agriculture

MDPI

Article

A Classification Model for Fine-Grained Silkworm Cocoon Images Based on Bilinear Pooling and Adaptive Feature Fusion

Mochen Liu [1], Xin Hou [1], Mingrui Shang [1], Eunice Oluwabunmi Owoola [1], Guizheng Zhang [2], Wei Wei [2], Zhanhua Song [1,3] and Yinfa Yan [1,4,*]

[1] College of Mechanical and Electrical Engineering, Shandong Agriculture University, Tai'an 271018, China; liu_mochen@sdau.edu.cn (M.L.); houxin0228@163.com (X.H.); smr13105488275@126.com (M.S.); bunmso@gmail.com (E.O.O.); songzh6688@163.com (Z.S.)
[2] Sericulture Technology Extension Station of Guangxi Zhuang Autonomous Region, Nanning 530000, China; zhangdoudou1999@163.com (G.Z.); gxcanyeweiwei@126.com (W.W.)
[3] Shandong Engineering Research Center of Agricultural Equipment Intelligentization, Tai'an 271018, China
[4] Shandong Key Laboratory of Intelligent Production Technology and Equipment for Facility Horticulture, Tai'an 271018, China
* Correspondence: yanyinfa@sdau.edu.cn

Abstract: The quality of silkworm cocoons affects the quality and cost of silk processing. It is necessary to sort silkworm cocoons prior to silk production. Cocoon images consist of fine-grained images with large intra-class differences and small inter-class differences. The subtle intra-class features pose a serious challenge in accurately locating the effective areas and classifying silkworm cocoons. To improve the perception of intra-class features and the classification accuracy, this paper proposes a bilinear pooling classification model (B-Res41-ASE) based on adaptive multi-scale feature fusion and enhancement. B-Res41-ASE consists of three parts: a feature extraction module, a feature fusion module, and a feature enhancement module. Firstly, the backbone network, ResNet41, is constructed based on the bilinear pooling algorithm to extract complete cocoon features. Secondly, the adaptive spatial feature fusion module (ASFF) is introduced to fuse different semantic information to solve the problem of fine-grained information loss in the process of feature extraction. Finally, the squeeze and excitation module (SE) is used to suppress redundant information, enhance the weight of distinguishable regions, and reduce classification bias. Compared with the widely used classification network, the proposed model achieves the highest classification performance in the test set, with *accuracy* of 97.0% and an *F1-score* of 97.5%. The *accuracy* of B-Res41-ASE is 3.1% and 2.6% higher than that of the classification networks AlexNet and GoogLeNet, respectively, while the *F1-score* is 2.5% and 2.2% higher, respectively. Additionally, the *accuracy* of B-Res41-ASE is 1.9% and 7.7% higher than that of the Bilinear CNN and HBP, respectively, while the *F1-score* is 1.6% and 5.7% higher. The experimental results show that the proposed classification model without complex labelling outperforms other cocoon classification algorithms in terms of classification accuracy and robustness, providing a theoretical basis for the intelligent sorting of silkworm cocoons.

Keywords: silkworm cocoon; bilinear pooling; fine-grained image classification; adaptive spatial feature fusion; feature enhancement

Citation: Liu, M.; Hou, X.; Shang, M.; Owoola, E.O.; Zhang, G.; Wei, W.; Song, Z.; Yan, Y. A Classification Model for Fine-Grained Silkworm Cocoon Images Based on Bilinear Pooling and Adaptive Feature Fusion. *Agriculture* **2024**, *14*, 2363. https://doi.org/10.3390/agriculture14122363

Academic Editor: Roberto Alves Braga Júnior

Received: 11 November 2024
Revised: 17 December 2024
Accepted: 21 December 2024
Published: 22 December 2024

1. Introduction

Cocoons are divided into reelable cocoons and waste cocoons based on whether they can be reeled for silk production. The quality of silkworm cocoons is one of the decisive factors for the quality of silk. It is necessary to sort silkworm cocoons before reeling. Traditional cocoon sorting mainly relies on manual subjective judgment, making it a labor-intensive task. In addition, cocoon classification is a delicate and repetitive task, where the skill and focus of workers can significantly impact the accuracy of cocoon sorting. In the context of workforce reduction, companies are gradually increasing their investment in

labor costs, resulting in decreased profits, which in turn requires the sericulture industry to transition towards mechanization and automation. Therefore, it is necessary to use modern technology to solve the problem of cocoon classification through intelligence and automation.

With the development of machine vision, digital image processing technology has been gradually applied to the recognition and classification of silkworm cocoons. Chen et al. [1] analyzed the stain area on the surfaces of silkworm cocoons by image processing and the numerical calculation of the color images of cocoons, which provided a certain theoretical basis for the sorting of silkworm cocoons. Liu et al. [2] used the FCM segmentation method and the color H component ratio threshold to detect the yellow spotted cocoon in the mountage, and the accuracy rate was 81.2%. Zhang et al. [3] used the K-means algorithm based on the YOLO v4 target detection model to perform cluster analysis on a dataset of four types of inferior cocoons to preset the candidate anchor parameters and improve the model accuracy. Jiang et al. [4] used a two-parameter multi-level threshold to detect double cocoons based on the ratio of the long and short axes of silkworm cocoons and area parameters, and the classification accuracy for double cocoons and reelable cocoons was 98.6%. Guo et al. [5] extracted the regions of a silkworm cocoon image and then extracted the S-channel in the HSV color model of the cocoon. They identified yellow spotted cocoons by calculating the area proportion and average saturation of the yellow spot area in the S-channel. The above methods using traditional image processing can only identify cocoons with significant differences, and they struggle to distinguish those with smaller variations in color, shape, and other subtle characteristics.

In recent years, artificial neural networks have been applied to the field of image classification due to their advantages of capturing spatial local and deep features from images, and deep learning technology provides a new solution for cocoon image classification [6,7]. Sun et al. [8] proposed an improved YOLOv3 algorithm to achieve the classification of group silkworm cocoons, with average accuracy of 85.52%. Vasta et al. [9] used multiple cameras to capture images of silkworm cocoons and designed a multi-step approach and machine learning algorithms to identify the shape, size, and external stains of silkworm cocoons, respectively. Compared with traditional machine learning, the deep learning method does not require a large amount of feature engineering, so it has stronger portability. However, silkworm cocoon images consist of fine-grained images with small inter-class differences and large intra-class differences, requiring a deep learning model to ensure strong positioning accuracy for detailed features. Both Liu and Wu used feature fusion to improve the characterization ability of the model for waste cocoon features, which improved the classification accuracy of waste cocoons occurring at low frequencies [10], and they achieved the classification of cocoons in the mountage [11]. Existing work usually focuses on model improvements for specific types of waste cocoons. Although the detection accuracy has been improved, the classification ability of the network is still unsatisfactory when the types of cocoons exceed the targeted optimization of the network.

Fine-grained image classification [12], also known as sub-category image classification, is a more detailed type of division on the basis of distinguishing basic categories. It is more difficult to classify fine-grained images than to perform ordinary classification tasks due to the small differences between subcategories. In the task of the classification of silkworm cocoons, the main characteristics of various waste cocoons consist of different colors or differences in shape compared to reelable cocoons, holes in the cocoons, etc. These features are often very subtle, increasing the classification difficulty. Therefore, with respect to the classification of silkworm cocoons, a fine-grained image classification algorithm could enable the model to pay more attention to the distinguishable regions and reduce the influence of irrelevant information on the classification of cocoon images.

To enhance the perception of intra-class features and improve the classification accuracy, a B-Res41-ASE model suitable for cocoon image classification is proposed in this paper. Using ResNet41 as the basic framework, bilinear pooling, ASFF, and SE attention are sequentially introduced in the proposed network to improve the feature extraction,

aggregation, and enhancement, respectively. The main contributions of this paper include the following:

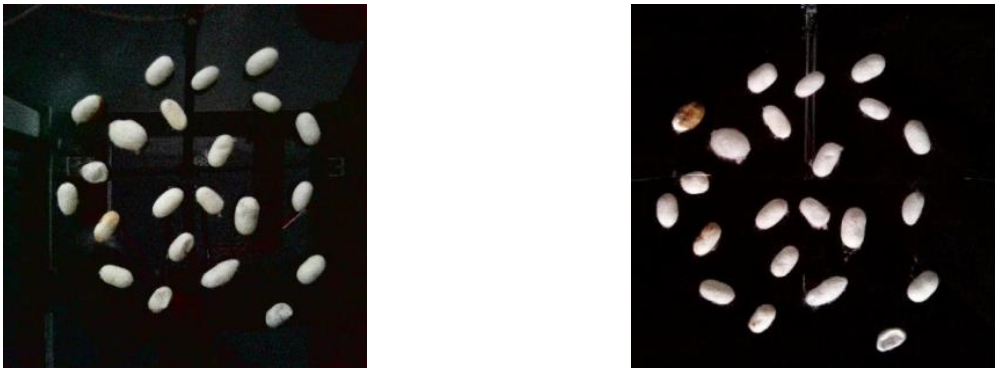
(1) An end-to-end trainable fine-grained image classification network is constructed with the bilinear pooling algorithm;
(2) The ASFF module is introduced at the end of the feature extraction network to fuse the high-level information, intermediate information, and shallow features in the feature extraction network;
(3) The feature enhancement module is integrated into the end of the fusion module, enhancing the discriminability of distinguishable regions of waste cocoon.

