MULTI-SCALE CROSS-FUSION FINE-GRAINED NETWORK FOR IDENTIFYING INVASIVE PLANTS

Hang Sun,* Yuting Zang,* Lu Wang,* Shun Ren,* Xidong Wang,* and Xiaolin Chen**

Abstract

The invasion of alien plants has resulted in detrimental impacts on the ecological, economic, and societal aspects. The current fine-grained identification methods have not considered features of different scales, which is crucial for extracting more discriminative features. Additionally, the presence of small inter-class differences and large intra-class differences in fine-grained classification tasks significantly heightens the challenges of accomplishing accurate fine-grained classification. To address these issues, we have designed a dual-branch cross-fusion fine-grained network to integrate different-scale features and enhance the discriminative ability of deep learning features. Specifically, we have developed a multi-scale cross-fusion attention module to fuse features of different scales and retain the most important areas for classification and recognition. Moreover, we have employed a simple and effective centre loss on the dual-branch network to obtain deep features with key learning objectives, namely, inter-class distribution and intra-class compactness. Experimental results on the iNat2021-Plants, iNat2018-Plants, and FGVC-Aircraft datasets demonstrate that the proposed method achieves recognition accuracies of 76.8%, 73.8%, and 93.6%, respectively. This indicates that the method is capable of more precise fine-grained recognition and provides new insights for AI-assisted invasive plant identification systems.

Key Words

Fine-grained recognition, multi-scale cross-fusion, centre loss, invasive plant identification

1. Introduction

Alien species invasion refers to the introduction of non-native species from other regions, either naturally or through human activities, that poses threats and cause harm to ecosystems, habitats, and species. The high similarity among alien invasive plants make the use of a fine-grained algorithm highly effective in identifying more discriminative regions from similar images. Consequently, this recognition method finds widespread application in the study of alien invasive plants. Figure 1 illustrates the remarkable resemblance in the appearance, shape, and colour distribution among various plant categories, leading to minimal differences between classes. Furthermore, within the same category, plants exhibit noticeable intra-class variations at different growth stages, primarily due to changes in flower colour and morphology.

Fine-grained classification aims to differentiate specific subclasses within broader categories, such as different subclasses of plants [inaturelist2018] [1], different types of flowers [flowers102] [2], and different types of birds [CUB-200] [3]. Fine-grained recognition, in comparison to general classification, enables more accurate differentiation among different types of invasive plants, thus making it highly relevant in invasive plant management research and practice. Moreover, the subtle variations within the same class, observed across different subclasses, pose additional challenges to fine-grained classification in contrast to traditional classification methods.

Earlier fine-grained localisation methods commonly relied on bounding boxes and local annotation information. However, the use of candidate bounding boxes required larger bounding boxes, resulting in potential confusion due to the inclusion of more foreground objects. This time-consuming and expensive approach is not widely used in practice. Consequently, the research focus in fine-grained recognition has gradually shifted towards training models that primarily rely on weakly supervised methods [4]–[6]. By relying primarily on locating more discriminative regions for classification, weakly supervised methods play a significant role. However, these models rarely discuss the effective integration of information from different granularities to improve classification accuracy.

The fusion of information from different granularities aids to mitigate the impact of large intra-class variations. Invasive plants exhibit different morphological changes

* Hubei Engineering Technology Research Center for Farmland Environmental Monitoring, China Three Gorges University, Yichang 443002, China; e-mail: sunhang@ctgu.edu.cn; 202108540021116@ctgu.edu.cn; wanglul@ctgu.edu.cn; renshun@ctgu.edu.cn; xdwang@ctgu.edu.cn
** Yichang Agricultural Ecology and Resource Conservation Station, China; e-mail: 106117549@qq.com
Corresponding author: Lu Wang

Recommended by Jiankun Wang
(DOI: 10.2316/J.2024.206-1123)

during different growth cycles. Merely recognising distinctive features is often insufficient, and this often requires combining the growth morphology of plants at different stages and the overall shape of the plant to achieve more accurate plant identification. Additionally, the distinct inter-class differences and subtle intra-class differences in fine-grained tasks necessitate the learned features to be both separable and discriminative. However, the softmax loss [7] solely promotes feature separability, potentially leading to suboptimal effectiveness in fine-grained tasks.

Based on the above analysis, this paper introduces cross top-K attention (CTKA), a multi-scale cross-attention fusion module based on the analysis conducted above. CTKA is built upon the CrossViT model for feature extraction. CTKA combines features from two scales, guiding the network to fuse varied granularities and learn enriched local and detailed features. CTKA achieves adaptive selection of contribution scores by masking out elements with lower weights. This process retains the most important local regions while removing irrelevant background regions, thus filtering out redundancies and selecting distinctive key areas Furthermore, the dual-branch structure incorporates the central loss. This involves the construction of a unified category center for clustering, which aims to mitigate the problem of large disparities between features belonging to the same class and the proximity of features belonging to different classes.

The main contributions of this study are as follows:

1. We introduce a multi-scale cross-attention fusion module known as CTKA. This module merges feature information from various scales and utilises mask operations to select important local regions.
2. In our dual-branch structure, we incorporate centre loss and construct unified clustering centers for both the coarse-grained and fine-grained branches. This approach is designed to minimise the distances between features belonging to the same class while maximising the distances between features belonging to different classes by bringing them closer to their respective centre points.
3. We conduct comprehensive experiments on two extensive plant datasets, namely, iNat2021-Plants, iNat2018-Plants, and FGVC-Aircraft to showcase the efficacy of our method. Additionally, through visualisation results, we demonstrate how our network model accurately identifies and localises crucial local regions, thereby enhancing our understanding of its ability to make accurate predictions.

