene images trained in a batch, and $1\{\cdot\}$ denotes the indicator function. Obviously, all operations of multi-scale and multi-level spatial features extraction, the learning of global aggregate features and local aggregate features, the adaptive fusion of global-local features and softmax classifier in the proposed GLFAF are optimized under the supervision of cross-entropy loss J .

4 Experimental analyses

In this section, extensive experiments are conducted to evaluate the effectiveness of the proposed GLFAF. First, the data sets to be used for our experiments are introduced. Second,

Table 1 Scene dataset information

Datasets	Images per class	Total images	Images size	Scene classes
IMS	242-605	2066	640*512	5
UIUC	137-250	1579	different size	8
UCM	100	2100	256*256	21

the experimental settings are then described. Third, the performance of GLFAF is compared with other state-of-the-art methods. Next, we study model ablation analysis to evaluate the contribution of each model component. Finally, the running time and the weight parameters of different network are reported and discussed.

4.1 Dataset description

We use three datasets from different fields, including the infrared maritime scene (IMS) dataset [14], UIUC-Sports(UIUC) dataset [24] and UC Merced Land-Use (UCM) dataset [69] to conduct a series of experiments. As shown in Table 1, the basic information of the three datasets is introduced. The IMS dataset contains 5 marine environment categories such as backlit sea, calm sea, foggy sea, rough sea and sea-sky-line. And it has a total of 2066 images with at least 200 images in each category. Scene examples of IMS dataset are shown in Fig. 8. The UIUC dataset contains 1579 images covering 8 categories of sports activity scenes, and each category contains 137-250 images. Scene examples of UIUC dataset are shown in Fig. 9. The UCM dataset contains 2100 images belonging to RGB images of 0.3 m spatial resolution. It covers 21 land-use categories, and each category contains 100 scene images. Scene examples of UCM dataset are shown in Fig. 10.

4.2 Experimental settings

All images are adjusted to the same size of $224 \times 224 \times 3$. In particular, VGG19-Net [44] is selected as the backbone network, and the subsequent global and local aggregation features are learned based on the 3-5 blocks in VGG19-Net. We adjust the output of the first convolution feature and the fourth convolution feature in these blocks to a feature map with the same channel size using a convolution operation with the kernel size of 3×3 and the filters number of 128, respectively. The stochastic gradient descent (SGD) method with small batch sizes and weight decay is used to optimize the entire network, where the weight decay value is 0.0001, the momentum value is 0.9, and the learning rate in the training phase is 0.001. The batch-size of the proposed GLFAF network is 16, and we exploit real-time data

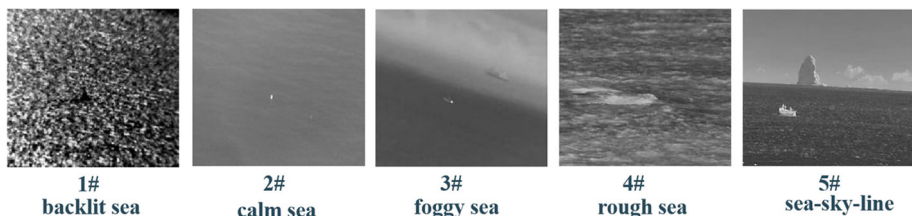


Fig. 8 Sample diagram of the scene classification in IMS dataset

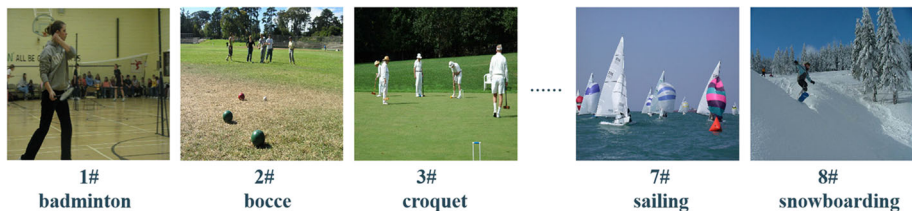


Fig. 9 Sample diagram of the scene classification in UIUC dataset

augmentation(e.g., random rotation, flip, and cropping) on the training dataset. The computer configuration is as follows: RAM: 32GB; processor: Intel (R) Core(TM)i7-9750H CPU @2.60 GHz; GPU: Nvidia GeForce GTX 1660Ti.

For the IMS dataset, 50% and 90% of the images in each category are randomly selected for training, and the remaining 50% and 10% are used for testing. For each train/test set in the UIUC dataset, 70 images per category are involved in the train set and 60 images are involved in the test set, as done by previous studies [45–47, 65]. We randomly selected 50% of the images in each scene class from the UCM dataset as the training set, and the rest are divided into the test set, as employed in previous studies [5–7, 31, 58].

We select the Precision, Recall, F1 score, and Accuracy as indicators to analyze and compare the experimental results, which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{44}$$

$$Recall = \frac{TP}{TP + FN} \tag{45}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{46}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{47}$$

the TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent the relationship between the predicted authenticity and the actual scene, respectively. It is known from equation (46) that F1 score is related to recall and precision, and is their harmonic mean. The higher the F1 score, the higher the precision and recall, and the better the performance.

In addition, the confusion matrix is also used to evaluate classification performance quantitatively. It is a specific table layout that allows to visualize directly the performance of each category, which reports errors and confusions among different categories by calculating the



Fig. 10 Sample diagram of the scene classification in UCM dataset

correct and incorrect classification of the test images for each category and accumulating the results in a table.

4.3 Comparison with state-of-the-arts

This section will compare the performance of the proposed GLFAF method with other recently proposed image scene classification methods on the three different datasets.