2. Materials and Methods

2.1. Materials

2.1.1. Dataset

The images of the silkworm cocoons were taken at the Biomass Energy Laboratory of Shandong Agricultural University. The cocoon variety was "Jingsong × Haoyue", the most widely bred silkworm in the northern region of China. The image acquisition equipment was the industrial camera WSD-P8002-V1.0.(1/3.06 inch CMOS-IMX258, Shenzhen Weishida Technology Co., Ltd., Shenzhen, China) All images were collected on a black background in an indoor environment, with an average of 35 silkworm cocoons randomly placed on a 50 cm × 50 cm transparent board. A light strip was used as a constant-brightness light source, and the acquisition height was set at 40 cm from the top and bottom of the transparent board, with a resolution of 3840 px × 2880 px. The sample images are shown in Figure 1. The collected images were segmented using area thresholding and edge detection methods, resulting in a total of 3500 images of individual cocoons, including 1500 images of reelable cocoons and 2000 images of waste cocoons.



(**a**) (**b**)

Figure 1. Images of silkworm cocoons captured by top and bottom camera. (**a**) Cocoon image captured by camera (top). (**b**) Cocoon image captured by camera (bottom).

Taking into account the actual cocoon sorting rules and the national standard (GB/T9111-2015) [13], we classified the silkworm cocoon images into two main types: reelable cocoons and waste cocoons. Additionally, the waste cocoons were further divided into 8 types based on the characteristics of defects: cocoons pressed by the cocooning frame, double cocoons, yellow spotted cocoons, malformed cocoons, cocoons polluted by oil, crushed cocoons, stained cocoons, and decayed cocoons, as shown in Figure 2.

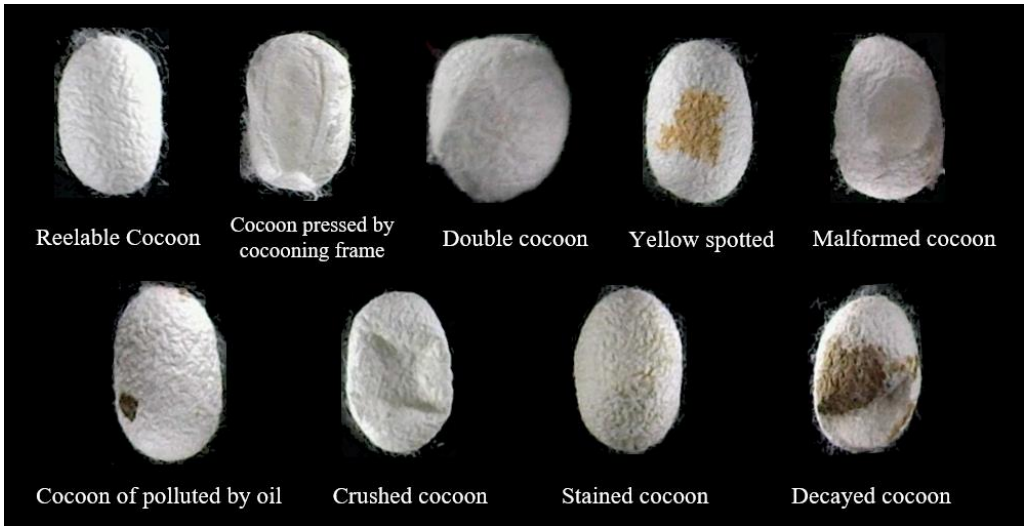


Figure 2. Images of reelable cocoons and different types of waste cocoons.

2.1.2. Data Augmentation

Influenced by the breeding environment, the probabilities of different types of waste cocoons are different, which will cause an imbalance in the samples in the set. In order to sufficiently learn the feature information of various cocoons and improve the generalization of the model, data augmentation techniques, including blurring, mirroring, brightness enhancement, brightness attenuation, and 180° rotation, were used to expand the cocoon images in the training set. The training set was augmented to a total of 4000 (including 2000 reelable cocoons and 2000 waste cocoons). The test set consisted of 1000 images (600 reelable cocoons and 400 waste cocoons) and was not augmented, maintaining an 8:2 ratio between the training set and the test set, as detailed in Table 1.

Table 1. Silkworm cocoon dataset.

Type of Cocoon		Training Set	Test Set
Reelable Cocoon	Reelable Cocoon	2000	600
Waste cocoon	Cocoon Pressed by Cocooning Frame	271	38
	Double Cocoon	312	51
	Yellow Spotted Cocoon	380	100
	Malformed Cocoon	159	21
	Cocoon Polluted by Oil	300	27
	Crushed Cocoon	256	69
	Stained Cocoon	222	71
	Decayed cocoon	100	23
	total	2000	400

2.2. Methods

2.2.1. Model Overview

In order to improve the accuracy of silkworm cocoon image classification, we propose a bilinear pooling model, and the overall framework is shown in Figure 3. It is mainly composed of four parts: a feature extraction module, feature fusion module, feature enhancement module and bilinear pooling algorithm. The proposed model utilizes the shallow, middle, and deep features of cocoon images and fuses different semantic information through the feature fusion module. After reducing redundant information using the feature enhancement module, the Hadamard product is used to improve the representation ability of the model, and finally a more accurate silkworm cocoon image classification is realized.

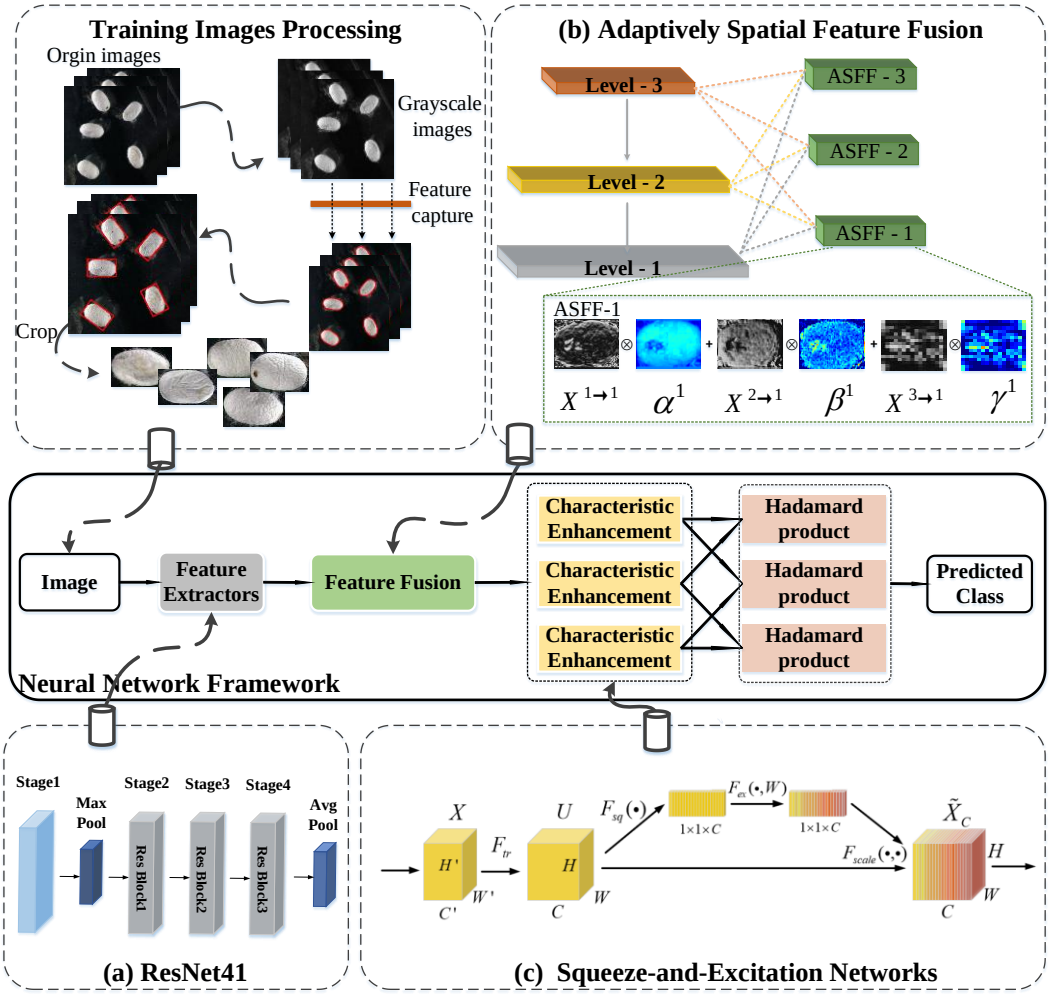


Figure 3. Cocoon image classification model architecture.

2.2.2. Bilinear Pooling Algorithm

The bilinear model proposed by Lin et al. [14] is the most representative weakly supervised network model of fine-grained image classification methods, achieving classification accuracy of 84.1% on the bird dataset (CUB200-2011). However, the complex structure of the model requires high computing power. We use the Hadamard product [15] to construct a bilinear pooling algorithm to learn multimodal joint representations by multiplication. Finally, a linear map with a bias term is used to project the multimodal joint representation to the output vector of a given output dimension. The low-rank bilinear pooling achieved with this method can effectively reduce the computational complexity.