2. Related Work

In this section, we will briefly review previous works in three aspects: multi-scale fusion, attention mechanisms, and vision transformer (ViT).

2.1 Multi-Scale Fusion

Multi-scale fusion is a common practice in computer vision, widely used in tasks, such as object detection, semantic segmentation, and image dehazing. In fine-grained recognition, PMG [8] employs a progressive training strategy, incorporating additional network layers in each iteration to refine the existing smaller-scale features. Training images are constructed through block mixing and stitching. The original image is shuffled using a jigsaw generator according to various specified scales, and training occurs in stages. Outputs are obtained after each layer of the backbone network, with different training processes corresponding to different token sizes. Parameters are updated at each stage. In RA-CNN [9], the network is divided into three subnetworks, all sharing the same structure but with distinct parameters. Convolutional features from the preceding subnetworks pass through an attention proposal network to capture region attention. The attention regions are subsequently scaled and interpolated to serve as inputs for the next subnetwork, generating convolutional features recursively for all three networks, which are then fused.

2.2 Attention Mechanism

The key to fine-grained recognition methods lies in finding critical local regions [10], [11]. Attention-based methods can automatically detect the discriminative regions of an image through self-attention mechanisms. This approach eliminates the need for manual annotation of discriminative regions, making learning attention distinctions one of the most popular directions. Currently, various attention mechanisms are also applied in fine-grained methods [12]. MA-CNN [13] consists of convolution, channel grouping, and component classification sub-networks. Generally, each feature channel corresponds to a type of visual pattern, and the network clusters spatially correlated patterns. Then, it weights these patterns into the attention maps of neighbouring externally peak-responding channels. Different response positions generate different attention maps, and through cropping, it further extracts target boxes of different components. The classification network categorises the image based on the features of these components. Finally, through two optimised loss functions, the network is forced to learn critical local regions and enhance the learning of more locally fine-grained feature areas through mutually reinforcing optimisation. However, the number of generated component attentions is limited, which restricts the flexibility of the network and hinders the learning of more component areas. TASN [14] expresses rich fine-grained features through a convolutional neural network, which includes a trilinear attention module for fine-grained localisation, an attention-based sampling module for detail extraction, and a feature distiller for detail optimisation. By employing weight sharing and feature preservation strategies, partial features are distilled into a global feature, enabling the learning of hundreds of proposed boxes. This is a common teacher-guided approach. However, variations in lighting, object background, and pose make it challenging for the network to learn a consistent attention map. Moreover, if each component is recognised individually, this method does not offer significant efficiency advantages.

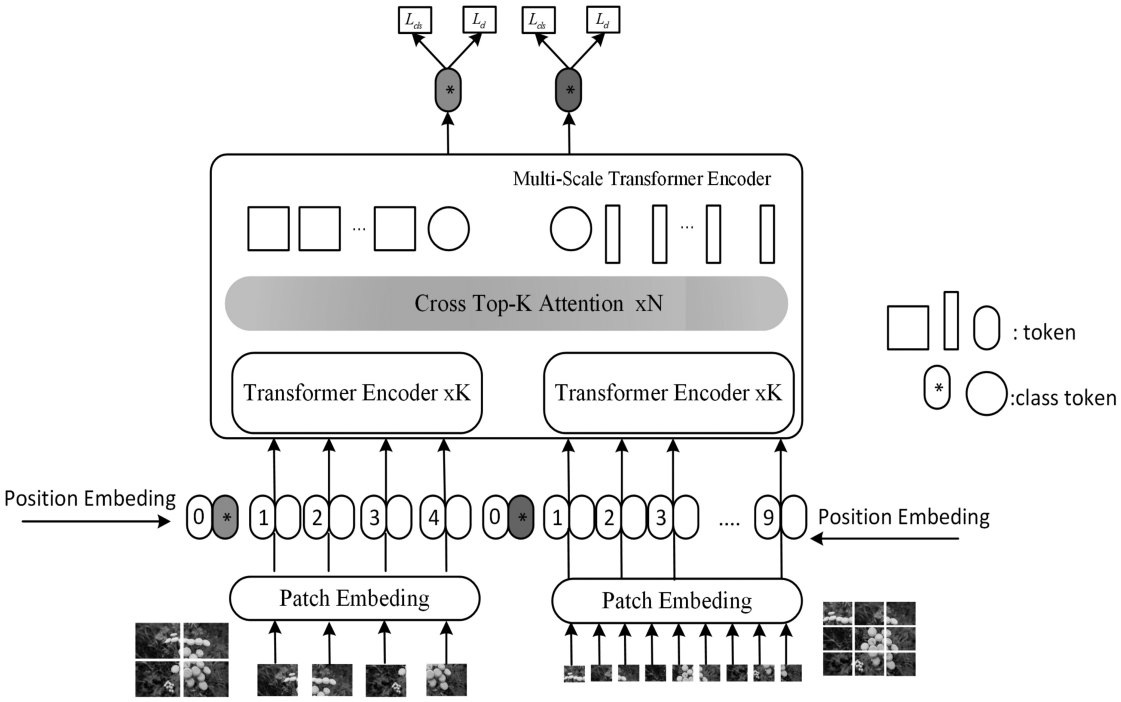

Figure 1. Overall framework diagram.