4.3.1 Experiment 1: IMS dataset

In this experiment, we use the IMS dataset under the training ratios of 50% and 90% to evaluate the effectiveness of the proposed method. Table 2 lists the overall accuracy results of the comparative experiment of the respective methods. As shown in Table 2, the proposed GLFAF method achieves overall accuracy of 90.61%, 91.58%, 90.51%, 90.80% under training ratios of 50%, and 94.69%, 96.14%, 94.20%, 95.65% under training ratios of 90%, respectively. Obviously, the classification performance of the proposed GLFAF method is higher than other comparison methods. Among them, the overall accuracy of the proposed methods are better than those of classical networks such as VGG-16, VGG-19 and ResNet50. Moreover, the proposed GLFAF method also shows better overall accuracy compared to the current state-of-the-art methods (e.g., FACNN, VGG16-CapsNet and GBNet+global feature). Compared with the multi-branch feature fusion network LCNN-BFF, the performance of the proposed GLFAF-NetFV-w and GLFAF-NextFV-w is roughly comparable to the LCNN-BFF method, while the proposed GLFAF-NetFV-b and GLFAF-NextFV-b significantly outperform the LCNN-BFF method. In terms of adaptive fusion of global and local features of GLFAF, our bidirectional adaptive fusion method works slightly better than the weighted adaptive fusion method. In addition, for the two different configurations of the fisher vector layer, the effect of using the NetFV configuration is slightly better than that of NextFV. Overall, the effect of the approach based on NetFV and the effect of the approach based on NextFV is almost equivalent. Moreover, with the increase of trainable

Table 2 Overall accuracy of different methods on IMS dataset

Methods	T.R.=50%(OA)	T.R.=90%(OA)
VGG-16 [44]	89.13%	92.79%
VGG-19 [44]	89.25%	93.24%
ResNet50 [17]	88.58%	92.01%
CNN+VLAD [26]	90.35%	93.04%
FACNN [30]	89.84%	94.21%
VGG16-CapsNet [73]	89.94%	92.96%
GBNet [51]	89.69%	92.92%
GBNet+global feature [51]	90.42%	93.85%
LCNN-BFF [42]	90.78%	94.18%
Ours		
GLFAF-NetFV-w	90.61%	94.69%
GLFAF-NetFV-b	91.58%	96.14%
GLFAF-NextFV-w	90.51%	94.20%
GLFAF-NextFV-b	90.80%	95.65%

Table 3 Performance of GLFAF-NetFV-b in different classes at 90% training rate on the IMS dataset

	Precision	Recall	f1-score
backlit sea	90.91%	93.02%	91.95%
calm sea	93.33%	95.45%	94.38%
foggy sea	100%	95.24%	97.56%
rough sea	100%	97.14%	98.55%
sea-sky-line	98.44%	98.44%	98.44%

infrared maritime data, the recognition accuracy of scene classification is higher. It shows that a larger number of training data sets are more conducive to the accurate classification of infrared maritime scenes.

Table 3 shows the performance of the GLFAF-NetFV-b method for each class in the IMS dataset under the 90% training rate. It can be seen that the proposed GLFAF-NetFV-b method achieves more than 90% performance on the three evaluation metrics (i.e., precision, recall and F1 score) for all categories. In addition, to more intuitively visualize the performance of the proposed GLFAF-NetFV-b method, the confusion matrix of our method on different categories is further plotted. Figure 11 shows the confusion matrix of GLFAF-NetFV-b at a training ratio of 90% on the IMS dataset. As shown in Fig. 11, the proposed method achieves more than 95% classification accuracy on the four categories of 'calm sea', 'foggy sea', 'rough sea', and 'sea-sky-line', while 93% classification accuracy in the category of 'backlit sea'. It can be seen that our global-local feature adaptive fusion method based on multi-scale spatial features can mine discriminative infrared sea surface features

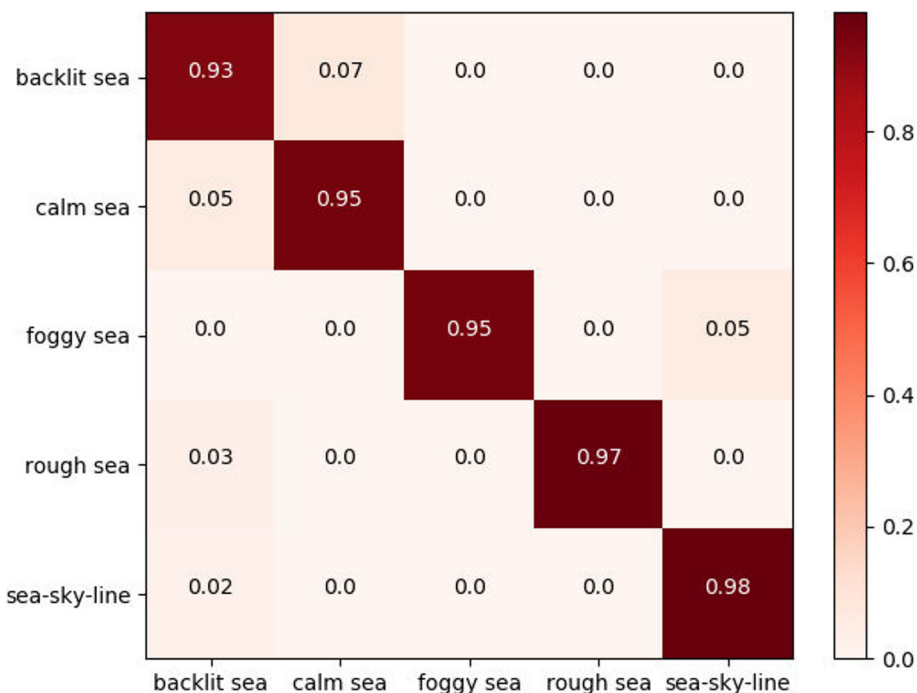


Fig. 11 Confusion matrix of GLFAF-NetFV-b under the 90% training ratio on the IMS dataset

in the category-unbalanced infrared maritime scene dataset. Of course, the value of the diagonal line in Fig. 11 also represent recall of each scene category. The appearance of tiny ripples in the calm sea is similar to texture in backlit sea, which causes confusion between the 'backlit sea' category and the 'calm sea' category. Meanwhile, the rough textures in the 'rough sea' category are similar to the 'backlit sea' category, which makes some samples in it be incorrectly classified as 'backlit sea'. Furthermore, the background of the sea and fog segmentation generated by the foggy sea surface is similar to the scene of the sea and sky segmentation, which makes some samples of the 'foggy sea' category be incorrectly identified as the 'sea-sky-line' category.

