agriculture

MDPI

Article

Attention-Based Fine-Grained Lightweight Architecture for Fuji Apple Maturity Classification in an Open-World Orchard Environment

Li Zhang [1], Qun Hao [1,2,3] and Jie Cao [1,2,*]

[1] Key Laboratory of Biomimetic Robots and Systems, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[2] Yangtze Delta Region Academy, Beijing Institute of Technology, Jiaxing 314003, China
[3] School of Opto-Electronic Engineering, Changchun University of Science and Technology, Changchun 130013, China
* Correspondence: caojie@bit.edu.cn

Abstract: Fuji apples are one of the most important and popular economic crops worldwide in the fruit industry. Nowadays, there is a huge imbalance between the urgent demand of precise automated sorting models of fruit ripeness grades due to the increasing consumption levels and the limitations of most existing methods. In this regard, this paper presents a novel CNN-based fine-grained lightweight architecture for the task of Fuji apple maturity classification (FGAL-MC). Our proposed FGAL-MC architecture has three advantages compared with related previous research works. Firstly, we established a novel Fuji apple maturity dataset. We investigated the Fuji apple's different growth stages using image samples that were captured in open-world orchard environments, which have the benefit of being able to guide the related methods to be more suitable for the practical working environment. Secondly, because maturity grades are difficult to discriminate due to the issues of subtle expression differences, as well as the various challenging disadvantages for the unstructured surroundings, we designed our network as a fine-grained classification architecture by introducing an attention mechanism to learn class-specific regions and discrimination. Thirdly, because the number of parameters of an architecture determines the time-cost and hardware configuration to some extent, we designed our proposed architecture as a lightweight structure, which is able to be applied or promoted for actual agriculture field operations. Finally, comprehensive qualitative and quantitative experiments demonstrated that our presented method can achieve competitive results in terms of accuracy, precision, recall, F1-score, and time-cost. In addition, extensive experiments indicated our proposed method also has outstanding performance in terms of generalization ability.

Keywords: fine-grained; attention mechanism; maturity classification; open-world

Citation: Zhang, L.; Hao, Q.; Cao, J. Attention-Based Fine-Grained Lightweight Architecture for Fuji Apple Maturity Classification in an Open-World Orchard Environment. *Agriculture* **2023**, *13*, 228. https://doi.org/10.3390/agriculture13020228

Academic Editors: Muhammad Sultan, Redmond R. Shamshiri, Md Shamim Ahamed and Muhammad Farooq

Received: 10 November 2022
Revised: 11 January 2023
Accepted: 12 January 2023
Published: 17 January 2023

1. Introduction

The Fuji apple is one of the most popular kinds of fruits worldwide [1]. Different ripeness levels have great differences in taste and usage [2]. For instance, at an early stage of maturity, apples can be used for cider vinegar or cider brewing, whereas apples are ready to be eaten or used for bread and pies at a later stage of ripeness. Moreover, there is a direct relationship between maturity level and the quality or price sold on the market [3]. In addition, different maturity levels have different storage condition requirements, such as temperature, humidity, duration, and so on. Due to this, agricultural managers may formulate a specific harvesting strategy or configure different quantitative environmental factors according to their current maturity stage, such as sunlight, soil, nutrients, etc. Therefore, an accurate and effective distinction of the maturity level is crucial to its management [4–7]. When the Fuji apples gradually mature, the content of internal chlorophyll decreases while the epidermal red pigment increases, and as the pulp begins to soften, the starch begins to

convert into sugar and the acidity decreases. Therefore, the indicators of apple ripeness can be determined by the fruit's color, firmness, starch content, soluble solid content, and sugar content. Basically, the task of apple maturity classification can be divided into destructive and non-destructive methods. The destructive method of grading the ripeness of fruit is performed by quantitatively testing the content of various components in the fruit by chemical means [8,9], which usually causes certain damage to the fruit and makes it impossible to subsequently sell [10]. Computer vision-based apple maturity classification methods are very meaningful and hopeful for tasks in the practical agricultural working system [11]. In recent years, fruit maturity grades sorting or classification research has been widely carried out, such as citrus fruits maturity [12], tomato [13,14], strawberry [15,16], mango [17], and specifically for apples [2,8,9]. With the rapid development of apple harvesting robots in orchard environments, there is urgent demand for visual systems with diverse functions to meet the different custom requirements. Unlike operating in the indoor environment, where many uncontrollable influencing factors may be mitigated compared with the wild, open-world scenarios. In an outdoor environment, fruits are easily occluded by nearby branches, leaves, or even weeds [18]. Moreover, algorithms or models are also easily affected by different sunlight conditions and the background. Despite all these challenges, distinguishing different maturity levels of fruits effectively is one of the most important tasks to achieve intelligent orchard management. Moreover, orchard managers may arrange certain harvesting tasks or the numbers of employees for different maturity periods accordingly. In this study, we presented a novel way to classify different maturity grades of Fuji apples. Our proposed method has many significant differences compared with most related published work. The overall flowchart of our proposed method is shown in Figure 1.

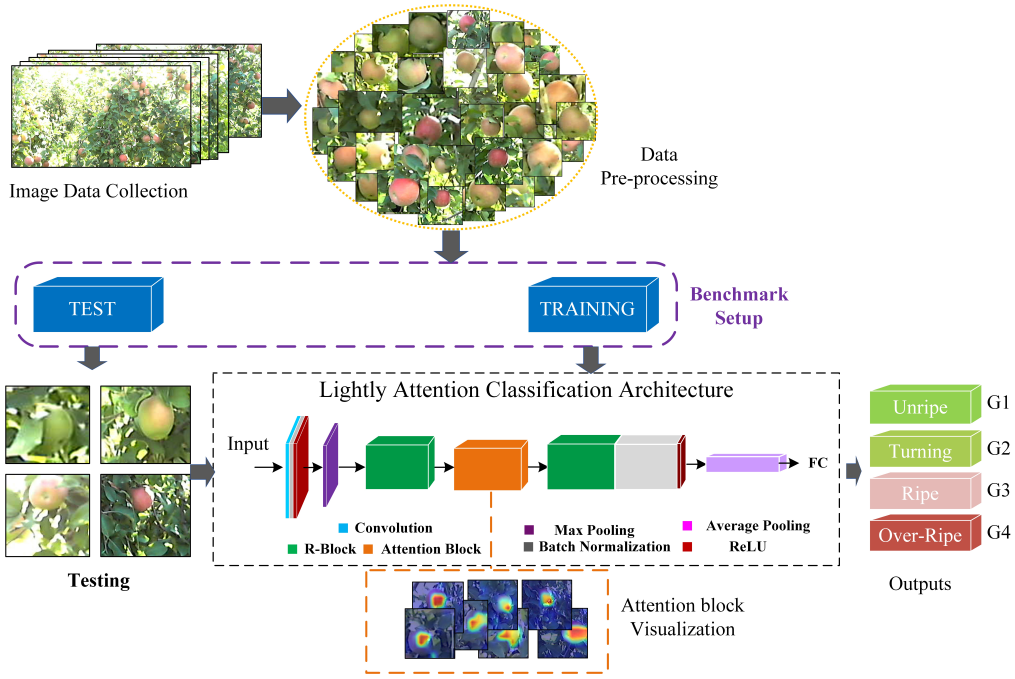

Figure 1. The overall flowchart of our Fuji apple maturity classification task.

