Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

A Cervical Lesion Recognition Method Based on ShuffleNetV2-CA

Chunhui Liu[1], Jiahui Yang[2], Ying Liu[3*], Ying Zhang[2], Shuang Liu[2], Tetiana Chaikovska[4], Chan Liu[1]

[1] Affiliated Hospital of Hebei University, 071002 Baoding, China
[2] College of Quality and Technical Supervision, Hebei University, 071002 Baoding, China
[3] Bournemouth University, BH12 5BB Bournemouth, United Kingdom
[4] Department of Medical Imaging, Clinical Infectious Diseases Hospital N1, 02154 Kryvyi Rih, Ukraine

* Correspondence: Ying Liu (yliu@bournemouth.ac.uk)

Received: 03-02-2023 **Revised:** 03-12-2023 **Accepted:** 04-16-2023

Citation: C. H. Liu, J. H. Yang, Y. Liu, Y. Zhang, S. Liu, T. Chaikovska, C. Liu, "A cervical lesion recognition method based on ShuffleNetV2-CA," *Inf. Dyn. Appl.*, vol. 2, no. 2, pp. 77–89, 2023. https://doi.org/10.56578/ida020203.

Abstract: Cervical cancer is the second most common cancer among women globally. Colposcopy plays a vital role in assessing cervical intraepithelial neoplasia (CIN) and screening for cervical cancer. However, existing colposcopy methods mainly rely on physician experience, leading to misdiagnosis and limited medical resources. This study proposes a cervical lesion recognition method based on ShuffleNetV2-CA. A dataset of 6,996 cervical images was created from Hebei University Affiliated Hospital, including normal, cervical cancer, low-grade squamous intraepithelial lesions (LSIL, CIN 1), high-grade squamous intraepithelial lesions (HSIL, CIN 2/CIN 3), and cervical tumor data. Images were preprocessed using data augmentation, and the dataset was divided into training and validation sets at a 9:1 ratio during the training phase. This study introduces a coordinate attention mechanism (CA) to the original ShuffleNetV2 model, enabling the model to focus on larger areas during the image feature extraction process. Experimental results show that compared to other classic networks, the ShuffleNetV2-CA network achieves higher recognition accuracy with smaller model parameters and computation, making it suitable for resource-limited embedded devices such as mobile terminals and offering high clinical applicability.

Keywords: Cervical cancer screening; Colposcopy; Data augmentation; ShuffleNetV2; Attention mechanism

1 Introduction

Cervical cancer is the fourth leading cause of death among women worldwide [1]. Statistics from the World Health Organization (WHO) demonstrate that in 2020, there were approximately 600,000 new cases globally, accounting for 7.7% of all female cancer deaths. The cure rate for early cervical cancer is high, but the lack of signs and symptoms at this stage hinders early diagnosis. Thus, a mature cervical precancerous screening program can reduce the disease's incidence and prevent numerous cervical cancer-related deaths. In high-income countries, well-organized screening programs have significantly reduced the incidence of cervical cancer, which can be attributed to the effective prevention of cervical intraepithelial neoplasia (CIN) through initial screening, colposcopy, and treatment. Cervical cancer's precancerous lesions are cervical intraepithelial neoplasms (CINs), divided into low-grade squamous intraepithelial lesions (LSIL), such as CIN1, and high-grade squamous intraepithelial lesions (HSIL), such as CIN2 and CIN3 [2]. HSIL patients require surgical treatment, while LSIL patients need conservative observation. Presently, cervical cytology testing and human papillomavirus (HPV) testing are utilized for cervical cancer screening. If abnormalities are detected during screening tests, colposcopy using acetic acid and iodine solutions is performed to identify cervical lesions. Certain manifestations under colposcopy, such as dense acetowhite epithelium, coarse mosaicism, punctuation, and atypical vessels, are considered abnormal [3]. Since these symptoms often suggest HSIL or invasive cancer, colposcopy-guided biopsy is necessary. Colposcopy requires an experienced colposcopy physician. There are significant differences in the detection rates of lesions between different colposcopy physicians. Additionally, the lack of experienced medical personnel and insufficient funding for screening systems in developing countries contribute to a severe shortage of cervical cancer screening facilities. Therefore, digital colposcopy with high-definition imaging combined with deep learning for subjective diagnosis of images can provide opportunities for image-based automated diagnosis. However, the application of artificial intelligence currently faces

a series of problems, such as the scarcity of high-quality clinical data and the lack of model promotion in clinical applications. Consequently, a lightweight cervical cancer screening system based on deep learning can not only reduce the cost of early cervical cancer screening but also is suitable for deployment on limited hardware devices, aligning more with clinical application.

Deep learning is defined as a machine learning algorithm that attempts higher-level abstractions through a composition of multiple nonlinear transforms [4]. Artificial neural networks are used to implement deep learning, with the convolutional neural network (CNN) being the most representative example. CNN comprises several hidden layers, such as convolutional layers, pooling layers, and fully connected layers. Through numerous hidden layers, CNN can extract the features of the data. Some studies have employed CNN to classify colposcopy images or cervical radiographs. Buiu et al. [5] proposed an automated colposcopy image analysis framework for classifying cervical precancerous lesions and cancerous lesions. The framework is based on an ensemble of MobileNetV2 networks. The experimental results demonstrate that the method achieves an accuracy of 83.33% and 91.66% on four-class and binary classification tasks, respectively. Saini et al. [6] proposed a deep learning-based ColpoNet network using colposcopy images for cervical cancer classification. It was tested and validated on the dataset released by the National Cancer Institute of the United States and compared with other deep learning models AlexNet, VGG16, ResNet50, LeNet, and GoogleNet. The experimental analysis reveals that, compared to other state-of-the-art deep technologies, ColpoNet achieves an accuracy of 81.353% and exhibits the highest performance. Luo et al. [7] proposed a method using multiple CNN decision feature fusion to classify and diagnose cervical lesions. First, the k-means algorithm is utilized in the data preprocessing stage to aggregate the training data into specific classes. Then, DenseNet121 and ResNet50 are fine-tuned based on transfer learning, and the XGBoost algorithm is employed to integrate the decisions of different CNNs and optimize the final prediction. The experimental results indicate that the k-means data preprocessing method can improve the training effect of the neural network and the proposed multi-decision feature fusion strategy. Park et al. [8] compared the performance of machine learning and deep learning models, using the deep learning model ResNet-50 and machine learning models XGB, SVM, and RF. The final experimental results show that in identifying cervical cancer using cervical radiographs, the ResNet-50 deep learning algorithm can provide higher performance than current machine learning models. Liu et al. [9] proposed a new non-homogeneous bilinear pooling convolutional neural network model and combined it with an attention mechanism to further enhance the network's ability to extract specific features of images. The experimental results indicate thatthe proposed network model can significantly improve the prediction accuracy of the network while maintaining computational efficiency. In recent years, constructing deeper and larger neural networks has been the main trend in the development of major visual tasks [9–11], which requires billions of FLOPs for computation. The high cost limits the actual deployment of cervical cancer classification models. He et al. [11] proposed a lesion attention-aware convolutional neural network (CNN) model using ResNet-50 as the convolutional backbone and self-attention mechanism to locate lesion areas in WCE images. The experimental results demonstrate that their method accurately aggregates spatial features in the global context to locate lesion attention maps in WCE images. In this study, a lightweight deep learning-based cervical precancerous lesion classification method is proposed. The main contributions of this study are as follows:

