Mathematical Biosciences and Engineering

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

MBE, 20(9): 16744–16761.
DOI: 10.3934/mbe.2023746
Received: 09 June 2023
Revised: 17 July 2023
Accepted: 02 August 2023
Published: 23 August 2023

Research article

AGMG-Net: Leveraging multiscale and fine-grained features for improved cargo recognition

Aigou Li* **and Chen Yang**

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, China

* **Correspondence:** Email: liag@xust.edu.cn.

Abstract: Security systems place great emphasis on the safety of stored cargo, as any loss or tampering can result in significant economic damage. The cargo identification module within the security system faces the challenge of achieving a 99.99% recognition accuracy. However, current identification methods are limited in accuracy due to the lack of cargo data, insufficient utilization of image features and minimal differences between actual cargo classes. First, we collected and created a cargo identification dataset named "Cargo" using industrial cameras. Subsequently, an Attention-guided Multi-granularity feature fusion model (AGMG-Net) was proposed for cargo identification. This model extracts both coarse-grained and fine-grained features of the cargo using two branch networks and fuses them to fully utilize the information contained in these features. Furthermore, the Attention-guided Multi-stage Attention Accumulation (AMAA) module is introduced for target localization, and the Multi-region Optimal Selection method Based on Confidence (MOSBC) module is used for target cropping. The features from the two branches are fused using a fusion branch in a Concat manner for multi-granularity feature fusion. The experimental results show that the proposed model achieves an average recognition rate of 99.58, 92.73 and 88.57% on the self-built dataset Cargo, and the publicly available datasets Flower and Butterfly20, respectively. This is better than the state-of-the-art model. Therefore, this research method accurately identifies cargo categories and provides valuable assistance to security systems.

Keywords: computer vision; cargo identification; security system; attention guidance; multi-branch network

1. Introduction

High-accuracy image recognition tasks in specific scenarios have always been a prominent research topic in computer vision [1–4]. Railway stations, banks, airports and border controls are critical systems where deep learning technology has been widely applied [5, 6]. On the one hand, the extensive use of deep learning in these scenarios has significantly reduced human and material resources. On the other hand, the uniqueness of these scenarios imposes a high demand for accuracy in deep learning techniques. In particular, security systems necessitate real-time and efficient tracking, management and protection of valuable materials. In this context, cargo recognition plays a critical role. It enables automated and efficient management, enhancing both work efficiency and safety. Therefore, achieving high accuracy and reliability in cargo recognition within high-security systems has emerged as a pressing concern in the field of security technology.

Traditional methods for cargo identification require manual intervention to obtain additional information about the cargo [7–9]. Furthermore, the recognition results of traditional methods tend to be suboptimal in complex cargo environments. In contrast, deep learning-based cargo recognition methods exhibit stable performance in intricate environments and can be readily deployed in practical production settings. By employing neural network models and corresponding training techniques, cargo recognition can be automatically achieved in diverse environments, regardless of artificial or other external factors.

Zhu et al. [10] proposed an attribute-guided two-layer learning framework that can identify unknown image categories, thus improving the robustness and performance of few-shot image recognition. Zeng et al. [11] proposed a convolutional neural network model for classifying house styles and achieved reasonable classification results on a small sample dataset, which confirmed the possibility of house style recognition. Yi et al. [12] proposed an end-to-end trained superpixel convolutional neural network by treating irregular superpixel crystals as 2D point clouds and using PointConv layers instead of standard convolutional layers to process these point clouds, thereby learning advanced representations of image superpixel elements and improving superpixel efficiency while obtaining considerable image recognition effects. The aforementioned methods have achieved satisfactory recognition results in their respective scenarios, but they have also ignored the fine-grained features of the recognition objects while only focusing on global and coarse-grained features.

Koyun et al. [13] proposed a two-stage object detection framework named "Focus-and-Detect" for detecting small objects in aerial images and introduced the Incomplete Box Suppression (IBS) method to address the truncation effect of region search methods. This framework demonstrated the best performance in small object detection on the VisDrone validation dataset. Wang et al. [14] presented a small object detection method based on an enhanced Single Shot MultiBox Detector (SSD) algorithm. The method replaced the original VGG-16 with an improved dense convolutional network (C-DenseNet) and incorporated residual prediction layers and DIoU-NMS. This approach effectively resolves the issues of false detection and missed detection in small object detection for object detection algorithms. Dong et al. [15] proposed a new object detection method based on a feature pyramid network (FPN), which introduces a multi-scale deformable attention module (MSDAM) and a multi-level feature aggregation module (MLFAM) to enhance the performance of remote sensing object detection (RSOD), achieving accurate detection on optical remote sensing images (DIOR) and RSOD datasets. These object detection methods have achieved effective object localization and recognition, but at the cost of a large amount of manual annotation and more focused attention on the local and fine-grained features of the objects.

Addressing the aforementioned issues, we present an AGMG-Net to enhance the accuracy of cargo recognition in security scenarios. The proposed network can effectively capture the distribution of focused regions, accurately locate the target position without manual annotation, separate the target from the background and fuse multi-granularity features to achieve precise identification of cargo. The major contributions of this study are outlined as follows:

- We propose an AMAA method to solve the problem of difficulty in locating targets in complex security system environments.
- We also propose a multi-region confidence-dependent optimal selection method to reduce the dependency on the threshold of foreground-background segmentation.
- Building on these two methods, we present an attention-guided multi-granularity feature fusion network that effectively enhances the accuracy of cargo recognition in security systems.

2. Related works

This section provides a brief review of the most relevant work, encompassing multi-branch models, weakly supervised object localization (WSOL) and fine-grained visual classification.

2.1. *Multi-branch model*

Multi-branch networks as a fundamental structure in deep learning have found wide applications in various task domains including semantic segmentation [16, 17] and object detection [18–20], enabling the capture and learning of richer and more diverse features. For instance, Xie et al. [21] proposed a multi-branch network for disease detection in retinal images, enhancing the representation of disease-specific features through the fusion of multi-scale and spatial features. To overcome the limited capacity for extracting global spatial information, Xu et al. [22] introduced a dual-branch network composed of a grouped bidirectional LSTM (GBiLSTM) network and a multi-level fusion convolutional transformer (MFCT), generating distinct and robust spectral-spatial features for hyperspectral image classification with limited labeled samples. Addressing the challenge of small object detection in aerial images with limited samples, Zhang et al. [23] proposed a multi-branch network incorporating a transformer branch, leveraging the strengths of generative models and transformer networks to improve the robustness of small object detection in complex environments. To address the problem of significant non-linear differences between image blocks in image matching, in the context of image matching, where significant non-linear differences exist between image blocks, Yu et al. [24] presented a composite metric network comprising a main metric network module and multiple branch metric network modules to capture richer and more distinctive feature differences. Overall, the concept of multi-branch networks has emerged as a crucial research direction in deep learning, offering practical solutions for diverse task domains.

