An Interactive Deep Learning Method For Fine-grained Image Classification

Liumin Luo[1], Mingxia Wang[2], and Xiaoqing Liu[1*]

[1] School of Mechanical and Electrical Engineering, Zhoukou Normal University, Zhoukou 466000 China
[2] Department of Mechanical and Electrical Engineering, PLA Army Special Operations College, Guilin 541000, Guangxi Province, China
[*] Corresponding author. E-mail: 404736146@qq.com

Received: March 19, 2023; Accepted: Appril 24, 2024

Fine-grained image classification refers to the classification of subcategories based on the basic categories already divided. Fine-grained image classification is a very challenging research task because of the data characteristics of small inter-class differences and large intra-class differences. Based on the analysis and research of existing fine-grained image classification algorithms, a novel fine-grained image classification method based on an interactive deep learning is proposed. First, YOLOv5 is used as the backbone network to improve the classification performance, and a random elimination enhancement selection strategy is designed. The feature elimination branch and feature enhancement branch interactions promote the network to learn more relevant information and capture potential distinguishable features. Then, a global diversified module is proposed to model the feature maps of different levels to improve the ability of network comparison cues. Finally, the internal standard imprinting data set is established, and the fine-grained algorithm is applied to the authenticity identification work to realize the practical application of fine-grained image classification in natural scenes. Model training can be efficiently trained in an end-to-end manner without bounding boxes and comments. Experimental results show that the accuracy of the proposed algorithm on three fine-grained image datasets, namely, CUB-200-2011, Standford Cars and FGVC-Aircraft, reaches 90.6%, 95.9% and 95.8%, respectively.

Keywords: Fine-grained image classification; YOLOv5; interactive deep learning; feature enhancement

http://dx.doi.org/10.6180/jase.202504_28(4).0004

1. Introduction

Fine-grained image classification refers to the recognition of sub-classes under a broad category. The main task of fine-grained image classification is to distinguish some sub-categories of base classes, such as some specific categories for birds, cars and aircraft. It is a great challenge to effectively solve the problem of fine-grained image classification due to the large intra-class attitude variation, high similarity between classes and lack of annotation data. Fine-grained image classification also has a wide range of application needs in production and life, such as automatic identification of goods in the smart retail scene, identification of biological types in the field of biological protection, and identification of different models of cars in intelligent transportation systems [1–3].

Compared with traditional image classification, the difficulty of fine-grained image classification is that different categories have little difference in appearance under conditions such as similar attitude and similar viewing Angle, while the same category has great difference in appearance under conditions such as different attitude and different viewing Angle.

In order to solve these problems, many ideas have been

put forward. In the early stage, localization based methods are mainly adopted [4–6]. In addition to category labels of images, these methods also use additional annotation information such as object annotation frame and part annotation point. However, annotating the position of the object is a tedious and costly manual task, and it is also prone to manual annotation errors. Therefore, the current mainstream research focuses on the design of fine-grained models with only image-level labeling data, that is, weakly supervised fine-grained image classification methods. The key to weakly supervised fine-grained classification is how to mine subtle differences in images without a lot of manual annotation information [7]. The weak supervised fine-grained image classification mainly includes the localization of the recognition region and the learning of the recognition region features. Li et al. [8] proposed a two-level attention model: object-level attention filtered the relevant regions from the same object, and partial-level attention was used to locate the distinguished regions. Ke et al. [9] proposed a weak supervised fine-grained image classification method based on depth filter response selection. Guang et al. [10] proposed an object-part attention model for weakly supervised fine-grained image classification. The model showed that object attention could effectively locate the area where the object was located, and then located the distinguished area in the object by partial attention. Yang et al. [11] proposed a Destruction and Construction Learning (DCL) framework to realize fine-grained image recognition. In this model, the input image was first destroyed appropriately to emphasize distinguishable local details, and then reconstructed to model the semantic associations between the segmented regions. Different from previous methods, Wan et al. [12] realized sub-feature semantics by arranging the feature channels of Convolutional Neural Networks (CNNs) into different groups, and then learned fine-grained features by enhancing the sub-feature semantics of global features. Xian et al. [13] proposed a model that combined the target location module and the multi-layer bilinear pooling module. The model could accurately locate the object of interest through the object location module, and extract the identification feature representation through the multi-layer bilinear pooling module. Touvron et al. [14] proposed a channel interaction network that could model channel interactions within and across images. In a recent study, Zhu et al. [15] proposed a new Cross-layer Non-local (CNL) module based on the local module, which was used for fine-grained image recognition. The module learned multi-scale features by cross-layer association of features from a spatial perspective. However, the above methods either rely on complex component positioning modules or do not adequately consider the channel and spatial correlation of different layer features, that is, combining the characteristics of different layer features. Therefore, in this paper, convolution is used to realize the combination of different features of high and low level features, and semantic attention is obtained by associating high and low level features in the network to obtain the distinguishing region. In addition, there are no explicit detection or alignment sections, and important regional discriminant features are obtained in a multi-scale manner.