Firstly, we captured image samples according to the different growth stages of Fuji apples. After annotation, cropping, labeling, etc., we established a dataset for the Fuji apple maturity classification task. Moreover, we proposed an attention-based fine-grained lightweight architecture for Fuji apple maturity classification (AFGL-MC). Our proposed AFGL-MC is able to learn class-specific regions and discrimination by improved attentional mechanisms, which helped the model to achieve high accuracy outputs. Finally, the results of comprehensive and extensive experiments demonstrated that our presented

architecture provided a potential method for practical application and promotion. The main contributions of our proposed method can be summarized as follows:

- We captured images from outdoor orchard environments during various Fuji apple growth periods, and a Fuji apple maturity benchmark, which contains certain practical influencing factors, was established. Learning-based networks that are trained with diverse scenario data are more suitable to be applied and generalized for practical orchard working environments.
- We proposed a novel AFGL-MC architecture for Fuji apple maturity classification. To the best of our knowledge, it is the first time Fuji apple maturity classification has been applied as a fine-grained task. In addition, to reduce the probability of confusion caused by high similarity in different categories, an improved attentional mechanism was introduced to enhance the discriminative ability of our model. Moreover, we designed the structure as lightweight as possible to facilitate model promotion and practical use.
- Finally, comprehensive and extensive experiments were conducted to demonstrate that our proposed method not only has good performance for the task of Fuji apple maturity classification, but also has excellent performance in other fruit categories and quality classification tasks.

The rest of this paper is organized as follows. In Section 2, we introduce background information and related work. Section 3 shows how we collected and established the Fuji apple maturity dataset. Then, we present our proposed AFGL-MC architecture in Section 4, and describe our experiments and analysis in Section 5. Finally, in Section 6 we discuss the conclusions and plans for future work.

2. Related Work

2.1. Deep-Learning-Based Methods

Since color changes in the appearance of the fruit can be used as a key indicator for the stage of maturity [19], potential non-destructive automatic ways of assessing apple maturity are provided by computer vision techniques [20]. Ref. [21] proposed a method with two layers of a feed-forward back-propagation artificial neural network (ANN) to classify the maturity category of apples and count the number of ripe and unripe apples. The images used to evaluate this method were from online websites. Most of these images were captured under ideal conditions, having almost no background influence. Similarly, Ref. [22] graded Fuji apples into three stages, namely, immature, overripe, and mature according to the changes in apple appearance color. Furthermore, the ANN-based method was utilized to classify Fuji fruits. The methods based on machine vision need to extract features manually, then segment the fruit from the background before judging the level of standards, and such kinds of operations are quite complex. To improve this weakness, Ref. [23] exploited CNN-based architectures to achieve ripe, overripe, and unripe categorizations. In particular, Ref. [24] took Manalagi and Rome Beauty apples as research objects and presented two layers of CNN-based methods to classify the captured apple images into ripe, half-ripe, and raw categories. Such CNN-based methods are more robust in performance and limit the steps of the workflow. Ref. [16] proposed a strawberry ripeness detection system through a camouflage-based data augmentation technique to simulate the strawberry harvesting in natural environment conditions and achieved promising outputs.

All the above research works have inspired the vigorous development of the orchard management industry. However, when applied in working fields, the following shortcomings need to be addressed. First of all, all computer vision-based methods seldom consider the influence of the fruit background. The fruits are entirely in the middle of the image, with a pure different background color, and hardly occluded by leaves or branches. Such kinds of images are too idealistic to be suitable for actual use cases. Secondly, there is no set standard for the classification of maturity levels, so they are difficult to promote or apply practically. Finally, the robust performance of these methods is not strong enough in some regards, although such CNN-based methods reduce some manual operations and lead to a

more intelligent workflow, the designs of these models exploit the CNN to extract features, or try to use some complex architecture, that hardly translates to real-world environments. For all these reasons, the model is difficult to generalize and apply practically.

2.2. Fine-Grained Visual Categorization (FGVC)

In recent years, CNN-based methods ahve achieved great improvements in many fields [25–27]. FGVC refers to a more detailed classification and has become a hot topic in recent years [28–32]. Some existing FGVC methods had used category-level information intuitively and completed FGVC tasks by combining visual information and text descriptions [33]. Generally, such kinds of methods require high requirements for data collection and annotation, which leads to high costs. Some are using the transfer learning method by adding domain adaptation or instance-level-weighted mechanisms, as [34]. Although these methods have taken adaptability into account, they are not suitable in practice due to the complicated training steps involved in its application. Other classical algorithms have focused on the informative parts of an image by introducing an attentional mechanism [35]. Focusing on both the representations and regions used to distinguish categories, attentional mechanisms are widely used in FGVC tasks [36,37]. Therefore, FGVC-based classification tasks for fruits and vegetables have made great progress and have brought valuable and meaningful thoughts to researchers. However, most of the aforementioned methods were applied in laboratory settings; rare investigation on the portability of these methods is one of the most serious limitations.

3. Data Acquisition

First and foremost, we captured images from an open-world orchard environment in Bologna (11°21′ E, 44°30′ N), Italy, with a digital camera (Cannon EOS Kiss X5). More than sixteen weeks of images were captured from 5 July to 15 November once per week, and the distances between the camera and the Fuji trees were around one to one and a half meters. In the first four weeks, the fruits were quite small and totally immature, therefore the images from the last twelve weeks, from which we captured 9852 images, were kept as our original image dataset and processed with a resolution of 1280 by 720 pixels. Some samples are shown in Figure 2.