2.2. *WSOL*

Since Zhou et al. [25] introduced the use of Class Activation Maps (CAM) to characterize object locations, an increasing number of works have applied them to the field of WSOL [26–31]. CAM represents a feature map obtained from the global average pooling layer of a classification network and it is weighted before applying softmax to emphasize the position of the target object. Hwang et al. [32] introduced a target localization strategy using entropy regularization, which considers the one-hot labels and the entropy of predicted probabilities, thereby striking a balance between WSOL scores and classification performance. Zhang and Yang [33] proposed an adaptive attention enhancer to address the limitation of existing WSOL methods that lack modeling of the correlation between different regions of the target object. This enhancer supplements object attention by discovering the semantic correspondence between different regions. Gao et al. [34] presented a token semantics coupled attention map (TS-CAM) to tackle the challenge of learning object localization models given image category labels. The self-attention mechanism in vision transformers is utilized to extract long-term dependencies and compensate for the limitations of Convolutional Neural Networks (CNNs) in partial activations. These works demonstrate that WSOL can accomplish object localization tasks with only image annotations, eliminating the need for manual annotation in object detection tasks. Moreover, they hold significant value in image recognition by enabling networks to quickly and accurately identify recognized subjects through WSOL methods, while extracting fine-grained features of the image through the CG-Net branch.

2.3. Fine-grained visual classification

The closest approach to this paper is [35, 36], where Wang et al. [35] proposed an accurate semantic-guided discriminative region localization method for fine-grained image recognition methods that ignore the spatial correspondence between low-level details and high-level semantics. Du et al. [36] introduced a novel framework for fine-grained visual classification, addressing challenges in identifying discriminative granularities and fusing information. The framework includes a progressive training strategy and a jigsaw puzzle generator, achieving state-of-the-art performance on benchmark datasets. The difference between this paper and [35, 36] is that this paper introduces CAM, a classic practice in WSOL, to initially localize the target, and accomplishes richer feature extraction through operations such as accumulation and coarse- and fine-grained feature fusion, to adequately learn the multi-granularity feature information in the image. Wang et al. [37] proposed Prompting vision-Language Evaluator (PLEor), a novel framework for open set fine-grained retrieval, based on the Contrastive Language-Image Pretraining (CLIP) model. PLEor leverages the pre-trained CLIP model to infer category-specific discrepancies and transfer them to the backbone network trained in close set scenarios. Wang et al. [38] introduced a Fine-grained Retrieval Prompt Tuning (FRPT), and by utilizing sample prompting and feature adaptation, FRPT achieves state-of-the-art performance on fine-grained datasets with fewer parameters.

3. Methods

This paper introduces the AGMG-Net, which consists of the Fine-Grained Net (FG-Net), the Coarse-Grained Net (CG-Net) and the Multi-Granularity Fusion Net (MGF-Net). Cargo image information is complex and exhibits both coarse-grained and fine-grained features. Conventional deep learning approaches primarily focus on learning and extracting coarse-grained features, which limits their ability to capture fine-grained features and leads to inaccuracies and omissions in cargo recognition. To address these challenges, we propose the AGMG-Net. The FG-Net uses global attention to extract features and learn coarse-grained characteristics, such as color, shape and position. The CG-Net employs deep convolution to extract fine-grained features, including texture and structure. The MGF-Net conducts feature fusion and learning on multi-granularity features. The final cargo recognition result is obtained by applying a majority rule after classification using fully connected layers. Figure 1 illustrates the network structure.

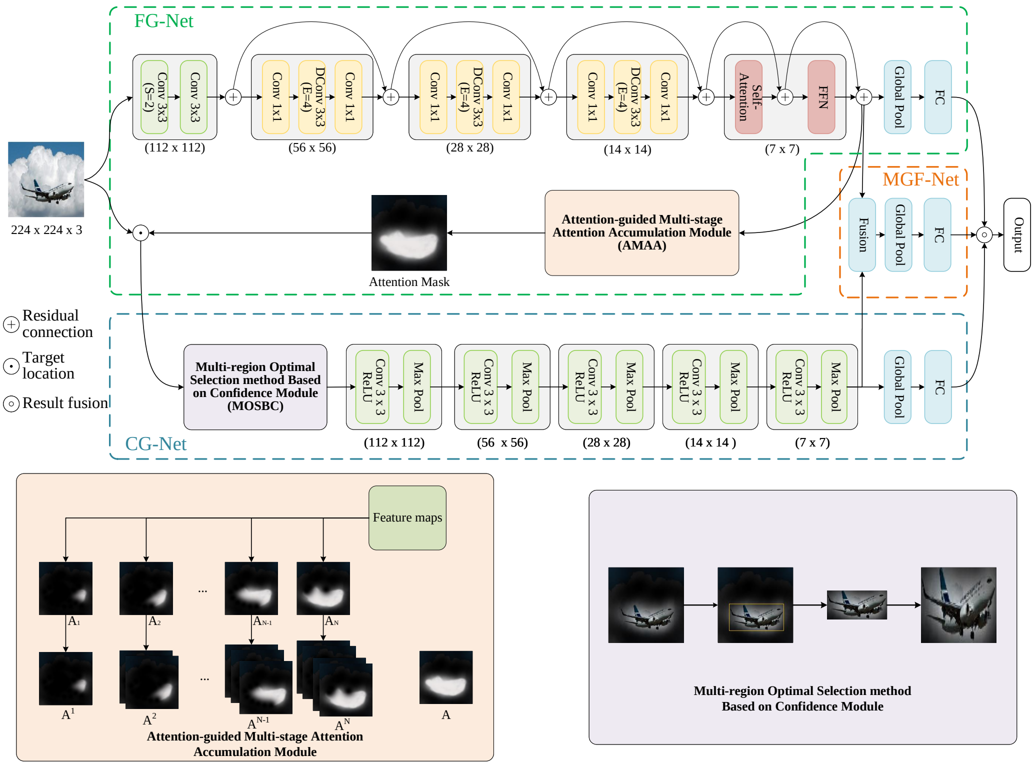


Figure 1. The structure of AGMG-Net, the dotted frame above contains the structure of FG-Net, CG-Net and MGF-Net, and below is the AMAA module and MOSBC module.