Because of the different shapes of fine-grained image objects, multi-scale features are very important for fine-grained classification. Different layers of CNN contain different feature information. High-level features contain rich semantic information, while low-level features pay more attention to detailed information such as contour, edge, color, texture and shape. However, high-level features lack information such as detail and location, and low-level features also have problems such as background confusion and semantic ambiguity. Therefore, if feature fusion is carried out at different levels [16], the advantages of multi-scale features can be exploited and a balance can be struck between high and low level features. It can be seen that the inclusion of less background noise is conducive to the acquisition of discriminative features. In order to obtain fine multi-scale discrimination features, it is necessary to remove meaningless background noise and semantic invalid features.

In addition, the challenge of fine-grained tasks also exists in the application side, and the fine-grained image databases currently used by mainstream research include dogs, airplanes, cars, and birds. Although these data sets have a certain number of categories and quality of annotation, the above data are not images obtained in real life scenes. For example, there are big differences between photos taken by mobile phones and images in the data set. As a result, the recognition technology obtained is also limited to the existing data sets, which affects the practical application of fine-grained classification tasks. Therefore, how to combine the fine-grained classification algorithm with practical application and closely serve the real life is a worthwhile work in this field.

Based on the above analysis, this paper proposes a weakly supervised random selection global diversity classification method. YOLOv5 is used as the backbone network [17], and a random elimination enhanced selection strategy is proposed in the network training stage to explore globally potentially distinguishable features by suppressing the most significant information and rewarding the most dis-

criminative part. Furthermore, the global diversification module is designed to establish the common relation of various features and improve the richness of its features. At the same time, according to the process of handbag authenticity identification, the data set of the image of the inner label stamping part of the handbag is collected and established. Through the data set and the proposed algorithm, an accurate classification model is constructed, which can screen fake products on a large scale and assist the appraiser to carry out efficient identification.

2. Proposed fine-grained image classification

The interactive deep learning classification network framework includes backbone network YOLOv5, feature enhancement selection strategy and interactive feature fusion module.

2.1. Backbone network

For fine-grained tasks, the choice of backbone network is very important if you want to construct a powerful feature representation, which determines the ability to extract fine-grained features. As shown in below Table, Mask-CNN uses the same annotation information and algorithm to carry out comparative experiments in Alex-Net, VGG and Resnet backbone networks respectively. The experimental results show that when the classification capability of the backbone network is good enough, the downstream tasks will also achieve good performance, and the references [18, 19] even take ViT as the backbone to obtain strong performance. According to the above analysis, how to choose an excellent backbone network and build a strong feature representation is one of the keys to fine-grained tasks.

ConvNeXt was proposed by the Facebook AI Research Institute and was built entirely from standard convolutional modules, bringing together the special designs in Swin transformer and ViT. Starting from macro design, deep separable convolution, inverse bottleneck layer, large convolution kernel, and other details, the ResNet architecture has been upgraded to have faster inference speed and higher accuracy than Swin transformer. Therefore, this paper chooses YOLOv5 as the backbone network, as shown in Figure 1, which contains five stages. Where stage 1 has a simple structure and can be regarded as the preprocessing of the input image; Stages 2∼4 are composed of ConvNeXt block stacks with similar structures. The depth of the network deepens with the increase of the stage, and the information it contains becomes richer. When the input image passes through different stages, the feature map $X \in F^{C\times W\times H}$ at different scales can be obtained, where C, W and H are the channel number, width and height of the feature map respectively.

