mathematics

MDPI

Article

MFANet: A Collar Classification Network Based on Multi-Scale Features and an Attention Mechanism

Xiao Qin [1,2,3], Shanshan Ya [2,3], Changan Yuan [2,4,*], Dingjia Chen [2,3], Long Long [2,3] and Huixian Liao [2,3]

[1] Department of Software Engineering, College of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China
[2] Guangxi Key Lab of Human–Machine Interaction and Intelligent Decision, Guangxi Academy of Science, Nanning 530007, China
[3] Center for Applied Mathematics of Guangxi, Nanning Normal University, Nanning 530100, China
[4] Department of Mathematics and Computer Science, College of Education, Guangxi College of Education, Nanning 530023, China
* Correspondence: yuanchangan@126.com

Abstract: The collar is an important part of a garment that reflects its style. The collar classification task is to recognize the collar type in the apparel image. In this paper, we design a novel convolutional module called MFA (multi-scale features attention) to address the problems of high noise, small recognition target and unsatisfactory classification effect in collar feature recognition, which first extracts multi-scale features from the input feature map and then encodes them into an attention weight vector to enhance the representation of important parts, thus improving the ability of the convolutional block to combat noise and extract small target object features. It also reduces the computational overhead of the MFA module by using the depth-separable convolution method. Experiments on the collar dataset Collar6 and the apparel dataset DeepFashion6 (a subset of the DeepFashion database) show that MFANet is able to perform at a relatively small number of collars. MFANet can achieve better classification performance than most current mainstream convolutional neural networks for complex collar images with less computational overhead. Experiments on the standard dataset CIFAR-10 show that MFANet also outperforms current mainstream image classification algorithms.

Keywords: deep learning; image classification; collar classification; attention mechanism; multi-scale

MSC: 68T07

Citation: Qin, X.; Ya, S.; Yuan, C.; Chen, D.; Long, L.; Liao, H. MFANet: A Collar Classification Network Based on Multi-Scale Features and an Attention Mechanism. *Mathematics* **2023**, *11*, 1164. https://doi.org/10.3390/math11051164

Academic Editor: Jakub Nalepa

Received: 17 January 2023
Revised: 16 February 2023
Accepted: 21 February 2023
Published: 27 February 2023

1. Introduction

The collar classification task is a data processing task based on clothing images, and its main purpose is to accurately identify the collar types in clothing images. At present, domestic and foreign scholars' research on clothing images mainly focus on the classification tasks [1–3] of clothing style, pattern and other aspects. Liu et al. established the DeepFashion dataset and proposed a deep model FashionNet [4] for clothing style and multi-attribute classification based on VGGNet [5]. Other studies based on the DeepFashion dataset also classify the garment as a whole and do not involve the classification task of collars as well. However, the collar, as a part of garment, is one of the important factors affecting the style and fashion of the garment. The collar classification task can provide powerful technical support for garment retrieval. Since the collar is only a small part of the garment and fits around the neck of the body, the collar part has a small area in the garment image and is easily deformed, which poses a challenge to the collar classification task. Huang et al. collected and organized 18847 garment images to build the collar classification dataset Collar6, and proposed the EMRes-50 [6] network for collar image classification. The network used the attention method combined with MC-Loss (mutual-channel loss) to make

the neural network focus on the more discriminative collar part and obtained a classification accuracy of 73.6% on the Collar6 dataset. However, this study neglected the nature of containing multiple targets at different scales in clothing images. Therefore, the proposed network is less capable of multi-scale feature extraction, which leads to less superior classification results. In recent years, works have effectively improved model performance by means of designing excellent multi-scale feature extraction methods [7,8] that can effectively model global information and capture semantic relationships between different objects by using different perceptual fields to extract features of objects at different scales. The application of multi-scale feature extraction in collar classification algorithms would be a good choice, but when using large convolutional kernels to obtain large receptive fields, it will lead to an increase in convolutional computation. Due to the small collar area and small differences between classes, it is also necessary to use attention mechanisms to purposely guide the network to focus more on collar features when modelling global information. The starting point of this paper is to reduce the computational overhead incurred when performing multi-scale feature extraction and to use the attention mechanism to weigh the important features so that the neural network focuses more on the task-relevant regions. An effective way of combining multi-scale feature extraction methods with attention mechanisms is investigated to enable the network to perform better and thus be applied in collar classification tasks.

Based on the above research, this paper combines multi-scale features with attention mechanisms and proposes a new module named MFA (multi-scale feature attention). Specifically, the input feature map is first subjected to a depth-wise separable convolution operation using sets of convolution operators of different sizes to separately extract feature information at different scales for modelling contextual information, thus capturing global dependencies while also reducing the computational overhead. Then, the channel attention weights of the multi-scale feature mappings are learned and recalibrated by aggregating the channel attention weights of different dimensions through a non-linear activation function to obtain the attention weights that include long-range dependencies. Finally, the obtained attention weights are dot-producted with the input feature map to obtain the attention feature map. A novel network, MFANet, is constructed by replacing the 3×3 convolutional blocks in the residual blocks of ResNet [9] with MFA modules. The proposed MFANet not only outperforms existing techniques in terms of accuracy, but also requires fewer parameters. The main contributions of this work are summarized as follows:

1. A novel module MFA is proposed, which is able to efficiently extract multi-scale feature information while reducing the computational overhead, incorporating a lightweight attentional approach to highlight the feature representation of key components and regions.
2. A new network architecture, MFANet, is built on the basis of the MFA module, which inherits the advantages of the MFA module and can better handle image classification problems that contain multiple objects, multiple noises and a small percentage of recognition targets.
3. Extensive experimental results show that, compared with the current mainstream network structure, the proposed MFANet obtains significant gains in classification accuracy on the collar image dataset Collar6 with fewer parameters and computations, achieving better classification results on the fashion dataset DeepFashion6, similarly obtaining accuracy gains on the standard classification dataset CIFAR-10.