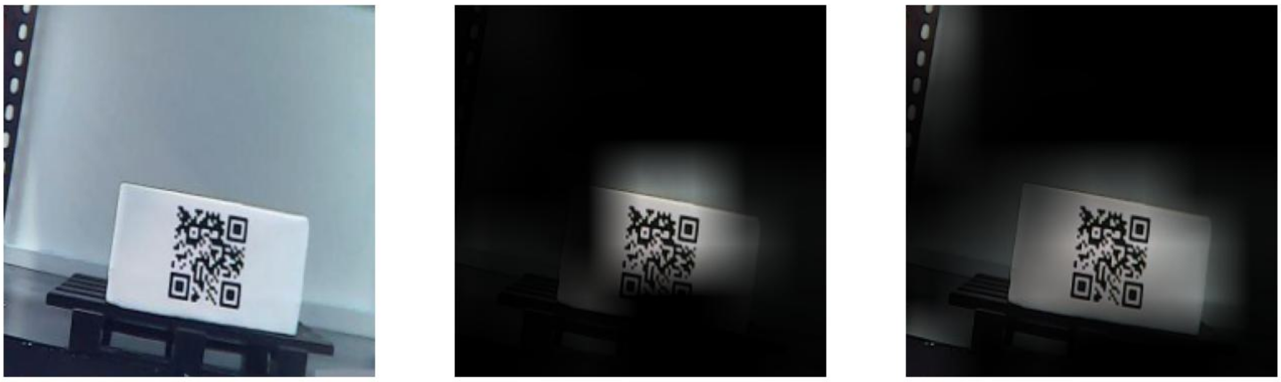
3.1. FG-Net

The input image undergoes initial processing by the FG-Net module to extract coarse-grained features. The FG-Net module comprises three components: a feature extractor, a classifier and an AMAA module. The feature extractor consists of convolutional and global self-attention blocks, which enable the learning of global features and the extraction of coarse-grained features. The classifier generates the cargo classification result using a global average pooling layer and a fully connected layer. During the early-to-mid training phase, the AMAA module accumulates the multi-stage attention map, directing the CG-Net to focus on the target object for extracting fine-grained features. As training progresses, the AMAA module guides the CG-Net to pay more attention to the overall image, to a certain extent, facilitating effective fusion of the cargo's coarse-grained features.

Suppose the original image is shown in Figure 2(a), the AMAA module generates an attention map using the Class Activation Mapping (CAM) method, as illustrated in Figure 2(b). Let $I \in \mathbb{R}^{C\times W\times H}$ denote the input. The attention map acquired from each iteration is denoted as $A_t \in \mathbb{R}^{w\times h}$. By setting the initial condition as $M_0 = A_0$, the cumulative attention map M^c can be calculated using Eq (3.1).

$$M_t^c = max(M_{t-1}^c, A_t^c), \quad t = 1, 2, \ldots \tag{3.1}$$

Here C represents the number of channels, H and W represent the width and height of the image, h and w represent the width and height of the attention map, and c stands for the category.



(a) Image **(b)** Attention Map **(c)** Integrated Attention Map

Figure 2. Attention Map in the training stage of Cargo dataset.

Different positions are emphasized at each stage of the network during training, and the resulting cumulative attention map M_t^c reflects the attention region distribution for a given category. In the early and middle stages of training, M_t^c exhibits more accurate localization ability than the attention map A_t^c. However, due to the gradual accumulation of maximum values during the calculation of the attention map M, regions with excessively high attention values can lead to inaccurate target localization. To address this issue, we propose the AMAA module in this paper. The AMAA module retains the attention maps from the previous $k-1$ stages along with the cumulative attention map M_t, forming an attention sequence denoted as ML, which has a length of k. The specific formulation of ML is presented in Eq (3.2). By utilizing the attention sequence, which comprehensively considers the cumulative attention map M and the recent $k-1$ attention maps, and performing a weighted summation of ML using Eq (3.3), we can prevent misleading final localization caused by abnormal attention maps.

$$ML = [M_t, A_t, A_{t-1}, \ldots, A_{t-k+1}] \tag{3.2}$$

$$A = \sum_{i=0}^{k} \alpha_i ML_i \tag{3.3}$$

where $\alpha \in \mathbb{R}^k$. Since taking the simple average of the attention sequence ML would diminish the impact of the most recent attention map on the overall attention map A, this paper adopts the approximate forgetting function [39] $L(x, k)$ to compute the initial value of α. Subsequently, α undergoes normalization using Eq (3.5).

$$L(x, k) = \frac{184k}{125x + 1.84k} \tag{3.4}$$

$$\alpha_i = softmax(L(i,k)) \tag{3.5}$$

The AMAA module not only considers the cumulative attention map but also incorporates the attention maps computed in the previous $k-2$ iterations to generate the comprehensive attention map A for the current iteration. As the model undergoes continuous training, AMAA gradually converges towards the actual distribution of the target position, i.e., the true location of the target object. This enables AMAA to efficiently accomplish target localization, consequently guiding CG-Net in performing localization and recognition.

3.2. CG-Net

The comprehensive attention map generated by FG-Net and the original input are both fed into CG-Net for fine-grained feature extraction. CG-Net consists of three components: the MOSBC module, the feature extractor and the classifier. As shown in Figure 3, the MOSBC module first performs image fusion on the input and then selects the target region. The feature extractor has 11 convolutional layers, which enable the capture of detailed information and local features in the image, facilitating the extraction of fine-grained image features. The classifier consists of one global average pooling layer and two fully connected layers, enabling precise image classification.

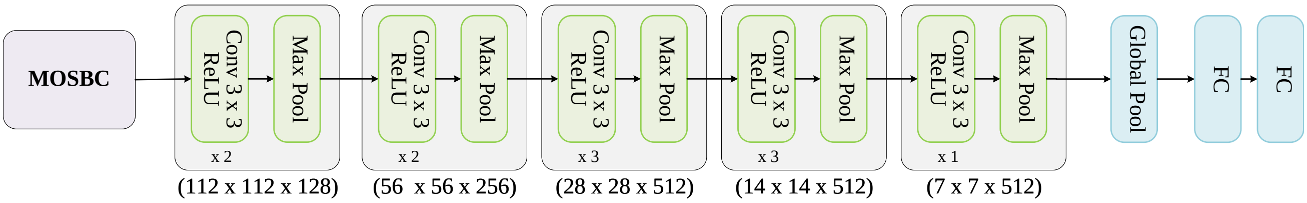


Figure 3. Detailed structure of CG-Net.