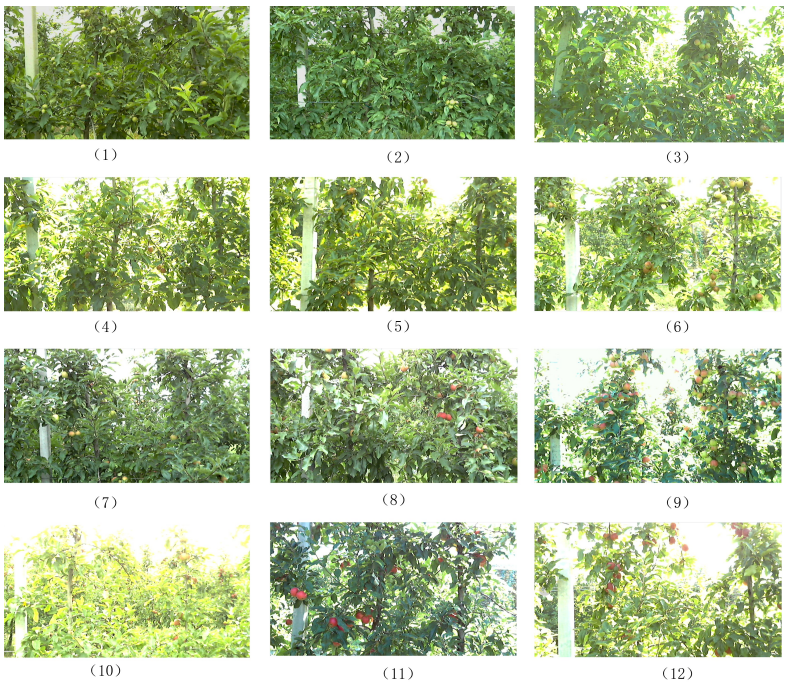

Figure 2. Collected samples from a variety of growth periods. From item (1) to (12), it represents the process of changing the appearance of the Fuji fruit from immature to totally ripe.

As samples shown in Figure 2, when Fuji apples are in the early immature stages, their appearance is almost entirely turquoise. Gradually, the green part is lightly reduced, yellows, and slowly turns red during the progression from the immature to the mature stage. Moreover, images collected in the outdoor environment are easily affected by natural sunlight conditions. For instance, when images were captured in the case of strong sunlight, the color of the fruit's appearance is prone to a certain degree of whitening. For such kinds of images, traditional image pre-processing methods may meet great challenges. Therefore, a more robust method is urgently needed for real surroundings. Due to the crucial factor of establishing an accurate and appropriate dataset for the maturity model learning and prediction, we labeled the images according to the period of image collection and the appearance of color changes. Depending on the corresponding relationship between the changes in fruit epidermis and ripeness, the fruit ripeness grade was divided into four categories according to the United States Standards for Grades of Apples [38]. Ref. [38] indicated as a mature apple becomes overripe it will show varying degrees of firmness, depending upon the stage of the ripening process, hence "Hard", "Firm", "Firm ripe", and "Ripe" were the four terms used for describing the different stages. Due to the appearance of Fuji apples being highly related to its attributes, we invited experts to label these images. According to the captured image date and the appearance of each fruit, the experts classified these images into unripe, turning, ripe, and overripe. These four categories were respectively marked as grade one (G1), grade two (G2), grade three (G3), and grade four (G4) for short. Partial samples from our proposed benchmark are shown in Figure 3.

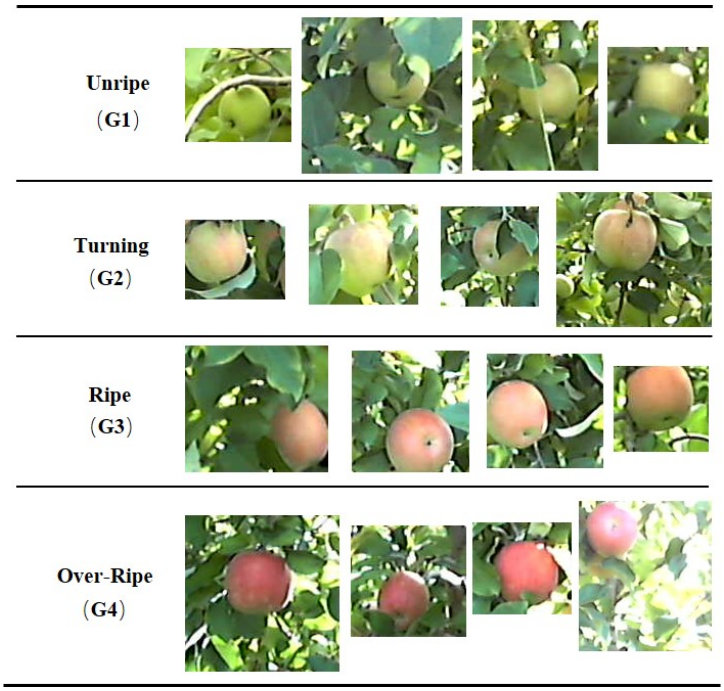

Figure 3. The partial samples of our proposed Fuji apple maturity benchmark.

As the samples show, when the fruits transition from G1 to G2, to G3, and to G4, the appearance color of the fruits change from green (G1) to light pink in some areas of the fruits (G2). By G3, most of the fruit has turned to pink, and finally most of the fruit's appearance presents a relatively dark red area (G4). This process takes an apple from unripe to fully ripe. Moreover, we collected the images totally from an outdoor orchard environment under natural sunlight and surroundings. Therefore, Fuji apples have a variety of poses and degrees of illumination, such as over-light, normal and insufficient. Ideally, if we completely ignore or eliminate the background factor when designing the technical algorithms for fruit classification, detection, counting, or depth estimation, the methods indeed have

certain limitations for practical use. Therefore, we took unfavorable background factors into account and established our benchmark for the Fuji apple's maturity classification. The samples of Fuji apple objects from our proposed benchmark are affected by varying degrees of occlusion due to leaves, branches, or weeds, as shown in the image samples. In fact, the unfavorable influential factors of complex orchard background and diverse severity of occlusion posed a greater challenge to the robustness of the performance in the design of these methods. In summary, more serious challenges to models or methods have to be faced in realistic open-world environment conditions. Nevertheless, we hope to train and evaluate the model based on the challenging data obtained from the outdoor environments, which are likely to guide our model closer to practical application. Finally, 1740 images were obtained and the number of each grade was confirmed to be the same. We then split the images of each grade according to the ratio of the training set and the test set to 4:1. Compared with the other public datasets used for classification tasks, such as ImageNet [39] and MS COCO [40], our dataset has a smaller set of samples with lower resolution. In fact, such kinds of tiny-scale datasets are more prevalent in practical fields due to the high time and monetary costs for data collection in practical use cases. Training with such a tiny-scale dataset means that the network model with deep CNN layers is extremely prone to the phenomenon of overfitting, such as VGG16 [41] or AlexNet [42]. On the contrary, networks with very few CNN layers may be able to mitigate the issue of overfitting effectively. However, the accuracy of the prediction results is hard confirm, especially in an open-world environment, such as the influences of complex background information, sunlight conditions, and occlusions, etc. Therefore, designing a CNN-based architecture suited for this tiny-scale dataset meets even more challenges compared with large-scale datasets.