Next, we elaborate on our work in four aspects. Firstly, Section 2 focuses on the discussion of the relevant algorithms used in the text. Secondly, Section 3 elaborates the main work of this paper, i.e., the construction of the MFA module and MFANet and the related algorithms. Subsequently, our experimental results are described in detail in Section 4. Finally, we give a concise summary of the work in this paper and present the outlook.

2. Related Work

The MFANet proposed in this paper is primarily driven by the ideas of multi-scale feature extraction methods and attention mechanisms. In this section, the related work of previous researchers concerning multi-scale feature extraction methods as well as attention mechanisms are discussed.

2.1. Multi-Scale in Computer Vision

Convolutional neural networks perceive different scale features by the receptive field. If the receptive field is too small only local features can be observed, and if the receptive field is too large too much noisy information will be acquired. Therefore, acquiring information from different scales is crucial for vision tasks that require understanding parts and objects, such as fine-grained classification [10], object detection [11], and semantic segmentation [12]. Using convolution to vary the perceptual field size to obtain multi-scale information is one of the mainstream approaches. Related studies such as InceptionNet [13] perform multi-scale feature extraction by constructing four parallel branching structures using convolutional operators of different sizes at different receptive fields, and Res2Net [14] constructs layered residual connections within a single residual block using a single convolutional operator, enabling it to vary the receptive field at a fine level to capture details and global features. However, the above-mentioned neural networks require convolutional computations using larger convolutional kernels when expanding the receptive field, leading to an increase in network computation.

2.2. Channel Attention

The attention mechanism [15] has been proven to be an effective way to improve the performance of neural networks and is widely used in computer vision tasks such as image classification [16–18], object detection [19], semantic segmentation [20], face recognition [21,22], person re-identification [23], etc. SENet [16] first proposed an effective mechanism for learning channel attention, which can weigh the degree of importance of each channel to improve the sensitivity of the model to channel information, thus reinforcing important features to suppress non-important features, achieving satisfactory performance at that time. Subsequently, a large amount of research on channel attention has been mainly conducted based on SENet [16]. Wang et al. found that SENet [16] performs a dimensionality reduction operation to reduce the complexity of the model, but it destroys the direct correspondence between channels and their weights. To overcome this shortcoming of SENet [16], Wang et al. proposed a strategy to achieve local cross-channel interactions by one-dimensional convolutions without dimensionality reduction, named ECA [17] (efficient channel attention), which performs well in various vision tasks.

Our work addresses the characteristics of collar image data with a lot of noise and small targets, and provides an effective combination of multi-scale feature extraction and channel attention methods, using multi-scale feature extraction methods to obtain the features of the image and attention mechanisms to guide the network to pay more attention to the collar features. Meanwhile, the depth-separable convolution is used to reduce the computational overhead caused by multi-scale feature extraction operations (See Table 1).

Table 1. The similarities and differences between our work and related multi-scale feature extraction methods and attention mechanisms.

Network	Includes Attention Mechanism	Includes Multi-Scale Feature Extraction	Computational Volume Scale	Embeddable Modules
InceptionNet	✗	✓	the larger	✗
Res2Net	✗	✓	the larger	✗
SENet	✓	✗	small	✓
ECANet	✓	✗	small	✓
Ours	✓	✓	small	✗

3. Proposed Method

To construct an efficient collar classification network, we propose a MFA module that combines multi-scale feature extraction and attention mechanisms and refers to the network structure of ResNet50 [9] to construct bottleneck blocks with the MFA module, as shown in Figure 1 (right). By stacking MFA bottleneck blocks, we build the collar classification network, MFANet. The detailed structure and related algorithms of the MFA module will be described in the first part of this section and the network structure and specific implementation of MFANet will be described in the second part of this section.

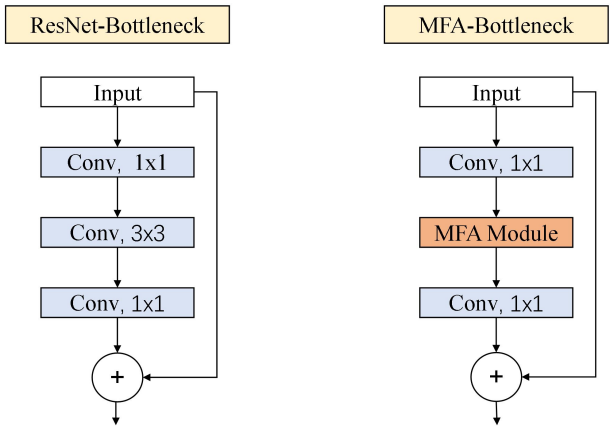

Figure 1. Illustrations and comparison of the ResNet block and MFANet block proposed in this paper.

3.1. MFA Module

The core function of the MFA module proposed in this paper is to extract multi-scale feature information and weigh the channels, and its structure is shown in Figure 2. The workflow of the MFA module is divided into 3 steps. Firstly, the input feature map is convolved with convolution operators of different sizes to extract multi-scale feature information; Secondly, the ECA [17] module is used to extract the channel attention weights from the feature maps at different scales and then calibrated by the non-linear activation function to obtain the channel attention weights with multi-scale information. Finally, the obtained channel attention weights are multiplied with the corresponding feature maps to obtain a finer feature map as the output. The above process will be explained in detail in the following section.