The comprehensive attention map A, as illustrated in Figure 2(c), still includes non-target areas and exhibits blurred edges within the target region, despite attention accumulation. To address this issue, this paper proposes a region localization method called MOSBC, which aims to identify the most probable area where the target is located. Initially, a two-dimensional bilinear interpolation is employed to transform the comprehensive attention map A into an output of the same width and height, denoted as $A' \in \mathbb{R}^{H\times W}$. Subsequently, a threshold $\gamma \in (0,1)$ is utilized to delineate the foreground and background in the attention map A, as depicted in Eq (3.6).

$$A^* = \begin{cases} 0 & A' \leq \gamma \\ 1 & otherwise \end{cases} \tag{3.6}$$

After dividing the foreground and background, the Euclidean distance from each foreground position to the nearest background is first calculated. Then, the peak distance values are found from the image, and the connected components are analyzed using eight connectivity for local peaks. The watershed algorithm [40] is then used for image segmentation to obtain the target region set, generating N candidate regions $R_n = (x_n, y_n, h_n, w_n), n = 1, \ldots, N$, where x_n, y_n is the center coordinate of the candidate region R_n and h_n, w_n is the height and width. Then, Eq (3.7) is used to calculate the confidence score ρ_n for each region and used Eq (3.8) to select the region with the highest confidence score as the finally target *Region*.

$$\rho_n = \frac{1}{h_n \times w_n} \sum_{i=0}^{h_n} \sum_{j=0}^{w_n} R_n \tag{3.7}$$

$$Region = argmax\left\{R_i | \rho_i = max\left\{\rho_j\right\}, 1 \leq i, j \leq N\right\} \tag{3.8}$$

It is important to note that the *Region* obtained at this stage represents a localization box, which requires overlaying it onto the input I to extract the fine-grained image of the target through cropping. Subsequently, the fine-grained image is resized using two-dimensional bilinear interpolation to match the size of I and then fed into the feature extractor of CG-Net for fine-grained feature extraction.

3.3. MGF-Net

MGF-Net is responsible for multi-scale feature fusion and classification. It consists of a feature fusion layer and a classifier. The feature fusion layer uses the concatenation operation to achieve multi-scale feature fusion at the channel level. This is in contrast to other feature fusion methods (such as addition or multiplication), which do not ensure the diversity and completeness of features. Concatenation can handle feature maps of multiple scales simultaneously, which makes it more flexible. The MGF-Net classifier consists of one global average pooling layer and three fully connected layers. This enables multi-scale image classification. The structure of MGF-Net is depicted in Figure 4 and the network's layer parameters are detailed in Table 1.

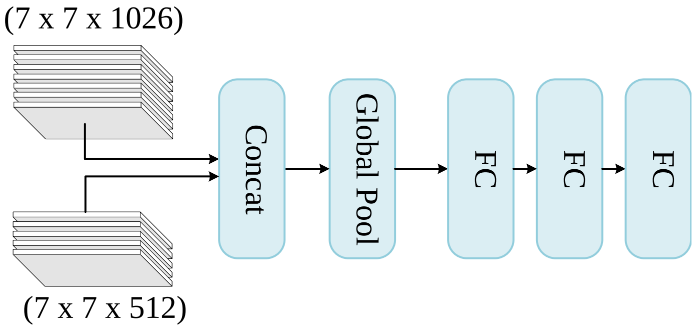


Figure 4. Detailed structure of CG-Net MGF-Net.

Table 1. MGF-Net parameters.

Layers	Resolution	Description
Input	$(7, 7, 1026), (7, 7, 512)$	-
Concat	$(7, 7, 1026 + 512)$	Concat operation
Global Pool	$(7, 7, 1538)$	Global average pooling
FC	-	4096-dimensional FC layer
FC	-	4096-dimensional FC layer
FC	-	k-dimensional FC layer

The feature map with 1026 channels is derived from FG-Net, while the feature map with 512 channels originates from CG-Net. The number of channels obtained is 1026 after the input image is convolved by 4 layers of FG-Net and 1 layer of transformer, and the number of channels obtained is 512 after the image processed by the MOSBC module is input to CG-Net and convolved by 11 layers as shown in Figure 3.

4. Experiments and datasets

To evaluate the effectiveness of our proposed method, AGMG-Net was compared against state-of-the-art methods, and ablation experiments were conducted on three publicly available and self-built image recognition datasets.

4.1. Datasets

To simulate real-world scenarios, the experiments were performed on the publicly available Flower and Butterfly datasets. The Flower dataset, obtained from http://download.tensorflow.org/example_images/flower_phones.tgz, comprises 4323 images of flowers from 5 categories, each with random resolutions. The original butterfly dataset [41] consists of 200 categories, from which 20 categories were selected to create a smaller dataset named butterfly20. This subset contains 2066 images with random resolutions. Sample images from both datasets are depicted in Figure 5, exhibiting diverse lighting angles, shooting angles, distances and backgrounds, resembling the conditions and environments encountered in security systems for cargoes. Utilizing these datasets in the experiments enables a more accurate reflection of practical scenarios, enhancing the experiment's reliability and generalization ability. Evaluating the performance of AGMG-Net under different conditions using these datasets allows for a more comprehensive assessment, thereby improving the algorithm's robustness.