4.3.2 Experiment 2: UIUC dataset

Table 4 tabulates the comparison results on the UIUC dataset. The UIUC dataset is sport event scenes, which is different from the general scenes. It is an understanding of sports scenes, and it pays more attention to the action of the prominent target and semantic context in the image. Since UIUC dataset is a class imbalanced dataset, we choose 70 samples from each class for training and 60 samples from each class for testing. As shown in Table 4, the overall accuracy of the four variants of the proposed GLFAF method has reached more than 95%. Moreover, the overall accuracy of GLFAF-NetFV-w is the highest, which is better than that of the WS-AM method that focuses on spatial relationship aggregation, and is equal to

Table 4 Overall accuracy of different methods on UIUC dataset

Methods	OA
RSP [22]	79.6%
BOW [54]	83.5%
CNN [9]	85.9%
LPR-RBF [38]	86.2%
Hybird Parts+GIST+SP [76]	87.2%
VC+VQ [25]	88.4%
ISPR [27]	89.5%
s-CNN(max) [54]	90.9%
s-CNN(avg) [54]	91.2%
s-CNNC(max) [54]	91.5%
EISR [72]	92.7%
WS-AM [65]	93.07%
TSF [45]	94.4%
VGG [77]	95.6%
HDF [46]	96.2%
TF [47]	95.8%
DF [47]	97.5%
Ours	
GLFAF-NetFV-w	97.29%
GLFAF-NetFV-b	96.88%
GLFAF-NextFV-w	96.46%
GLFAF-NextFV-b	96.25%

Table 5 Performance of GLFAF-NetFV-W in different classes on the UIUC dataset

	Precision	Recall	f1-score
RockClimbing	100%	100%	100%
badminton	100%	100%	100%
bocce	88.89%	93.33%	91.06%
croquet	93.33%	93.33%	93.33%
polo	100%	96.67%	98.31%
rowing	96.72%	98.33%	97.52%
sailing	100%	100%	100%
snowboarding	100%	96.67%	98.31%

the overall accuracy of the DF method that focuses on deep content features. As far as the four variants of our model are concerned, the effect based on NetFv configuration is better than that based on NextFv configuration, and the effect based on weighted adaptive feature fusion is also significantly better than that based on bidirectional adaptive feature fusion.

Table 5 shows the performance of GLFAF-NetFV-w method in each category on UIUC dataset. It can be seen that the proposed method achieves better than 90% performance on precision, recall and F1 score on all categories except 'bocce'. The confusion matrix of the UIUC dataset is indicated in Fig. 12. It can be seen that the classification performance of six

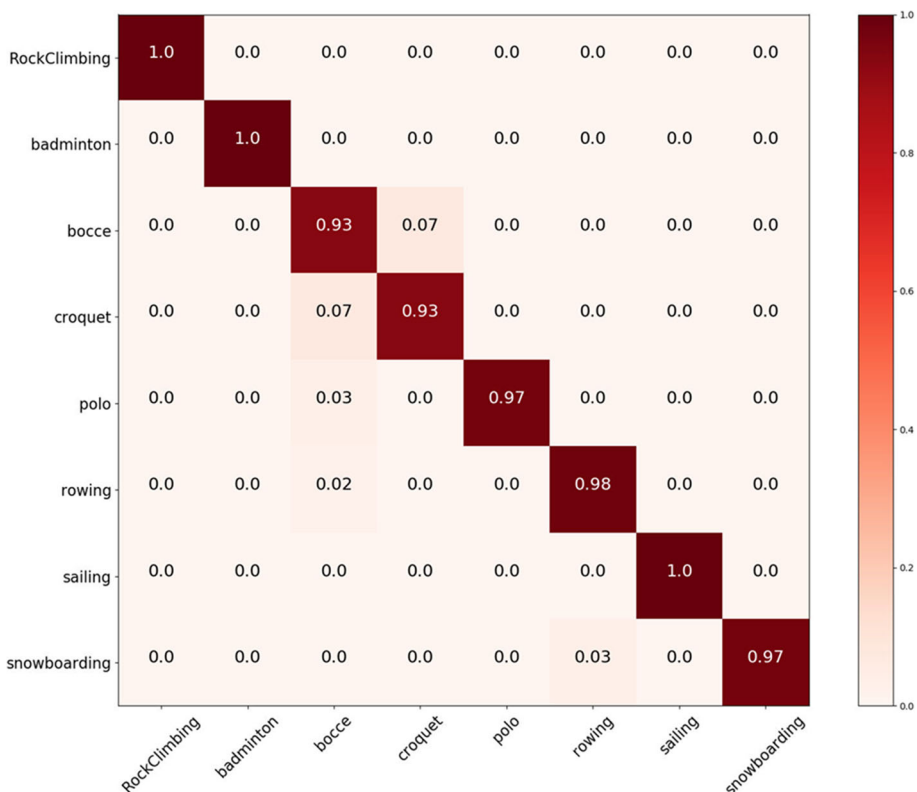


Fig. 12 Confusion matrix of GLFAF-NetFV-w on the UIUC dataset

different categories (e.g., RockClimbing, badminton, polo, rowing, sailing, and snowboarding) can reach more than 95%, and even the classification performance of three categories such as 'RockClimbing', 'badminton' and 'sailing' reaches 100%. The distinction between the 'bocce' category and the 'croquet' category is the most confusing, especially for the 'bocce' category, due to the similar body movements and spheres between the 'bocce' category and the 'croquet' category.