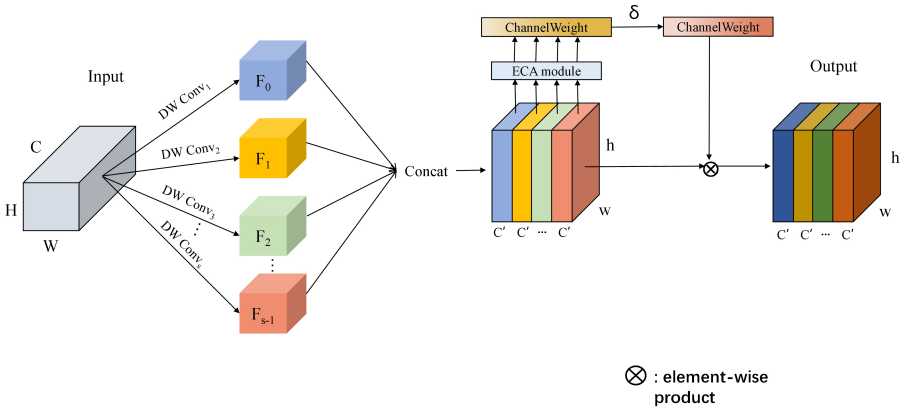

Figure 2. MFA module.

3.1.1. Multi-Scale Feature Information Extraction

In order to obtain the multi-scale features of the input image, it is necessary to convolute the images with different sizes of convolution. Let $X \in \mathbb{R}^{H,W,C}$ be the input feature map, and after conducting parallel convolution of the input feature map using s convolution kernels of different sizes, we can obtain s outputs, each with a channel $C' = \frac{C}{s}$ of the output feature map, and for each output, the module learns the feature map of a single convolution operator. If feature extraction is performed on X using standard convolution, assuming a step size of 1 and a fill of 1, the output feature map for one standard convolution is computed as:

$$F_{h',w',n} = \sum_{k_h,k_w,m} K_{k_h,k_w,m,n} \cdot X_{h'+k_h-1,w'+k_w-1,m} \tag{1}$$

where F is the output feature map; h', w', and n denote the height, width, and number of channels of the output feature map; K is the convolution kernel; k_h and k_w denote the height and width of the convolution kernel; and $h' = [0, H - Dk_i + 3]$, $w' = [0, W - Dk_i + 3]$, $n = [0, C']$, $m = [0, C]$, $k_h = k_w = [0, Dk_i]$, and Dk_i denote the height and width of the convolution kernel needed for the i-th convolution. The computational cost required for one standard convolution operation is:

$$Dk_i \cdot Dk_i \cdot C \cdot C' \cdot H \cdot W \tag{2}$$

where the computational cost is multiplicatively related to the number of input channels C, the number of output channels C', the kernel size Dk_i and the input feature map size. The computational cost will undoubtedly increase when using multiple convolution 160 kernels for convolution operations.

In order to reduce the computational effort caused by multiple standard convolutions while being able to better extract the feature information, we use the depth-separable convolution to convert part of the multiplicative relationship in Equation (2) into an additive relationship to achieve the purpose of reducing the computational effort. We denote the depth-wise separable convolution [24,25] operation as DW, then the steps for performing one DW operation on the feature map $X \in \mathbb{R}^{H,W,C}$ are as follows: First, the depth-wise convolution operation is performed. The convolution kernel is split into a single channel using depth-wise convolutions, and the convolution operation is performed on each channel without changing the depth of the input feature image. This operation can be expressed as:

$$\hat{F}_{h',w',m} = \sum_{k_h,k_w} \hat{K}_{k_h,k_w,m} \cdot X_{h'+k_h-1,w'+k_w-1,m} \tag{3}$$

where $\hat{F}$ is the feature map output from the first operation of the depth-wise convolution and $\hat{K}$ is the depth-wise convolutional kernel of size $Dk_i \times Dk_i \times 1$. The m_{th} filter in $\hat{K}$ is applied to the m_{th} channel in X to produce the channel of the filtered output feature map $\hat{F}$.

The range of values of h', w', m, k_h, and k_w is consistent with the description in Equation (1). Second, point-wise convolutions are performed. After the depth-wise operation, the feature channels are aggregated and the number of output channels is changed using a convolution operator of size 1×1. This operation can be expressed as:

$$F'_{h',w',n} = \sum_m K'_{1,1,m,n} \cdot \hat{F}_{h',w',m} \tag{4}$$

where F' is the output feature map after point-wise convolutional, and K' is the pointwise convolution kernel of size $1 \times 1 \times C$. The range of values of h', w', n, and m is consistent with the description in Equation (1). The computational cost required for one depth-wise separable convolution operation is:

$$Dk_i \cdot Dk_i \cdot C \cdot H \cdot W + C \cdot C' \cdot H \cdot W \tag{5}$$

By expressing convolution as a two-step process of filtering and combining we obtain a reduction in computation of:

$$\frac{Dk_i \cdot Dk_i \cdot C \cdot H \cdot W + C \cdot C' \cdot H \cdot W}{Dk_i \cdot Dk_i \cdot C \cdot C' \cdot H \cdot W} = \frac{1}{C'} + \frac{1}{Dk_i^2} \tag{6}$$

The DW process is shown in Algorithm 1.

Algorithm 1 DW algorithm

Input: Feature map X, K_{size}
Output: Feature map F'

1: $\hat{F}_{h',w',m} = \sum_{k_h,k_w} \hat{K}_{k_h,k_w,m} \cdot X_{h'+k_h-1,w'+k_w-1,m}$
2: $F'_{h',w',n} = \sum_m K'_{1,1,m,n} \cdot \hat{F}_{h',w',m}$

3.1.2. Attention Weight Computing

In order to enable the network to adaptively weigh the importance of the channels of each input feature map and output more important information to improve the overall performance of the MFA module, after obtaining the features, we use the attention method ECA [17] to obtain the channel attention. The global average pooling operation is first performed on the input feature map F', to compress the spatial information and apply it to the channel information. The global average pooling operation can be calculated as:

$$y = g(F') = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} F'_{ij} \tag{7}$$

where y is the feature mapping output after global average pooling and $F' \in \mathbb{R}^{h,w,C'}$ is the output feature map after depth-wise separable convolution; and h and w denote the height and width of F', respectively. The attention weights w_i for each channel y_i can be learned by

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w^j y_i^j\right), y_i^j \in \Omega_i^k \tag{8}$$

where σ is a sigmoid function, Ω_i^k denotes the k neighbouring channels of y_i, w^j is the parameter matrix, and all channels share the same learning parameters. The way ECA [17] obtains cross-channel attention weights can be readily implemented by a fast 1D convolution with a kernel size of k using Equation (9) to obtain the channel attention mapping.

$$Z' = \sigma\big(C1Dk\big(GAP\big(F'\big)\big)\big) \tag{9}$$

where Z' is the channel weight vector obtained after ECA module processing, C1D denotes the 1D convolution, and k denotes the range of channel interactions. The value of k is analysed in detail in the experiments in Section 4. GAP indicates the global average pooling operation. With the channel attention module ECA, it is possible to use as little computation as possible to obtain cross-channel attention weights without dimensionality reduction.