2.3 Vision Transformer

Due to the powerful expressive capabilities of transformers [15] in the field of natural language processing (NLP), they have gradually been migrated to the domain of computer vision, giving rise to the application of ViT [16]. Transformers have been widely applied in tasks, such as image retrieval [17]–[19], image dehazing [20]–[22], and face recognition [23], [24]. However, their application in fine-grained tasks is not as common. Considering that transformers' self-attention mechanism aggregates and weights information from all patches onto the classification token, and given their superior performance on large-scale datasets, they are well-suited for fine-grained recognition tasks. TransFG [25] is the first method to apply transformer to fine-grained visual classification, providing an alternative to dominant CNN backbones with RPN model designs. By utilising multi-head self-attention (MSA) mechanism, TransFG proposes a part selection module to calculate discriminative regions and remove redundant information. The integrated tokens and global classification tokens are connected and input into the last transformer layer. In addition, to enlarge the distance between feature representations of different categories and reduce the distance between feature representations of the same category, contrastive loss is introduced. Since the block operations of ViTs may result in the loss of local information, which is crucial for fine-grained tasks, and selecting critical regions for precise classification is of great importance to our work, we attempt to design a CTKA mechanism for multi-scale fusion in the dual-branch ViT. This is done to pinpoint key regions in the image and enhance the network's classification accuracy.

3. Method

In this paper, we adopt CrossVit as the baseline model. CrossVit is a dual-branch image classification algorithm based on the ViT model. The model consists of K multi-scale encoders, with each encoder comprising a coarse-grained branch and a fine-grained branch. Each branch takes image patches of different sizes as input. The two branches cross-fuse features of different granularities, and the category scores are computed using two classification tokens.

Our design is based on the baseline model and incorporates the CTKA attention module and introduces centre loss. The CTKA is introduced in the multi-scale fusion module to filter out non-essential background information and focus on subtle differences in local regions, thereby preserving local regions with more informative details for fine-grained recognition. The centre loss, introduced for handling challenging samples, helps to increase the feature distance between different categories of images and decrease the feature distance within the same category, thus making samples of the same category closer in the same feature space. This section provides a detailed description of the overall framework of the proposed method, as shown in Fig. 1.

3.1 Basic Architecture of ViT

The ViT divides an image x with a size of $x \in H \times W \times C$ into image patches of size $P \times P \times C$, resulting in $N = HW/P^2$ image patches with dimensions $N \times P \times P \times C$. Each image patch is then flattened, resulting in a dimension of $N \times P^2 \times C$, where W is the height of the input

image, H is the width of the input image, N is the length of the input sequence, C is the number of input channels, and P is the size of the image patch. The image patches are then embedded, and a linear transformation is applied to each flattened image patch vector to output token embeddings $x_p^i E \in R^D$, where $(i = 1, 2, \ldots, L)$. Among them, $E \in R^{p^2 \times D}$. In addition, an additional category token x_{class} is added for category prediction, which is inputted into the transformer encoder together with other tokens for feature extraction. To maintain the coherence between images, corresponding positional vectors $E_{\text{POS}} \in R^{(L+1)\times S}$ need to be added to the embedding of image blocks, resulting in the input token $x_0 \in R^{(L+1)\times D}$. The obtained tokens are processed by stacked transformer encoders, and finally classified using the classification token. The transformer encoder consists of a MSA layer, layer normalisation (LN) layer, and feed-forward network (FFN). During the normalisation layer of the input sequence, residual connections are used for each LN layer. The input of vit and the processing of the k-th block are shown in (1)–(3):

$$x_0 = \left[x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^l E\right] \tag{1}$$

$$y_k = x_{k-1} + \text{MSA}\left(\text{LN}\left(x_{k-1}\right)\right) \tag{2}$$

$$x_k = y_k + FFN\left(\text{LN}\left(y_k\right)\right) \tag{3}$$

3.2 Cross Top-K Attention

In this section, we will introduce the basic idea of the proposed CTKA module, which takes the query-key-value matrices generated by the transformer encoder from the coarse-grained branch and fine-grained branch as input, and fuses different scales of image features through permutation classification heads. The fused classification token is then returned to the original branch as the output of the CTKA module. Compared to the original classification head, the permutation classification token has learned abstract information in its own branch patches, so interacting with the patches tokens of other branches contributes to, including more information at different scales. After fusion and interaction with tokens from other branches, the classification token interacts again with its own branch tokens, enriching the representation of each patch token. The core of fine-grained recognition tasks is to locate local features with discriminative and diverse characteristics, so the fusion of image features of different granularities is crucial for learning and locating key local regions.

Figure 2 depicts the fundamental concept of our CTKA module, which involves the integration of the classification token from the coarse-grained branch and the patch tokens from the fine-grained branch. Subsequently, the classification head, containing abstract information at various scales, is fed back to the coarse-grained branch for training. In the subsequent sections, we predominantly utilise the coarse-grained branch as an illustrative example to elucidate the multi-scale cross-attention module, while the fine-grained branch follows an analogous process. In particular, subsequent to the dual-branch network performing MSA using the transformer