A deep learning-based method for classifying cervical precancerous lesions is proposed, with the main contributions being the development and evaluation of a novel lightweight cervical cancer screening system that can be deployed on limited hardware devices, aligning more with clinical applications.

(1) This study investigates automatic analysis methods for colposcopy medical images based on deep learning to classify different levels of cervical precancerous lesions, cervical tumors and cervical cancer.

(2) An improved deep inverted residual network with additional coordinate attention based on ShuffleNetV2 is proposed. The deep inverted residual network automatically completes the self-learning of the data feature-to-expression relationship. By introducing the coordinate attention mechanism, a larger range of feature extraction is obtained. The improved network was trained on the self-made dataset, achieving an accuracy and precision of 82.48% and 82.53% respectively.

(3) Compared with traditional residual networks, the improved network can ensure automatic feature extraction of images while reducing computational complexity and improving the computing speed of the model.

2 Methodology

The overall process of this method primarily comprises image preprocessing, network architecture, and model evaluation. Specifically, the five class labels of the training set are initially processed using a single channel. Subsequently, the training samples are scaled proportionally. An improved network is then trained with the training images enhanced by various transformations. Finally, to evaluate the classification performance, the test samples are fed into the pre-trained model, and the performance of the network is analyzed and evaluated by comparing the evaluation metrics. Each step will be discussed in detail in the following sections.

2.1 Basic Network Architecture

This section systematically introduces the structural model of ShuffleNetV2 and expands the description of the channel random mixing operation and the ShuffleNetV2 unit contained in the network structure. Due to the low contrast between different levels of cervical lesions in colposcopy images, especially HSIL and LSIL lesions, the five classifications of colposcopy images present a challenging task. In order to address these challenges, a high-performance network is required to extract rich features from the images. The front-end CNN networks include MobileNet [12, 13], ShuffleNet [14, 15], VGGNet [16], GoogleNet [17], DenseNet [18], and ResNet [19]. Among them, although deep convolutional networks such as ResNet and DenseNet can significantly improve the accuracy of image classification, they increase the computational complexity. The increase in parameters makes the network deployment in clinical applications difficult. Therefore, designing a lightweight and efficient classification network is essential. In view of the excellent performance of the ShuffleNetV2 [14] model in biomedical image classification, a lightweight feature attention network based on ShuffleNetV2 is proposed.

Based on the ShuffleNetV2 unit, the overall architecture of ShuffleNetV2 is shown in Figure 1. The proposed network consists of a stack of ShuffleNetV2 units grouped into three levels. The first block in each stage applies stride = 2. Other hyperparameters remain unchanged for the first stage, while the number of output channels doubles for the next stage. The number of bottleneck channels is set to 1/4 of the output channels of each ShuffleNet unit. The ShuffleNet unit mainly adopts two technologies: depth separable convolution (Depthwise Separable Convolution) and channel random mixing (channel shuffle) to substantially reduce the computational overhead while retaining the precision of the model.