4.3.3 Experiment 3: UCM dataset

We further perform experiment 3 on the UCM dataset to evaluate the performance of our proposed method. As discussed in Section 4.1, this dataset is challenging due to its small data size and many categories. As shown in Table 6, four variants of proposed GLFAF method is compared with some novel networks based on VGG16, ResNet50, and attention mechanism. The best performance is 97.52% obtained by our method GLFAF-NetFV-w, which is 0.56% higher than the original best performance acquired by VGG-VD16 + RIR. Moreover, the overall accuracy of the GLFAF-NetFV-w method is also better than the global-local feature fusion method ResNet_LGFFE with resnet50 as the backbone network. Compared with the multi-scale and multi-level feature aggregation network EFPN-DSE-TDFF, our methods GLFAF-NetFV-w and GLFAF-NetFV-b also improve by 1.33% and 0.95%, respectively. This is because our method not only extracts multi-scale and multi-level features, but also enriches the feature representation of images based on global and local multi-modal learning. From the perspective of spatial local aggregation, the NetFv-based configuration performs better than the NextFv-based configuration. From the perspective of multimodal fusion, the method based on weighted adaptive feature fusion is

Table 6 Overall accuracy of different methods on UCM dataset

Methods	T.R.=50%(OA)
VGGNet-16 [64]	94.14%
TEX-Net with VGG [1]	94.22%
ARCNet-VGG16 [57]	96.81%
ResNet50 [31]	92.43%
ResNet_LGFFE [31]	95.36%
MIDCCNN [7]	94.93%
APDC-Net [5]	95.01%
RADC-Net [6]	94.79%
MIDC-Net_CS [4]	95.41%
LCNN-BFF [42]	94.64%
EFPN-DSE-TDFF [58]	96.19%
AlexNet + RIR [36]	95.91%
VGG-VD16 + RIR [36]	96.96%
Ours	
GLFAF-NetFV-w	97.52%
GLFAF-NetFV-b	97.14%
GLFAF-NextFV-w	95.37%
GLFAF-NextFV-b	94.84%

Table 7 Performance of GLFAF-NetFV-W in different classes on the UCM dataset

	Precision	Recall	f1-score
agricultural	100%	100%	100%
airplane	100%	100%	100%
baseballdiamond	96.15%	100%	98.04%
beach	100%	100%	100%
buildings	92.45%	98.00%	95.15%
chaparral	100%	100%	100 %
denseresidential	95.74%	90.00%	92.78%
forest	100%	96.00%	97.96%
freeway	100%	100%	100%
golfcourse	98.00%	98.00%	98.00%
harbor	100%	100%	100%
intersection	88.89%	96.00%	92.31%
mediumresidential	92.00%	92.00%	92.00%
mobilehomepark	96.15%	100%	98.04%
overpass	100%	96.00%	97.96%
parkinglot	100%	100%	100%
river	98.00%	98.00%	98.00 %
runway	100%	100%	100%
sparseresidential	92.31%	96.00%	94.12%
storagetanks	100%	88.00%	93.62%
tenniscourt	100%	100%	100%

also slightly better than the method based on bidirectional adaptive feature fusion, but the performance of these two fusion methods is generally comparable on the UCM dataset.

Table 7 shows the performance of the GLFAF-NetFV-w method on each scene category in the UCM dataset. It can be seen that the proposed method achieves better than 90% performance on precision, recall and F1 score on all categories except 'intersection' and 'storagetanks'. Figure 13 shows the confusion matrix of GLFAF-NetFV-w on the UCM dataset. When the training rate is 50%, the classification accuracy of 20 scene categories reaches more than 90%. The 'storagetanks' category has the worst classification performance compared to the other 20 scene categories. This is due to the fact that the distribution of circular landmark buildings in the 'storagetanks' category is very similar to that of street intersections, rooftops, and round courts, which makes the samples in the 'storagetanks' category prone to misclassification. Besides, some images of the 'mediumresidential' category may be incorrectly identified as 'denseresidential' and 'sparseresidential'. The main features of the three scene categories ('denseresidential', 'mediumresidential', 'sparseresidential') are buildings, the only difference is the density of buildings in the image.

4.4 Ablation experiments

In order to study the impact of different modules in the proposed method on scene classification, we use the proposed GLFAF-NetFV-w as the benchmark and Accuracy as the evaluation metric to conduct ablation experiments on the IMS dataset, UIUC dataset and UCM dataset, as shown in Table 8. Furthermore, different methods designed in our ablation experiments exploit the same parameter settings, optimization strategies, and training