3.1.3. Attentional Calibration and Feature Aggregation

After s times of the above operations, we will obtain s output feature maps containing different scale information, and s channel attention weight mappings. The description is as follows:

$$F_i' = DW(X, Dk_i) \quad i = 0, 1, 2 \cdots s-1 \tag{10}$$

$$Z_i' = \sigma\big(C1Dk\big(GAP\big(F_i'\big)\big)\big) \quad i = 0, 1, 2 \cdots s-1 \tag{11}$$

where F_i' is the feature map output from the i-th depth-wise separable convolution operation and Z_i' is the attention weight vector output from F_i' after the ECA module. Next, we first aggregate the different attention mappings and recalibrate them using a non-linear operation. This operation can be expressed as follows:

$$Z = \delta\big(\big[Z_0' \oplus Z_1' \oplus \cdots Z_{S-1}'\big]\big) \tag{12}$$

where Z is the aggregated attention weight vector, which aggregates the output vectors of several different scale feature maps after the ECA module, δ represents the non-linear activation function softmax, and $\oplus$ represents the concat operation. Finally, the feature maps containing information at different scales are aggregated, and the attention weight mapping after recalibration is applied to the aggregated feature maps to obtain the final output. This process can be expressed as:

$$X_{\text{out}} = Z \odot \big(F_0' \oplus F_1' \oplus \cdots F_{s-1}'\big) \tag{13}$$

where $\odot$ represents the channel-wise multiplication, X_{out} is the final output after the MFA module. To date, the output feature maps after going through the MFA module both combine background information at different scales and produce better pixel-level attention to the channels containing important features. The overall flow of the MFA module algorithm is shown in the Algorithm 2.

Algorithm 2 MFA algorithm

Input: Feature map X
Output: Feature map X_{out}

for i = 1 to n
    $F_i = DW(X, K_{size})$
    $Z_i = \text{Sigmoid}(\text{Conv 1D}k(GAP(F_i)))$
$F = \text{Concat}(F_1, \ldots, F_n)$;
Feats $\leftarrow F$
$Z = \text{Concat}(Z_1, \ldots, Z_n)$
Attention_Vectors = softmax (Z)
X_{out} = Feats $\odot$ Attention_Vectors
return X_{out}

3.2. Network Design

ResNet [9] was proposed by Kaiming He et al. in 2016 and is still used as a backbone network for many computer vision tasks. The residual structure used by ResNet [9] can effectively solve the vanishing gradient (exploding gradient) of neural networks due to layers that are too deep. A large part of the subsequent research on neural network structures is based on the ResNet [9] structure, extending or improving it.

The bottleneck block of ResNet [9] mainly uses a set of 3×3 convolution operators to extract features. In this paper, we seek a higher performance architecture to replace the 3×3 convolution to accomplish the extraction of image features, while also ensuring a smaller computational load. The MFA module proposed in Section 3.1 of this paper possesses an efficient multi-scale feature extraction capability and an attention mechanism, both of which have been shown to be effective ways to improve the performance of neural networks. Therefore, based on the above-proposed MFA module, we improved the bottleneck structure of ResNet-50 [9]. Since the multi-scale feature extraction operation leads to inconsistent sizes of the output feature map, we take the size of the output feature map after 3×3 convolution in each bottleneck block of ResNet-50 [9] as the standard, and fill the convolution operations of different scales according to the convolution calculation formula to ensure that the output feature map size is consistent and conforms to the output of the ResNet [9] bottleneck block, so that the bottleneck of ResNet [9] block 3×3 convolution is replaced with the MFA module.

The MFA module with multi-scale feature extraction capabilities as well as channel attention is embedded in the residual blocks, compensating the shortage of feature extraction abilities of the single convolution operator and reduce the computational effort to a great extent by using depth-wise separable convolution operations instead of normal convolution operations. These residual blocks are stacked in the style of ResNet [9] to construct a new simple and effective neural network MFANet, whereby MFANet can extract finer multi-scale feature information and obtain the importance of each channel without increasing the amount of computation. It effectively solves the collar classification problem that contains multiple objects, noises and small percentage of recognition targets. The layer structure of the MFANet proposed in this paper is shown in Table 2.

Table 2. MFANET network design.

Output	ResNet-50	MFANet
$112 \times 112 \times 64$	7×7, 64, stride = 2	
$56 \times 56 \times 64$	3×3, max pool, stride = 2	
$56 \times 56 \times 256$	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, & 64 \\ MFA, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$
$28 \times 28 \times 512$	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, & 128 \\ MFA, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$
$14 \times 14 \times 1024$	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, & 256 \\ MFA, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$
$7 \times 7 \times 2048$	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, & 512 \\ MFA, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$
1×1	7×7, global average pool, n-d fc	

4. Experiments

Experimental section elaborates conducted experiments, including the introduction of the experimental environment, datasets, and most importantly, the accuracy of MFANet on three classification datasets, Collar6, DeepFashion6, and CIFAR-10, compared with the current mainstream classification networks. Meanwhile, the ablation experiment of the hyperparameter k in Equation (9) is conducted so as to choose the value that makes the performance of MFANet optimal.