4. AFGL-MC Architecture

In recent years, CNN-based computer vision classification methods have achieved encouraging performance in many fields, for example, VGG16/19, InceptionNet [43,44], ResNet [45], AlexNet, etc. These methods require a large number of training epochs and a relatively large-scale dataset with diverse image samples. Generally, the model has deeper layers, and hence the time cost is higher for the final prediction and training. Therefore, a model has with a deep architecture is difficult to combine with agricultural robots and promote in practical fields. However, while improvements such as simply reducing or cutting the number of layers from the architecture, can overcome the high time cost and the requirements for large-scale and diverse samples to a certain extent, the model is still prone to encounter problems related to insufficient information learned from the samples of the dataset and the accuracy of the final prediction cannot be guaranteed. Furthermore, as mentioned previously, the acquisition and setup of a large-scale and diverse dataset is a time-consuming and labor-intensive task generally, hence the models that rely on such large-scale datasets may have limitations in their ability to be generalized. The ability of generalization is one of the most important indicators in evaluating whether the model has the potential to be applied in practice in an outdoor orchard environment. To strengthen its generalization, we consider designing a lightweight network model by introducing attentional mechanisms to lead the model to focus on areas with important features rather than them all. In trying to use the limited number of parameters effectively, it is possible to ensure the model has a lightweight architecture and excellent outputs at the same time. Based on the above considerations, we designed the model as shown in Figure 4.

Compared with the classical network models, such as VGG, InceptionNet, ResNet, and AlexNet, as well as in the case of MobileNet-Tiny, AFGL-MC has the notable characteristic of being lightweight. Specifically, the inputs to AFGL-MC are resized RGB color images with both the values of width and height being 64 pixels, and the output is a one-hot encoder that presents the classification results in four kinds of categories. We also detail the AFGL-MC architecture in the following section.

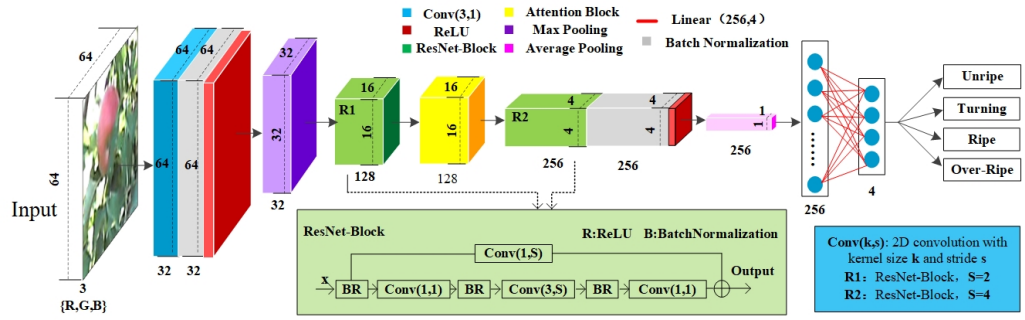

Figure 4. The architecture of our proposed AFGL-MC.

4.1. Layer Unit

The layer unit of our proposed architecture includes the layer of convolution, batch normalization, activation function, pooling, and fully connected layer. The convolution (CONV) layer plays a vital role in extracting image features. Each convolution kernel has the attributes of size, stride, and depth. Generally, the larger the kernel size, the better the quality of the bandpass filters learned [46]. However, for our Fuji apple maturity classification task, the distinction between different levels is so close, and the larger kernel size may lead the model to focus on the large shape or structure of the image, but not on the attributes of local characteristics. Based on such considerations, we chose 1 or 3 as our CONV kernel size. Therefore, the convolution in our architecture only has one or three of these two kinds of kernel sizes. Another benefit is that the number of parameters is reduced as much as possible, which allows it to be lightweight. Regarding the attributes of stride and depth, we designed the model according to the principle that the feature size is gradually reduced. Meanwhile, depth is gradually increased. The high degree of correlation and coupling between layers in the network is one of the factors that causes difficulties with deep neural network convergence and achieving good performance [47]. We exploited the batch normalization (BN) layer after each CONV layer to improve the internal covariate shift phenomenon. We chose ReLU as the activation function [48], shown as equation:

$$f(x) = max(0, x) \tag{1}$$

such that when $x \geq 0$, the output is a linear function. Otherwise, the output value is zero. Additionally, the ReLU has faster convergence than sigmoid and tanh when the outputs are generated using a linear function. We applied the fully connected (FC) layer [49,50] to connect every neuron from the previous layer to the last layer, adn finally the flattened matrix goes through a fully connected layer to classify the images. The FC is presented as equation:

$$u_l = \omega^l x^{(l-1)} + b^l \tag{2}$$

where u_l is the FC layer and $x^{(l-1)}$ is the output of the previous layer. In addition, ω^l and b^l are the weight and bias coefficients, respectively. The output vector u_l goes through softmax activation to obtain the classification likelihood for each potential category. The max-pooling and average-pooling layers provide a typical down-sampling operation that reduces the in-plane dimensionality of the feature maps. The advantages of a pooling [50] layer are to introduce a translation invariance for small shifts and distortions and to decrease the number of subsequent learnable parameters. For the shallow layer of the architecture, we applied max-pooling to filter out features with large amounts noise information, whereas an average pooling layer that is arranged before a fully connected layer based on the consideration of the features from deeper layers could help the final classification results in general.

4.2. Block Unit

The block unit of our presented architecture includes both ResNet [44] and attention blocks. Benefiting from the shortcut structure, the residual network is quite popular as a

unit for the backbone in feature extraction, as applied to a large amount of CNN-based network models. The operation expression is shown as equation:

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{3}$$

here, we denote the output of the ℓth layer as x_l, and $H_l(\cdot)$ represents a nonlinear transformation. The final feature map is the element-wise addition of $H_l(x_{l-1})$ and x_{l-1}. According to its characteristics, we applied the idea of the ResNet block as our attention and the backbone module. To design our attention model, we kept the dimensions of the input and output the same. The improved module is presented as equation:

$$x_l = H_l(x_{l-1}) + Conv_l(x_{l-1}) \tag{4}$$

here, $Conv_l(\cdot)$ represents a convolution operation, and the element-wise addition of $H_l(x_{l-1})$ and $Conv_l(x_{l-1})$ is the output of the attention module. For the backbone network, we have optimized and reduced the module as much as possible, through means of the size of the input samples or in the number of layers of the network model. Intuitively, our model has several stack layers, thus we strove to reduce the values for height and width of the output features from residual blocks by improving its structure. Next, we discuss the attention block [51]. It is hard to guarantee accurate indicators for the final outputs of a CNN-based network with very shallow layers. We proposed an attentional mechanism to strengthen our proposed AFGL-MC to obtain important information from learned samples. Our improved attention structure is shown in Figure 5.