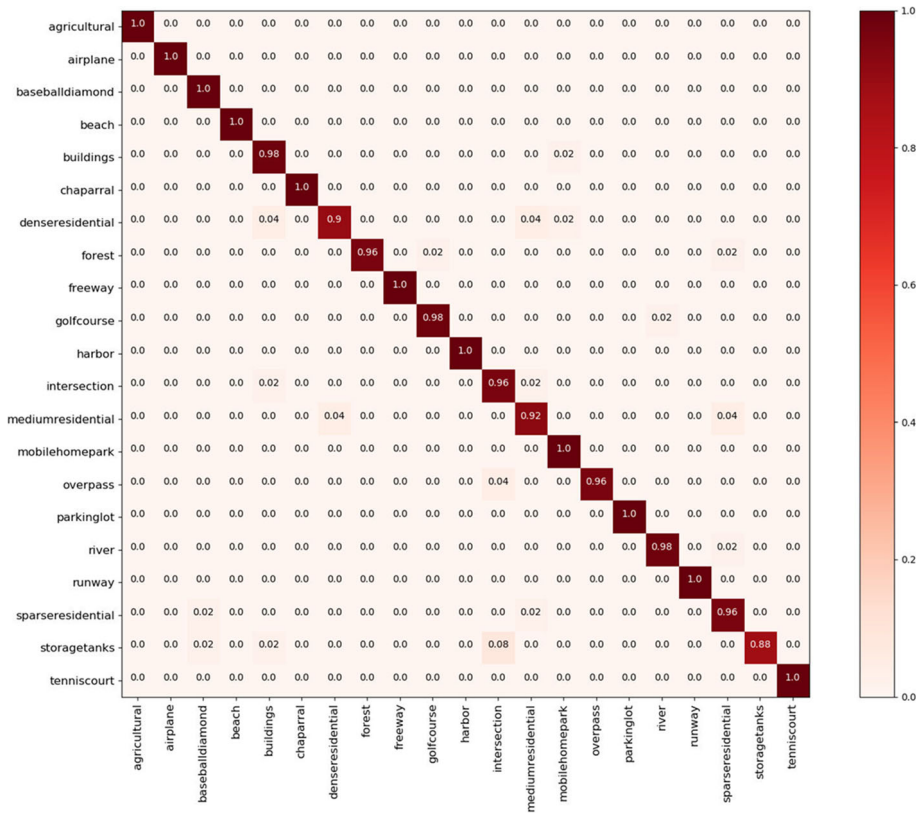


Fig. 13 Confusion matrix of GLFAF-NetFV-w on the UCM dataset

settings. To simplify the representation, we call the global feature aggregation learning branch as GFA_Net, the local feature aggregation learning branch as LFA_Net, and the GLFAF-NetFV-w as GLFA_Net. As can be seen from Table 8, when only GFA_Net or LFA_Net is used, the overall accuracy on the IMS dataset are 93.75% and 94.12%, respectively; the overall accuracy on the UIUC dataset are 96.46% and 96.88%, respectively; the overall accuracy on the UCM dataset is 96.95% and 96.38%, respectively. GLFA_Net combines the advantages of both, and the overall accuracy on the three datasets

Table 8 Ablation experiment on the different dataset

	IMS	UIUC	UCM
GFA_Net	93.75%	96.46%	96.95%
LFA_Net	94.12%	96.88%	96.38%
GLFA_Net	94.69%	97.29%	97.52%
GLFA_Net_without_channel attention	93.96%	96.04%	94.76%
GLFA_Net_without_spatial attention	93.42%	96.47%	95.33%
GLFA_Net_without_channel&spatial attention	92.65%	95.66%	94.67%

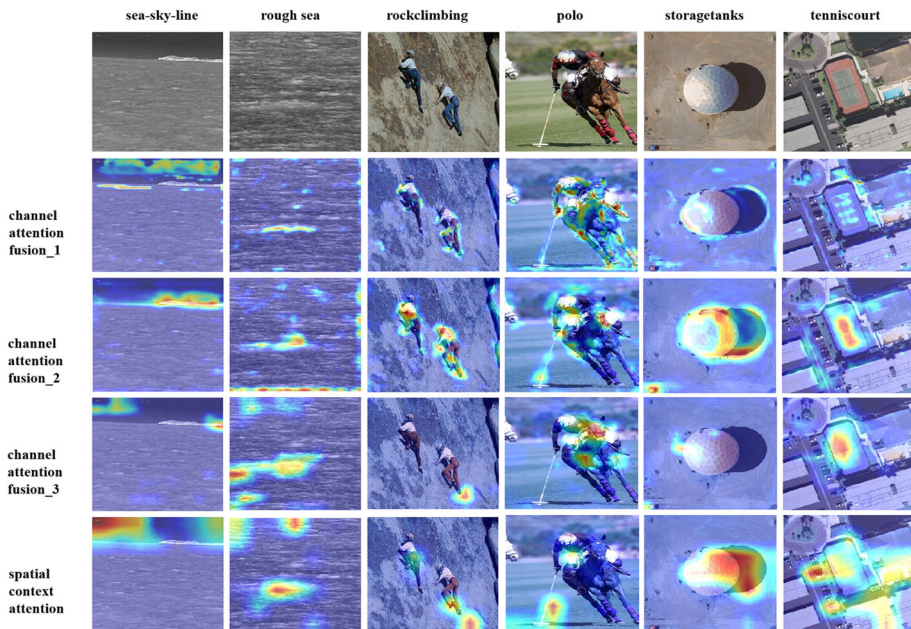


Fig. 14 The visualization results of different attention modules in the proposed method

reach 94.69%, 97.29%, and 97.52%, respectively. Additionally, to verify that the channel attention module (as shown in Fig. 2) and spatial contextual attention module (as shown in Fig. 3) in the proposed GLFA_Net method are effective, we remove different types of attention modules and verify their performance. As can be seen from Table 8, the effects of removing channel attention (GLFA_Net_without_channel attention), removing spatial attention (GLFA_Net_without_spatial attention), and removing channel-spatial attention (GLFA_Net_without_channel&spatial attention) have all decreased to varying degrees, which indicates that the attention module in the proposed method is effective in mining discriminative features.