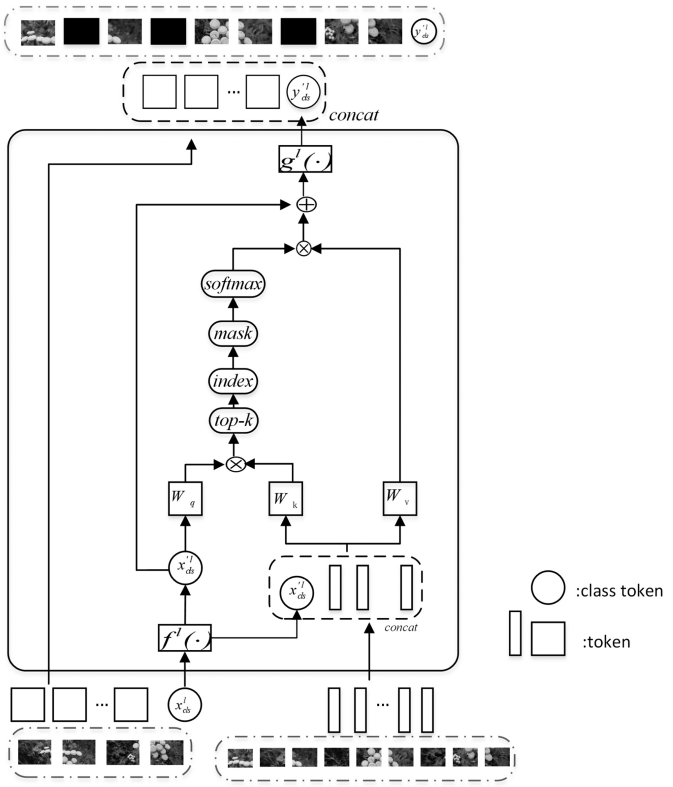

Figure 2. Structure diagram of the CTKA module.

encoder, the coarse-grained branch generates query-key-value matrices W_L^Q, W_L^K, W_L^V, and the fine-grained branch generates query-key-value matrices W_S^Q, W_S^K, W_S^V. The query-key-value matrices of the dual-branch network serve as the input for the CTKA module. Initially, we extract the classification head of the coarse-grained branch as the query key, ensuring its dimensionality aligns with the image tokens from the fine-grained branch, and subsequently concatenate them, as denoted in (2).

$$x^{'l} = \left[\ f^l\left(x_{\text{cls}}^l\right) \middle\| x_{\text{patch}}^s\right] \tag{4}$$

In this equation, $f^l(\cdot)$ represents the projection function. x_{cls}^l is the class token extracted from the coarse-grained branch, and x_{patch}^s represents the image block token from the fine-grained branch. The tokens are then concatenated and subjected to cross-attention fusion operations, as shown in (5)–(7):

$$Q = x_{\text{cls}}^{'l} W_S^Q, K = x^{'l} W_S^K, V = x^{'l} W_S^K \tag{5}$$

$$\delta = \frac{QK^T}{\sqrt{\frac{c}{h}}} \tag{6}$$

$$A = \text{softmax}\left(\theta_k \delta\right) \tag{7}$$

Where Q, K are all learnable parameters, c, h are the dimensional parameters of the classification head and individual head. Based on the similarity between Q and K, the element area with top-k scores is retained, and for other elements with scores lower than top-k, their probabilities are replaced with 0 using a scatter function. θ_k is a learnable

top-k selection operator, as shown in the (8):

$$[\theta_k(S)]_{\mathrm{ij}} = \begin{cases} S_{\mathrm{ij}} \;\; S_{\mathrm{ij}} \in \mathrm{top}-k\,(\mathrm{row}\;j) \\ 0 \quad \mathrm{otherwize.} \end{cases} \tag{8}$$

Then, the top-k values within the selected range are normalised, while computing the softmax to obtain weights A. Finally, the weights are multiplied with matrix V, as shown in (9):

$$\mathrm{CA}\left(X^{'l}\right) = \mathrm{AV} \tag{9}$$

Where $V = W_v X^{'l}$, after cross-attention module, we do not use feedforward network. Therefore, the output z^l of a given multi-scale cross fusion module can be represented as (11).

$$y^l_{\mathrm{cls}} = f^l\left(x^l_{\mathrm{cls}}\right) + \mathrm{MSA}\left(\mathrm{LN}\left(x_{k-1}\right)\right) \tag{10}$$

$$z^l = \left[\; g^l\left(y^l_{\mathrm{cls}}\right) \middle\| \, x^l_{\mathrm{patch}}\right] \tag{11}$$

Among them, $g^l(\cdot)$ is the projection function in the alignment dimension. In the Section 4.2 experimental analysis, we further prove the effectiveness of CTKA module through ablation experiments. Finally, the classification tokens that have fused information from different scales are returned to their original branches as the output of the CTKA module. Since the classification tokens have learned abstract information in their own branches after multiple iterations of transformer encoder, they can learn feature information at different scales during interaction with other branches, thus more effectively fusing features at different scales.

3.3 DC Loss

Centre loss [26] was originally designed as a loss function to address the issue of high similarity between different class features in face recognition. Similar problems also exist in fine-grained recognition tasks, where images within the same category exhibit significant variations, while the differences between images from different categories are relatively small. Therefore, in this study, the centre loss is applied to fine-grained recognition tasks to enhance the network's classification ability. This is achieved by making samples from the same category more compact in the feature space while increasing the feature distance between different categories. In classification tasks, cross-entropy loss is commonly used to calculate the classification loss, as shown in (12):

$$L_{\mathrm{cl}} = L_{\mathrm{ce}}\left(\mathrm{softmax}\left(\mathrm{FC}\left(x_{\mathrm{class}}\right)\right), y\right) \tag{12}$$

In this formula, FC is a fully connected layer that maps x_{class} to the label space, softmax transforms the output of FC into a probability distribution over the classes, and y represents the true class label of the sample. L_{ce} (cross-entropy) is used to calculate the error between the predicted probability values and the true label values. However, as the number of required classes increases, the linear matrix of the classification layer also grows. Cross-entropy loss can only reduce the differences between different class features but it cannot effectively increase the differences within classes. Therefore, we incorporate centre loss into fine-grained classification tasks to address the issues of feature diversity within the same class and subtle differences between different class images.