To improve the ability to capture fine-grained features, the 3×3 small convolutional kernel is used in the shallow layer of the feature extraction module. The residual structure [16] is employed to construct the middle and deep modules of the ResNet41 model, avoiding gradient vanishing caused by network deepening. As shown in Figure 4, the cocoon image $\boldsymbol{I}$ is filtered by ResNet41 to produce feature maps $\boldsymbol{X} \in \mathbf{R}^{H \times W \times C}$, where $\boldsymbol{H}$ represents the height, $\boldsymbol{W}$ represents the width, and $\boldsymbol{C}$ represents the number of channels. The $\boldsymbol{C}$-dimensional descriptor of a spatial position on $\boldsymbol{X}$ is represented as $\boldsymbol{x} = [x_1, x_2 \cdots x_C]^{\mathrm{T}}$. The complete bilinear model is defined as $y_i = x^{\mathrm{T}} W_i x$, where $W_i \in \mathbf{R}^{C \times C}$ is a projection matrix, and y_i is a bilinear model output. By learning the spatial information $\boldsymbol{W} = [W_1, W_2, \ldots W_O] \in \mathbf{R}^{C \times C \times O}$, the trained weight information provides an output $\boldsymbol{Y}$ in O-dimensional space. According to the projection matrix method, the projection matrix $\boldsymbol{W_i}$ can be decomposed into two one-rank vectors $\boldsymbol{U}_i \in \mathbf{R}^C$ and $\boldsymbol{V}_i \in \mathbf{R}^C$, leading to the output

being expressed as $y_i = x^T W_i x = x^T U_i \circ V_i^T x$. The computation of the model's output features is as follows:

$$Y_{B-Res41} = P^T(U^T x \circ V^T x) \quad (1)$$

where $U \in R^{C \times d}$ and $V \in R^{C \times d}$ are projection matrices, $P \in R^{C \times O}$ is the classification matrix, x is the local descriptor of the convolutional layer, and o is the Hadamard product.

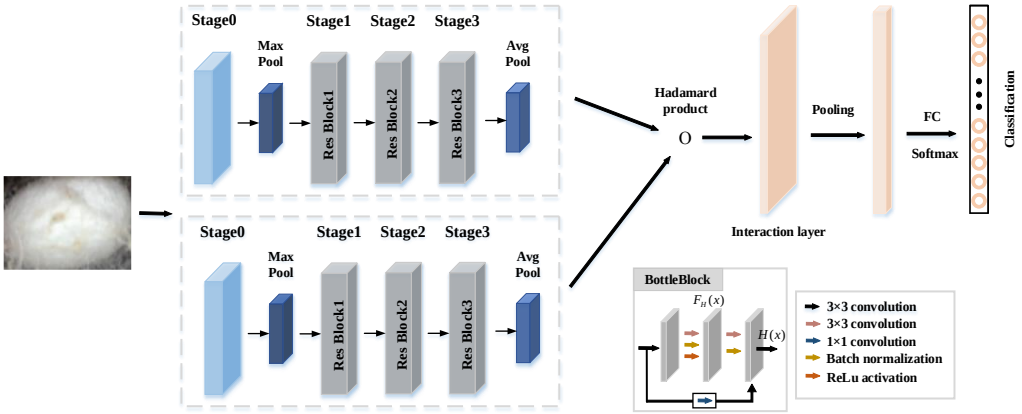


Figure 4. Bilinear pooling-based image classification model for silkworm cocoon images.

2.2.3. Feature Fusion Module

The convolutional neural network extracts the target features through layer-by-layer learning. However, in deep layers, the strong semantic information loses the fine-grained details present in shallow layers, making it challenging to achieve a balance between semantic and visual features. In the silkworm cocoon image classification model, adaptive spatial feature fusion (ASFF) [17] is introduced to improve the classification ability and robustness, which weights and fuses high-level semantic information with low-level semantic information to enhance the inter-layer feature interaction.

ASFF fuses the cocoon image features of Stage 1, Stage 2, and Stage 3 in ResNet41. The three cocoon feature maps of different scales are uniformly scaled into three corresponding scale feature maps by the interpolation method and pooling method. The network adaptively learns the spatial weight coefficients α_{ij}^l, β_{ij}^l, γ_{ij}^l of the feature maps of different scales and adds them together to obtain the new fused feature y_{ij}^l. The formula is as follows:

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1-l} + \beta_{ij}^l \cdot x_{ij}^{2-l} + \gamma_{ij}^l \cdot x_{ij}^{3-l} \quad (2)$$

Here α is computed by using a softmax function with λ_α as the control parameter, which can be learned by the standard back-propagation(BP). Similarly, β and γ are defined by using another parameters λ_β and λ_γ, respectively. We force $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$ and $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0, 1]$ [18], and we define

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \quad (3)$$

where, y_{ij}^l represents the vector (i, j) of the output feature y^l; x_{ij}^{n-l} indicates that the feature vector adjusts the feature map from level n to level l at position (i, j); $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l$ denotes the feature weights of three different scale feature layers.

The cocoon features of each level are aggregated with different weights at each scale, increasing the richness of the semantic features through this method. The output feature $\{y^1, y^2, y^3\}$ is obtained, as shown in Figure 5. The features of different scales are pooled and

compressed into compact features through the Hadamard product to obtain the B-Res41-A model. The output result of this step is

$$Y_{\mathrm{B-Res41-A}} = P^{\mathrm{T}} concat[(y^1)^{\mathrm{T}} x \circ (y^2)^{\mathrm{T}} x, (y^2)^{\mathrm{T}} x \circ (y^3)^{\mathrm{T}} x, (y^1)^{\mathrm{T}} x \circ (y^3)^{\mathrm{T}} x] \quad (4)$$

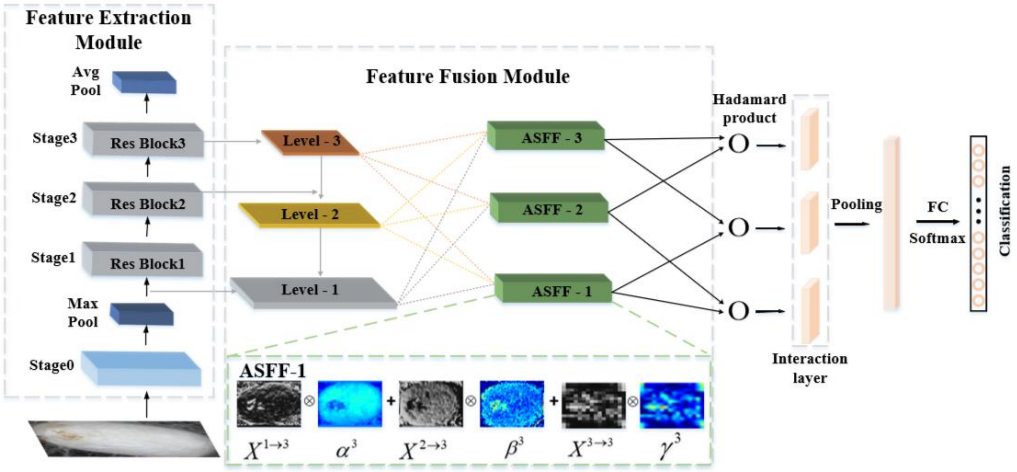


Figure 5. Silkworm cocoon image classification model based on bilinear pooling with feature fusion.

2.2.4. Feature Enhancement Module

Due to the highly similar characteristics of silkworm cocoon images and the limitation of the local receptive field of the network, it is challenging to distinguish the discriminable regions of the cocoons. Therefore, the squeeze and excitation (SE) [19] module is introduced into the network to guide the network to focus on the features that are helpful for classification. The $W \times H \times C$ feature vectors of the silkworm cocoon images are compressed into a $1 \times 1 \times C$ channel descriptor by global average pooling (GAP) for the squeeze operation. Then, the feature data obtained by the extruded part are passed through the two fully connected layers to output the feature map, resulting in a feature vector. Finally, the weights obtained from the excitation module are used to weight the feature map, resulting in a new silkworm cocoon image $\tilde{x}_C$. The feature map is calculated as follows:

$$\tilde{x}_C = F_{\mathrm{scale}}(u_C, s_C) = s_C u_C \quad (5)$$

In the formula: $F_{\mathrm{scale}}(u_C, s_C)$ multiplies each value in the matrix u_C with a scalar s_C; s_C is the weight of the C channel learned from the squeeze and excitation operation, and u_C is the output that maps the input. The final eigenvector is $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_C]$. The squeeze and excitation module helps the network to enhance the pixel region that is rich in semantic information by controlling the size of the convolutional layer to enhance the important features of the image and attenuate the unimportant features, as shown in Figure 6. The bilinear pooling model achieves better inter-layer feature interaction by fusing and concatenating three squeeze and excitation nets, and it obtains the final output through Equation (6).

$$Y_{\mathrm{B-Res41-ASE}} = P^{\mathrm{T}} concat(\tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \quad (6)$$

where P is the classification matrix; *concat* is the merging of the feature vectors; and $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3$ represent the feature vectors with enhanced significant differences.