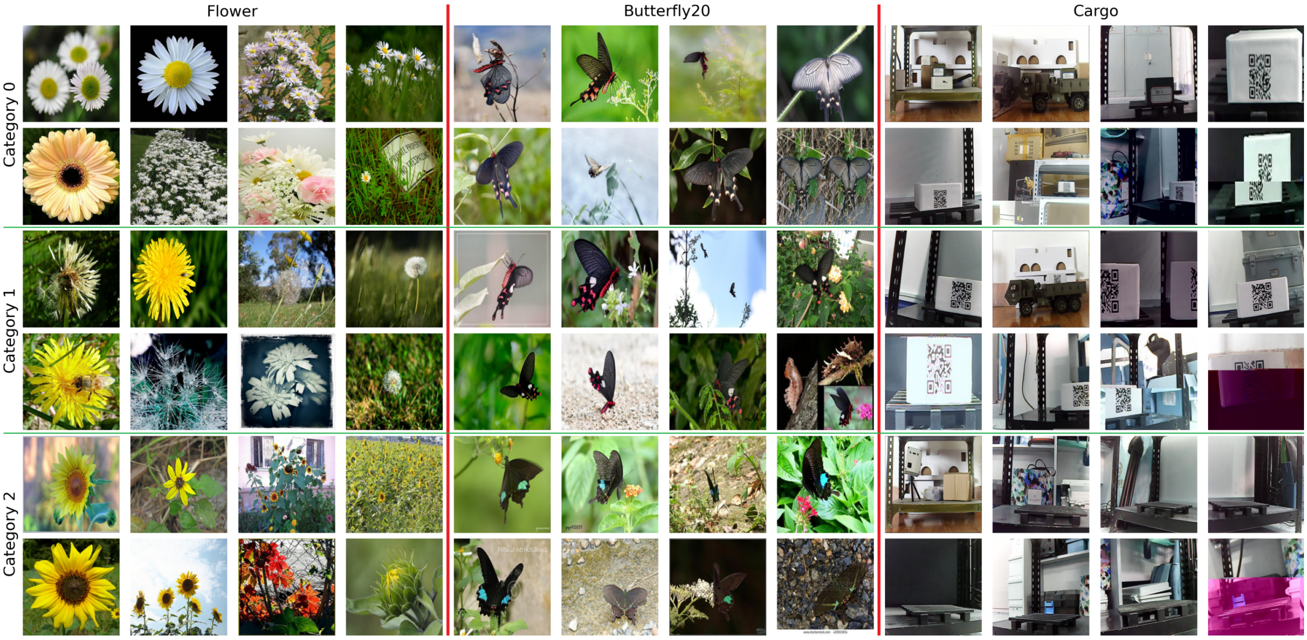


Figure 5. Sample sample of each dataset.

In addition, we created a self-built cargo recognition dataset named "Cargo" specifically for recognizing cargoes in security systems. The dataset comprises 3 categories and 4715 images with random resolutions, organized in a structure identical to the aforementioned publicly available datasets. An overview of the Cargo dataset is provided in Table 2.

Table 2. Data specification.

Item	Values
Modalitites	RGB
Total number of images	4715
Number of classes	3
Number of angle classes	6
Number of distance classes	5
Number of background classes	7

4.2. *Experimental environment and parameter settings*

The hardware environment used for this experiment is Intel® Xeon® Platinum 8255C CPU @ 2.50GHz with 12 CPU cores, 32GB DDR4 memory, and NVIDIA GeForce RTX 3090 graphics card. The software platform is Ubuntu 20.02-LTS operating system, Python version 3.7, Pytorch version 1.12.1, CUDA version 11.3 and cuDNN version 8.2.1.

The hyperparameters were set as follows: 100 training iterations were performed for the Flower dataset, 200 for the Butterfly20 dataset and 120 for the Cargo dataset. The Adam optimizer was utilized with a batch size of 16 and a learning rate of 0.001. The training set and testing set were randomly split in a ratio of 8:2, ensuring an equal number of samples for each class.

4.3. Evaluation metrics

To objectively evaluate the classification performance of different models, two commonly used metrics in classification tasks, namely accuracy and F1-score, were adopted. The F1-score is a statistical measure of classification model precision, as expressed mathematically in Eq (4.1).

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.1}$$

where $Precision = TP/(TP + FP)$, $Recall = TP/(TP + FN)$, TP represents the number of correctly classified positive samples, TN represents the number of correctly classified negative samples, FP represents the number of incorrectly classified positive samples and FN represents the number of incorrectly classified negative samples.

4.4. Robustness of γ and ML

In this section, we first conducted test experiments on CG-Net with the pruning threshold γ, while setting the attention sequence length *ML* to 4. The performance of AGMG-Net on the Cargo dataset is presented in Table 3.

Table 3. Accuracy of different clipping thresholds in Cargo dataset.

Threshold γ	Accuracy (%)	F1-score (%)
0.1	98.81	94.26
0.2	99.12	95.33
0.4	99.22	96.14
0.6	99.22	96.09
0.8	99.01	95.51

As shown in Table 3, the accuracy of AGMG-Net is the best when the cropping threshold γ is 0.2 to 0.6. The accuracy is the worst when it is 0.1, and there is a significant decrease in accuracy when it is 0.8. This shows that γ in the range of 0.2 to 0.6 can help the model to complete the target cropping and effectively improve the recognition performance of the model. When γ is set to 0.4, the F1-score of AGMG-Net reaches the maximum value of 96.14%. Therefore, all subsequent experiments set γ to 0.4.

Furthermore, we conducted additional experiments on the attention sequence length *ML* with γ set to 0.4. The performance of AGMG-Net on the Cargo dataset is presented in Table 4.

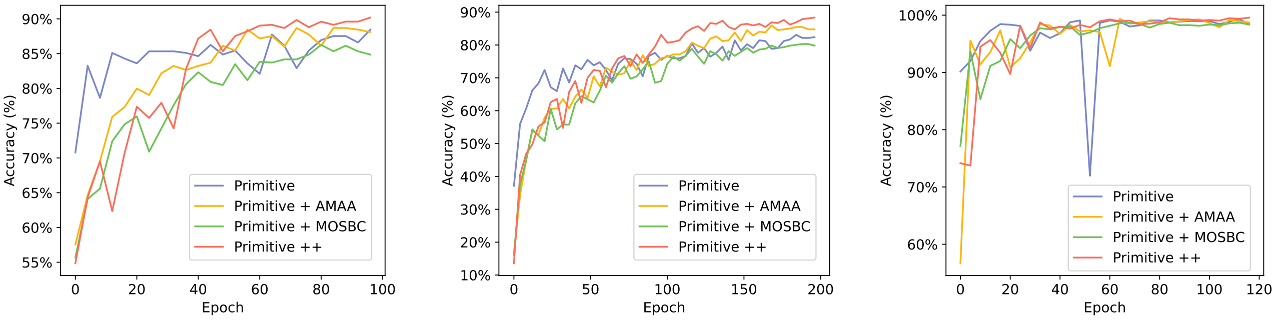
Table 4 shows that AGMG-Net achieves the highest accuracy when *ML* is between 6 and 8. The accuracy is significantly lower when *ML* is between 2 and 4, and it decreases significantly when *ML* is set to 10. These findings suggest that maintaining multiple adjacent attention maps between 6 and 8 helps mitigate the influence of outliers during training and improves recognition performance. The model achieves the highest recognition performance when *ML* is set to 8. Therefore, we set *ML* to 8 in subsequent experiments.

Table 4. Accuracy of different attention sequence lengths in Cargo dataset.