Due to the high confusion between images of invasive plant species and the small differences between images of different categories and large differences between images of the same category, it is challenging for cross-entropy loss alone to optimise the network to an ideal state. Centre loss addresses this issue by increasing the similarity between deep features and their class centres, making them more compact in the feature space. This can be represented by (13):

$$L_{\mathrm{cl}} = \frac{1}{2}\left\|x_{\mathrm{class}} - C\right\|^2 \tag{13}$$

In this formula, C represents the class centres of the samples, which have the same dimension as the x_{class} feature. As the centre loss gradually decreases, the feature mapping of each sample gets closer to its corresponding class centre. The class centres are continuously updated as the deep features change. In this paper, by separately constructing clustering centres on the dual-branch network, the dual-branch centre loss is averaged with weights, resulting in the final center loss of the network, denoted as L_{dc}. This can be represented by (14):

$$L_{\mathrm{dc}} = \frac{1}{2}\left(L_{cl1} + L_{cl2}\right) \tag{14}$$

In the formula, L_{cl1} represents the centre loss of the coarse-grained branch, L_{cl2} represents the centre loss of the fine-grained branch. Therefore, the overall loss function of the network is represented by (15):

$$L = L_{\mathrm{cl}} + \;\beta * L_{\mathrm{dc}} \tag{15}$$

Where β is a hyperparameter which is set to 0.001 in our experiments to give more weight to the classification loss in the optimisation process.

4. Experiment

4.1 Experimental Background

4.1.1 Dataset

Since this paper mainly focuses on the recognition of invasive species in Yichang City, Hubei Province, a fine-grained plant dataset is chosen. iNat2021-Plants is a "mini" training dataset derived from iNaturalist2021 [27], which is one of the widely used benchmarks in the field of fine-grained image analysis. The goal of iNaturalist2018 is to advance the latest technologies in image classification and detection of wildlife data with a large amount of imbalanced, fine-grained, and diverse categories. We selected plant sample data iNat2018-Plants from it for model validation. Extensive experiments have demonstrated the effectiveness of methods on this dataset.

Table 1
Distribution Information of the Datasets

Datasets	Category	Train	Test
iNat2021-Plants	10,000	500,000	118,800
iNat2018-Plants	2,917	118,800	8,751
FGVC-Aircraft	100	6,667	3,333

FGVC-Aircraft is an aircraft classification dataset, commonly used for fine-grained classification tasks. We use this dataset to evaluate the robustness of our model. The training set contains 6,667 images, and the test set contains 3,333 images, with a total of 100 aircraft categories. The main information of the datasets will be presented in Table 1 below:

4.1.2 Implementation Details

The fine-grained recognition method proposed in this paper is based on the PyTorch framework and utilises the NVIDIA GeForce RTX3090. The network is trained using the datasets mentioned in Table 1. A dual-branch structure is employed to construct the dual-branch fine-grained network. The input size for the coarse-grained branch is 224×224. The image is divided into 16×16 image blocks, resulting in N, P, and C values of 196, 16, and 384, respectively. The input size for the fine-grained branch is 240×240. The image is divided into 12×12 image blocks, resulting in N, P, C values of 400, 12, and 192, respectively. The training utilises a batch size of 16 and the SGD optimiser with a momentum of 0.9 and a weight decay factor of 0.0005. The initial learning rate is set to 0.0001 and a warm-up method is employed. The training process consists of 300 epochs, incorporating cosine annealing for updates.

4.2 Experimental Results and Analysis

4.2.1 Comparative Experiment

To validate the correctness and effectiveness of our method, we compare it with current state-of-the-art fine-grained recognition methods. The accuracy on the iNat2021-Plants and iNat2018-Plants datasets are shown in Table 2. The accuracy on the FGVC-Aircraft datasets are shown in Table 3.

As shown in Table 2, the proposed method in this paper outperforms the other six methods in terms of recognition results. Our algorithm utilises ViT_16 as the backbone network. The accuracy improvements on the iNat2021-Plants dataset compared to TASN, PMG, PCA-Net, P2P-Net, MHEM, and ViT are 4.1%, 1.5%, 3.8%, 2.0%, 3.7%, and 8.2%, respectively. On the iNat2018-Plants dataset, the accuracy improvements are 4.5%, 4.9%, 3.2%, 1.9%, 3.4%, and 7.5%, respectively. These results demonstrate the rationality of the proposed improvement method in this paper and also validate the effectiveness of our approach.

Table 2
Comparison of Different Fine-Grained Algorithms

Method	Backbone	iNat2021-Plants	iNat2018-Plants
$TASN(CVPR19)$[14]	Resnet50	72.7	69.3
$PMG(ECCV20)$[8]	Resnet50	75.3	68.9
$PCANet(VCIP21)$[28]	Resnet50	73.0	70.6
$P2PNet(CVPR22)$[29]	Resnet50	74.8	71.9
$MHEM(TNNLS22)$[30]	Resnet50	73.1	70.4
$ViT(ICLR21)$[16]	ViT_16	68.6	66.3
Ours	**ViT_16**	**76.8**	**73.8**

Table 3
The Comparative Results on FGVC-Aircraft

Model	Backbone	Classification accuracy (%)
PMG (20-ECCV)	ResNet50	93.4
PCA-Net (21-VCIP)	ResNet101	92.8
MHEM (22-TNNLS)	ResNet50	92.9
P2P-NET (22-CVPR)	ResNet50	94.2
$MetaFormer$ $(22-CORR)$[31]	MetaFormer-2	92.8
Ours	**ViT_16**	**93.6**

The outcomes reveal that our approach outperforms PCA-Net and PMG even on public datasets, thereby demonstrating its capability to acquire more refined discriminative information. "Metaformer: A unified meta framework for fine-grained recognition" is a unified meta framework designed to enhance fine-grained recognition tasks, showcasing commendable performance in FGVC-Aircraft classification. The integration of multi-scale feature structures enriches the network and fosters diverse feature learning, leading to superior performance compared to Metaformer. Moreover, in comparative experiments utilising various backbone networks and fine-grained classification methods, our approach exceeds these cutting-edge methods, yielding the optimal performance on the FGVC-Aircraft dataset.