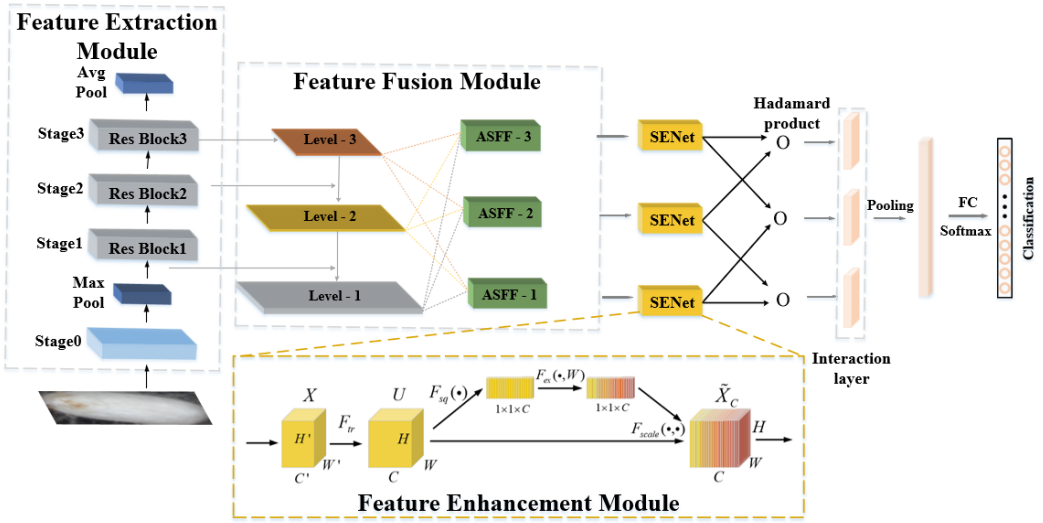


Figure 6. Bilinear pooling classification model for silkworm cocoon images based on feature fusion and enhancement.

2.2.5. Model Training Parameters

The network is trained based on the Windows system and PyTorch framework. The workstation is mainly configured with an Intel(R) Xeon(R) Gold 5218R CPU and NVIDIA GeForce RTX 3090 GPU and installed with CUDA 11.0, CUDNN 8.0.1, and Python 3.8.

When the training curve converges, the model is considered to be in a good fitting state, and the training can be terminated. Based on this, the number of iterations is set to 800, and the remaining training parameters are set as follows: the model batch size is set to 8, the learning rate is 0.0001, the momentum parameter is 0.9, the optimizer is Adam, and the loss function is the cross-entropy loss function.

2.2.6. Evaluation Metrics

In this study, the *accuracy*, *precision*, *recall*, and *F1-score* are used to evaluate the model's performance, with the formulas shown in (7)–(10). The higher the metrics, the better the classification performance of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

where TP is the number of correctly classified cocoons; TN is the number of correctly classified waste cocoons; FP is the number of waste cocoons incorrectly classified as reelable cocoons; and FN is the number of reelable cocoons incorrectly classified as waste cocoons.

3. Results

3.1. Comparison with Different Fusion Modules

In order to verify the performance of the ASFF fusion module, we conducted ablation experiments under the same experimental parameter conditions, selecting three feature fusion methods: ASFF, feature pyramid networks (FPN) [20], and the CONCAT function. These methods were named B-Res41-A, B-Res41-F, and B-Res41-C, respectively.

As shown in Figure 7a, the accuracy of each model increases rapidly during the early stages of training. The accuracy of B-Res41 starts to increase from around 50%, while B-Res41-A, B-Res41-F, and B-Res41-C show a rapid increase starting from about 67%. After approximately 200 iterations, the training accuracy begins to plateau, and the models reach a stable state after 600 iterations. Notably, the training accuracy curve of B-Res41-A converges the fastest and it achieves the highest accuracy after stabilization.

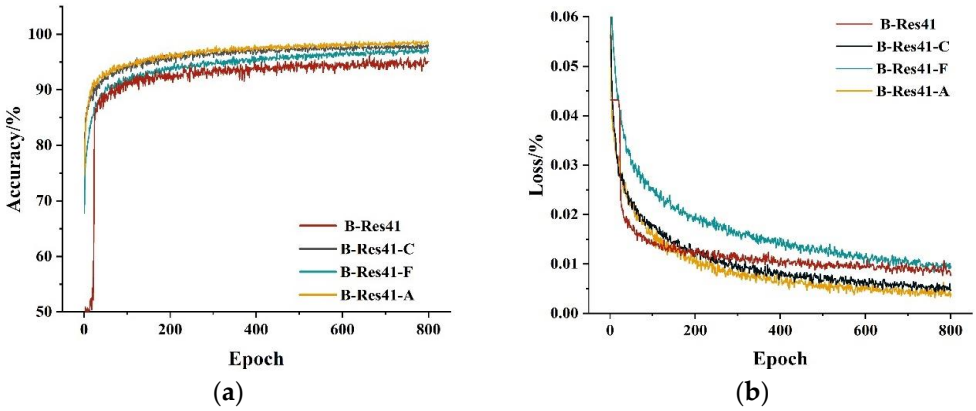


Figure 7. The training accuracy curves of different fusion algorithms. (**a**) The training accuracy curves of different fusion algorithms. (**b**) The training loss curves of different fusion algorithms.

From the training loss curve in Figure 7b, all four models experience a rapid decrease during the initial iterations and tend to stabilize after 700 iterations. Among them, B-Res41 reaches stability first, but its final loss value is only slightly lower than that of B-Res41-F. Ultimately, B-Res41-A has the lowest loss value. Both the training accuracy and loss curves indicate that B-Res41-A demonstrates excellent performance.

The data in Table 2 show that, for the B-Res41 model without feature fusion, the *accuracy*, *precision*, *recall*, and *F*1-*score* are 94.4%, 95.3%, 95.3%, and 95.3%, respectively. This indicates that the model has good classification capabilities on the silkworm cocoon dataset, providing a solid foundation for performance improvement. The network B-Res41-A integrated with ASFF could achieve better performance than the original B-Res41. However, due to the addition of the feature fusion module, the time required for model inference will inevitably increase. The inference time of B-Res41-A, B-Res41-F, and B-Res41-C increased by 40 ms, 27 ms, and 11 ms, respectively. An increase in the inference time of milliseconds will not have a substantial impact on the cocoon classification task, but the accuracy of cocoon classification directly affects the quality of silk reeling, so we are more concerned about the classification accuracy.

Table 2. Performance comparison of different fusion models.

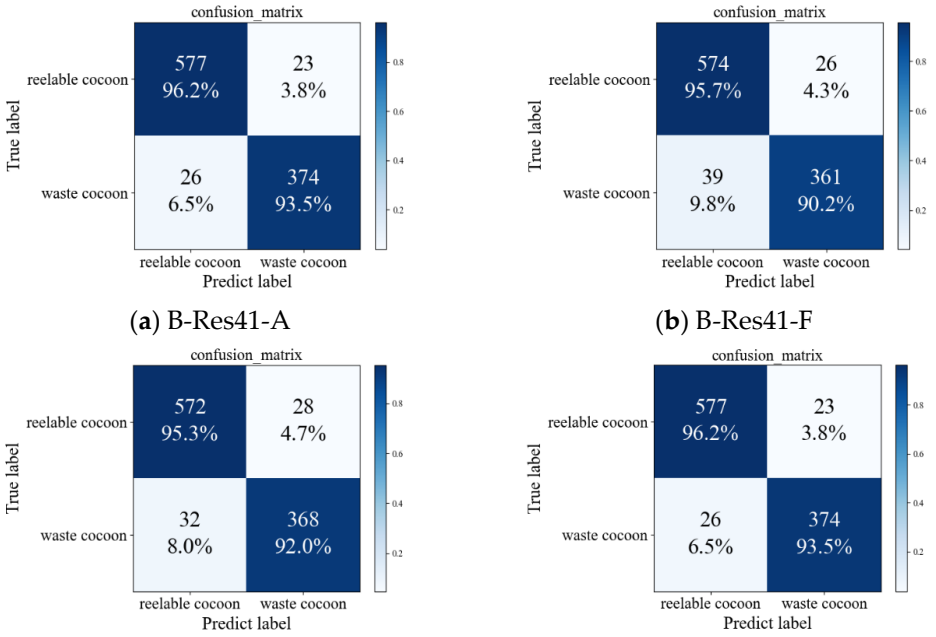
Algorithm	*Accuracy*/%	*Precision*/%	*Recall*/%	*F*1-*Score*/%	*Params*/*M*	*FPS*/*ms*
B-Res41	94.4	95.3	95.3	95.3	49.1	129
B-Res41-A	95.1	96.2	95.7	95.9	67.3	169
B-Res41-F	93.5	95.7	93.6	94.6	65.9	156
B-Res41-C	94.0	95.3	94.7	95.0	53.2	140

When ASFF is integrated to the model, the *accuracy* is increased by 0.7%, *precision* by 0.9%, *recall* by 0.4%, and the *F*1-*score* by 0.6%. The reason is that ASFF adaptively learns feature mappings at different scales through the same scaling operation and generates new feature maps using weighting parameters, thereby retaining more detailed information.

When using FPN, we can theoretically enhance the model's utilization of features at different scales; however, the overall performance of B-Res41-F compared to B-Res41-C did not show significant improvement and even declined. This may be due to the fact that the

additional parameters introduced by FPN became redundant information in the model and did not provide corresponding performance enhancements. Similarly, the B-Res41-C model attempts to enhance the performance through the concatenation of parameters; however, the results show that this method did not enable the network to effectively learn the features of the images, resulting in only a modest overall performance improvement.