ML	Accuracy (%)	F1-score (%)
2	98.28	96.02
4	99.22	96.14
6	99.43	97.21
8	99.58	97.35
10	99.36	97.13

4.5. *Ablation*

The ablation experiment was conducted to investigate the impact of the AMAA and MOSBC modules in AGMG-Net on network performance. We removed the AMAA and MOSBC modules from FG-Net and CG-Net and conducted ablation experiments on the three datasets mentioned in Section 4.1. The experimental results, depicted in Figure 6, show that the AMAA and MOSBC modules have a significant impact on the recognition accuracy of the AGMG-Net model.



(a) Flower dataset

(b) Butterfly20 dataset

(c) Cargo dataset

Figure 6. The comparison effect of adding and removing AMAA and MOSBC modules in the three datasets. “Primitive” stands for after removing the modules, “Primitive ++” stands for both AMAA and MOSBC in effect.

It is evident from Figure 6(c) that the addition of the AMAA and MOSBC modules increases the number of parameters that AGMG-Net needs to learn. In the early training process, only FG-Net is trained, resulting in relatively low accuracy. However, as the test accuracy reaches a threshold of 0.95, CG-Net starts participating in the training process, leading to rapid object localization and multi-scale feature fusion in 63 rounds. This can be observed in Figure 6(a) and (b), where AGMG-Net surpasses the performance of other models on the Flower (threshold: 0.8), Butterfly20 (threshold: 0.8) and Cargo dataset, with increasing margins at rounds 43, 81 and 63. The optimal values are achieved at rounds 97, 105 and 143, with lead values of 3.7, 5.48 and 0.29%. These experiments demonstrate that the AMAA and MOSBC modules positively impact the recognition accuracy of AGMG-Net in image recognition.

Table 5 shows the results of the quantization of the ablation experiment in Figure 6. The results in Table 5 using only the AMAA module are higher than the accuracy of Primitive. This demonstrates that using only the CAM as a basis for localization does not learn more features, and can even be misled by the anomalous CAM. After AMAA and MOSBC work together, the model shows the highest accuracy and the lowest loss. This indicates that the MOSBC module can perform proper target acquisition after effectively localizing the target, providing more complete target information for AGMG-Net.

Table 5. Impact of AMAA and MOSBC modules on three datasets.

	Flower		Butterfly20		Cargo	
Methods	Accuracy(%)	Loss	Accuracy(%)	Loss	Accuracy(%)	Loss
Primitive	89.03	0.7017	83.10	0.8899	99.29	0.0698
Primitive + AMAA	89.27	0.4241	85.95	0.7994	99.47	0.0762
Primitive + MOSBC	86.72	0.6617	81.01	1.4036	98.79	0.1127
Primitive ++	**92.73**	**0.3697**	**88.57**	**0.3751**	**99.58**	**0.0211**

4.6. Comparative analysis

To evaluate the effectiveness of the proposed model, this section conducts comparative experiments on the Flower, Butterfly20 and Cargo datasets. Comparing AGMG-Net with representative traditional convolutional networks (VGG [42]), residual networks (ResNeSt [43]), visual transformers (ViT [44]) and CoAtNet [45] models that integrate the advantages of both convolution and transformer. The experiment maintains consistent training parameters, including batch size, learning rate, number of iterations and weight decay. The results are presented in Tables 6–8.

VGG represents traditional convolutional networks and performs excellently in image classification tasks due to its regular network structure, which comprises convolutional and pooling layers. This structure enables VGG to quickly capture image features within a limited number of training iterations. ResNeSt represents residual networks, which incorporate residual blocks that facilitate easier model training and enable capturing deep features of cargoes in complex environments. ViT is a transformer-based visual model that partitions the image into small blocks and processes them through multi-head self-attention mechanisms. This allows ViT to capture global information and the relationships between image blocks, resulting in high recognition accuracy on the ImageNet dataset. CoAtNet combines deep convolutional networks with self-attention mechanisms, enabling it to efficiently extract target features and perform well on both small-scale and large-scale datasets.

Table 6. Comparison of Classification models in Flower dataset. “Train” stands for training speed and “Test” stands for testing speed in Seconds Per Image (SPI).

Methods	Accuracy (%) ↑	F1-score (%) ↑	Loss↓	Train (SPI)↓	Test (SPI)↓
VGG	74.13	37.84	0.8189	**0.0066**	**0.0020**
ResNeSt	86.49	44.01	0.6651	0.0075	**0.0020**
ViT	69.86	34.08	0.8850	**0.0066**	0.0025
CoAtNet	88.50	54.11	0.3550	0.0103	0.0031
DeepMAD (SOTA)	90.57	61.33	**0.3368**	0.0226	0.0068
AGMG-Net	**92.73**	**63.71**	0.3697	0.0212	0.0127

Table 7. Comparison of Classification models in Butterfly20 dataset.

Methods	Accuracy (%) ↑	F1-score (%) ↑	Loss↓	Train (SPI)↓	Test (SPI)↓
VGG	62.62	34.11	1.1135	**0.0066**	**0.0020**
ResNeSt	77.62	41.45	0.8012	0.0076	0.0021
ViT	50.00	15.52	1.7806	**0.0066**	0.0026
CoAtNet	85.00	52.60	0.6337	0.0101	0.0031
DeepMAD (SOTA)	88.41	71.62	0.4246	0.0232	0.0081
AGMG-Net	**88.57**	**73.45**	**0.3751**	0.0217	0.0131

Table 8. Comparison of Classification models in Cargo dataset.

Methods	Accuracy (%) ↑	F1-score (%) ↑	Loss↓	Train (SPI)↓	Test (SPI)↓
VGG	99.15	94.69	0.0421	0.0067	**0.0019**
ResNeSt	99.21	95.97	0.0332	0.0075	0.0020
ViT	92.48	69.16	0.2025	**0.0066**	0.0025
CoAtNet	99.23	96.02	0.0244	0.0103	0.0032
DeepMAD (SOTA)	99.53	97.04	0.0256	0.0225	0.0065
AGMG-Net	**99.58**	**97.35**	**0.0211**	0.0194	0.0111

The results of ViT in Table 7 show that visual transformers excel at learning global and coarse-grained features, but they struggle to extract effective features from a limited number of samples. On the other hand, convolutional models like VGG and ResNeSt perform relatively poorly at learning fine-grained features of the target, but they can still achieve decent results on small-scale datasets. CoAtNet combines the strengths of both approaches, exhibits powerful feature extraction capabilities and achieves an impressive accuracy of 85%. AGMG-Net, with its CG-Net subnetwork that extracts finer-grained features and combines multiscale features, surpasses CoAtNet in accuracy, demonstrating the effectiveness of combining coarse and fine-grained features.