For fine-grained tasks, networks often focus only on the most significant parts and ignore other potentially discernible parts. In order to avoid the network focusing only on the most significant local features and ignoring the global features of the whole, this paper assumes that after the feature map is sliced in the training process, the network is forced to learn more relevant information by suppressing the most significant part of each slice, so as to promote the network to pay attention to the global information. However, if suppression is used throughout the training process, the network will completely ignore the most significant features, resulting in reduced accuracy. Therefore, an enhancement of the most discriminative part of the operational reward is also needed to improve the predictive power of the model. Based on the above analysis, this paper proposes a feature enhancement selection strategy, which forces the network to learn more comprehensive effective features by evenly slicing the feature map and randomly performing the above two operations in each slice.

The specific structure of the feature enhancement selection strategy (FESS) is shown in Figure 1. The policy input can be the output feature graph F of any layer of the backbone network. In this paper, the output of stage 3 and stage 4 will be used as the input of FESS. First, the feature graph $\mathrm{F} \in F^{C\times W\times H}$ is uniformly sliced n times along the width dimension to obtain $F_{(k)} \in R^{C\times (W/n)\times H}$. The feature elimination operation or feature enhancement operation is then randomly performed for each slice $F_{(k)}$. That is, the strategy provides two candidate branches, namely the feature elimination branch and the feature enhancement branch. Each slice has a 50% probability of performing a feature elimination operation or a feature enhancement operation. FESS uses 0 and 1 to represent two branches, and decides the selection of branches by random extraction, so as to realize the randomized execution of the two types of operations of slice $F_{(k)}$.

For the feature elimination branch, the channel average pooling operation is performed on the input feature map using the following formula to obtain $F_{P(k)} \in R^{(W/n)\times H}$.

$$F_{P(k)} = CAP\left(F_{(k)}\right) \in R^{(W/n)\times H} \tag{1}$$

CAP is channel-wise average pooling.

The value range of each pixel of $F_{P(k)}$ is the same as that of the input feature map, representing the key feature expression obtained by the classification model. Since the FESS classification network is trained for the classification task, $F_{P(k)}$ can approximately reflect the spatial distribution

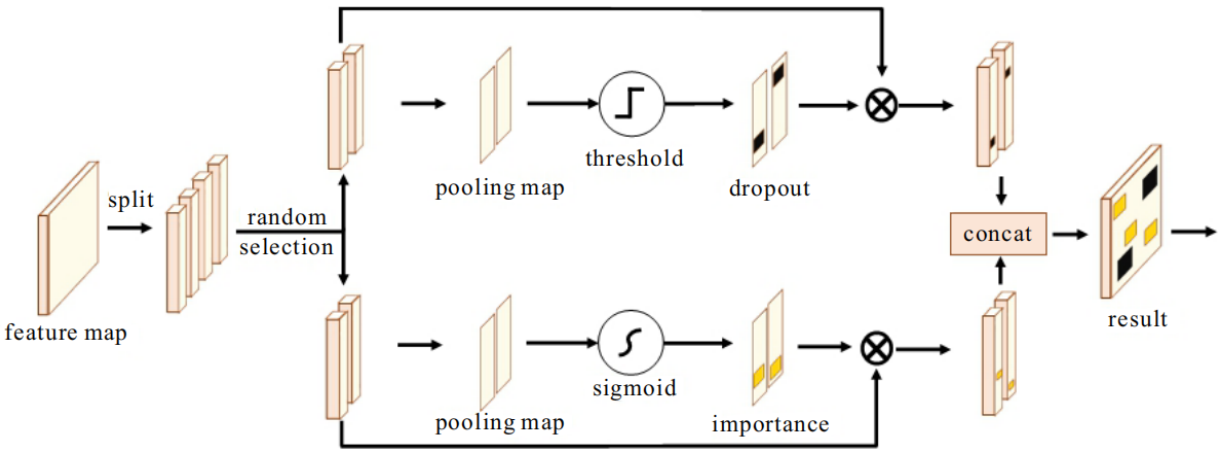

Fig. 1. FESS.

of the most discriminating part. The higher element value denotes the stronger discriminating power. That is, for the classification task, the intensity of each pixel in $F_{P(k)}$ represents its ability to discriminate. To eliminate the most discriminating part, set the threshold rate δ based on the pixel value of maximum intensity in $F_{P(k)}$ to generate the elimination mask $P(k)_{\text{drop}}$, setting the portion of pixels larger than the threshold to 0 . Conversely, the part of the pixel that is less than the threshold value is set to 1, as shown below:

$$P(k)_{\text{drop}} = \begin{cases} 0, F_{P(k)}(i,j) > \delta \cdot \max\left(F_{P(k)}\right) \\ 1, \text{ otherwise} \end{cases} \quad (2)$$