In order to better demonstrate the intermediate process of the attention module in GLFA_Net, we select images of six scenes of rough sea, sea-sky-line, polo, rockclimbing, storagetanks and tenniscourt as experimental data to visualize the output of the attention model. In Fig. 14, channel attention fusion_1, channel attention fusion_2, and channel attention fusion_3 respectively represent the visualization results of multi-level channel attention fusion in the three scale spaces of Block3-Block5, while spatial context attention represents the visualization results of spatial context attention. The visualization experiments on the six scene images demonstrated that different attention modules in the proposed method could extract different detailed features in scene images, which is helpful to solve the problems of within-class differences and between-class similarities in the classification of scene images.

4.5 Computational efficiency analysis

The training and testing time can directly reflect the computational efficiency of our proposed method. The backbone network of our proposed GLFAF model is modified based

Table 9 Computational efficiency analysis of different networks on three different scene data sets

	VGG19		GLFAF-NetFV-w		GLFAF-NetFV-b		GLFAF-NextFV-w		GLFAF-NextFV-b	
	Train.(m)	Test.(s)	Train.(m)	Test.(s)	Train.(m)	Test.(s)	Train.(m)	Test.(s)	Train.(m)	Test.(s)
IMS(50%/50%)	71.00	11.52	72.50	12.12	76.90	12.85	72.59	12.24	74.33	12.49
UCM(50%/50%)	74.53	10.39	76.25	12.38	81.91	13.02	74.76	12.38	75.93	12.70
UIUC(70/60)	41.54	7.13	38.66	6.98	40.80	7.40	38.69	7.03	39.20	7.22
parameters	139.6M		43.7M		44.3M		31.8M		32.5M	

on the VGG-19 network. Therefore, the VGG-19 network is exploited as the model benchmark to analyze the running time of four variants of GLFAF, namely GLFAF-NetFV-w, GLFAF-NetFV-b, GLFAF-NextFV-w and GLFAF-NextFV-b. The training time, test time and weight parameters of different networks are shown in Table 9. The data divided by different configurations in each data set are trained on their respective models for 150 epochs. As shown in Table 9, the running time of different variants of GLFAF is very close to the running time of VGG-19 network. Among them, the VGG-19 network contains approximately 139.6 million weight parameters, most of which are introduced by the FC layers. In GLFAF, the FC layers of VGG-19 network are removed, but the global and local invariant features in multi-scale and multi-level spatial maps are learned. Next, global and local features are aggregated in an adaptive manner for image scene classification. Specifically, GLFAF-NetFV-w has 43.7 million weight parameters, GLFAF-NetFV-b has 44.3 million weight parameters, GLFAF-NextFV-w has 31.8 million weight parameters and GLFAF-NextFV-b has 32.5 million weight parameters. Obviously, the weight parameter of the four variants of GLFAF is significantly smaller than that of VGG-19 network. In GLFAF network, the weight parameters of our methods based on NetFV are larger than our methods based on NextFV, and the weight parameters of our methods using bidirectional adaptive feature fusion are slightly larger than our methods using weighted adaptive feature fusion under the same configuration of NetFV or NextFV.

Although the weight parameters of the GLFAF network are significantly smaller than those of the VGG-19 network, the running time of the GLFAF network is not significantly reduced. That is, compared to the VGG-19 network, the GLFAF network requires slightly less running time on the UIUC dataset, but slightly more running time on the IMS and UCM datasets. In terms of the neural networks, the computational efficiency of FC layers is very high, while the computational efficiency of convolutional layers is relatively low. Compared with the VGG-19 network, some additional convolutional layers, recursive learning mechanism, different attention mechanisms, different association aggregation mechanisms of spatial region features, and different feature adaptive fusion mechanisms are introduced in our GLFAF network. Although these added components make the GLFAF network richer in feature representation, these different mechanisms reduce its computational efficiency to varying degrees. More specifically, NextFV-based methods have significantly smaller weight parameters than NetFV-based methods, but the running time of NextFV-based methods is not reduced compared to NetFV-based methods. It is due to the fact that grouped attention mechanism in NextFV is more complex than the attention mechanism in NetFV, which greatly reduces the computational efficiency of the NextFV-based method. Meanwhile, the running time of the bidirectional adaptive feature fusion method based on the same NetFV or NextFV configuration is slightly longer than that of the weighted adaptive

feature fusion method. This is because the gated attention mechanism in bidirectional adaptive feature fusion is more complex than that in weighted adaptive feature fusion, which reduces the computational efficiency of the model.

5 Conclusions

In this paper, an end-to-end global-local feature adaptive fusion network is proposed for image scene classification. The global aggregate features and the local aggregate features can be extracted respectively based on the multi-scale and multi-level spatial features acquired from the same backbone network. Specifically, the global aggregate features are generated by exploring the multiple relationships of global features at different scales. Meanwhile, the local aggregate features are generated by gradually fusing spatial features of different scales, and mining the relationship of spatial local features based on multiple attentions. Furthermore, assuming the different roles of global aggregate features and local aggregate features, two different adaptive feature fusion strategies are proposed to merge global-local features together. Overall, our approach can explore the complementary nature of global and local features to comprehensively describe the image scene. Finally, the proposed method is comprehensively evaluated on three datasets with different characteristics. The experimental results demonstrate that the proposed method achieves better image scene classification performances than other related methods.

We show that there is a best-performing value for our proposed approach. As a final remark we would like to note that through the experimental results that discussed in Section 4.3 we also observed that our method still suffers from some misclassifications for scene images that are highly similar in appearance. This is due to the fact that although our method studies the high-order correlation of spatial local features, it does not focus on the spatial positional relationship of local features. In the future, we will incorporate local positional relationships in higher-order associations of local features, and explore metric learning method to integrate local and global information.

Funding This paper was supported in part by the Fundamental Research Funds for the Central Universities of China under Grant 3132019340 and 3132019200. This paper was supported in part by high tech ship research project from ministry of industry and information technology of the people's republic of China under Grant MC-201902-C01.