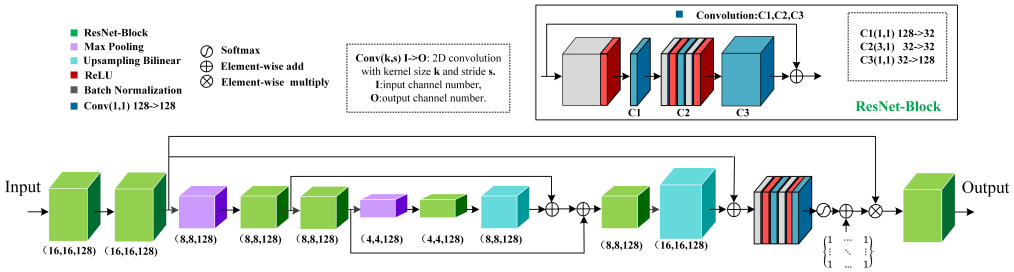

Figure 5. The architecture of our proposed attention module.

Specifically, our proposed attention module was inspired from a residual attention network [51] as multi-scale, and the attention module H is presented as equation:

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * F_{i,c}(x) \tag{5}$$

where i values represent spatial positions and c is the index of the channel. $M(x)$ arranges the values from $[0, 1]$. By element-wise addition of a constant matrix and $M(x)$, we can get the equation $H_{i,c}(x) = F_{i,c}(x)$, when $M(x)$ is zero. As for our proposed attention block, the $M(x)$ includes three branch features that are concatenated to improve the utilization of the extracted features. Moreover, instead of the form of stacked attention blocks, we applied only one attention block by element-wise addition with different feature scales. In addition, we added residual modules at the head and tail to strengthen the functionality of our attention block.

4.3. Loss Function

To mitigate large changes in parameters that may be caused by one or a few noisy samples, we calculated the loss value over the whole batch. During the training processing, we smoothed the loss value by calculating a certain number of batch sizes to mitigate large changes in parameters that may be caused by one or a few noisy samples. The loss value L is presented as equation:

$$L = \frac{1}{N}\sum_{i}^{N} L_i \tag{6}$$

where N is the number of samples in a batch size for one training epoch. L_i is the loss value for each sample, and is shown as equation:

$$L_i = \sum_{c=1}^{M} y_{ic} log(p_{ic}) \tag{7}$$

where M is the number of categories. The probability that sample i belongs to category c is presented as p_{ic}, and the sign function y_{ic} is denoted as equation:

$$y_{ic} = \begin{cases} 1 & i = C \\ 0 & otherwise \end{cases} \tag{8}$$

when the class label i is equal to the true label C, then $y_{ic} = 1$. Otherwise, $y_{ic} = 0$.

4.4. Training and Evaluation

The main hardware configurations of our experiments were conducted on one Intel i5-7500 CPU @ 3.20 GHz processor with 8GB of memory. The GPU is a NVIDIA GeForce GTX 1080 with 11 GB of memory. The operating system is Ubuntu 18.04. We use Python as a programming language and pytorch as our framework. Moreover, due to the learning rate (LR) settings playing a crucial role during the network training, we set the LR as dynamic with the training epochs, correspondingly. In general, a larger LR is suitable for helping the network adjust more quickly and effectively in the beginning stage. As the number of epochs increases, the value of the parameter gets close to the optimal value, hence, if we maintain a large LR throughout the process, it is prone to lead the network to skip its optimal value. Based on the above considerations, we designed a dynamic LR that can be updated as the number of epochs increases with flexibility (Table 1).

Table 1. The relationship between the LR and the number of epochs.

Epochs	LR
$[1 \sim N \times 30\%)$	0.1
$[N \times 30\% \sim N \times 60\%)$	0.01
$[N \times 60\% \sim N \times 90\%)$	0.001
$[N \times 90\% \sim N)$	0.0001

Here, N is the number of total epochs, which was set to 200 in this experiment. The AFGL-MC was trained with a mini-batch size of 64 and tuned with 200 epochs. All the hyper-parameters were optimized with 60 update epochs at learning rates of 10^{-1}, 60 to 120, and 120 to 180, whereas the last 20 update epochs were optimized at lower learning rates of 10^{-2}, 10^{-3}, and 10^{-4}. The terms of precision, recall, F_1-score, and accuracy were taken as the evaluation indexes to verify our proposed network extensively and quantitatively. The corresponding formulas are:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{11}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{12}$$

Specifically, the *TP*, *TN*, *FP*, and *FN* represent the relationship between the observed value and the predicted value, as shown in Table 2.

Table 2. Instructions for the elements of TP, TN, FP and FN.

Label Name	Predicted Positive	Predicted Negative
True Positive	TP	FN
True Negative	FP	TN

5. Experiments and Results Analysis

5.1. Fuji Apple Ripeness Results

We trained the AFGL-MC from scratch, setting the number of training epochs and batch sizes to 200 and 64, respectively. For our specific Fuji apple maturity classification task, we applied the test dataset to verify the accuracy for each ripeness grade after every epoch. These evaluation results are shown in Figure 6.