4.1. Experimental Configuration

The main hardware device used in the experiments is an NVIDIA GeForce RTX 2080Ti. The size of the input tensor is randomly cropped to 224 × 224 by random horizontal flipping and normalization. The batch size is set to 64, and a total of 120 epochs are iteratively trained. Optimization is performed by using an adaptive moment estimation (Adam) with a weight decay of 1×10^{-4}. The initial learning rate is set to 0.01, and the cosine annealing learning rate is implemented using a custom learning rate adjustment function (LambdaLR). The label-smoothing regularization is used with the coefficient value as 0.1 during training.

4.2. Datasets

4.2.1. CIFAR-10 Dataset

The CIFAR-10 dataset is a public dataset for identifying pervasive objects and is widely used in image classification tasks for deep learning. It consists of 60,000 colour RBG images covering 10 categories (aircraft, cars, birds, cats, deer, dogs, frogs, horses, boats, and trucks). There are 6000 images for each type with a 32 × 32 image size, of which 50,000 images are used for training and 10,000 for testing. Figure 3 shows the classes in CIFAR-10 and 10 random images for each class.

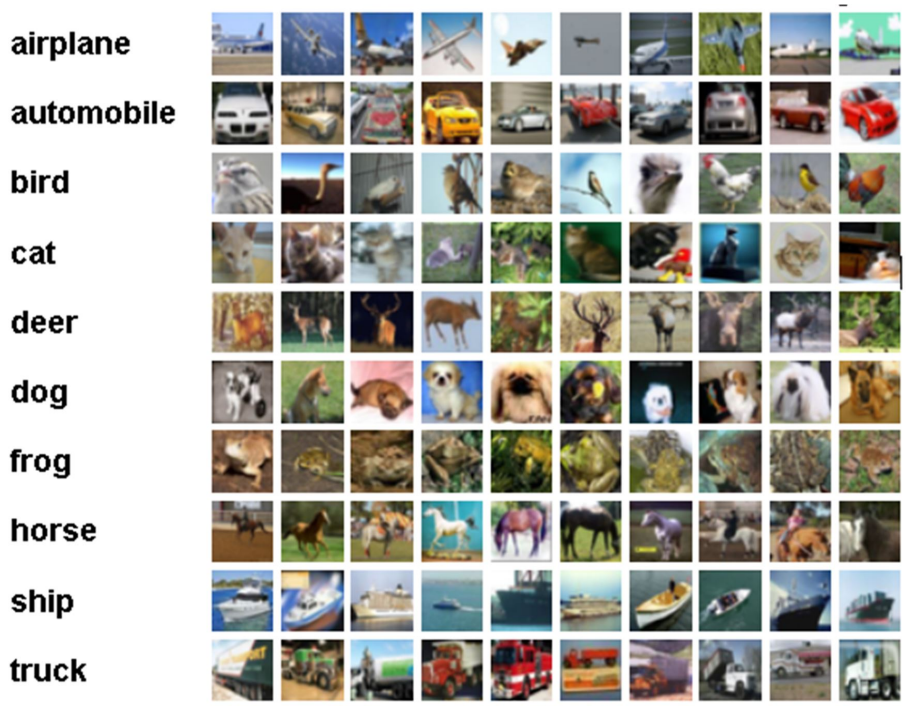

Figure 3. CIFAR-10 partial image data.

4.2.2. Collar6 Dataset

The Collar6 dataset are collected and sorted by manual collection and crawlers from all major domestic e-commerce platforms, including round collar, lapel, stand-up collar, hood, V-neck and fur collar, with a total of 18,847 images. The collar part, which plays a key role in the classification task, only accounts for a small part of the image, which poses a considerable challenge to the classification model and tests the ability of the classification model to extract key features. The detailed data distribution of the Collar6 dataset is shown in Table 3, and the presentation of some of the image data is shown in Figure 4.

Table 3. Data distribution for the Collar6 dataset.

Type	The Number of Training Set Pictures	The Number of Test Set Pictures	Total
Crew neck	2480	620	3100
Lapel	2608	652	3260
Stand collar	2464	616	3080
Hoodie	2560	640	3200
V-neck	2468	617	3085
Fur lapel	3122	625	3122

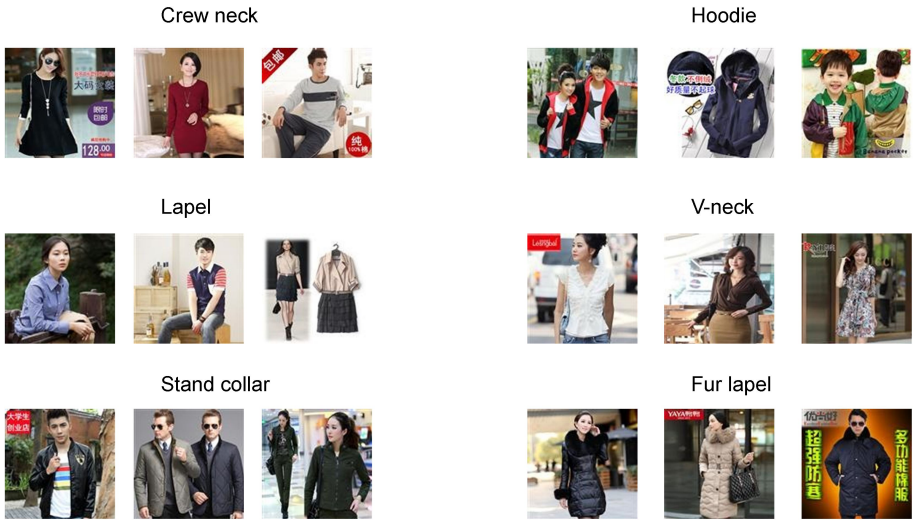

Figure 4. Data distribution for the Collar6 dataset.