Figure 8 shows the confusion matrices for the classification of reelable and waste cocoons by the four models. The B-Res41-A misclassified the smallest number of cocoons, totaling 49, compared to B-Res41-F 65 B-Res41-C 60. This indicates that the B-Res41-A model significantly outperforms the other two models in terms of the overall classification accuracy.

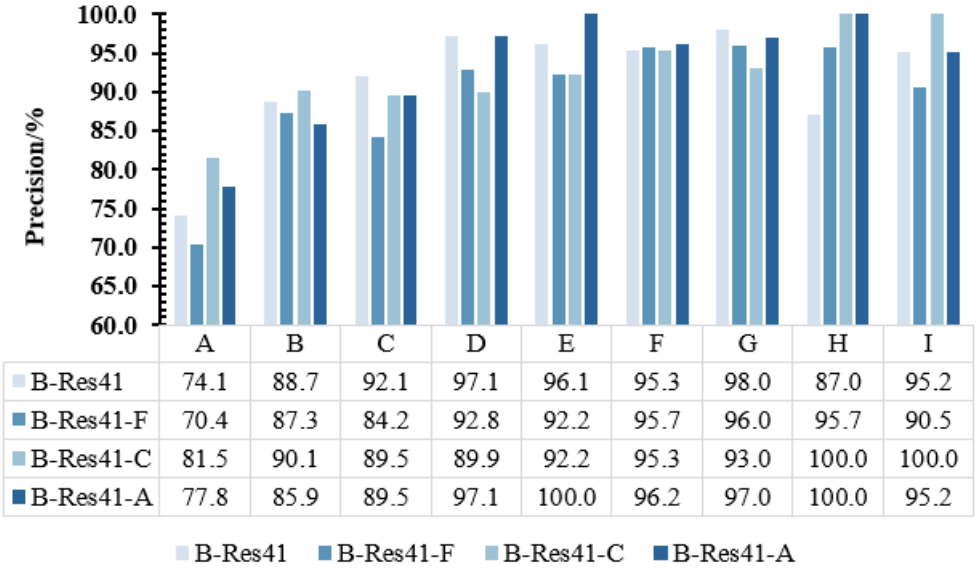


(**a**) B-Res41-A (**b**) B-Res41-F

(**c**) B-Res41-C (**d**) B-Res41

Figure 8. Confusion matrix for different fusion algorithms.

Figure 9 shows the fine-grained classification *precision* of the silkworm cocoons. With respect to the classification *precision* of reelable cocoons, which is the most important metric for silkworm cocoon sorting, the *precision* of B-Res41-A is 96.2%, which is 0.5% higher than that of B-Res41-F and 0.9% higher than that of B-Res41-C. Yellow spotted cocoons account for the highest percentage of waste cocoons. The *precision* of B-Res41-A reaches 97.0%, which is 1.0% higher than B-Res41-F and 4.0% higher than B-Res41-C. In addition, B-Res41-A has classification *precision* of 100.0% for the double cocoons and decayed cocoons. This shows that the model can effectively capture the feature information of these two types. However, the *precision* of B-Res41-A for stained cocoons is lower than that of the other two methods. The reason may be that ASFF integrates new feature maps through weight parameters, which interferes with the model's learning of stained cocoon feature information. Overall, B-Res41-A is more suitable for the fusion and extraction of silkworm cocoon features.



	A	B	C	D	E	F	G	H	I
B-Res41	74.1	88.7	92.1	97.1	96.1	95.3	98.0	87.0	95.2
B-Res41-F	70.4	87.3	84.2	92.8	92.2	95.7	96.0	95.7	90.5
B-Res41-C	81.5	90.1	89.5	89.9	92.2	95.3	93.0	100.0	100.0
B-Res41-A	77.8	85.9	89.5	97.1	100.0	96.2	97.0	100.0	95.2

Figure 9. The fine-grained classification *precision* of the silkworm cocoon for different fusion algorithms. A. Cocoon polluted by oil. B. Stained cocoon. C. Cocoon pressed by cocooning frame. D. Crushed cocoon. E. Double cocoon. F. Reelable cocoon. G. Yellow spotted cocoon. H. Decayed cocoon. I. Malformed cocoon.

3.2. Comparison with Different Fusion Enhancements

While the model's classification accuracy is improved by integrating feature maps from different layers, it also increases the redundant information. We introduce a feature enhancement mechanism to reduce the interference of redundant information in the feature maps, allowing the model to focus on distinguishable feature regions. The comparative experiments were conducted using a convolutional block attention module (CBAM) [21], efficient channel attention (ECA) [22], and SE, and the models were named B-Res41-AB, B-Res41-AE, and B-Res41-ASE, respectively. Figure 10 shows the training curves for the accuracy and loss of each model, and all models converged after 800 epochs. Among them, B-Res41-ASE had the highest accuracy and the lowest loss.

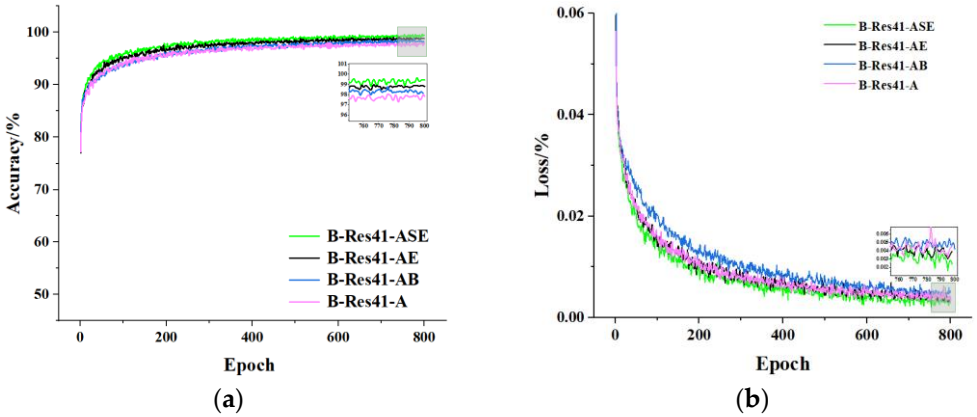


Figure 10. The training curves of different feature fusion and enhancement methods. (**a**) The training accuracy curves. (**b**) The training loss curves.

Table 3 shows the test results of the methods using different feature enhancement mechanisms. Adding feature enhancement modules also increases the inference time. Our model is deployed on a cocoon sorting machine that we designed. The time taken for the robot arm to pick a single cocoon is 1 s. Therefore, the increased inference time of the model is within an acceptable range. B-Res41-AB employs the CBAM, which focuses on both channel information and the spatial information of features. Compared with

B-Res41-A, there is a slight improvement in various metrics, but the improvement effect is not significant. Using the one-dimensional convolutional operation of ECA to allocate weights, B-Res41-AE improves the *precision* of B-Res41-A by 0.5%, but the *accuracy*, *recall*, and *F1-score* all decrease slightly. This may result from the insufficient ability of ECA to handle global context dependencies. In contrast, B-Res41-ASE, which uses the SE method, automatically learns feature weights through a fully connected network based on loss, increasing the weights of effective feature channels. Its evaluation metrics are all improved compared to the models using other feature enhancement mechanisms.

Table 3. Comparison of test results for different feature enhancements.

Algorithm	*Accuracy*/%	*Precision*/%	*Recall*/%	*F1-Score*/%	*Params*/*M*	*FPS*/*ms*
B-Res41-A	95.1	96.2	95.7	95.9	67.3	169
B-Res41-AB	95.2	96.5	95.5	96.0	82.6	174
B-Res41-AE	94.9	96.7	94.9	95.8	75.8	161
B-Res41-ASE	97.0	97.7	97.3	97.5	77.5	186

Figure 11 shows the confusion matrices of four models for the classification of reelable and waste cocoons, while Figure 12 presents the *precision* for the fine-grained classification of silkworm cocoons. From the two figures, it can be seen that all three improved methods enhance the *precision* for reelable cocoons, but not all of them are effective for the classification of waste cocoons. Compared to the original model, B-Res41-A, B-Res41-AB improved the classification *precision* for five types of cocoons: reelable cocoons, cocoons polluted by oil, stained cocoons, cocoons pressed by the cocooning frame, and malformed cocoons; however, the *precision* for decayed cocoons decreased by 14.3%. B-Res41-AE improved the *precision* for cocoons polluted by oil and stained cocoons, while, the for other types of waste cocoon, it decreased. As for B-Res41-ASE, the *precision* for double cocoons, decayed cocoons, and malformed cocoons was 100.0%, while that for yellow spotted cocoons was 98.0%. The *precision* for other types of cocoons has also been improved, with a significant reduction in the misclassification rate of cocoons. Therefore, B-Res41-ASE exhibits the best overall classification performance, particularly in classifying waste cocoons.