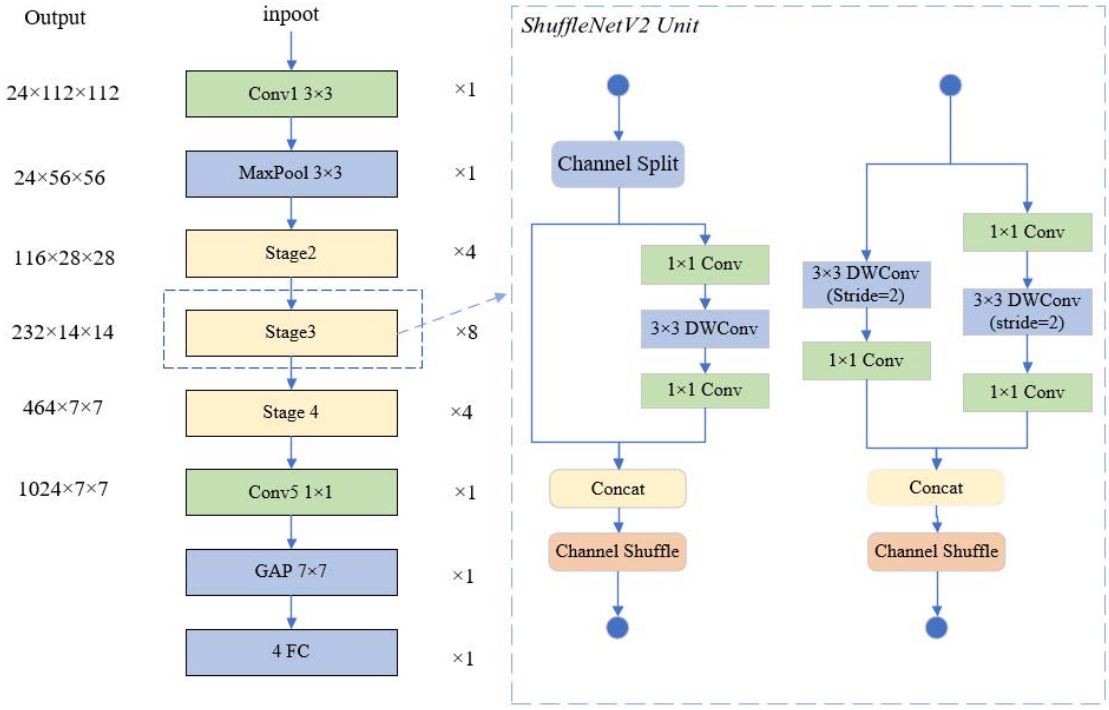

Figure 1. ShuffleNetV2 structure model

2.1.1 Channel random mixing

Group convolution also leads to the problem that different groups cannot share information. Therefore, in order to achieve circulation in different groups, ShuffleNet performs channel reordering operations on the output elements. Figure 2(a) shows the related situation of two stacked group convolutional layers. Obviously, the output of a group is only related to the input of that group. It prevents the transmission of information between different channel groups and weakens its representation. As shown in Figure 2(b), group convolution (group convolution) can obtain input information from different groups, so the input and output channels will be fully related. This process can be efficiently achieved by the channel shuffling operation, as shown in Figure 2(c): assuming a convolutional layer has g groups, its output is g×n channels; the output channel dimension is first reorganized into (g, n), then exchanged and flattened as the input of the next layer.

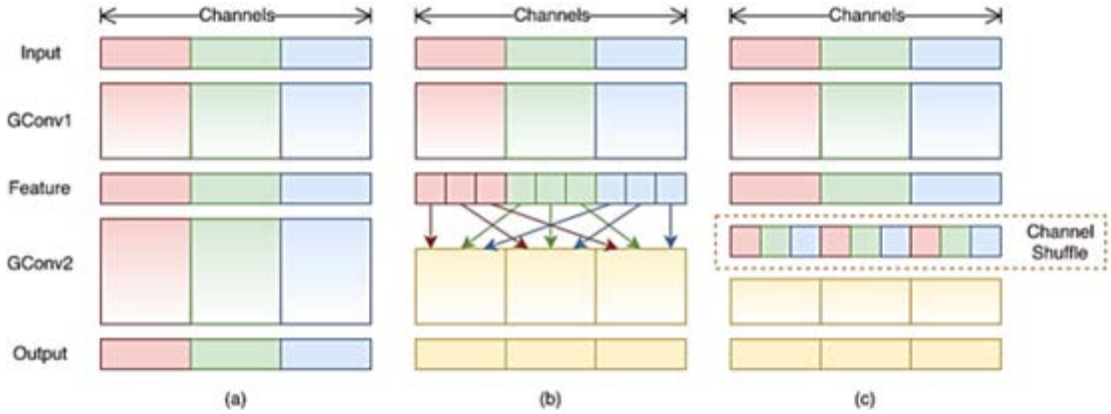

Figure 2. Channel random mixing (channel shuffling)

2.1.2 ShuffleNetV2 unit

Using the channel random mixing operation, a new ShuffleNetV2 specifically designed for small networks is proposed, starting from the bottleneck unit, as shown in the ShuffleNetV2 unit in Figure 1. It is a residual block (residual block). The two branches are concatenated by Concat to minimize arithmetic complexity (MAC) when the number of input and output feature channels of the convolutional layer is equal, at which point the model speed is fastest. When Stride=1, the left module is used. Since the residual edge has no convolution, the width and height remain unchanged, which is mainly used to deepen the network layers. When Stride=2, the right module is used. Since the residual edge has convolution, the width and height can change, which is mainly used to compress the width and height of the feature layer for downsampling.

In order to make the classification more accurate, attention mechanisms are added to the model. The attention mechanism is one of the ways to achieve network adaptive attention. When convolutional neural networks are used to process images, it is expected that convolutional neural networks will focus on the places that need attention. At this time, attention mechanisms are needed to adaptively pay attention to important attention objects. Four different attention mechanisms are added to the model to compare and analyze the effects of improving network performance.

2.2 Attention Mechanism

This section presents the structures of four attention mechanisms. The attention mechanism is a technique in artificial neural networks that emulates human cognitive attention. This mechanism can automatically allocate different weights based on the importance of various parts of the feature matrix, enabling the network to pay more attention to significant local features. Currently, attention mechanisms mainly comprise the (SE) attention mechanism, CBAM attention mechanism, ECA attention mechanism, and CA attention mechanism. The SE attention mechanism can learn adaptive channel weights, allowing the model to focus more on valuable channel information. However, the SE attention mechanism only takes into account the channel dimension's attention and cannot capture the spatial dimension's attention. It is suitable for scenarios with a higher number of channels but may not perform as well as other attention mechanisms in situations with fewer channels.

The CBAM attention mechanism combines convolution and attention mechanisms, enabling it to focus on images in both spatial and channel dimensions. Nevertheless, it demands more computing resources and has a higher computational complexity. The ECA attention mechanism considers the attention of both channel and spatial dimensions simultaneously. It offers high computational efficiency for feature maps with larger sizes. However, extra calculation is necessary, which results in considerable computational overhead for smaller feature maps. CA attention mechanism not only acquires information between channels but also takes into account direction-related position information, aiding the model in better locating and recognizing targets. Additionally, it is flexible and lightweight enough to be easily integrated into the core structure of mobile networks. Ultimately, the CA attention mechanism module is incorporated into the network to more effectively allocate resources and process crucial information in the feature map.

2.2.1 SE attention mechanism

The SE network [20] focuses on the relationship between channels and can automatically learn the importance of different features. The structure is depicted in Figure 3. Before entering the SE attention mechanism (left figure C), the importance of each channel of the feature map is the same. After passing through SENet (right colored figure C), different colors represent different weights, making the importance of each feature channel distinct. This enables the neural network to concentrate on channels with larger weights.