4.2.2 Ablation Experiment

We conducted ablation experiments on the iNat2021-Plants, iNat2018-Plants, and FGVC-Aircraft datasets. To further validate the effectiveness of the proposed method. As shown in Table 4, our method achieved a 0.8% and 1.2% improvement on the iNat2021-Plants and iNat2018-Plants datasets, respectively, compared to the baseline model. This indicates that CTKA selects the most effective regions from a large number of background areas, forcing

Table 4
Ablation Experiments

	iNat2021-Plants	iNat2018-Plants
Crossvit	74.5	71.4
CTKA+Crossvit	75.3	72.6
dcloss+CrossVit(single)	75.4	72.6
dcloss+CrossVit(double)	75.8	72.8
dcloss+CTKA	76.8	73.8

Table 5
Ablation Experiments on FGVC-Aircraft

Model	Classification accuracy (%)
Crossvit	91.5
CTKA+Crossvit	92.7
dcloss+CrossVit(single)	92.2
dcloss+CrossVit(double)	92.8
dcloss+CTKA	93.6

the model to learn useful information from these selected regions for fine-grained classification tasks. The results in Table 4 confirm the effectiveness of our method.

To evaluate the performance of the centre loss function, we conducted training on the iNat2021-Plants, iNat2018-Plants and designed two different ablation experiments. Table 4 summarises the classification accuracy, precision in fine-grained recognition, and overall accuracy of the two experiments on the iNat2021-Plants and iNat2018-Plants datasets. Centre loss improved the accuracy in all aspects. Particularly, the effect was most significant on the iNat2021-Plants dataset. In experiment 1, after adding centre loss on a single branch, the accuracy increased by 0.9%. In experiment 2, after adding centre loss on both branches, the accuracy increased by 1.3%. There was also a noticeable effect on the iNat2018-Plants dataset. In experiment 1, after adding centre loss on a single branch, the accuracy increased by 1.2%. In experiment 2, after adding centre loss on both branches, the accuracy increased by 1.4%.

To evaluate the effectiveness of our proposed CTKA and dcloss, we conducted an ablation study on the FGVC-Aircraft dataset, as shown in Table 5. Apart from the proposed modules, both the baseline model and our method utilised the same data augmentation methods and hyperparameter settings, including learning rate, training epochs, and other parameters. This ensured the consistency of the experimental setup and facilitated fair comparisons between the proposed modules and the baseline. From the table, it can be observed that our method demonstrated improved accuracy compared to the baseline on the iNat2018-Plants and FGVC-Aircraft datasets.

Table 7
The Impact of β on Classification Accuracy (%) on the iNat2018-Plants and iNat2018-Plants Datasets

Method	iNat2021-Plants	iNat2018-Plants
Crossvit	74.5	71.4
dcloss+CrossVit(β=0.0001)	76.4	73.6
dcloss+CrossVit(β=0.001)	76.8	73.8
dcloss+CrossVit(β=0.01)	76.7	73.5
dcloss+CrossVit(β=0.1)	75.1	73.2
dcloss+CrossVit(β=1)	58.5	54.3

Table 8
The Impact of β on Classification Accuracy on the FGVC-Aircraft Datasets

Method	Classification accuracy (%)
Crossvit	91.5
dcloss+CrossVit(β=0.0001)	93.3
dcloss+CrossVit(β=0.001)	93.6
dcloss+CrossVit(β=0.01)	93.5
dcloss+CrossVit(β=0.1)	92.9
dcloss+CrossVit(β=1)	85.5

4.2.3 The Impact of β on Classification Accuracy

By establishing a class centre for each category, the objective of the centre loss is to minimise the distance between each sample and the class centre, thereby significantly enhancing intra-class feature differences. Introducing the centre loss based on the cross-entropy loss allows the model to focus on intra-class loss while learning inter-class features, thereby improving the balanced representation of inter-class and intra-class features. β was set as 0.0001, 0.001, 0.01, 0.01, 0.1, and 1.

we conducted experiments on iNat2018-Plants, iNat2018-Plants, and FGVC-Aircraft dataset, the experimental results are as follows:

As can be seen from the Tables 7 and 8, with the continuous increase of over parameter β, the accuracy rate shows a trend of gradual decline. We believe that the increase of β will interfere with the optimisation direction of the model to a certain extent, resulting in the loss function failing to converge to the minimum value in a short period of time. Therefore, after several experimental analyses, to improve the accuracy rate of model classification, we set β to 0.001 in our experiment.