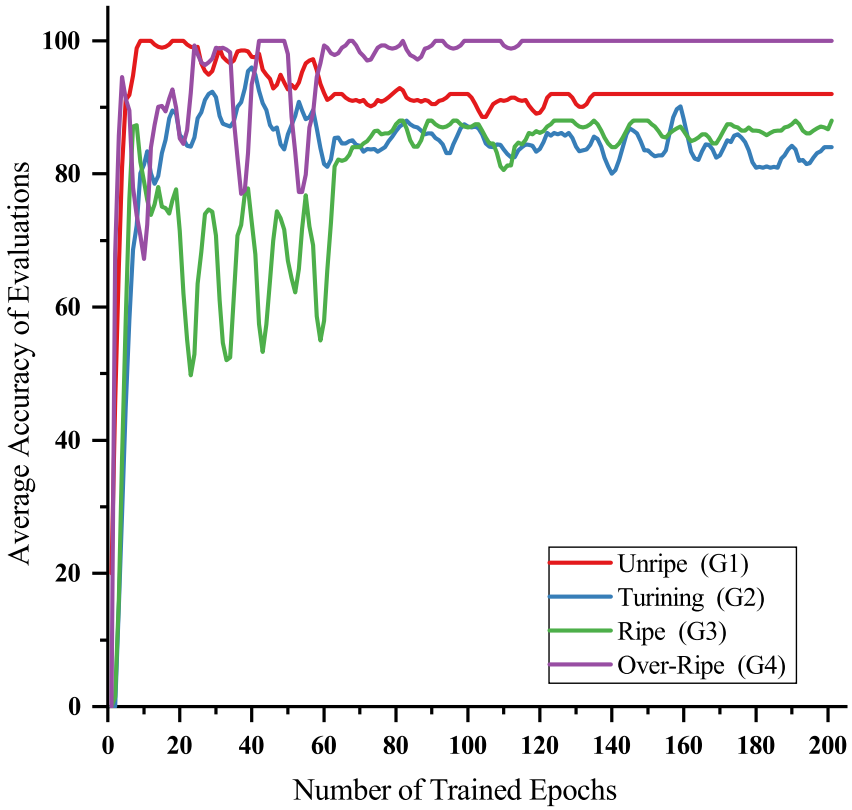

Figure 6. The relationship between the accuracy of each maturity grade and the different training epochs.

As the curves show, when the epochs progress from 1 to 60, each maturity category varies in accuracy severely, which is due to the set of the large learning rate values and the adjustment of the initialization parameters. In this range, G3 and G4 tend to have opposite results. Such a phenomenon is reflected in the practical situation, where fruits at the G3 and G4 levels have extremely similar features in terms of appearance. With the number of iterations increasing, the accuracy of G4 and G1 tends to be stable at in the early ranges compared with the other two grades, while G2 and G3 gradually smooth out and the accuracy tends to a more stable value. This phenomenon may be caused by the obvious appearance feature characteristics in G1 (totally green) and G4 (totally red). Whereas for a fruit in G2 or G3, the changing stage includes some influence factors from appearances that may cause a lower accuracy and a slower change to stable values. Moreover, we visualized

the feature map and focused on the part that was predicted by our proposed attention module, as shown in Figure 7.

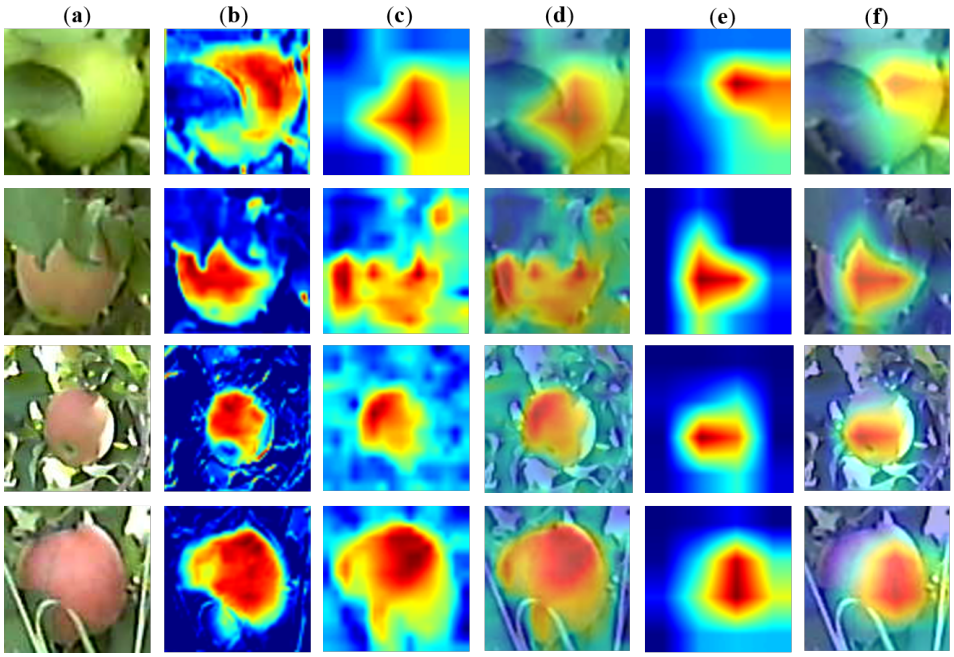

Figure 7. Visualizations of the feature map, highlighting the class-specific discriminative regions for the predicted examples from our proposed network. (**a**) Original four grades' images from our test dataset. (**b**) Visualization feature map by our trained attention block. (**c**,**d**) Localization of class-discriminative regions by the CAM technique in our attention layer, overlapped with the original image. (**e**,**f**) Grad_CAM visualizations of our attention block and overlapped with the original image. Note that in (**c**–**f**), red regions correspond to high class scores.

These samples are from our test dataset, which comprise four maturity grades, as well as images of the Fuji apples occluded by leaves, weeds, or partial appearance whitening due to over-sunlight. These unfavorable factors present great challenges for our fine-grained classification tasks. From the visualization feature map, the features predicted by our attention module can distinguish the background from the fruit accurately, even from the whitening part. As the highlighted region map shown by CAM [52] and Grad-CAM [53], our network evidently could discriminate specific classes for fine-grained maturity classification.

5.2. Model Comparison

In order to verify our proposed network comprehensively, we compared our proposed AFGL-MC with ResNet-18 [45], DenseNet-121 [54], MobileNet-Tiny [55], AlexNet [42], and VGG16 [41] on the task of Fuji apple ripeness classification. We trained all these network models independently on our tiny-scale dataset with 200 training epochs and confirmed the other conditions were as close as possible. We found AlexNet and VGG16 were hard to converge on such a limited number of training samples and number of iterations, therefore we only compared the AFGL-MC with ResNet-18, DenseNet-121, and MobileNet-Tiny. We saved the parameters for each trained epoch and predicted the images from the test dataset, and the results of which are shown in Figure 8.