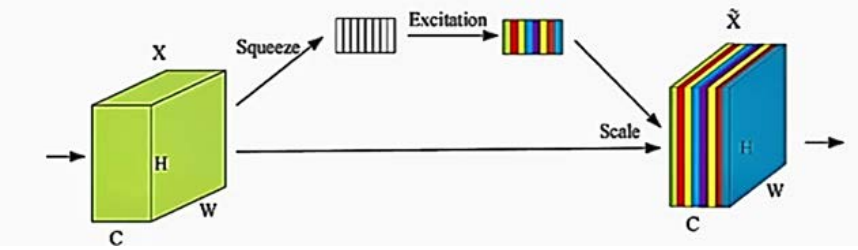

Figure 3. SE attention mechanism

CBAM [21] combines channel attention mechanism and spatial attention mechanism, which can achieve improved results compared to SENet, which only focuses on the channel attention mechanism. Its implementation is illustrated in the schematic diagram. CBAM processes the input feature layer separately with the channel attention mechanism and the spatial attention mechanism.

The specific implementation of the channel attention mechanism and spatial attention mechanism in the CBAM attention mechanism is shown in Figure 4. The upper half of the image is the channel attention mechanism. The implementation of the channel attention mechanism can be divided into two parts. Global average pooling and global maximum pooling are performed on the input single feature layer respectively. After processing the results of average pooling and maximum pooling with a shared fully connected layer, the two results are added and a sigmoid is applied, obtaining the weight value of each channel of the input feature layer (between 0-1). After obtaining this weight, this weight is multiplied by the original input feature layer.

The lower half of the image is the spatial attention mechanism. The maximum value and average value of each channel at each feature point of the input feature layer are taken. Then these two results are stacked, the number of channels is adjusted with a 1-channel convolution, and a sigmoid is applied, obtaining the weight value of each feature point of the input feature layer (between 0-1). After obtaining this weight, this weight is multiplied by the original input feature layer.

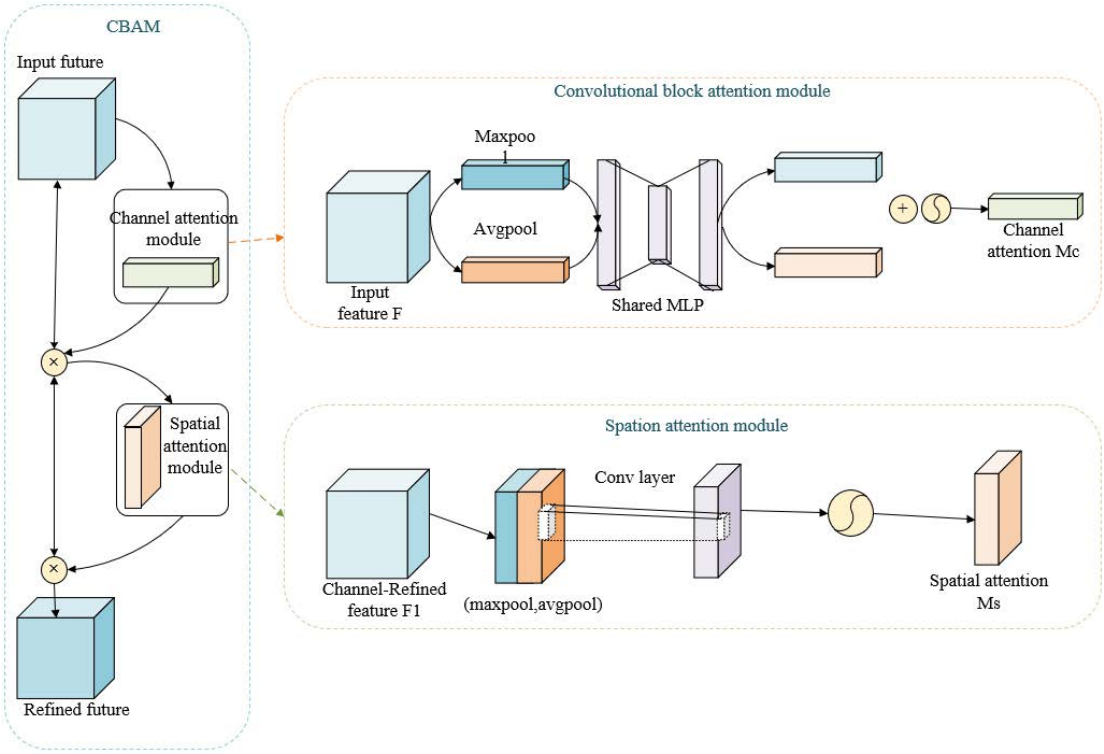

Figure 4. CBAM attention mechanism

2.2.2 ECA feature extraction module

ECA [22] is also an implementation of the channel attention mechanism, which can be considered as an improved version of SE. Its implementation is illustrated in Figure 5. The ECA module replaces two fully connected layers

with one-dimensional convolutions. The authors of ECA argue that capturing the dependence of all channels in the SE module is inefficient and unnecessary. Convolution has a strong ability to obtain cross-channel information.

Therefore, the ECA module removes the original fully connected layer in the SE module and directly learns on the globally averaged pooled features through a one-dimensional convolution. This module avoids dimension reduction and effectively captures cross-channel interactions with few parameters. The size of the convolution kernel is adaptively changed through a function, allowing layers with more channels to interact more across channels. The adaptive function is shown in Eq. (1), where γ=2 and b=1.