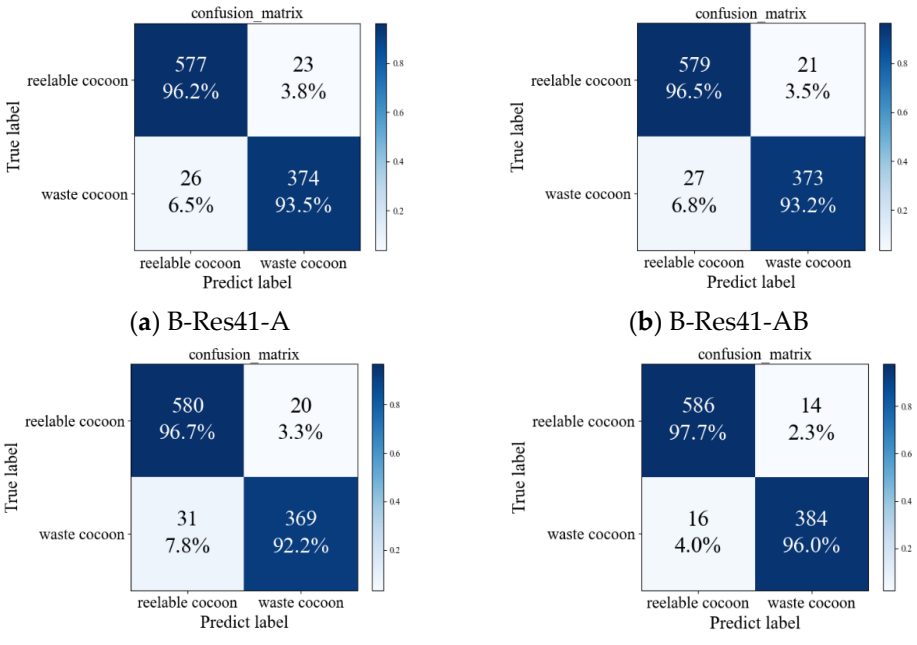
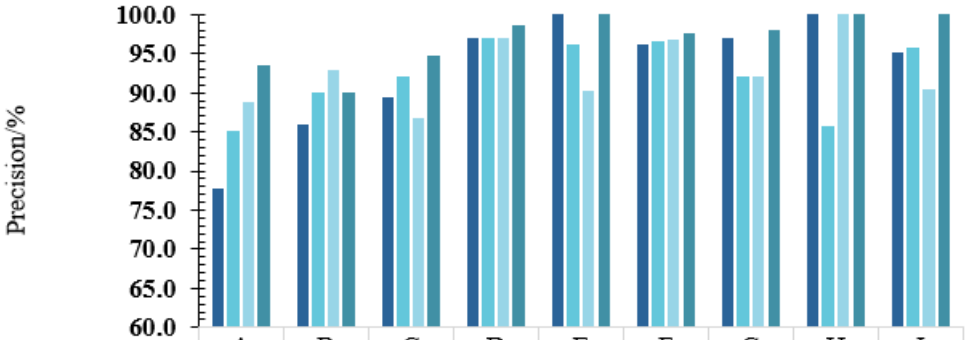


Figure 11. The confusion matrix for different feature enhancements.



	A	B	C	D	E	F	G	H	I
B-Res41-A	77.8	85.9	89.5	97.1	100.0	96.2	97.0	100.0	95.2
B-Res41-AB	85.2	90.1	92.1	97.1	96.1	96.5	92.0	85.7	95.7
B-Res41-AE	88.8	93.0	86.8	97.1	90.2	96.7	92.0	100.0	90.5
B-Res41-ASE	93.6	90.1	94.7	98.6	100.0	97.7	98.0	100.0	100.0

Figure 12. The fine-grained classification precision of the silkworm cocoon for different feature enhancements. A. Cocoon polluted by oil. B. Stained cocoon. C. Cocoon pressed by cocooning frame. D. Crushed cocoon. E. Double cocoon. F. Reelable cocoon. G. Yellow spotted cocoon. H. Decayed cocoon. I. Malformed cocoon.

3.3. Ablation Experiments

To verify the classification effectiveness of the combined models using different methods, B-Res41 is used as the basic feature extraction network, and the ASFF module and SE module are introduced for ablation experiments.

As shown in Table 4, the ASFF module increases the *accuracy* by 0.7% and the *F*1-*score* by 0.6%. The SE module reduces the *accuracy* by 1.0% but increases the *F*1-*score* by 0.7%. When the two methods are combined, the *accuracy* of the model B-Res41-ASE is increased by 2.6% and the *F*1-*score* is increased by 2.2%. The results show that the combined effect of the two improvement methods is more effective in enhancing the classification performance than a single algorithm.

Table 4. Ablation experiments on different method combination models.

Algorithm	ASFF	SE	*Accuracy*/%	*F*1-*Score*/%	*Params*/*M*	*FPS*/*ms*
B-Res41	×	×	94.4	95.3	49.1	129
B-Res41-A	√	×	95.1	95.9	67.3	169
B-Res41-SE	×	√	93.4	96.0	53.2	157
B-Res41-ASE	√	√	97.0	97.5	77.5	186

4. Discussion

4.1. Visual Analysis Based on Grad-CAM

To investigate the relationship between the feature maps and their class weights in various models, we produced visual explanations for the base model B-Res41 and two improved models using Grad-CAM [23]. As shown in Figure 13, the darker the red color, the higher the attention paid to these features.

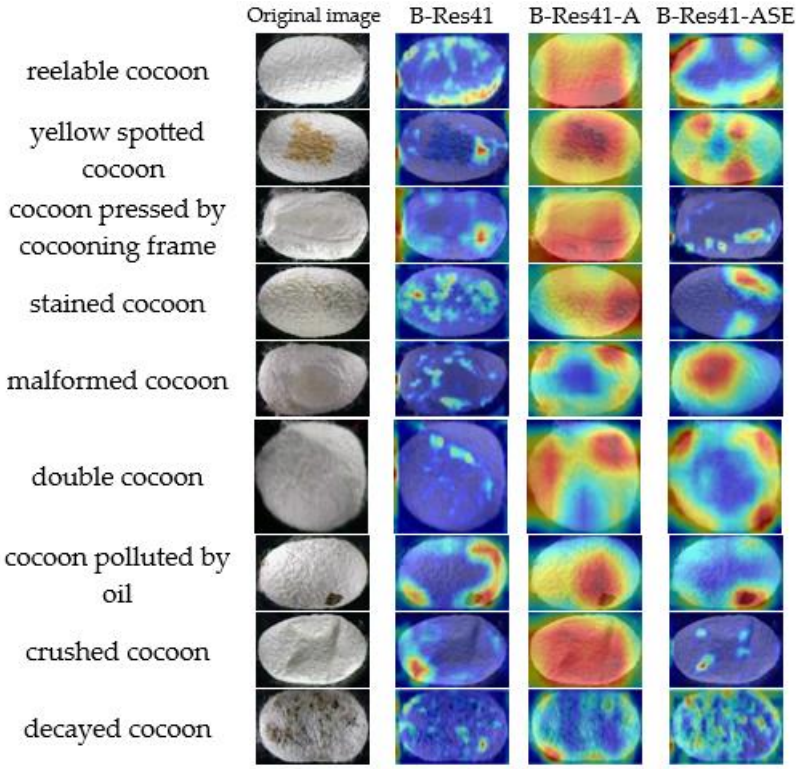


Figure 13. Comparison of different models with Grad-CAM visualization.

The results show that B-Res41 has a relatively weak positioning ability for the discriminable regions of the silkworm cocoon, e.g., it does not pay attention to the yellow spot area of the yellow spotted cocoon and only extracts texture for the double cocoon, without focusing on its edges. Both B-Res41-A and B-Res41-ASE can locate discernible areas of the cocoons; however, B-Res41-A has a broader focus on cocoon features, while B-Res41-ASE shows more concentrated attention on defects in the cocoons due to the introduction of the SE module. Wu et al. [10] employed the regional significant information suppression and feature fusion module (RSIS-FFM) for the feature fusion of silkworm cocoons. By comparing the feature extraction results in the Grad-CAM maps, it was observed that both methods showed similar attention to the features of the stained cocoons. However, our algorithm provides the more comprehensive extraction of contour features for the double cocoons. Additionally, B-Res41 focuses its attention on areas with high lighting intensity in most cases, while B-Res41-ASE enhances the learning of edge features and reduces the impact of illumination. Overall, B-Res41-ASE is more aligned with the classification logic of the human brain and is less influenced by the light intensity, which effectively enhances the classification accuracy and robustness.

4.2. Visual Analysis Based on Feature Maps

To validate the effectiveness of ASFF [17] for silkworm cocoon images, we extracted the feature maps before and after fusion, as shown in Figure 14. It can be seen that the features learned by Level-1, Level-2, and Level-3 are progressively coarsened from color to texture to semantic layer by layer. In contrast, ASFF-1, ASFF-2, and ASFF-3 fuse the features across the layers, enabling information to be exchanged across the layers. As a result, the cocoon feature maps no longer focus solely on the features from a specific layer. For instance, in the case of yellow spotted cocoons, every fused feature map not only retains detailed information on the edge features and the overall global features, but also accurately represents the local yellow spot features. In comparison to the improved YOLOv8 model that utilizes BiFPN for multi-scale feature fusion [11], the proposed B-Res41-ASE demonstrates superior performance in silkworm cocoon classification, with an *F1-score*

improvement of 2.7%. Therefore, the introduction of ASFF significantly improves the performance of the model, enhances the network's ability to integrate features at different levels, and helps to improve the accuracy of silkworm cocoon image classification.