Data Availability The UC Merced Land-Use dataset that support the findings of this study are available in the ucmcered repository: <http://weegeevision.ucmerced.edu/datasets/landuse.html>. The UIUC Sports dataset that support the findings of this study are available in the stanford repository: http://vision.stanford.edu/lijjiali/event_dataset/. The infrared maritime scene dataset that support the findings of this study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare no conflicts of interest.

References

1. Anwer RM, Khan FS, van de Weijer J et al (2018) Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J Photogrammetry Rem Sens* 138:74–85
2. Basiri ME, Nemati S, Abdar M et al (2021) ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Futur Gener Comput Syst* 115:279–294

3. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features, European conference on computer vision. Springer, Berlin, pp 404–417
4. Bi Q, Qin K, Li Z et al (2019) Multiple instance dense connected convolution neural network for aerial image scene classification. In: 2019 IEEE International conference on image processing (ICIP). IEEE, pp 2501–2505
5. Bi Q, Qin K, Zhang H et al (2019) APDC-Net: attention pooling-based convolutional network for aerial scene classification. *IEEE Geosci Rem Sens Lett* 17(9):1603–1607
6. Bi Q, Qin K, Zhang H (2020) RADC-Net: a residual attention based convolution network for aerial scene classification. *Neurocomputing* 377:345–359
7. Bi Q, Qin K, Li Z et al (2020) A multiple-instance densely-connected ConvNet for aerial scene classification. *IEEE Trans Image Process* 29:4911–4926
8. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J Mach Learn Res* 3:993–1022
9. Chen Y (2015) Convolutional neural network for sentence classification. University of Waterloo
10. Cheng G, Ma C, Zhou P et al (2016) Scene classification of high resolution remote sensing images using convolutional neural networks. In: 2016 IEEE International geoscience and remote sensing symposium (IGARSS). IEEE, pp 767–770
11. Cheng G, Xie X, Han J et al (2020) Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. *IEEE J Selected Topics Appl Earth Observ Rem Sens PP(99):1–1*
12. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. IEEE, pp 886–893
13. Ding C, Tao D (2015) Robust face recognition via multimodal deep face representation. *IEEE Trans Multimed* 17(11):2049–2058
14. Dong L, Zhang T, Ma D et al (2020) Maritime background infrared imagery classification based on histogram of oriented gradient and local contrast features. *Journal of Infrared and Millimeter Waves* 39:5
15. Dosovitskiy A, Beyer L, Kolesnikov A et al (2020) An image is worth 16x16 words: transformers for image recognition at scale, arXiv:2010.11929
16. Feng Y, Chen F, Ji Y et al (2021) Efficient cross-modality graph reasoning for RGB-infrared person re-identification. *IEEE Signal Process Lett* 28:1425–1429
17. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
18. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 42(1):177–196
19. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
20. Hu X, Yang K, Fei L et al (2019) Acnet: attention based network to exploit complementary features for rgbd semantic segmentation. In: IEEE International conference on image processing (ICIP). IEEE, pp 1440–1444
21. Huang H, Xu K (2019) Combing triple-part features of convolutional neural networks for scene classification in remote sensing. *Remote Sens* 11(14):1687
22. Jiang Y, Yuan J, Yu G (2012) Randomized spatial partition for scene recognition, European conference on computer vision. Springer, Berlin, pp 730–743
23. Jgou H, Douze M, Schmid C et al (2010) Aggregating local descriptors into a compact image representation. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 3304–3311
24. Li LJ, Li FF (2007) What, where and who? Classifying events by scene and object recognition Computer Vision. In: Proc.of IEEE International conference on computer vision, pp 1–8
25. Li Q, Wu J, Tu Z (2013) Harvesting mid-level visual concepts from large-scale internet images. In: 2013 IEEE Conference on computer vision and pattern recognition, pp 851–858
26. Li Q, Peng Q, Yan C (2018) Multiple VLAD encoding of CNNs for image classification. *Comput Sci Eng* 20(2):52–63
27. Lin D, Lu C, Liao R et al (2014) Learning important spatial pooling regions for scene classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3726–3733
28. Liu Z, Lin Y, Cao Y et al (2021) Swin transformer: hierarchical vision transformer using shifted windows. *International Conference on Computer Vision*, 10012–10022
29. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60(2):91–110
30. Lu X, Sun H, Zheng X (2019) A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Trans Geosci Remote Sens* 57(10):7894–7906