4.3 Visualisation Analysis

As shown in Fig. 3, the following images display the visual effects of six selected images from the two datasets. Among

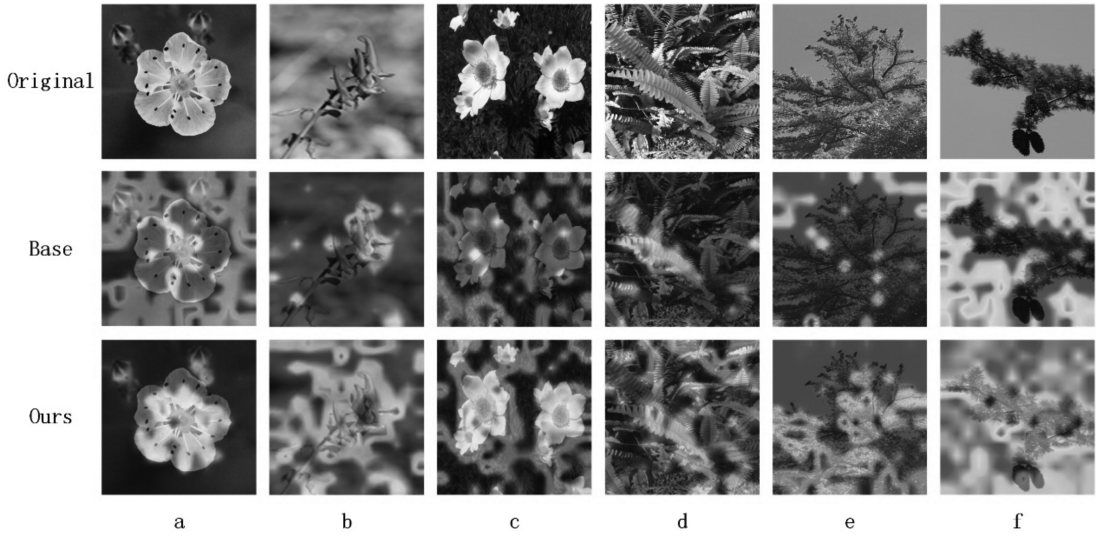

Figure 3. Visualisation results.

them, *a*, *b*, *c* are images from the iNat2018-Plants dataset, primarily focusing on flowers as the main objects. *d*, *e*, *f* are images from the iNat2021-Plants dataset, primarily focusing on trees as the main objects. The images in the first, second, and third rows represent the original input images, attention heatmaps of the baseline network, and attention heatmaps of the improved network, respectively. From the images, it can be observed that our method, by fusing features from different scales, can locate more distinctive and crucial local regions, while disregarding the background noise. This further enhances the fine-grained recognition accuracy of invasive plant identification methods.

5. Conclusion

In the research of invasive plant identification, the subtle differences between different category images and the diversity within the same category images remain highly challenging. In this work, we proposed a multi-scale fusion attention mechanism to address the problem of localising crucial regions in fine-grained tasks. By reconstructing the similarity of pixel pairs between queries and keywords, we aimed to preserve the most important components, thus improving the method's performance. The results show that the CTKA module achieves adaptive selection of top-k contribution scores by masking unnecessary elements with lower attention weights, effectively localising crucial local regions. The centre loss enhances the similarity between deep features and their class centres, making the distances in the feature space tighter. The proposed method achieved promising performance on two plant datasets, and the experimental results demonstrate that it improved the classification accuracy of the standard ViT at different granularities.

5.1 Contribution

To strengthen the prevention and control system for invasive alien species and enhance the comprehensive prevention and control capabilities, it is effective to improve the accurate identification of invasive alien plants by utilising artificial intelligence platforms, deep learning algorithms, and big data analysis techniques. This can effectively solve the problem of difficulty in identification in the current prevention and control work for invasive alien species.

5.2 Limitations

Due to the adoption of a dual-branch structure in this article, as well as the fact that the transformer network will partition the images, this will exacerbate the generation of redundant data, resulting in the problem of the network model having excessively large network parameters.

Acknowledgement

This work was supported by the Natural Science Research Project of Yichang (Grant No. A23-2-018).

References

[1] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A.Shepard, H. Adam, P. Perona, and S. Belongie, The inaturalist species classification and detection dataset, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, 8769–8778.

[2] Y. Wu, X. Qin, Y. Pan, and C. Yuan, Convolution neural network based transfer learning for classification of flowers, Proc. *2018 IEEE 3rd International Conf. on Signal and Image Processing (ICSIP)*, Shenzhen, 2018, 562–566.

[3] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Be- Longie, The caltech-UCSD birds-200-2011 dataset, 2011.

[4] Y. Wang, V.I. Morariu, and L.S. Davis, Learning a discriminative filter bank within a CNN for fine-grained recognition, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* Salt Lake City, UT, 2018, 4148–4157.

[5] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, Learning to navigate for fine-grained classification, *Proc. of the European Conf. on Computer Vision (ECCV),* Cham, 2018, 420–435.

[6] Y. Chen, Y. Bai, W.W. Zhang, and T. Mei, Destruction and construction learning for fine-grained image recognition, *Proc.*

of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, CA, 2019, 5157–5166.

[7] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, Learning representations by back-propagating errors, *Nature*, *323*(6088), 1986, 533–536.

[8] R. Du, D. Chang, A.K. Bhunia, J. Xie, Z. Ma, Y.Z. Song, and J. Guo, Fine-grained visual classification via progressive multi-granularity training of jigsaw patches, *Proc. European Conf. on Computer Vision, Cham: Springer International Publishing*, 2020, 153–168.

[9] J. Fu, H. Zheng, and T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, 2017, 4438–4446.

[10] H. Tang, J. Liu, S. Yan, R. Yan, Z. Li, and J. Tang, M^3Net: Multi-view encoding, matching, and fusion for few-shot fine-grained action recognition, *Proc. of the 31st ACM International Conf. on Multimedia*, Ottawa ON, 2023, 1719–1728.

[11] Z. Zha, H. Tang, Y. Sun, and J. Tang, Boosting few-shot fine-grained recognition with background suppression and foreground alignment, *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(8), 2023, 3947–3961.