For the feature enhancement branch, the enhancement mask $P(k)_{imp}$ is generated using the sigmoid activation function for $F_{P(k)}$ using the following formula.

$$P(k)_{imp} = \text{sigmoid}\left(F_{P(k)}\right) \in [0,1]^{(W/n)\times H} \quad (3)$$

When slice $F_{(k)}$ selects the feature elimination branch, the feature elimination map $F_{(k)drop}$ is obtained by the following formula.

$$F_{(k)\text{ drop}} = F_{(k)} \times F_{(k)\text{ drop}} \quad (4)$$

Correspondingly, if the feature enhancement branch is selected, it is important to obtain the enhanced feature graph $F(k)_{imp}$ through the following formula.

$$F(k)_{imp} = F(k) \times F(k)_{imp} \quad (5)$$

Finally, $F_{(k)\text{ drop}}$ and $F(k)_{imp}$ are concatenated by width dimension to get $F_{\text{result}} \in R^{C\times W\times H}$.

$$F_{\text{result}} = \text{concat}\left(F_{(k)\text{ drop}}, F(k)_{\text{imp}}\right) \quad (6)$$

Where *concat* refers to the concatenation of each segmented feature map in the width dimension.

2.2. Interactive feature fusion

The potential global features are obtained by the above method, but the direct output limits the ability of the model to compare cues from different features. This paper argues that these different global characteristics should not be treated in isolation. A more reasonable approach is to model the feature maps of different levels, forcing different layers of the network to share the mined information and enhance the semantic complementary information.

First, the feature maps of the backbone network at stage3 3stage 5 after ConvBlock1 ConvBlock3 are taken as an image pair $(F_1, F_2, F_3) \in \left(R^{C\times W_1\times H_1}, R^{C\times W_2\times H_2}, R^{C\times W_3\times H_3}\right)$, and the widths and heights of these three feature maps are compressed by $W_1 \times H_1, W_2 \times H_2, W_3 \times H_3$ into L_1, L_2, L_3, and it can get $\left(F_1', F_2', F'_3\right) \in \left(R^{C\times L_1}, R^{C\times L_2}, R^{C\times L_3}\right)$. Then, the similarity matrices M_1, M_2 and M_3 are obtained by the inner product operations of $F_1'^T$ and F'_2, $F_2'^T$ and F'_1, and $F_1'^T$ and F'_3. Element $M_{i,j}$ in the similarity matrix is the similarity of pixels of different feature maps. The lower the similarity of two pixel values, the stronger the complementarity between them. Therefore, it takes $-M_1, -M_2$, and $-M_3$ as interaction matrices and normalizes their rows and columns according to the following formula to get W_{12}, W_{23}, W_{13}.

$$W_{12} = \text{softmax}\left(-M_1^T\right) \in [0,1]^{L_1\times L_2}, M_1 = F_1'^T F'_2 \quad (7)$$

$$W_{23} = \text{softmax}\left(-M_1^T\right) \in [0,1]^{L_2\times L_3}, M_2 = F_2'^T F'_3 \quad (8)$$

$$W_{13} = \text{softmax}\left(-M_3^T\right) \in [0,1]^{L_1\times L_3}, M_3 = F_1'^T F'_3 \quad (9)$$

Then, $W_{F_1}, W_{F_2}, W_{F_3}$ are obtained by weighting the interaction feature graphs W_{12}, W_{23}, W_{13} obtained from the above normalization operation to F_1', F_2', F'_3 using the following formula. Their dimensions L_1, L_2, L_3 are converted

back to $W_1 \times H_1, W_2 \times H_2, W_3 \times H_3$, and $(W_{F_1}, W_{F_2}, W_{F_3}) \in \left(R^{C\times W_1\times H_1}, R^{C\times W_2\times H_2}, R^{C\times W_3\times H_3}\right)$ are obtained.

$$W_{F_1} = F_2' \times W_{12}^T + F_3' \times W_{13}^T \tag{10}$$

$$W_{F_2} = F_1' \times W_{12} + F'_3 \times W_{23}^T \tag{11}$$

$$W_{F_3} = F_1' \times W_{13} + F_2' \times W_{12} \tag{12}$$

Finally, the feature map with rich semantic information is obtained by down-sampling fusion of $W_{F_1}, W_{F_2}, W_{F_3}$.

In the part of classifier, this paper maps the fused feature map to one-dimensional feature vector, and uses softmax logistic regression to achieve the final classification of the image.