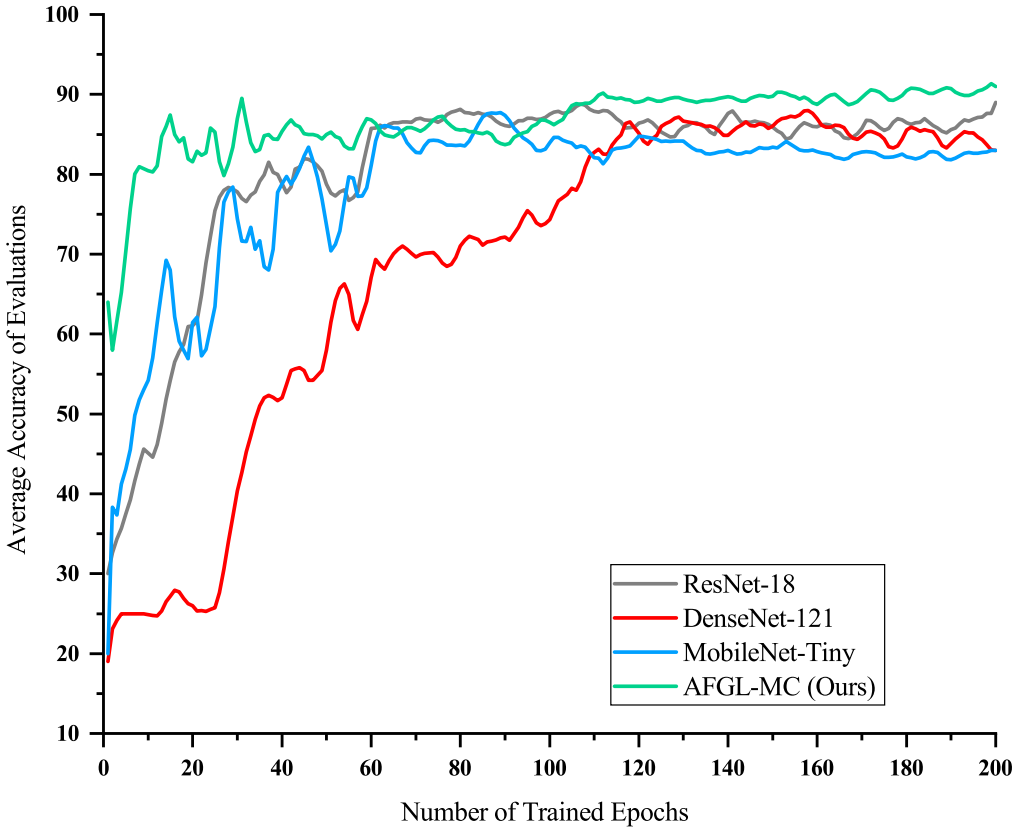

Figure 8. The evaluation results between accuracy and the saved trained models.

As can be seen in the curves shown in Figure 8, our proposed AFGL-MC has the fastest speed and tends to stabilize, followed by ResNet-18 and MobileNet-Tiny, respectively, with the slowest being DenseNet-121. Moreover, our proposed AFGL-MC achieved the best accuracy after 200 epochs of training compared with the others. For a more in-depth analysis, we took precision and recall as evaluation indicators to verify the performance of trained models for each maturity grade. To confirm the fairness of the experiments, we uniformly saved the parameters of the 200th epoch as the final evaluated pre-trained model. The results for each maturity level are shown in Table 3.

Table 3. The evaluation results of four comparison models on each maturity grade.

Level	Indicator	ResNet-18	Dense-121	Mobile-Tiny	Ours
G1	Precision	0.960	0.880	0.80	0.92
G1	Recall	0.960	0.846	0.833	1.0
G1	F_1	0.960	0.863	0.816	0.958
G2	Precision	0.840	0.760	0.80	0.840
G2	Recall	0.840	0.731	0.833	0.875
G2	F_1	0.840	0.745	0.816	0.875
G3	Precision	0.76	0.80	0.76	0.88
G3	Recall	0.86	0.870	0.731	0.846
G3	F_1	0.809	0.833	0.745	0.863
G4	Precision	1.0	0.88	0.88	1.0
G4	Recall	0.893	0.88	0.786	0.926
G4	F_1	0.943	0.88	0.830	0.962
Average	–	0.89	0.83	0.83	0.91

From the table, the results obtained by our proposed model at each ripeness level are generally good, and most of the indicators have achieved optimal values, indicating the effectiveness of our proposed model. In addition, our network achieved notable

prediction results for G1 and G4 levels, with F1 scoring at 0.958 and 0.962, respectively. Although G2 and G3 have many appearance influences, as were explored previously, our proposed network obtained comparatively superior F1 values of 0.88 and 0.863, respectively. Moreover, we took the confusion matrix to more intuitively analyze the performance of each network at different maturity levels, as is shown in Figure 9.

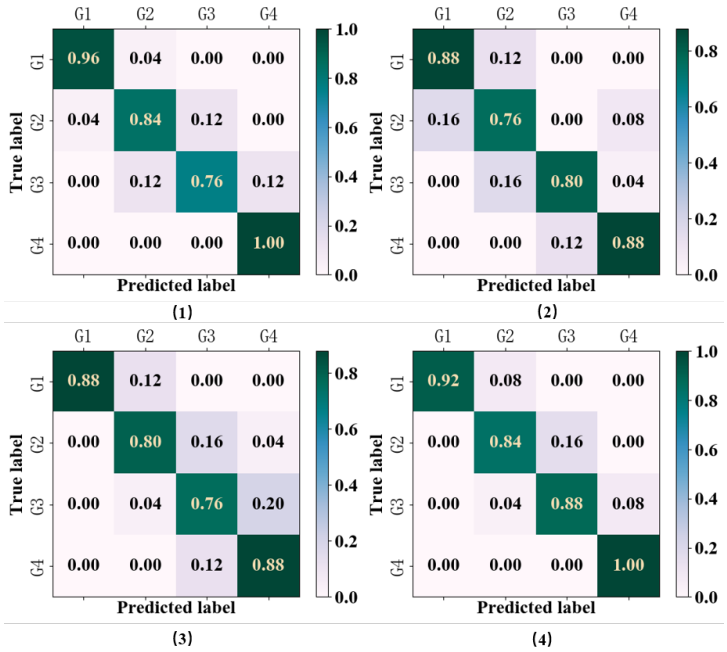

Figure 9. The confusion matrix results of comparison model. (1), (2), (3) and (4) are the results of the comparison model ResNet-18, Dense-121, Mobile-Tiny and our proposed AFGL-MC on each maturity grade, respectively.

From the confusion matrix, it can be seen that the model we proposed and the results obtained by ResNet-18, DenseNet-121, and MobileNet-Tiny on G1 and G4 have obvious advantages compared to the other two levels. However, our proposed network obtained comparable results with the other networks on G2 and G3. It may have benefited from the increased attention mechanism of our proposed network, which may lead our model to pay more attention to the fruit regions, while mitigating the influence of occlusion, sunlight conditions, etc. To analyze the performance of our proposed model under different occlusion and illumination conditions, we set up a experiment for the four previously compared models under normal light without occlusion, normal light with occlusion, and non-integrated light without occlusion. The final results are shown in Table 4.

Table 4. The accuracy results of the four comparison models under different Occlusion and illumination conditions.