$$K = \left| \frac{\log_2(c)}{\gamma} + \frac{b}{\gamma} \right| \tag{1}$$

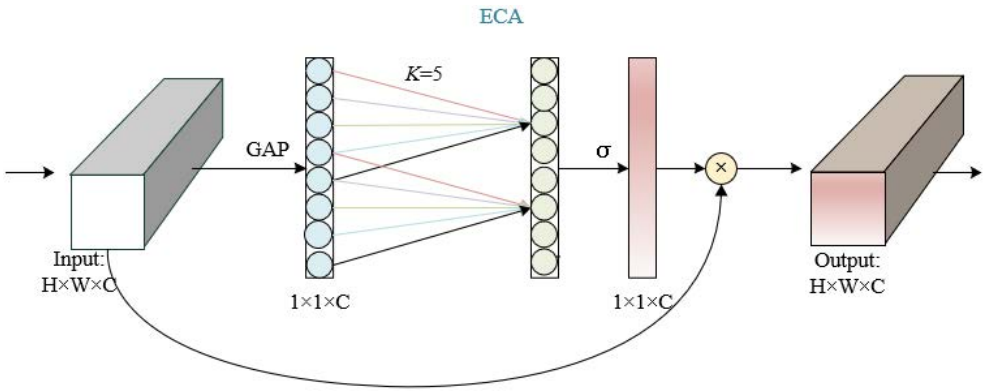

Figure 5. ECA feature extraction module

2.2.3 CA feature extraction module

CA [23] is an efficient coordinate attention mechanism for mobile devices, which can consider both channel relationship and long-distance position information. It can effectively improve the accuracy of the model with little additional computation. The CA module encodes channel relationships and long-range dependencies through precise position information in two steps: coordinate information embedding and coordinate attention generation. Its specific structure is shown in Figure 6.

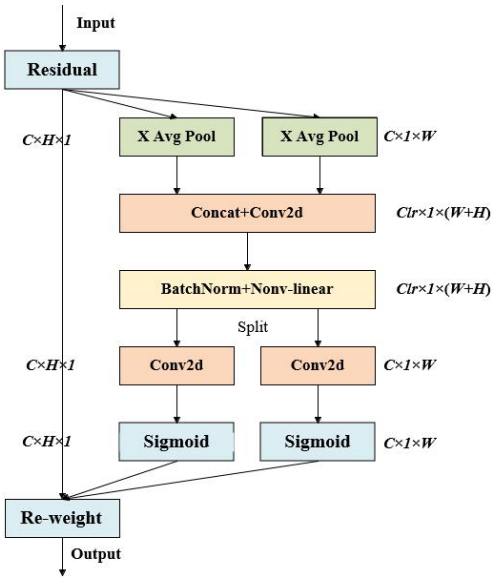

Figure 6. CA feature extraction module

Coordinate information embedding: The input feature map of C×H×W shape is averaged pooled channel by channel. Using pooling kernels of (H,1) and (1,W) respectively along the X and Y axes for each channel encoding, generating C×H×1 and C×1×W shaped feature maps. The output of the c-th channel at height h is shown in Eq. (2):

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le w} x_c(h, i) \tag{2}$$

The output of the c-th channel with width W is shown in Eq. (3):

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \le i \le w} x_c(w, i) \tag{3}$$

The above two transformations aggregate features along two different directions to obtain a pair of directional perception directional and position-sensitive feature maps. This is different from SE, which generates a single feature vector in the channel attention method. This can capture long-range dependencies along one spatial direction and retain precise position information along the other spatial direction. It helps the network accurately locate targets of interest.

The coordinate attention generation stage: After the transformation in the information embedding, this part performs a concatenate operation on the above transformations, and then uses a 1*1 convolution transformation function F_1,

$$\delta(F_1([z^h, z^W])) \tag{4}$$

where, [,] is the concat operation along the spatial dimension, δ is a nonlinear activation function, and f is an intermediate feature mapping that encodes spatial information in the horizontal and vertical directions. Using another two 1*1 convolution transformation functions F_1, f_h and f_w are transformed into tensors X with the same number of channels, to obtain:

$$g^h = \sigma(F_h(f^h)) \tag{5}$$

$$g^w = \sigma(F_w(f^w)) \tag{6}$$

Output Y:

$$y_c(i, j) = x_c(i, j) * g_c^h(i) * g_c^w(j) \tag{7}$$

After inserting SE, CBAM, ECA, and CA channel attention into the depth separable convolution of ShuffleNetV2 respectively, the effects of four feature attention mechanisms on network performance were compared and analyzed. The experimental structure shows that compared with the module integrated with SE, CBAM, and ECA attention mechanisms, the accuracy, precision, and floating point operations (FLOPs) of the network integrated with the CA module are 82.48%, 82.53%, and 0.157G, respectively. The accuracy is the highest while the FLOPs is the lowest. Finally, ShuffleNetV2-CA network is selected. The inverted residual network structure model integrated with the CA module is shown in Figure 7.

3 Data Source and Processing

High-quality datasets are crucial for cervical cancer identification tasks based on deep learning. To complete the five-category task of hysteroscopy, our experiment was conducted on a dataset of 1189 cases of vaginal speculum (Leisegang vaginal speculum 3ML) examinations from Hebei University Affiliated Hospital from July 2019 to February 2023. All hysteroscopy images were taken as part of routine clinical practice for patients. No exclusion criteria based on age or ethnicity were used. Each case contained 5 consecutive images of acetic acid white test with sequence labels, images taken with a green lens, and images taken after application of compound iodine solution. Vaginal speculum images 2 minutes after acetic acid application were selected, labeled by experienced gynecologists, and finally 6996 vaginal speculum images of different stages of precancerous lesions (normal, CIN 1 and CIN 2/3), cervical tumors and cervical cancer were generated using image enhancement techniques.

The steps to obtain preprocessed images of five categories are as follows: First, each original image was cleaned by manually cropping to retain information about the cervix. Second, the original images were classified into five categories according to the case report information, as shown in Figure 3, namely cervical cancer, low-grade