3. Experimental description and analysis

3.1. Experimental data set

In this paper, three data sets commonly used in the field of fine-grained image classification are used to evaluate the effectiveness of the proposed method, namely, CUB-200-2011, Standford Cars, and FGVC-Aircraft.

The CUB-200-2011 dataset is one of the most challenging in the field of fine-grained image classification, with 11,788 images of 200 different bird species. The difference of birds' flying posture and illumination conditions increases the difficulty of identification. The Standford Cars dataset has 16,185 images of 196 different makes and models of cars. Each category contains about 80 images of cars from different angles and perspectives, and the FGVCAircraft dataset has a total of 10,200 images of 100 different models of aircraft. Each category contains about 100 images of aircraft from different perspectives and attitudes.

3.2. Evaluation index

In this paper, classification accuracy is used as the evaluation index of the final classification accuracy. Classification accuracy is defined as follows:

$$Acc = RA/R \tag{13}$$

Where RA represents the number of correct predictions by the model in the test set, and R represents the number of all images in the test set. The classification accuracy is used as the experimental evaluation index in this paper to compare three commonly used fine-grained image classification data sets. In the experiment, the method of this paper adopts the training set and test set partitioning method and data preprocessing method, which are commonly used in the field of fine-grained image classification. The experimental results of other methods are all derived from the experimental results given in the original paper.

3.3. Experimental environment and parameter setting

The proposed model in this paper uses YOLOv5 as the backbone network and loads the weight of the pre-trained model. The server hardware used in the experiment was configured with an i9 12900K CPU and an Nvidia RTX 3090 GPU. The software is configured for the Windows10 operating system and builds a deep learning framework based on Python 3.7, PyTorch 1.11.0 and TorchVision 0.12.0.

Training parameters. Uniformly resize the input image to 550×550 pixels. The image is then randomly cropped to a size of 448×448 pixels. This paper uses stochastic gradient descent (SGD) to optimize the network model by training 200 maximum iterations with a batch size of 16. Set the learning rate to 0.0002 for the convolution layer using the pre-training weight, and 0.002 for the newly added convolution layer and the fully connected layer. In the training, the cosine annealing strategy is used to optimize the learning rate. The SGD optimizer sets momentum and weight decays to 0.9 and 0.0005.

Test parameters. Adjust the input image to 550×550 pixels and crop the center to a size of 448×448 pixels. The rest does not mention that the parameter settings are consistent with the training parameter settings.

3.4. Experimental results and analysis

In order to fully verify the effectiveness of the proposed method, experiments are performed on three commonly used fine-grained image classification datasets, including CUB200-2011, Standford Cars, and FGVC-Aircraft.

Table 1 shows the performance comparison of different models on the CUB-200-2011 dataset. The PPS achieved 88.9% accuracy by breaking the global structure of the image and disrupting local areas to force the network to discover potentially subtle features. LECR improves the performance of the classification model through separation and smooth sampling operations, and can better deal with intra-class differences and inter-class similarities, achieving an accuracy of 90.1%. Compared with their method, the proposed method achieves an accuracy of 90.6% by fusing synergistic attention features between different network layers and using feature grouping attention.

Table 1. Performance comparison of different models on the CUB-200-2011 dataset.

Method	Backbone network	Acc/%
TransIFC [20]	YOLOv5	87.8
PPS [21]	YOLOv5	88.9
LECR [22]	YOLOv5	90.1
Proposed	YOLOv5	**90.6**

The experimental results on Standford Cars in the

dataset are shown in Table 2. The proposed method in this paper is superior to most other methods. By distinguishing the features of different channels and restricting their distribution through the loss function, PPS makes the features belonging to the same category have discriminative power, and the accuracy rate is 94.8%. The LECR method uses different images and different network layers to learn multi-scale features and obtains 95.7% accuracy. The accuracy of this method is 95.9%.

Table 2. Performance comparison of different models on the Standford Cars dataset.

Method	Backbone network	Acc/%
TransIFC	YOLOv5	87.8
PPS		
LECR	YOLOv5	88.9
		90.1
Proposed	YOLOv5	**90.6**

Table 3 shows that the proposed method achieved the best accuracy of 95.8% on the FGVCAircraft dataset. PPS uses the contrast method to capture fine-grained details and distinguishing information through paired interactions between different image areas to pay attention to local relationships and small differences among object parts, achieving an accuracy rate of 94.5%. LECR improves the classification accuracy and achieves 95.2% accuracy by constructing the relationship diagram between objects to explain the complex relationships and subtle differences between objects.