4.2.3. DeepFashion6

DeepFashion is an open and large clothing dataset collected and organized by the Chinese University of Hong Kong. It contains 800,000 images, including different angles, different scenes, buyers' shows, and so on. The subset "Category and Attribute Prediction Benchmark" is used to perform the classification and attribute prediction, which contains 50 kinds of classification tags and 1000 kinds of attribute tags, with 289,222 images. In order to verify that our method is also effective on the DeepFashion dataset, we extracted clothing images of six kinds of collars, and selected more than 3000 images of each collar to form the DeepFashion6 dataset. The dataset composition is shown in Table 4.

Table 4. Data distribution for the DeepFashion6 dataset.

Type	The Number of Training Set Pictures	The Number of Test Set Pictures	Total
Dress	2555	639	3194
Jacket	2505	627	3132
Jeans	2412	603	3015
Shorts	2541	636	3177
Tank	2528	632	3160
Tee	2439	610	3049

4.3. Comparative Experiments

4.3.1. Experimental Results and Analysis of Comparison with Mainstream Neural Networks on the Collar6 Dataset

To verify the performance of the MFANet on realistic collar image classification tasks, we conducted experiments comparing MFANet with various types of mainstream neural networks on the Collar6 dataset, and the experimental results are shown in Table 5. Compared with the backbone networks ResNet-50 [9], Res2Net [14], DenseNet [26], etc.,

our MFANet shows a large improvement in accuracy. Compared with MobileNet_v3 [24] and ShuffleNet_v2 [27], which are lightweight networks designed for mobile, they have improved by 6.9% and 5.3% in accuracy, respectively. Compared with the networks EPSANet [28] and SKNet [29] with attention mechanisms and multi-scale design, our MFANet improves by 2.3% and 24.3% in accuracy, respectively. The same networks containing multi-scale design and attention mechanisms, SKNet [29] and EPSANet [28] both use SE [16] to obtain channel attention, and where the dimensionality reduction operation causes some information loss. SKNet [29] uses summation for multi-scale feature fusion before passing through the attention module, which leads to multi-scale information loss to some extent. Our MFANet performs the attention weighing operation on its output feature map separately after extracting multi-scale features, which better preserves the representation of multi-scale features by introducing the attention method without dimensionality reduction to ensure that channel attention is not lost. Compared with the latest research results in collar classification EMRes-50 [6], MFANet achieves an effective combination of attention and multi-scale feature extraction, improves the classification accuracy by 6.8%, and is more capable of accurately classifying complex collar images.

Table 5. Comparative experimental results of MFANet and various neural networks on the Collar6 dataset.

Network	Parameters	FLOPs	Top-1 Accuracy (%)
EMRes-50 [6]	28.02 M	4.34 G	73.6
ResNet-50 [9]	23.52 M	4.12 G	66.5
ResNeXt-50 [30]	22.99 M	4.26 G	75.7
Res2Net [14]	23.66 M	4.29 G	74.8
DenseNet-161 [26]	26.49 M	7.82 G	72.3
Xception [25]	20.82 M	4.58 G	76.3
EPSANet [28]	20.53 M	3.63 G	78.1
SKNet [29]	25.44 M	4.51 G	56.1
MobileNet_v3_small [24]	1.52 M	58.79 M	73.1
MobileNet_v3_large [24]	4.21 M	226.4 M	73.5
ShuffleNet_v2 [27]	1.26 M	149.58 M	75.1
SqueezeNet [31]	0.73 M	2.65 G	57.1
Ours	13.81 M	2.61 G	**80.4**

4.3.2. Experimental Results and Analysis of Comparative Experiments with Mainstream Attention Networks on the Collar6 Dataset

Table 6 shows the comparison results of the proposed MFANet on the Collar6 dataset with the mainstream classification network after adding the attention method. In order to be able to make a full comparison, we embed multiple attention methods in different networks for comparison. These networks include the basic residual network ResNet-50 [9], the aggregated residual network ResNeXt-50 [30], the Res2Net [14] with multi-scale design, and the dense connections network DenseNet [26]. It can be seen from Table 6 that the accuracy rate of MFANet reaches 80.4%, which is higher than other networks, and obtains very competitive performance at a lower cost. The experimental results in Table 6 show that our network design is reasonable.

Table 6. Comparative experimental results of MFANet and various neural networks with added attention methods on the Collar6 dataset.

Method	Backbone Models	Parameters	FLOPs	Top-1 Accuracy (%)
SENet [16]	ResNet-50	26.05 M	4.12 G	67.9
CBAM [32]		26.05 M	4.12 G	67.3
ECANet [17]		23.52 M	4.12 G	68.8
CANet [33]		25.43 M	4.14 G	66.7
FcaNet [34]		26.03 M	4.12 G	69.7
Ours		13.81 M	2.61 G	**80.4**
SENet [16]	ResNeXt-50	25.51 M	4.27 G	74.3
CBAM [32]		25.52 M	4.27 G	71.8
ECANet [17]		22.99 M	4.26 G	75.7
CANet [33]		24.91 M	4.29 G	69.9
FcaNet [34]		25.51 M	4.27 G	71.8
SENet [16]	Res2Net	26.18 M	4.29 G	73.9
CBAM [32]		26.20 M	4.29 G	70.3
ECANet [17]		23.66 M	4.29 G	72.2
CANet [33]		25.58 M	4.31 G	70.8
FcaNet [34]		26.18 M	4.29 G	71.6
SENet [16]	DenseNet-161	27.14 M	7.82 G	70.4
CBAM [32]		27.14 M	7.82 G	73.2
ECANet [17]		26.49 M	7.82 G	72.9
CANet [33]		26.98 M	7.83 G	71.9

4.3.3. Comparative Experimental Results and Analysis on the DeepFashion6 Dataset

We extracted some data from the large dress public dataset DeepFashion, collated a dress classification subset DeepFashion6, and verified the effectiveness of our method on this dataset for the dress style classification task, and the experimental results are shown in Table 7. The models compared include the basic backbone networks ResNet [9], ResNeXt [30] and DenseNet [26]; the attention networks SENet [16], ECANet [17], and CANet [33]; networks containing multi-scale designs Res2Net [14], and Xception [25]; and the attention mechanism and multi-scale design networks EMRes-50 [6], EPSANet [28], SKNet [29], etc. The MAFNet proposed in this paper achieves an accuracy of 87.7% on the DeepFashion6 dataset, higher than all the above-mentioned networks. It can be seen that rich multi-scale feature information as well as attention information are equally effective in improving the accuracy of clothing style classification.