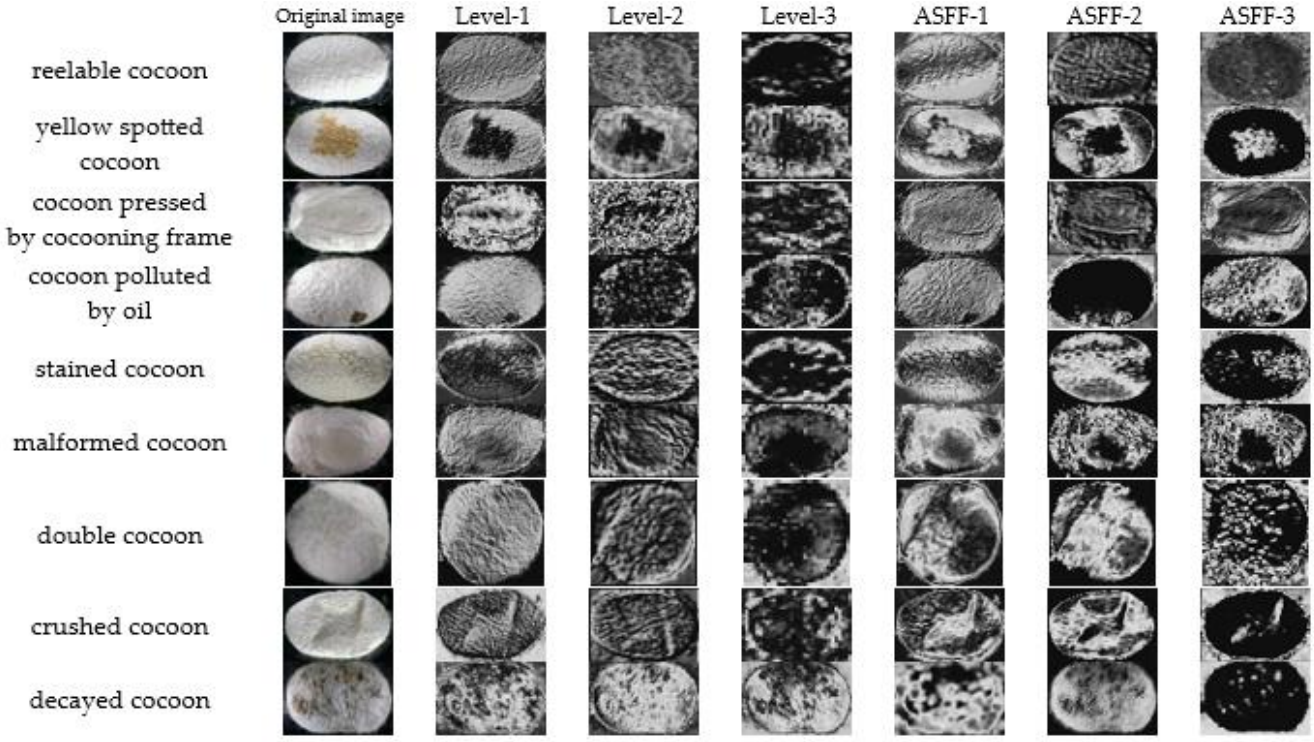


Figure 14. Adaptive spatial feature map before and after fusion visualization.

4.3. Comparison with Different Classification Models

To test the performance of our model B-Res41-ASE, it was compared with models including AlexNet [24], GoogLeNet [25], Bilinear CNN [14], Hierarchical Bilinear Pooling (HBP) [26], and MC loss [27]. The training curves are shown in Figure 15. All networks started to converge after around 400 epochs, with HBP showing greater fluctuations in the training curve compared to the other models. AlexNet had the lowest accuracy and the highest loss after convergence, indicating that models with shallower network depths are not suitable for classifying silkworm cocoon datasets. In contrast, B-Res41-ASE showed the most stable performance in the training process, and the accuracy and loss after convergence were significantly better than those of other models.

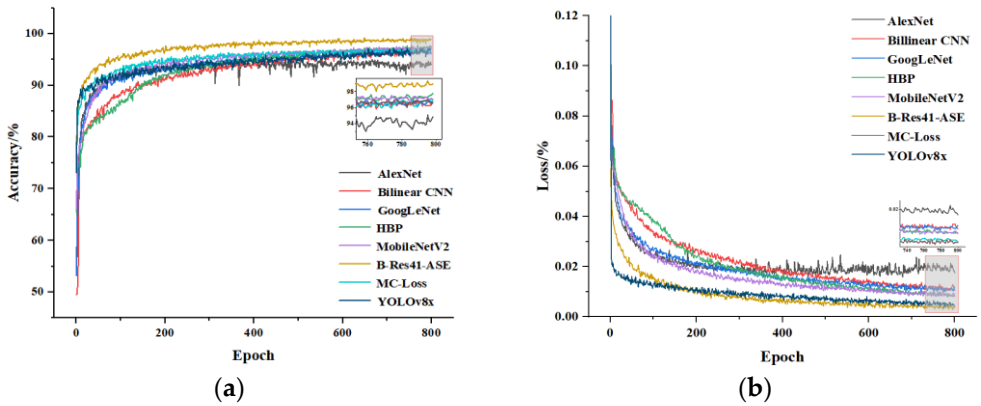


Figure 15. Accuracy and loss value change curve for each model. (**a**) The training accuracy curves of different algorithms. (**b**) The training loss curves of different algorithms.

The evaluation metrics for the different models are shown in Table 5, and B-Res41-ASE achieved *accuracy*, *precision*, *recall*, and an *F*1-*score* of 97.0%, 97.7%, 97.3%, and 97.5%. Compared with the commonly used classification networks AlexNet and GoogLeNet, B-Res41-ASE shows improvements in all evaluation metrics, with the *accuracy* and *F*1-*score* exceeding 3.1%, 2.6% and 2.5%, 2.2%, respectively. This improvement may be attributed to the fact that AlexNet and GoogLeNet only take activations of the last convolution layer as the representation of an image, which is insufficient to describe various semantic parts of cocoons. In contrast, B-Res41-ASE adopts bilinear pooling and ASFF to integrate multiple cross-layer features, minimizing the loss of discriminative information of fine-grained categories. Compared with the fine-grained classification models Bilinear CNN, HBP, and MC loss, the improvement in B-Res41-ASE is mainly reflected in the increased *recall*, which enhances the model's equilibrium performance. Compared with YOLOv8x and MobileNetV2, B-Res41-ASE is slower, which is related to the lightweight characteristics of the model itself, but the *F*1-*score* of B-Res41-ASE is 4.5% higher than that of YOLOv8x and 2.4% higher than that of MobileNetV2. The reason is the feature enhancement module that we adopt, which dynamically calibrates the channel features and guides the model to pay more attention to the discriminative regions of the silkworm cocoon.

Table 5. Performance comparison of different algorithms.

Algorithm	*Accuracy*/%	*Precision*/%	*Recall*/%	*F*1-*Score*/%	*Params*/*M*	*FPS*/*ms*
AlexNet	93.9	97.2	93.0	95.0	30.7	101
GoogLeNet	94.4	95.7	95.0	95.3	44.6	122
Bilinear CNN	95.1	96.7	95.2	95.9	58.1	147
HBP	89.3	100.0	84.9	91.8	80.9	222
MC Loss	94.8	99.2	92.7	95.8	79.2	207
MobileNetV2	95.1	96.8	93.4	95.1	43.3	157
YOLOv8x	94.1	81.8	98.0	93.0	68.2	161
B-Res41-ASE	97.0	97.7	97.3	97.5	77.5	186

The current cocoon classification methods mostly focus on detecting specific types of cocoons. Zhang et al. [3] detected four types of cocoons: yellow spotted cocoons, crushed cocoons, thin-shelled cocoons, and small cocoons. Jiang et al. [4] studied double cocoons and reelable cocoons, obtaining classification accuracy of 98.6%. Guo et al. [5] proposed a model for the identification of yellow spotted cocoons. Our method identifies eight types of waste cocoons and reelable cocoons, achieving *precision* of 98.0% for yellow spotted cocoons and 100.0% for double cocoons and decayed cocoons, with an overall *F*1-*score* of 97.5%. Comparatively, our method is more suitable for application in the actual cocoon sorting processes in factories.

4.4. Comparison with Varieties of Silkworm Cocoons

In order to evaluate the robustness of the model, we selected two spring silkworm varieties (Guangyou No. 1, Jingsong × Haoyue) and two summer–autumn silkworm varieties (Jinxiu No. 2 and Liangguang No. 2) from the most widely cultivated varieties in China [28]. The experimental images are shown in Figure 16, and it can be seen that there are slight variations in size among the different varieties of white cocoons, but no significant differences in their feature characteristics and defect manifestations. The experimental results, as shown in Table 6, indicate that the classification accuracy for different varieties of cocoons is highly similar. This demonstrates that the model B-Res41-ASE is suitable for classifying the cocoons of the main mulberry silkworm varieties in China.