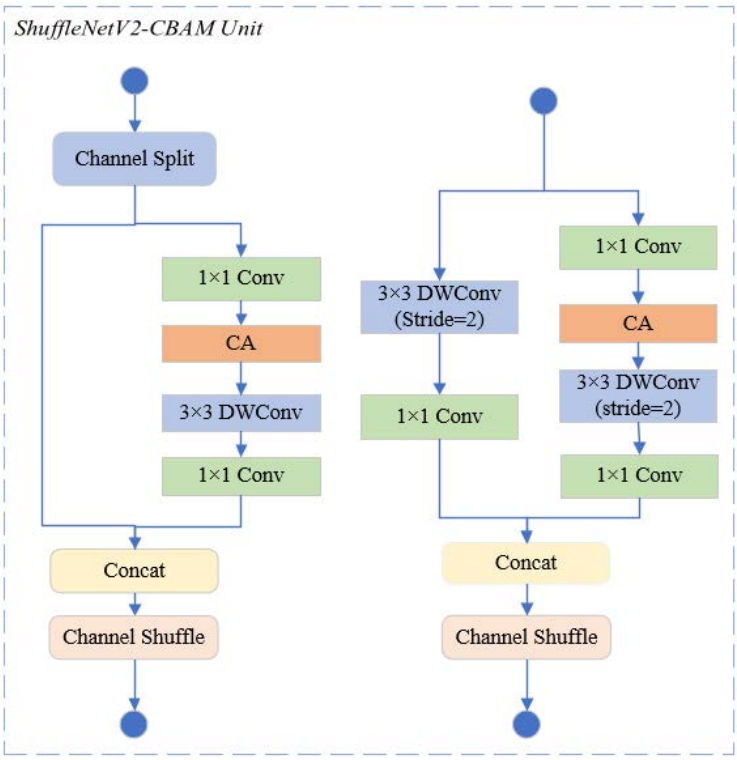

Figure 7. ShuffleNetV2-CA unit

squamous intraepithelial lesion (LSIL), high-grade squamous intraepithelial lesion (HSIL) and cervical tumors. Then, the classified data were checked and adjusted by experienced doctors. Afterwards, as is well known, convolutional neural networks require the same image size to work properly. Therefore, we uniformly adjusted the training set to a size of 224 × 224 by affine transformation, which can keep the original aspect ratio of the input image unchanged and avoid changes in the size of the traditional convolutional neural network training set. Finally, due to the uneven distribution of the datasets provided in each category and the small number of samples, we used random cropping, random horizontal and vertical flipping data augmentation techniques to create more available data and improve the overall safety of the trained model.

For the five-category cervical dataset, we selected vaginal speculum images 2 minutes after saline application. Before training the proposed network, the dataset was split 9:1. Specifically, to ensure consistent results from multiple runs of our network, these images were grouped separately according to the five types in the dataset so that cervical images of all categories could be sampled for training and testing. The distribution of cervical image data is shown in Table 1.

Table 1. Distribution of cervical dataset

Category	Data amount	Training set	Validation set
Normal	2352	2117	235
Low-grade squamous intraepithelial lesion (LSIL)	780	702	78
High-grade squamous intraepithelial lesion (HSIL)	2532	2279	253
Cervical cancer	408	367	41
Cervical tumor	924	832	92

4 Experiments

In this section, we performed extensive experiments to test and evaluate the feature representation ability of the proposed method. Specifically, the description of the dataset, evaluation indicators, implementation details and comparative experiments are the main contents discussed in the following sections.

4.1 Evaluation Indicators

To effectively evaluate the algorithm, training loss and model accuracy were used as metrics during the training phase. In the testing phase, this study introduces a confusion matrix as the basic evaluation criterion. The four parts of information in the confusion matrix include: true negatives (TN), indicating the number of negative samples predicted as negative; true positives (TP), indicating the number of positive samples predicted as positive; false negatives (FN), indicating the number of positive samples predicted as negative; and false positives (FP), indicating the number of negative samples predicted as positive.

For the multi-classification cervical dataset, accuracy, precision, recall, F1 score, and floating point operations (FLOPs) were used to evaluate the classification performance of the network. FLOPs was used to measure the complexity of the model. The lower the FLOPs, the easier it is to port the model to mobile devices. The formulas for classification accuracy, precision, recall, and F1 are Eqns.(8)-(11).

$$\text{Accurary}(\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \tag{8}$$

$$\text{Precision}(\%) \frac{TP}{TP + FP} \times 100 \tag{9}$$

$$\text{Recall}(\%) \frac{TP}{TP + FN} \times 100 \tag{10}$$

$$F1 - \text{score}(\%) = \frac{2 \times \text{ Recall } \times \text{ precision}}{\text{Recall } + \text{ Precision}} \times 100 \tag{11}$$

4.2 Implementation Details

To ensure iteration efficiency, improve model stability and generalization ability, the Stochastic Gradient Descent (SGD) algorithm was used to optimize network parameters. The Nesterov gradient descent method was used with a weight decay of 1e-4, a momentum of 0.9, and a batch size of 32. Each model was trained for 100 epochs, and the initial learning rate was set to 0.05. The CNN algorithm was implemented in the PyTorch coding framework. Intel® Xeon® Gold 6240 CPU@2.60 GHz and NVIDIA RTX 2080 ti GPU were used for model training and evaluation. All programs were run on Ubuntu 18.04.5 LTS.

4.3 The Influence of Attention Mechanism on Model Performance

To verify the improvement effect of the CA attention module compared with other attention modules on the model, comparative experiments were conducted under the same experimental conditions by replacing the CA module used in this study with the SE attention module and CBAM attention module. The network model parameters and identification results are shown in Table 2. The performance comparison of different attention modules is shown in Figure 8, and the floating point operations of different attention mechanisms are shown in Figure 9.

Table 2. Performance verification test results of different attention modules

Attention mechanism	Accuracy (%)	Precision (%)	Recall (%)	F1- Score (%)	FLOPs(G)
ShuffleNetV2–SE	81.38	81.76	80.74	81.16	0.1830
ShuffleNetV2-ECA	81.18	81.46	81.34	81.39	0.1642
ShuffleNetV2-CBAM	81.91	81.91	81.56	81.73	0.1597
ShuffleNetV2-CA	82.48	82.53	82.36	82.44	0.1570

As can be seen from Table 2, compared with other attention modules, the CA attention module added in this study achieved an Accuracy, Precision, Recall, and F1-Score of 82.48%, 82.53%, 82.36%, and 82.44%, respectively. It can also be intuitively seen from Figure 10 that ShuffleNetV2-CA performs best. Figure 11 shows that it has the lowest