Table 3. Performance comparison of different models on the FGVCAircraft dataset.

Method	Backbone network	Acc/%
TransIFC	YOLOv5	93.2
PPS	YOLOv5	94.5
LECR	YOLOv5	95.2
Proposed	YOLOv5	**95.8**

3.5. Ablation experiment

In order to verify whether each module can effectively improve the performance of the model, ablation experiments are conducted on three data sets for the algorithm model proposed in this paper. The experimental results are shown in Table 4. Where A stands for cross-layer cooperative attention module and B stands for channel grouping attention module. The accuracy values of the benchmark model on the CUB-200-2011, Standford Cars and FGVC-Aircraft datasets are 90.2%, 95.6% and 95.3%, respectively, and the introduction of cross-layer collaborative attention module has been improved by 0.3%, 0.1% and 0.2%, respectively. This is due to the fact that the model effectively integrates shallow and deep features, and effectively finds and focuses on information that helps to classify. By introducing the channel packet attention module, the accuracy has been improved by 0.1%, 0.2% and 0.3%, respectively. This shows that channel grouping can effectively improve the semantic feature learning ability of the model. Through the joint action of channel grouping attention and cross-layer cooperative attention, the accuracy of the proposed method is improved by 0.4%, 0.3% and 0.5% compared with the benchmark model, respectively. The effectiveness of the proposed method on fine-grained image classification is demonstrated, and the efficiency of channel grouping attention by integrating shallow and deep features is demonstrated.

Table 4. Ablation experiments of the proposed method on 3 commonly used data sets (%).

Method	CUB-200-2011	Standford Cars	FGVC -Aircraft
YOLOv5	90.2	95.6	95.3
YOLOv5+A	90.5	95.7	95.5
YOLOv5+A+B	**90.6**	**95.9**	**95.8**

4. Conclusions

In this paper, a new fine-grained classification network is proposed, and a feature enhancement selection strategy is proposed in the training process to promote the network to learn more relevant information through the interaction between feature elimination branches and feature enhancement branches. At the same time, in order to further enhance the semantic complementary information, the interactive feature fusion strategy is proposed, so that different layers of the network can share the mined information and compare different features to promote the classification performance of the network. The proposed method does not require boundary frame or location labeling information, and can be used for weakly supervised end-to-end training. Experiments show that the proposed method can exceed the accuracy of most mainstream methods on many common fine-grained image classification datasets. In the following work, we will continue to optimize the network structure, so as to further improve the classification performance of the model.

Acknowledgments

This work was supported by the following project: Training Program for Young Backbone Teachers in Colleges and Uni-

versities of Henan Province (2020GGJS216); Research and Practice of Higher Education Teaching Reform in Henan Province (2021SJGLX200, 2021SJGLX629); Zhoukou Teachers College Education Reform Project (J2022071).

References

[1] M. Tan, F. Yuan, J. Yu, G. Wang, and X. Gu, (2022) *"Fine-grained image classification via multi-scale selective hierarchical biquadratic pooling"* **ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) *18***(1s): 1–23. DOI: 10.1145/3492221.

[2] C. Zhang, H. Bai, and Y. Zhao, (2022) *"Fine-grained image classification by class and image-specific decomposition with multiple views"* **IEEE Transactions on Multimedia**: DOI: 10.1109/TMM.2022.3214431.

[3] X. Meng, X. Wang, S. Yin, and H. Li, (2023) *"Few-shot image classification algorithm based on attention mechanism and weight fusion"* **Journal of Engineering and Applied Science *70***(1): 14. DOI: 10.1186/s44147-023-00186-9.

[4] D. Liang, W. Xu, and X. Bai. "An end-to-end transformer model for crowd localization". In: *European Conference on Computer Vision*. Springer. 2022, 38–54. DOI: 10.1007/978-3-031-19769-7_3.

[5] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, and G. Pu, (2022) *"Fakelocator: Robust localization of gan-based face manipulations"* **IEEE Transactions on Information Forensics and Security *17***: 2657–2672. DOI: 10.1109/TIFS.2022.3141262.

[6] P. Wu, W. Zhai, and Y. Cao. "Background activation suppression for weakly supervised object localization". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2022, 14228–14237. DOI: 10.1109/CVPR52688.2022.01385.