31. Lv Y, Zhang X, Xiong W et al (2019) An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification. *Rem Sens* 2019 11(24):3006
32. Ma J, Ma Q, Tang X et al (2020) Remote sensing scene classification based on global and local consistent network, IGARSS 2020-2020. In: IEEE International geoscience and remote sensing symposium. IEEE, pp 537–540
33. Ni K, Liu P, Wang P (2021) Compact global-local convolutional network with multifeature fusion and learning for scene classification in synthetic aperture radar imagery. *IEEE J Selected Topics Appl Earth Observ Rem Sens* 14:7284–7296
34. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vis* 42(3):145–175
35. Perronnin F, Snchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification, European conference on computer vision. Springer, Heidelberg, pp 143–156
36. Qi K, Yang C, Hu C et al (2021) Rotation invariance regularization for remote sensing image scene classification with convolutional neural networks[J]. *Remote Sens* 13(4):569
37. Rublee E, Rabaud V, Konolige K et al (2011) ORB: an efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision. IEEE, pp 2564–2571
38. Sadeghi F, Tappen MF (2012) Latent pyramidal regions for recognizing scenes, European conference on computer vision. Springer, Berlin, pp 228–241
39. Satpathy A, Jiang X, Eng HL (2014) LBP-based edge-texture features for object recognition. *IEEE Trans Image Process* 23(5):1953–1964
40. Sheng G, Wen Y, Tao X et al (2012) High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int J Remote Sens* 33(8):2395–2412
41. Shen J, Zhang T, Wang Y et al (2010) A dual-model architecture with grouping-attention-fusion for remote sensing scene classification. *Remote Sens* 13(3):433
42. Shi C, Wang T, Wang L (2020) Branch feature fusion convolution network for remote sensing scene classification. *IEEE J Selected Topics Appl Earth Observ Rem Sens* 13:5194–5210
43. Shrinivasa SR, Prabhakar CJ (2022) Scene image classification based on visual words concatenation of local and global features. *Multimed Tools Appl* 81(1):1237–1256
44. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *Computer Science*
45. Sitaula C, Xiang Y, Basnet A et al (2019) Tag-based semantic features for scene image classification. In: International conference on neural information processing. Springer, Cham, pp 90–102
46. Sitaula C, Xiang Y, Basnet A et al (2020) Hdf: hybrid deep features for scene image representation. *International Joint Conference on Neural Networks (IJCNN) IEEE 2020:1–8*
47. Sitaula C, Aryal S, Xiang Y et al (2021) Content and context features for scene image representation[J]. *Knowl-Based Syst* 232:107470
48. Smeulders AWM, Worring M, Santini S et al (2000) Content-based image retrieval at the end of the early years. *IEEE Trans Pattern Anal Mach Intell* 22(12):1349–1380
49. Sun N, Li W, Liu J et al (2018) Fusing object semantics and deep appearance features for scene recognition. *IEEE Trans Circuits Syst Video Technol* 29(6):1715–1728
50. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
51. Sun H, Li S, Zheng X et al (2019) Remote sensing scene classification by gated bidirectional network. *IEEE Trans Geosci Rem Sens PP*(99):1–15
52. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need[J]. *Advances in Neural Information Processing Systems*, 30
53. Wang Y (2021) Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. *ACM Trans Multimed Comput Commun Appli (TOMM)* 17(1s):1–25
54. Wang D, Mao K (2019) Task-generic semantic convolutional neural network for web text-aided image classification. *Neurocomputing* 329:103–115
55. Wang Y, Zhang W, Wu L et al (2016) Iterative views agreement: an iterative low-rank based structured optimization method to multi-view spectral clustering. [arXiv:1608.05560](https://arxiv.org/abs/1608.05560)
56. Wang G, Fan B, Xiang S et al (2017) Aggregating rich hierarchical features for scene classification in remote sensing imagery. *IEEE J Selected Topics Appl Earth Observ Rem Sens* 10(9):4104–4115
57. Wang Q, Liu S, Chanussot J et al (2018) Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans Geosci Remote Sens* 57(2):1155–1167
58. Wang X, Wang S, Ning C et al (2021) Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. *IEEE Trans Geosci Rem Sens* 59(9):7918–7932
59. Wang W, Xie E, Li X et al (2021) Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. *International Conference on Computer Vision*, 568–578

60. Woo S, Park J, Lee JY et al (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
61. Wu J, Rehg JM (2010) Centrist: a visual descriptor for scene categorization. *IEEE Trans Pattern Anal Mach Intell* 33(8):1489–1501
62. Wu F, Jing XY, Dong X et al (2018) Intraspectrum discrimination and interspectrum correlation analysis deep network for multispectral face recognition. *IEEE Trans Cybern* 50(3):1009–1022
63. Wu F, Jing XY, Feng Y et al (2021) Spectrum-aware discriminative deep feature learning for multi-spectral face recognition. *Pattern Recogn* 111:107632
64. Xia GS, Hu J, Hu F (2017) AID: a benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans Geosci Remote Sens* 55(7):3965–3981
65. Xia S, Zeng J, Leng L et al (2019) Ws-am: weakly supervised attention map for scene recognition. *Electronics* 8(10):1072
66. Xiong Z, Yuan Y, Wang Q (2020) MSN: modality separation networks for RGB-D scene recognition. *Neurocomputing* 373:81–89
67. Xu K, Huang H, Deng P et al (2020) Two-stream feature aggregation deep neural network for scene classification of remote sensing images[J]. *Inform Sci* 539:250–268
68. Xu K, Huang H, Deng P (2021) Remote sensing image scene classification based on global-local dual-branch structure model. *IEEE Geoscience and Remote Sensing Letters*
69. Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, pp 270–279
70. Zeng D, Chen S, Chen B et al (2018) Improving remote sensing scene classification by integrating global-context and local-object features. *Remote Sens* 10(5):734
71. Zhang F, Du B, Zhang L (2015) Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans Geosci Remote Sens* 54(3):1793–1802
72. Zhang C, Zhu G, Huang Q et al (2017) Image classification by search with explicitly and implicitly semantic representations. *Inform Sci* 376:125–135
73. Zhang W, Tang P, Zhao L (2019) Remote sensing image scene classification using CNN-CapsNet. *Remote Sens* 11(5):494
74. Zhang J, Yang K, Constantinescu A et al (2021) Trans4Trans: efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. *International Conference on Computer Vision*, 1760–1770
75. Zhang C, Wang Y, Zhu L et al (2021) Multi-graph heterogeneous interaction fusion for social recommendation. *ACM Trans Inform Syst (TOIS)* 40(2):1–26
76. Zheng Y, Jiang YG, Xue X (2012) Learning hybrid part filters for scene recognition, *European conference on computer vision*. Springer, Berlin, pp 172–185
77. Zhou B, Khosla A, Lapedriza A et al (2016) Places: an image database for deep scene understanding, [arXiv:1610.02055](https://arxiv.org/abs/1610.02055)
78. Zhu Q, Zhong Y, Liu Y et al (2018) A deep-local-global feature fusion framework for high spatial resolution imagery scene classification. *Remote Sens* 10(4):568

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.