Table 7. Comparison experimental results on the clothing style classification dataset DeepFashion6.

Network	Parameters	FLOPs	Top-1 Accuracy (%)
EMRes-50 [6]	28.02 M	4.34 G	86.1
CANet [33]	23.52 M	4.12 G	86.4
ECANet [17]	22.99 M	4.26 G	86.3
SENet [16]	23.66 M	4.29 G	86.3
ResNet-50 [9]	26.49 M	7.82 G	85.8
ResNeXt-50 [30]	20.82 M	4.58 G	86.5
Res2Net [14]	20.53 M	3.63 G	87.0
DenseNet-161 [26]	25.44 M	4.51 G	87.3
EPSANet [28]	20.53 M	3.63 G	87.4
SKNet [29]	25.44 M	4.51 G	83.8
Xception [25]	20.83 M	4.58 G	87.3
Ours	13.81 M	2.61 G	**87.7**

4.3.4. Comparative Experimental Results and Analysis on the CIFAR-10 Dataset

To verify the generalization performance of MFANet, we conducted comparative experiments on the benchmark public dataset CIFAR-10 in image classification tasks. The training parameter settings of all networks were kept the same, and the experimental results are shown in Table 8. The experimental results show that the TOP-1 accuracy of MFANet on the CIFAR-10 dataset is slightly higher than the other neural network structures. It indicates that mainstream networks have been able to complete conventional image classification tasks well. On the other hand, MFANet uses multi-branch multi-scale convolution to widen the network, combining the advantages of the attention mechanism to possess a more powerful feature extraction ability, and the accuracy is slightly higher than other networks.

Table 8. Comparison experimental results of MFANet and various neural networks on the CIFAR-10 dataset.

Network	Parameters	FLOPs	Top-1 Accuracy (%)
ResNet50-CA [33]	25.45 M	4.14 G	91.2
ResNet50-ECA [17]	23.53 M	4.12 G	91.5
ResNet50-SE [16]	26.05 M	4.12 G	91.4
ResNet-50 [9]	23.53 M	4.12 G	91.2
ResNeXt-50 [30]	23.00 M	4.26 G	93.0
Res2Net [14]	23.67 M	4.29 G	93.1
DenseNet-161 [26]	26.49 M	7.82 G	92.2
EPSANet [28]	20.53 M	3.63 G	94.0
SKNet [29]	25.45 M	4.51 G	84.6
MobileNet_v3_small [24]	1.53 M	58. 80 M	92.2
MobileNet_v3_large [24]	4.21 M	226.44 M	92.6
ShuffleNet_v2 [27]	1.26 M	149.58 M	92.8
SqueezeNet [31]	0.73 M	2.65 G	82.3
Xception [25]	20.83 M	4.58 G	92.7
Ours	13.81 M	2.61 G	**94.4**

4.4. Ablation Experiments

4.4.1. Structural Ablation Experiments

The MFA block in the MFANet residual block contains two major functional modules, one is the multi-scale feature extraction module (multi-scale feature extraction); the other is the attention weighing module (attention). In this part of the experiments, we evaluate the importance of these two modules to the overall structure. Because our network is improved from ResNet-50 [9] as the basic architecture, we use the accuracy results of ResNet-50 [9] on the three datasets as a baseline, disassembled the MFA structure in the same environment, and trained it separately. Table 9 shows the results of this ablation experiment. On Collar6, DeepFashion6, and CIFAR-10 dataset, the accuracy of the model was 80.4, 87.7, and 94.4% using the full MFA module; when only the multi-scale feature extraction module was reserved, the accuracy of MFA on the three datasets was 79.6, 87.4, and 93.9%; and when only the MFA attention module was reserved, the accuracy was 68.8, 86.5, and 91.5%, respectively. Therefore, we prove that the two main functional modules that constitute MFA have a great impact on model performance. The combination of the two can bring richer and more effective feature information, thereby improving model performance.

Table 9. Results of the structural ablation experiments.

Dataset	Settings	Accuracy (%)
Collar6	Baseline(ResNet-50)	66.5
	+Multi-scale Feature Extraction	79.6
	+Attention	68.8
	+MFA	**80.4**
DeepFashion6	Baseline(ResNet-50)	85.8
	+Multi-scale Feature Extraction	87.4
	+Attention	86.5
	+MFA	**87.7**
CIFAR-10	Baseline(ResNet-50)	91.2
	+Multi-scale Feature Extraction	93.9
	+Attention	91.5
	+MFA	**94.4**

4.4.2. Hyperparameter Ablation Experiments

The hyperparameter K is involved in the channel attention module ECA, as shown in Equation (3), which has the kernel size of one-dimensional convolution, and its role is to determine the local cross-channel interaction range of attention. In this part of the experiments, we evaluate the impact of its value on the MFA module. To this end, we train the model by setting K to 3, 5, 7, and 9. The results are shown in Table 10. On the Collar6, DeepFashion6, and CIFAR-10 datasets the model performed best when K = 5, with accuracy of 80.4, 87.7, and 94.5%, respectively. When K = 3, the accuracy was 79.9, 87.0, and 94.5%, respectively. When K = 7, the accuracy weas 80.0, 87.4, and 94.3%, respectively. When K = 9, the accuracy was 79.7, 87.2, and 94.2%, respectively, and the model performed poorly. From this we conclude that for the MFANet, the ECA module achieves the best results when K = 5. When training the network, a small convolution kernel cannot capture enough feature information and an excessively large convolution kernel will bring noise and make the model performance worse.