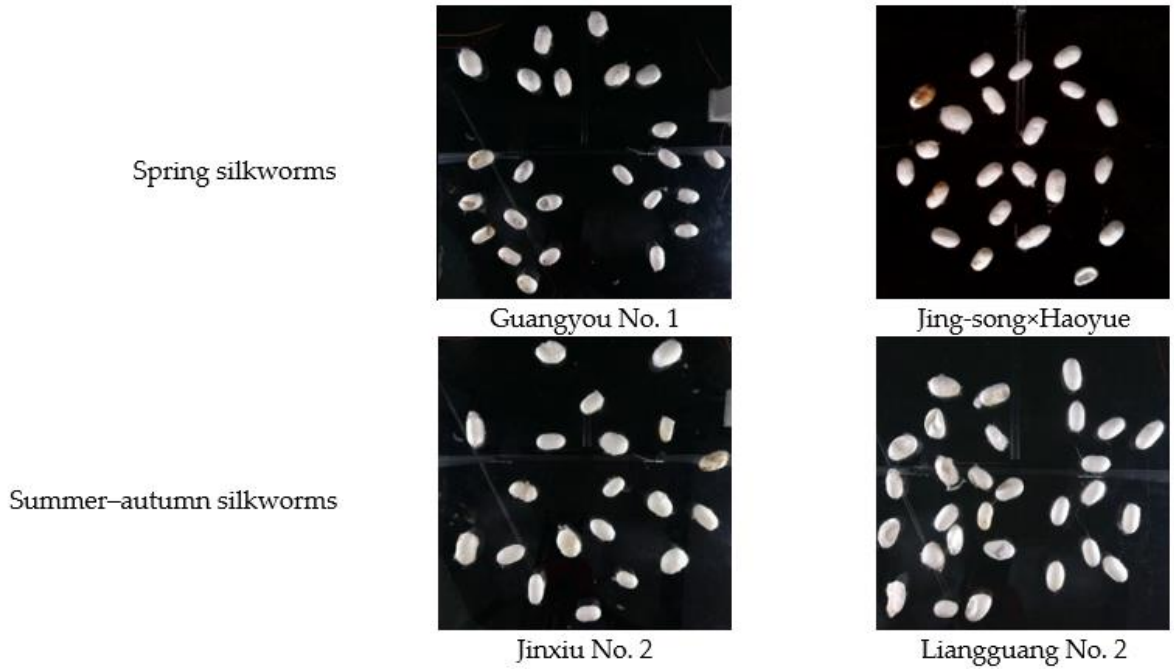


Figure 16. Experimental images of different varieties of silkworm cocoons.

Table 6. Performance comparison of different varieties of silkworm cocoons.

Cocoon Variety	*Accuracy*/%	*Precision*/%	*Recall*/%	*F*1-*Score*/%
Guangyou No. 1	97.1	97.7	97.3	97.4
Jingsong × Haoyue	96.9	97.8	97.2	97.5
Jinxiu No. 2	97.0	97.6	97.3	97.8
Liangguang No. 2	97.2	97.7	97.1	97.6

5. Conclusions

This paper proposes a bilinear pooling model based on adaptive multi-scale feature fusion and enhancement (B-Res41-ASE) to classify silkworm cocoon images. By continuously optimizing the model, the classification accuracy for reelable and waste cocoons is improved, and the classification performance of the model is enhanced gradually. Our model achieved *accuracy* of 97.0%, *precision* of 97.7%, *recall* of 97.3%, and an *F*1-*score* of 97.5% on the test set. The *precision* for each fine-grained category is as follows: cocoons polluted by oil (93.6%), stained cocoons (90.1%), cocoons pressed by the cocooning frame (94.7%), crushed cocoons (98.6%), double cocoons (100.0%), reelable cocoons (97.7%), yellow spotted cocoons (98.0%), decayed cocoons (100.0%), and malformed cocoons (100.0%). Compared with other algorithms, the algorithm proposed in this paper has the best performance in silkworm cocoon classification, which lays an important theoretical foundation for research on the automatic sorting of silkworm cocoon-sorting robots.

The model proposed in this study achieves high classification accuracy for cocoons with visible defects. However, it still faces challenges in distinguishing waste cocoons with internal contamination. In future research, we plan to introduce transmitted light into the classification model, which could improve the perception of invisible and internal features of waste cocoons, thereby enhancing the generalization and accuracy for silkworm cocoon classification.

Author Contributions: Conceptualization, M.L.; methodology, M.L.; software, X.H.; validation, X.H.; formal analysis, X.H.; investigation, M.L.; resources, W.W., Z.S., M.S., and G.Z.; data curation, M.L. and X.H.; writing—original draft preparation, X.H.; writing—review and editing, M.L., Y.Y., X.H.,

and E.O.O.; supervision, Y.Y.; funding acquisition, M.L. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Province Key Research and Development Plan Project (No. 2022TZXD0042); Special National Key Research and Development Plan (No. 2023YFD1600900); Shandong Province Modern Agricultural Industry Technology System, China (No. SDAIT-18-06); China Agriculture Research System of MOF and MARA (CARS-18); and National Natural Science Foundation of China (No. 32001419).

Institutional Review Board Statement: Not applicable.

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

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

References

1. Chen, H.; Yang, Z.; Liu, X.; Shao, L. Study on auxiliary testing method for mulberry silkworm cocoon sorting based on MATLAB. *J. Silk* **2016**, *53*, 32–36.
2. Liu, M.; Li, R.; Yan, X.; Yan, Y.; Li, F.; Liu, S. Detection and elimination of yellow spotted cocoon in mountage based on FCM algorithm and HSV color model. *Trans. Chin. Soc. Agric. Mach.* **2018**, *49*, 31–38. [CrossRef]
3. Zhang, Y.; Yang, H.; Zhu, S.; He, Z. Machine Vision Real Time Detection of Inferior Cocoons Based on Lightweight Manipulation Network. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 261–270.
4. Jiang, Z.; Ying, J.; Wang, C.; Song, L.; He, Q.; Guo, P.; Lin, X.G.; Deng, L.L. Research on cocoon segmentation and double cocoon recognition based on two-parameter thresholds. *J. Silk* **2023**, *60*, 18–25. [CrossRef]
5. Guo, D.; Li, Z.; Wang, X.; Ye, F.; Jun, J. Yellow-spotted cocoon recognition algorithm combined with deep learning and image processing. *Sci. Seric.* **2023**, *49*, 58–66. [CrossRef]
6. Zhou, X.; Han, Z.; Liu, C. Silkworm cocoon identification method based on improved convolution neural network and image processing. *J. Chin. Agric. Mech.* **2023**, *44*, 100–106+120. [CrossRef]
7. Li, Y.; Sun, W.; Liang, M.; Shao, T.; Sheng, J. Registration algorithm for cocoon defect image based on fusion featyres and FAST-SURFU. *Acta Sericologica Sin.* **2021**, *47*, 269–275.
8. Sun, W.; Yang, C.; Shao, T.; Liang, M.; Zheng, J. Intelligence recognition algorithm of group cocoons based on MSRCR and CBAM. *J. Silk* **2022**, *59*, 58–65. [CrossRef]
9. Vasta, S.; Figorilli, S.; Ortenzi, L.; Violino, S.; Costa, C.; Moscovini, L.; Tocci, F.; Pallottino, F.; Assirelli, A.; Saviane, A.J.S. Automated Prototype for Bombyx mori Cocoon Sorting Attempts to Improve Silk Quality and Production Efficiency through Multi-Step Approach and Machine Learning Algorithms. *Sensors* **2023**, *23*, 868. [CrossRef] [PubMed]
10. Wu, Y.; Li, Z.; Wang, X.; Ye, F.; Jin, J. Classification Algorithm of Class Imbalance Cocoon Images Based on Improved ResNet-50. *Acta Sericologica Sin.* **2024**, *50*, 116–126. [CrossRef]
11. Liu, M.; Cui, M.; Wei, W.; Xu, X.; Sun, C.; Li, F.; Song, Z.; Lu, Y.; Zhang, J.; Tian, F.J.A. Sorting of Mountage Cocoons Based on MobileSAM and Target Detection. *Agriculture* **2024**, *14*, 599. [CrossRef]
12. Qiu, C.; Zhou, W. A Survey of Recent Advances in CNN-Based Fine-Grained Visual Categorization. In Proceedings of the 2020 IEEE 20th International Conference on Communication Technology (ICCT), Nanning, China, 28–31 October 2020.
13. *G.T. 9111-2015*; Methods of Mulberry Silkworm Dried Cocoons. China National Standardization Administration, General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China: Beijing, China, 2015.
14. Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1449–1457.
15. Kim, J.; On, K.; Lim, W.; Kim, J.; Ha, J.; Zhang, B. Hadamard product for low-rank bilinear pooling. *arXiv* **2016**, arXiv:1610.04325.
16. 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]
17. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
18. Wang, G.; Wang, K.; Lin, L. Adaptively connected neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1781–1790.
19. Hu, J.; Shen, L.; Sun, G.; Albanie, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
20. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
21. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

22. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; 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.
23. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
24. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
25. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
26. Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; You, X. Hierarchical bilinear pooling for fine-grained visual recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 574–589.
27. Chang, D.; Ding, Y.; Xie, J.; Bhunia, A.K.; Song, Y.Z. The Devil is in the Channels: Mutual-Channel Loss for Fine-Grained Image Classification. *IEEE Trans. Image Process.* **2020**, *29*, 4683–4695. [CrossRef] [PubMed]
28. Feng, J. The Top of Some Silkworm Varieties in China(I). *Bull. Seric.* **2022**, *53*, 33–35+42.