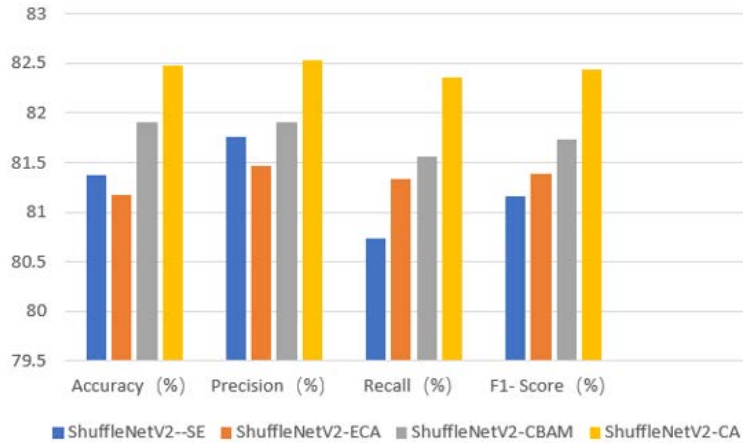

Figure 8. Performance comparison of different attention modules

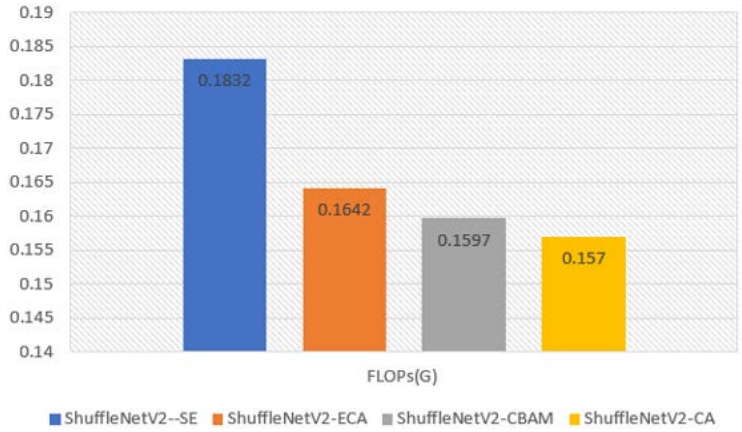

Figure 9. Floating point operations of different attention mechanisms

floating point operations, indicating that the complexity of the model is the smallest and most suitable for embedding in mobile devices. The CA attention mechanism can not only consider attention in the channel dimension and spatial dimension at the same time, but also learn adaptive channel weights to make the model pay more attention to useful channel information. This helps the model learn the interaction and dependence between the information of the characteristic channels of cervical lesions, further improving the recognition performance of the model. However, it cannot capture long-range dependencies. These features make it simple and lightweight enough for practical use, while being able to utilize the extracted position information.

4.4 Model Comparison

To evaluate the effectiveness of the designed model in classifying cervical cancer, VGG-16, ResNet 34, GoogleNet, DenseNet 121, and ShuffleNet were selected for comparison based on the competitiveness of the model. To compare the results more confidently, all models used the dataset in this study and were trained in the same training environment. As shown in Table 3, this study compared the accuracy, precision, recall, and F1 scores.

Table 3. Performance comparison of different network models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1- Score (%)	FLOPs(G)
VGG-16	50.72	45.63	45.67	45.07	15.47
ResNet34	83.95	84.88	81.28	82.81	6.428
GoogleNet	53.72	47.43	51.73	45.09	4.737
DenseNet121	86.39	87.00	83.95	85.17	2.869
MobileNet	54.30	65.12	44.60	43.45	0.1664
ShuffleNetV2	80.37	79.90	79.42	79.60	0.1568
ShuffleNetV2-CA	82.48	82.53	82.36	82.44	0.1570

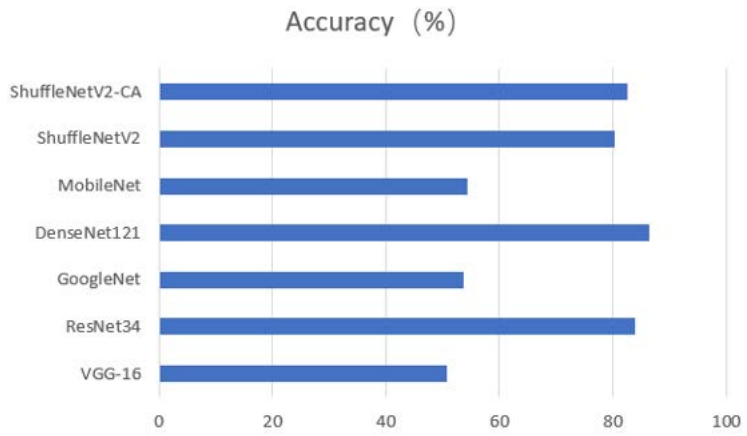

Figure 10. Accuracy comparison of different models

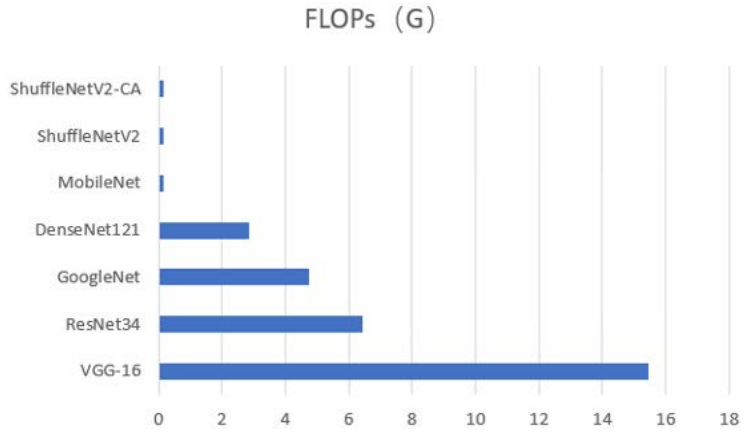

Figure 11. FLOPs comparison of different models

As can be seen from Table 3, different network architectures have different effects on the classification of the cervical dataset. The accuracy of the improved network ShuffleNetV2-CA reached 82.48%, slightly lower than ResNet34's 83.95% and DenseNet121's 86.39%, but the FLOPs of ShuffleNetV2-CA were much smaller than other networks. This is due to the use of grouped convolution in this network to reduce the number of parameters and channel reordering operations to enhance the interaction and fusion between different channels. Therefore, the network selected in this study achieved a good balance between recognition accuracy and model size, which can meet the requirements of mobile devices and other computing power limited devices for cervical lesion identification. Figures 10 and 11 also visually show that the improved network FLOPs are greatly reduced compared with traditional classification networks, and the classification accuracy is relatively high. Compared with the unimproved ShuffleNet, FLOPs increased slightly, but network performance improved significantly, and Accuracy increased by 2.11%. Therefore, the network selected in this study achieved a good balance between recognition accuracy and model size, which can meet the requirements of mobile devices and other computing power limited devices for cervical lesion identification.