The information presented in Tables 6 and 8 shows that the distinctive characteristics of the ViT model are effectively utilized on the medium-scale datasets Flower and Cargo, facilitating the extraction of features at a global scale. Furthermore, the VGG and ResNeSt models demonstrate improved learning capabilities in capturing intricate details within the datasets. However, it is noteworthy that the CoAtNet model surpasses both VGG and ResNeSt in terms of its remarkable feature extraction ability, achieving notably high accuracy rates of 88.5 and 99.23%, as well as F1-scores of 52.6 and 96.02%, respectively. Nevertheless, the proposed AGMG-Net model in this study, which incorporates more comprehensive coarse-grained features and richer fine-grained features, exhibits a slight performance improvement over the CoAtNet model, yielding gains of 3.57 and 0.35% in accuracy. Comparing the running speeds of the models in Tables 6–8, it can be seen that AGMG-Net is more time-consuming in the inference process, but faster than the state-of-the-art (SOTA) model in the training process. The multi-branch model AGMG-Net achieves 99.58% accuracy on the Cargo dataset using only a slightly longer time, which shows that the model proposed in this paper is more suitable for security systems that do not require high recognition speed but have higher recognition accuracy.

The above experiments demonstrate that AGMG-Net can effectively leverage the multi-granularity features of the data, resulting in higher classification accuracy and precision. Additionally, AGMG-Net has the potential to accurately classify cargo in complex environmental conditions, demonstrating its practicality and usability for cargo recognition tasks.

5. Conclusions

The AGMG-Net proposed in this paper has significant advantages in cargo recognition. Unlike existing methods, AGMG-Net specifically considers the fine-grained features of cargo and leverages multiscale features to enhance recognition accuracy, even when data is limited and there are minimal differences between cargo classes. AGMG-Net incorporates the coarse-grained branch's AMAA module for target localization and the fine-grained branch's MOSBC module for target cropping. It then combines the feature maps of both branches through a multiscale fusion branch in Concat mode. Furthermore, it improves prediction accuracy by employing the majority voting method in the prediction layer.

The average recognition rates on the self-built Cargo dataset, as well as the public datasets Flower and Butterfly20, are 99.58, 92.73 and 88.57%. Experimental results demonstrate that AGMG-Net outperforms VGG, ResNeSt, ViT and CoAtNet in terms of classification effectiveness.

In conclusion, AGMG-Net is an effective cargo recognition model that enhances recognition accuracy and classification effectiveness through the integration of attention-guided multiscale feature fusion. It can be successfully applied to cargo recognition tasks within complex security system environments, thereby significantly contributing to cargo safety.

Use of AI tools declaration

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

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. A. Khan, W. Jue, M. Mushtaq, M. U. Mushtaq, Brain tumor classification in MRI image using convolutional neural network, *Math. Biosci. Eng.*, **17** (2020), 6203–6216. https://doi.org/10.3934/mbe.2020328
2. S. Cao, B. Song, Visual attentional-driven deep learning method for flower recognition, *Math. Biosci. Eng.*, **18** (2021), 1981–1991. https://doi.org/10.3934/mbe.2021103
3. S. M. Zainab, K. Khan, A. Fazil, M. Zakwan, Foreign Object Debris (FOD) classification through material recognition using deep convolutional neural network with focus on metal, *IEEE Access*, **11** (2023), 10925–10934. https://doi.org/10.1109/ACCESS.2023.3239424
4. Z. Cao, Y. Qin, Z. Xie, Q. Liu, E. Zhang, Z. Wu, et al., An effective railway intrusion detection method using dynamic intrusion region and lightweight neural network, *Measurement*, **191** (2022), 110564. https://doi.org/10.1016/j.measurement.2021.110564

5. F. Azam, A. Rizvi, W. Z. Khan, M. Y. Aalsalem, H. Yu, Y. B. Zikria, Aircraft classification based on PCA and feature fusion techniques in convolutional neural network, *IEEE Access*, **9** (2021), 161683–161694. https://doi.org/10.1109/ACCESS.2021.3132062
6. F. Peng, L. Qin, M. Long, Face morphing attack detection and attacker identification based on a watchlist, *Signal Process. Image Commun.*, **107** (2022), 116748. https://doi.org/10.1016/j.image.2022.116748
7. A. S. Jaggi, R. S. Sawhney, P. P. Balestrassi, J. Simonton, G. Upreti, An experimental approach for developing radio frequency identification (RFID) ready packaging, *J. Cleaner Prod.*, **85** (2014), 371–381. https://doi.org/10.1016/j.jclepro.2014.08.105
8. L. Cui, L. Wang, J. Deng, RFID technology investment evaluation model for the stochastic joint replenishment and delivery problem, *Expert Syst. Appl.*, **41** (2014), 1792–1805. https://doi.org/10.1016/j.eswa.2013.08.078
9. L. Tarjan, I. Šenk, S. Tegeltija, S. Stankovski, G. Ostojic, A readability analysis for QR code application in a traceability system, *Comput. Electron. Agric.*, **109** (2014), 1–11. https://doi.org/10.1016/j.compag.2014.08.015
10. Y. Zhu, W. Min, S. Jiang, Attribute-guided feature learning for few-shot image recognition, *IEEE Trans. Multimedia*, **23** (2021), 1200–1209. https://doi.org/10.1109/TMM.2020.2993952
11. X. Zeng, W. Wu, G. Tian, F. Li, Y. Liu, Deep superpixel convolutional network for image recognition, *IEEE Signal Process. Lett.*, **28** (2021), 922–926. https://doi.org/10.1109/LSP.2021.3075605
12. Y. K. Yi, Y. Zhang, J. Myung, House style recognition using deep convolutional neural network, *Autom. Constr.*, **118** (2020), 103307. https://doi.org/10.1016/j.autcon.2020.103307
13. O. C. Koyun, R. K. Keser, I. B. Akkaya, B. U. Töreyin, Focus-and-Detect: A small object detection framework for aerial images, *Signal Process. Image Commun.*, **104** (2022), 116675. https://doi.org/10.1016/j.image.2022.116675
14. S. Wang, M. Xu, Y. Sun, G. Jiang, Y. Weng, X. Liu, et al., Improved single shot detection using DenseNet for tiny target detection, *Concurrency Comput.:Exper. Pract.*, **35** (2023), 7491. https://doi.org/10.1002/cpe.7491
15. X. Dong, Y. Qin, R. Fu, Y. Gao, S. Liu, Y. Ye, et al., Multiscale deformable attention and multilevel features aggregation for remote sensing object detection, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022), 1–5. https://doi.org/10.1109/LGRS.2022.3178479
16. H. Cao, H. Liu, E. Song, C. Hung, G. Ma, X. Xu, et al., Dual-branch residual network for lung nodule segmentation, *Appl. Soft Comput.*, **86** (2020), 105934. https://doi.org/10.1016/j.asoc.2019.105934
17. H. Shi, G. Cao, Z. Ge, Y. Zhang, P. Fu, Double-branch network with pyramidal convolution and iterative attention for hyperspectral image classification, *Remote Sens.*, **13** (2021), 1403. https://doi.org/10.3390/rs13071403
18. J. Wang, Y. Cui, G. Shi, J. Zhao, X. Yang, Y. Qiang, et al., Multi-branch cross attention model for prediction of KRAS mutation in rectal cancer with t2-weighted MRI, *Appl. Intell.*, **50** (2020), 2352–2369. https://doi.org/10.1007/s10489-020-01658-8
19. D. Zhang, M. Ye, Y. Liu, L. Xiong, L. Zhou, Multi-source unsupervised domain adaptation for object detection, *Inf. Fusion*, **78** (2022), 138–148. https://doi.org/10.1016/j.inffus.2021.09.011