[12] H. Tang, C. Yuan, Z. Li, and J. Tang, Learning attention-guided pyramidal features for few-shot fine-grained recognition, *Pattern Recognition*, *130*, 2022, 108792.

[13] H. Zheng, J. Fu, T. Mei, and J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, *Proc. of the IEEE International Conf. on Computer Vision,* Venice, 2017, 5209–5217.

[14] H. Zheng, J. Fu, Z.J. Zha, and J. Luo, Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition, *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition,* 2019, 5012–5021.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. Gomez, Attention is all you need, *Advances in Neural Information Processing Systems*, *30,* 2017.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, *arXiv:2010.11929*.

[17] K. Higuchi and K. Yanai, Patent image retrieval using transformer-based deep metric learning, *World Patent Information*, *74*, 2023, 102217.

[18] A. Baldrati, M. Bertini, T. Uricchio, and A. Bimbo, Composed image retrieval using contrastive learning and task-oriented CLIP-based features, *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[19] A. Venkataramanan, M. Laviale, and C. Pradalier, Integrating visual and semantic similarity using hierarchies for image retrieval, *Proc. International Conf. on Computer Vision Systems, Cham: Springer Nature Switzerland*, 2023, 422–431.

[20] H. Sun, Z. Luo, D. Ren, and L. Zhang, Partial siamese with multiscale Bi-codec networks for remote sensing image haze removal, *IEEE Transactions on Geoscience and Remote Sensing, 61,* 2023, 1–16. DOI: 10.1109/TGRS.2023.3321307.

[21] H. Sun, B. Li, Z. Dan*, W. Hu *, B. Du, W. Yang, and J. Wan, Multi-level feature interaction and efficient non-local information enhanced channel attention for image dehazing, *Neural Networks*, *163,* 2023*,* 10–27.

[22] H. Sun, Y. Zhang*, P. Chen, Z. Dan*, S. Sun, J. Wan, and W. Li, Scale-free heterogeneous cycleGAN for defogging from a single image for autonomous driving in fog, *Neural Computing and Applications*, *35,* 2023*,* 3737–3751.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, Facenet: A unified embedding for face recognition and clustering, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, 2015, 815–823.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019, 4690–4699.

[25] J. He, J.N. Chen, S. Liu, A. Kortylewski, C. Yang, Y.Bai, and C. Wang, Transfg: A transformer architecture for fine-grained recognition, *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(1), 2022, 852–860.

[26] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, A discriminative feature learning approach for deep face recognition, *Proc. Computer Vision–ECCV 2016: 14th European Conf., Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. Springer International Publishing*, 2016, 499–515.

[27] J.C. Su and S. Maji, The semi-supervised inaturalist challenge at the FGVC8 workshop, 2021, *arXiv:2106.01364*.

[28] T. Zhang, D. Chang, Z. Ma, and J. Guo, Progressive co-attention network for fine-grained visual classification, *Proc. 2021 International Conf. on Visual Communications and Image Processing (VCIP)*, IEEE, Munich, 2021, 1–5.

[29] X. Yang, Y. Wang, K. Chen, Y. Xu, and Y. Tian, Fine-grained object classification via self-supervised pose alignment, *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition,* 2022, 7399–7408.

[30] Y. Liang, L. Zhu, X. Wang, and Y Yang, Penalizing the hard example but not too much: A strong baseline for fine-grained visual classification, *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[31] Q. Diao, Y. Jiang, B. Wen, J. Sun, and Z. Yuan, Metaformer: A unified meta framework for fine-grained recognition, 2022, *arXiv:2203.02751*.

Biographies

Hang Sun received the Ph.D. degree from the School of Computer Science, Wuhan University, China, in 2017. He used to work as a Senior Engineer with Huawei Technologies Co., Ltd., responsible for the research and application of computer vision. He is currently an Associate Professor with the College of Computer and Information Technology, China Three Gorges University, China. His main research interests include computer vision and image generation. His works have been published in premier computer vision journals and conferences, including IEEE Transactions on Image Processing, IEEE Transactions on Cybernetics, *Neural Networks*, IEEE Geoscience and Remote Sensing Letters, *Neural Computing & Applications*, *Science China Information Sciences*, *Chinese Journal of Electronics*, Pacific-Rim Conference on Multimedia, International joint Conference on Neural Networks, etc.

Yuting Zang is currently pursuing the postgraduation degree with the College of Computer and Information Technology, China Three Gorges University.

Lu Wang received the masters degree from Harbin Institute of Technology. She is currently pursuing the doctoral degree with the College of Electrical Engineering and New Energy, China Three Gorges University, Yichang, China. Her research interests are artificial intelligence and the application of image processing in electric engineering.

Shun Ren received the Ph.D. degree from Jilin University. He is now working with the College of Computer and Information Technology, China Three Gorges University, Yichang, China. His research interests include artificial intelligence and Internet of Things.

Xidong Wang received the Ph.D. degree from Huazhong University of Science and Technology. He is now working with the College of Computer and Information Technology, China Three Gorges University. His research interests include artificial intelligence, embedded system, and sensors and systems.

Lilin Chen born in July 1979. She received the bachelor's degree and is a Senior Agricultural Technician. She works with the Yichang Agricultural Ecology and Resource Protection Station and has been engaged in agricultural technology promotion and agricultural environmental protection. Since joining the workforce, she has won three awards for scientific and technological achievements at the national, provincial, and municipal levels. She has also participated in the drafting of a local standard for Hubei Province and has published over 10 papers in national core journals and other publications.