5 Conclusion

For the identification of cervical lesions, this study took 5 different types of vaginal speculum images as the research object. A dataset of 6996 cervical lesion identification images was constructed from images collected from the hospital. On the basis of the ShuffleNetV2 model structure, this study embedded the CA attention module to adaptively extract channel information important for target identification and obtained the following conclusions: Compared with the modules embedded with SE, CBAM, and ECA attention mechanisms, the improved network in this study achieved an Accuracy, Precision, Recall, and F1-Score of 82.48%, 82.53%, 82.36%, and 82.44%, respectively, with the highest network performance and the lowest FLOPs of 0.157G. Compared with other classic VGG-16, ResNet 34, GoogleNet, DenseNet 121, and ShuffleNetV2 networks, ShuffleNetV2-CA obtained better classification effects while maintaining high resolution in feature extraction, smaller model parameters, and balanced the relationship between recognition accuracy and floating point operations. This is conducive to deploying the

convolutional neural network model on mobile terminals and other embedded resource-constrained devices to meet real-time needs and has high clinical application value. In addition, this study has limitations in the use of data diversity in datasets and does not make full use of different types of data in each case. At the same time, hardware deployment has not been implemented. Therefore, in the future, while continuing to expand high-quality datasets, multimodal data can be used to improve the accuracy of the model, and mobile hardware deployment can be achieved.

Funding

This study was supported by Baoding Science and Technology Planning Project (Grant No.: 2141ZF306 and 2141ZF135) and Youth Foundation of Affiliated Hospital of Hebei University (Grant No.: 2022QC54).

Data Availability

The data supporting our research results are included within the article or supplementary material.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] M. Brisson and M. Drolet, "Global elimination of cervical cancer as a public health problem," *Lancet Oncol.*, vol. 20, no. 3, pp. 319–321, 2019. https://doi.org/10.1016/S1470-2045(19)30072-5

[2] G. Peng, H. Dong, T. Liang, L. Li, and J. Liu, "Diagnosis of cervical precancerous lesions based on multimodal feature changes," *Comput. Biol. Med.*, vol. 130, p. 104209, 2021. https://doi.org/10.1016/j.compbiomed.2021.104209

[3] S. Kim, H. Lee, S. Lee, J. Y. Song, J. K. Lee, and N. W. Lee, "Role of artificial intelligence interpretation of colposcopic images in cervical cancer screening," *Healthcare*, vol. 10, no. 3, p. 468, 2022. https://doi.org/10.3390/healthcare10030468

[4] J. D. Waye, D. K. Rex, and C. B. Williams, "Colonoscopy: Principles and practice," 2008.

[5] C. Buiu, V. R. Dănăilă, and C. N. Răduţă, "MobileNetV2 ensemble for cervical precancerous lesions classification," *Processes*, vol. 8, no. 5, p. 595, 2020. https://doi.org/10.3390/pr8050595

[6] S. K. Saini, V. Bansal, R. Kaur, and M. Juneja, "ColpoNet for automated cervical cancer screening using colposcopy images," *Mach. Vision. Appl.*, vol. 31, p. 15, 2020. https://doi.org/10.1007/s00138-020-01063-8

[7] Y. M. Luo, T. Zhang, P. Li, P. Z. Liu, P. Sun, B. Dong, and G. Ruan, "MDFI: Multi-cnn decision feature integration for diagnosis of cervical precancerous lesions," *IEEE Access*, vol. 8, pp. 29 616–29 626, 2020. https://doi.org/10.1109/ACCESS.2020.2972610

[8] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, "Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images," *Sci. Rep.*, vol. 11, no. 1, p. 16143, 2021. https://doi.org/10.1038/s41598-021-95748-3

[9] P. Liu, X. Yang, B. Jin, and Q. Zhou, "Diabetic retinal grading using attention-based bilinear convolutional neural network and complement cross entropy," *Entropy*, vol. 23, no. 7, p. 816, 2021. https://doi.org/10.3390/e23070816

[10] P. Muruganantham and S. M. Balakrishnan, "Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization," *J. Med. Biol. Eng.*, vol. 42, no. 2, pp. 157–168, 2022. https://doi.org/10.1007/s40846-022-00686-8

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *arXiv preprint arXiv:1704.04861*, 2017.

[13] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, and R. Pang, "Searching for MobileNetV3," *In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South)*, pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

[14] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," *In Proceedings of the European conference on computer vision (ECCV), Munich, Germany*, pp. 116–131, 2018. https://doi.org/10.1007/978-3-030-01264-9_8

[15] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," *In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA*, pp. 6848–6856, 2018. https://doi.org/10.1109/CVPR.2018.00716

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA*, pp. 1–9, 2015. https://doi.org/10.1109/CVPR.2015.7298594

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, pp. 4700–4708, 2017. https://doi.org/10.1109/CVPR.2017.243

[19] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," *In Proceedings of the European conference on computer vision (ECCV), Munich, Germany*, pp. 19–34, 2018. https://doi.org/10.1007/978-3-030-01246-5_2

[20] J. In Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA*, pp. 7132–7141, 2018. https://doi.org/10.1109/CVPR.2018.00745

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *In Proceedings of the European conference on computer vision (ECCV), Munich, Germany*, pp. 3–19, 2018. https://doi.org/10.1007/978-3-030-01234-2_1

[22] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, pp. 11 534–11 542, 2020. https://doi.org/10.1109/CVPR42600.2020.01155

[23] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA*, pp. 13 713–13 722, 2021. https://doi.org/10.1109/CVPR46437.2021.01350