[7] S. Yin, (2023) *"Object Detection Based on Deep Learning: A Brief Review"* **IJLAI Transactions on Science and Engineering *1***(02): 1–6.

[8] F. Li, D. Yao, M. Jiang, and X. Kang, (2022) *"Smoking behavior recognition based on a two-level attention fine-grained model and EfficientDet network"* **Journal of Intelligent & Fuzzy Systems *43***(5): 5733–5747. DOI: 10.3233/JIFS-213042.

[9] X. Ke, Y. Huang, and W. Guo, (2022) *"Weakly supervised fine-grained image classification via two-level attention activation model"* **Computer Vision and Image Understanding *218***: 103408. DOI: 10.1016/j.cviu.2022.103408.

[10] J. Guang and J. Liang, (2022) *"Cmsea: Compound model scaling with efficient attention for fine-grained image classification"* **IEEE Access *10***: 18222–18232. DOI: 10.1109/ACCESS.2022.3150320.

[11] J. Yang, J. Duan, T. Li, C. Hu, J. Liang, and T. Shi, (2022) *"Tool wear monitoring in milling based on fine-grained image classification of machined surface images"* **Sensors *22***(21): 8416. DOI: 10.3390/s22218416.

[12] Y. Wan and J. Li, (2024) *"LGP-YOLO: an efficient convolutional neural network for surface defect detection of light guide plate"* **Complex & Intelligent Systems *10***(2): 2083–2105. DOI: 10.1007/s40747-023-01256-4.

[13] G. Xian, R. Tao, and G. Chen. "Combining spatial attention and cross-layer bilinear pooling for fine-grained image classification". In: *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*. IEEE. 2023, 271–276. DOI: 10.1109/ICPECA56706.2023.10075984.

[14] H. Touvron, P. Bojanowski, M. Caron, M. Cord, A. El-Nouby, E. Grave, G. Izacard, A. Joulin, G. Synnaeve, J. Verbeek, et al., (2022) *"Resmlp: Feedforward networks for image classification with data-efficient training"* **IEEE Transactions on Pattern Analysis and Machine Intelligence *45***(4): 5314–5321. DOI: 10.1109/TPAMI.2022.3206148.

[15] Q. Zhu, Z. Li, W. Kuang, and H. Ma, (2023) *"A multi-channel location-aware interaction network for visual classification"* **Applied Intelligence *53***(20): 23049–23066. DOI: 10.1007/s10489-023-04734-x.

[16] A. Jisi, S. Yin, et al., (2021) *"A new feature fusion network for student behavior recognition in education"* **Journal of Applied Science and Engineering *24***(2): 133–140. DOI: 10.6180/jase.202104_24(2).0002.

[17] R. Li and Y. Wu, (2022) *"Improved YOLO v5 wheat ear detection algorithm based on attention mechanism"* **Electronics *11***(11): 1673. DOI: 10.3390/electronics11111673.

[18] W. Chen, X. Du, F. Yang, L. Beyer, X. Zhai, T.-Y. Lin, H. Chen, J. Li, X. Song, Z. Wang, et al. "A simple single-scale vision transformer for object detection and instance segmentation". In: *European Conference on Computer Vision*. Springer. 2022, 711–727. DOI: 10.1007/978-3-031-20080-9_41.

[19] T. Li, Z. Zhang, L. Pei, and Y. Gan, (2022) *"HashFormer: Vision transformer based deep hashing for image retrieval"* **IEEE Signal Processing Letters *29***: 827–831. DOI: 10.1109/LSP.2022.3157517.

[20] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, and Y.-F. Li, (2023) *"TransIFC: invariant cues-aware feature concentration learning for efficient fine-grained bird image classification"* **IEEE Transactions on Multimedia**: DOI: 10.1109/TMM.2023.3238548.

[21] W. Zhang, Y. Zhao, Y. Gao, and C. Sun, (2024) *"Re-abstraction and perturbing support pair network for few-shot fine-grained image classification"* **Pattern Recognition *148***: 110158. DOI: 10.1016/j.patcog.2023.110158.

[22] X. Li, Q. Song, J. Wu, R. Zhu, Z. Ma, and J.-H. Xue, (2023) *"Locally-enriched cross-reconstruction for few-shot fine-grained image classification"* **IEEE Transactions on Circuits and Systems for Video Technology**: DOI: 10.1109/TCSVT.2023.3275382.