20. J. Cao, Y. Pang, S. Zhao, X. Li, High-level semantic networks for multi-scale object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **30** (2020), 3372–3386. https://doi.org/10.1109/TCSVT.2019.2950526
21. H. Xie, X. Zeng, H. Lei, J. Du, J. Wang, G. Zhang, et al., Cross-attention multi-branch network for fundus diseases classification using SLO images, *Med. Image Anal.*, **71** (2021), 102031. https://doi.org/10.1016/j.media.2021.102031
22. Q. Xu, C. Yang, J. Tang, B. Luo, Grouped bidirectional LSTM network and multistage fusion convolutional transformer for hyperspectral image classification, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–14. https://doi.org/10.1109/TGRS.2022.3207294
23. Y. Zhang, X. Liu, S. Wa, S. Chen, Q. Ma, GANsformer: A detection network for aerial images with high performance combining convolutional network and transformer, *Remote Sens.*, **14** (2022), 923. https://doi.org/10.3390/rs14040923
24. C. Yu, Y. Liu, C. Li, L. Qi, X. Xia, T. Liu, et al., Multibranch feature difference learning network for cross-spectral image patch matching, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 1–15. https://doi.org/10.1109/TGRS.2022.3176358
25. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319
26. M. Meng, T. Zhang, W. Yang, J. Zhao, Y. Zhang, F. Wu, Diverse complementary part mining for weakly supervised object localization, *IEEE Trans. Image Process.*, **31** (2022), 1774–1788. https://doi.org/10.1109/TIP.2022.3145238
27. F. Shao, L. Chen, J. Shao, W. Ji, S. Xiao, L. Ye, et al., Deep learning for weakly-supervised object detection and localization: A survey, *Neurocomputing*, **496** (2022), 192–207. https://doi.org/10.1016/j.neucom.2022.01.095
28. Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 12272–12281. https://doi.org/10.1109/CVPR42600.2020.01229
29. J. Choe, D. Han, S. Yun, J. Ha, S. J. Oh, H. Shim, Region-based dropout with attention prior for weakly supervised object localization, *Pattern Recognit.*, **116** (2021), 107949. https://doi.org/10.1016/j.patcog.2021.107949
30. B. Wang, C. Yuan, B. Li, X. Ding, Z. Li, Y. Wu, et al., Multi-scale low-discriminative feature reactivation for weakly supervised object localization, *IEEE Trans. Image Process.*, **30** (2021), 6050–6065. https://doi.org/10.1109/TIP.2021.3091833
31. Z. Ling, L. Li, A. Zhang, RSMNet: A regional similar module network for weakly supervised object localization, *Neural Process. Lett.*, **54** (2022), 5079–5097. https://doi.org/10.1007/s11063-022-10849-y
32. D. Hwang, J. Ha, H. Shim, J. Choe, Entropy regularization for weakly supervised object localization, *Pattern Recognit. Lett.*, **169** (2023), 1–7. https://doi.org/10.1016/j.patrec.2023.03.018
33. L. Zhang, H. Yang, Adaptive attention augmentor for weakly supervised object localization, *Neurocomputing*, **454** (2021), 474–482. https://doi.org/10.1016/j.neucom.2021.05.024

34. W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, et al., Ts-cam: Token semantic coupled attention map for weakly supervised object localization, in *2021 IEEE International Conference on Computer Vision (ICCV)*, (2021), 2866–2875. https://doi.org/10.1109/ICCV48922.2021.00288
35. S. Wang, Z. Wang, H. Li, J. Chang, W. Ouyang, Q. Tian, Semantic-guided information alignment network for fine-grained image recognition, in *IEEE Trans. Circuits Syst. Video Technol.*, (2023). https://doi.org/10.1109/TCSVT.2023.3263870
36. R. Du, D. Chang, A. K. Bhunia, J. Xie, Z. Ma, Y. Song, et al., Fine-grained visual classification via progressive multi-granularity training of jigsaw patches, in *European Conference on Computer Vision*, **12365** (2020). https://doi.org/10.1007/978-3-030-58565-5_10
37. S. Wang, J. Chang, H. Li, Z. Wang, W. Ouyang, Q. Tian, Open-set fine-grained retrieval via prompting vision-language evaluator, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 19381–19391.
38. S. Wang, J. Chang, Z. Wang, H. Li, W. Ouyang, Q. Tian, Fine-grained retrieval prompt tuning, in *AAAI Conference on Artificial Intelligence*, **37** (2023), 2644–2652. https://doi.org/10.1609/aaai.v37i2.25363
39. H. Ebbinghaus, Memory: a contribution to experimental psychology, *Ann. Neurosci.*, **20** (2013), 155. https://doi.org/10.5214/ans.0972.7531.200408
40. F. Meyer, Color image segmentation, in *1992 International Conference on Image Processing and its Applications*, (1992), 303–306.
41. T. Chen, W. Wu, Y. Gao, L. Dong, X. Luo, L. Lin, Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding, in *26th ACM international conference on Multimedia*, (2018), 2023–2031. https://doi.org/10.1145/3240508.3240523
42. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
43. H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, et al., ResNeSt: Split-attention networks, in *2022 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2022), 2735–2745. https://doi.org/10.1109/CVPRW56347.2022.00309
44. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
45. Z. Dai, H. Liu, Q. V. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, preprint, arXiv:2106.04803.