Table 10. The value of hyperparameter K affects the accuracy of the model.

Dataset	Number of K	Accuracy (%)
Collar6	3	79.9
	5	**80.4**
	7	80.0
	9	79.7
DeepFashion6	3	87.0
	5	**87.7**
	7	87.4
	9	87.2
CIFAR-10	3	94.5
	5	**94.5**
	7	94.3
	9	94.2

5. Conclusions

In this paper, we design a modular MFA including multi-scale feature extraction and attention mechanisms.The MFA module can extract multi-scale spatial information as well as finer cross-channel attention information. Based on the MFA module, we constructed a novel network structure MFANet, which inherited the advantages of the MAF module and can effectively integrate multi-scale contextual features and image-level classification information. Through extensive comparative experiments and ablation implementations, we demonstrated that our method MFANet can achieve more advanced

performance in general image classification tasks compared with various mainstream neural networks, and can also recognize small-scale targets in images. The state-of-the-art performance is achieved on small-scale image classification tasks. However, there are still some shortcomings in our work. First, our model is currently only useful for the image classification tasks and does not explore its possibilities in other vision tasks. Second, our use of only a single channel attention mechanism can result in loss of spatial information. In the future, we will investigate the application of MAFNet in more computer vision tasks and further investigate how to effectively embed spatial attention in the MFA module to combine it with channel attention for the purpose of improving model performance.

Author Contributions: Methodology, X.Q. and C.Y.; Software, S.Y., D.C. and H.L.; Validation, D.C. and H.L.; Formal analysis, S.Y. and L.L.; Investigation, S.Y. and H.L.; Resources, C.Y. and L.L.; Data curation, S.Y., D.C. and L.L.; Writing—original draft, S.Y.; Writing—review & editing, X.Q.; Visualization, S.Y.; Supervision, X.Q. and C.Y.; Project administration, X.Q. and C.Y.; Funding acquisition, C.Y. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Project of Science and Technology of Guangxi (Grant no. AA22068057), National Natural Science Foundation of China (Grant nos. 61962006) and supported by the Open Research Fund of Guangxi Key Lab of Human–Machine Interaction and Intelligent Decision (GXHIID2207).

Data Availability Statement: The data are available from the corresponding author on reasonable request.

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

References

1. Yuan, W. Computer-Aided Data Analysis of Clothing Pattern Based on Popular Factors. In Proceedings of the 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 January 2022; pp. 1543–1546. [CrossRef]
2. Rajput, P.S.; Aneja, S. IndoFashion : Apparel Classification for Indian Ethnic Clothes. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021; pp. 3930–3934. [CrossRef]
3. De Souza Inácio, A.; Lopes, H.S. EPYNET: Efficient Pyramidal Network for Clothing Segmentation. *IEEE Access* **2020**, *8*, 187882–187892. [CrossRef]
4. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1096–1104. [CrossRef]
5. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
6. Chengcheng, H.; Jian, Y.; Xiao, Q. Research and Application of Fine-Grained Image Classification Based on Small Collar Dataset. *Front. Comput. Neurosci.* **2022**, *15*, *121*. [CrossRef] [PubMed]
7. Fan, J.; Bocus, M.J.; Hosking, B.; Wu, R.; Liu, Y.; Vityazev, S.; Fan, R. Multi-Scale Feature Fusion: Learning Better Semantic Segmentation For Road Pothole Detection. In Proceedings of the 2021 IEEE International Conference on Autonomous Systems (ICAS), Montreal, QC, Canada, 11–13 August 2021; pp. 1–5. [CrossRef]
8. Yang, F.; Li, X.; Shen, J. MSB-FCN: Multi-Scale Bidirectional FCN for Object Skeleton Extraction. *IEEE Trans. Image Process.* **2021**, *30*, 2301–2312. [CrossRef] [PubMed]
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
10. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.Z.; Guo, J. Fine-Grained Visual Classification via Progressive Multi-granularity Training of Jigsaw Patches. In *Proceedings of the Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 153–168.
11. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 6053–6062. [CrossRef]
12. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 6230–6239. [CrossRef]
13. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]

14. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef] [PubMed]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
16. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
17. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539. [CrossRef]
18. Yang, M.; Wang, H.; Hu, K.; Yin, G.; Wei, Z. IA-Net: An Inception–Attention-Module-Based Network for Classifying Underwater Images From Others. *IEEE J. Ocean. Eng.* **2022**, *47*, 704–717. [CrossRef]
19. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
20. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149. [CrossRef]
21. Xu, S.; He, Q.; Tao, S.; Chen, H.; Chai, Y.; Zheng, W. Pig Face Recognition Based on Trapezoid Normalized Pixel Difference Feature and Trimmed Mean Attention Mechanism. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–13. [CrossRef]
22. Wang, Q.; Wu, T.; Zheng, H.; Guo, G. Hierarchical Pyramid Diverse Attention Networks for Face Recognition. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 8323–8332. [CrossRef]
23. Chen, B.; Deng, W.; Hu, J. Mixed High-Order Attention Network for Person Re-Identification. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 371–381. [CrossRef]
24. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 1314–1324. [CrossRef]
25. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 1800–1807. [CrossRef]
26. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 2261–2269. [CrossRef]
27. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 122–138.
28. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. EPSANet: An efficient pyramid squeeze attention block on convolutional neural network. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022; pp. 1161–1177.
29. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 510–519. [CrossRef]
30. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 5987–5995. [CrossRef]
31. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.
33. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 13708–13717. [CrossRef]
34. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency Channel Attention Networks. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 763–772. [CrossRef]