Level	ResNet-18	Dense-121	Mobile-Tiny	Ours
No Occlusion	0.927	0.85	0.842	0.930
Occlusion	0.710	0.653	0.620	0.72
Over-Illumination	0.807	0.760	0.759	0.782

From the compared results, we found that all the compared models achieved better accuracy results under natural illumination and no occlusion conditions. When fruits were occluded by objects, the accuracy of the output results were largely decreased, which is probably due to the fact that occlusion factors, in a serious way, will affect the judgment of the fruit's appearance. For the occlusion condition, our proposed model obtained better outputs than the other three compared models, which may be thanks to the attention

mechanism that focuses on the fruit region. As for the illumination changes, they are less affected than in the case of occlusions, and our proposed model handles this situation well.

5.3. Generalization Evaluation

In order to apply the AFGL-MC more prevalently, we took the Fruits Quality (FQ) dataset [56] and the Fruits and Vegetables Classification (FV) dataset [57] to verify the generalization performance of our proposed model. The FQ dataset is a public dataset that provides six fruit categories. Each category has a good, bad, and mixed class. The benchmark includes categories for apples, bananas, guava, lemon, orange, and pomegranate. We randomly chose a bad folder from each category as the benchmark. A selection of samples are shown in Figure 10.

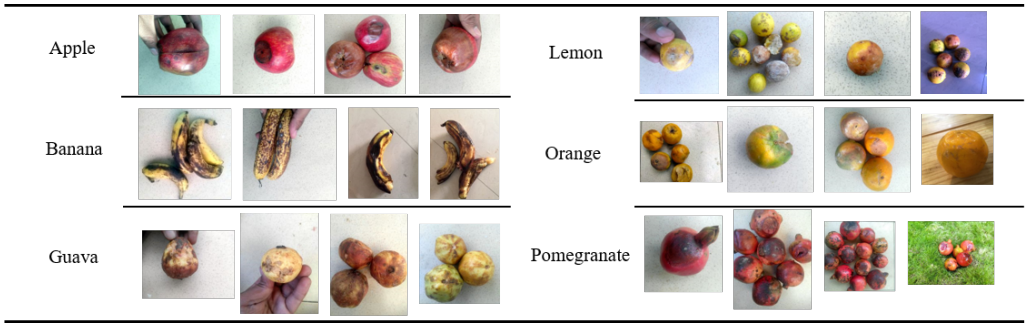

Figure 10. Examples of the FQ dataset on six categories.

In detail, all the samples from the FQ benchmark are image data with width and heights equal to 256 pixels, including the influences of occlusion, complex background, or varied illumination light. The number of samples in each category is around 1000. For example, there are 1104 and 419 samples as the training and test datasets for the apple category. We trained ResNet-18, DenseNet-121, MobileNet-Tiny, and our proposed network on this public dataset. The relationship between the training times and the average accuracy for the final prediction results is shown in Figure 11.

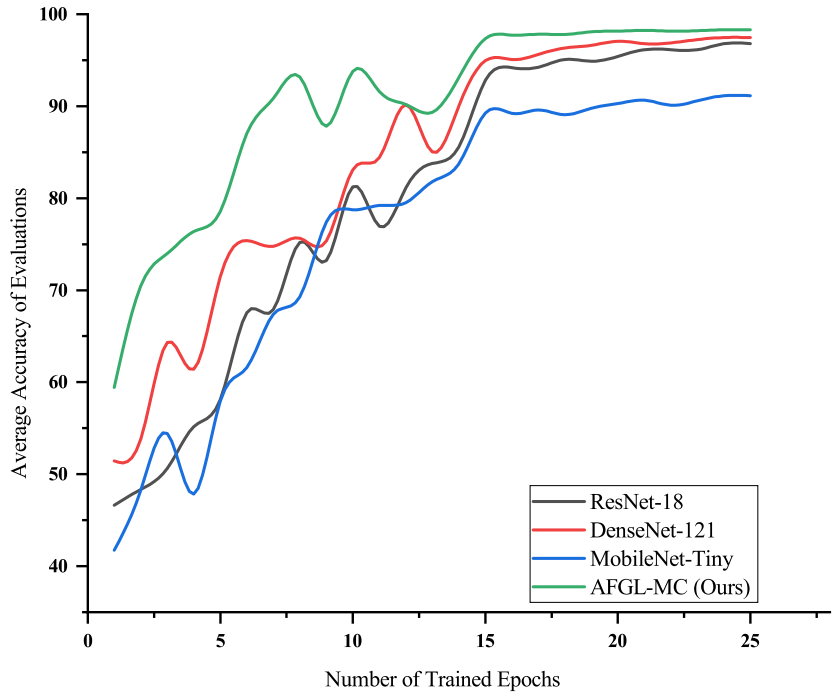

Figure 11. The evaluation results on FQ benchmark.

From the curves, the final prediction accuracy of ResNet-18, DenseNet-121, MobileNet-Tiny, and FGAL-MC are 0.97, 0.91, 0.96, and 0.98, respectively. The results evidently show that our proposed network has better average accuracy results at each epoch compared with the other three, which may reflect that the proposed method has good generalization ability in terms of this dataset. Similarly, we visualized the focused part using a proposed attention module with Grad-CAM highlight method, as shown in Figure 12.

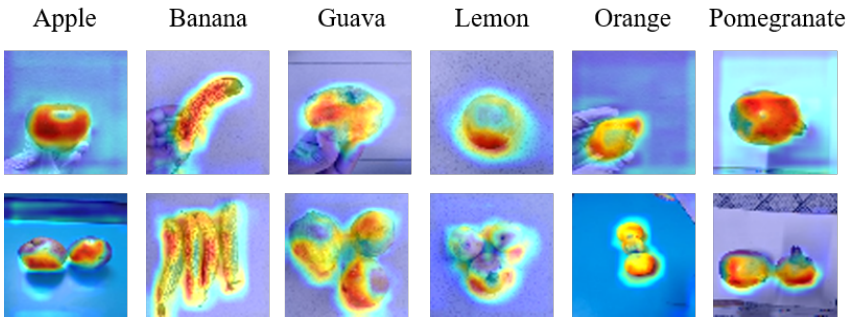

Figure 12. Highlights the class-specific discriminative regions for the predicted examples from our proposed network.

From the discriminative regions, it can be found that our attention model can effectively obtain regions with clear boundaries between objects and the background, which would be helpful for the classification task, thus improving the accuracy of the model. The FV dataset contains 36 categories, including fruits and vegetables. Here we randomly chose 5 fruit classes for our experiment dataset. A selection of samples are shown in Figure 13.

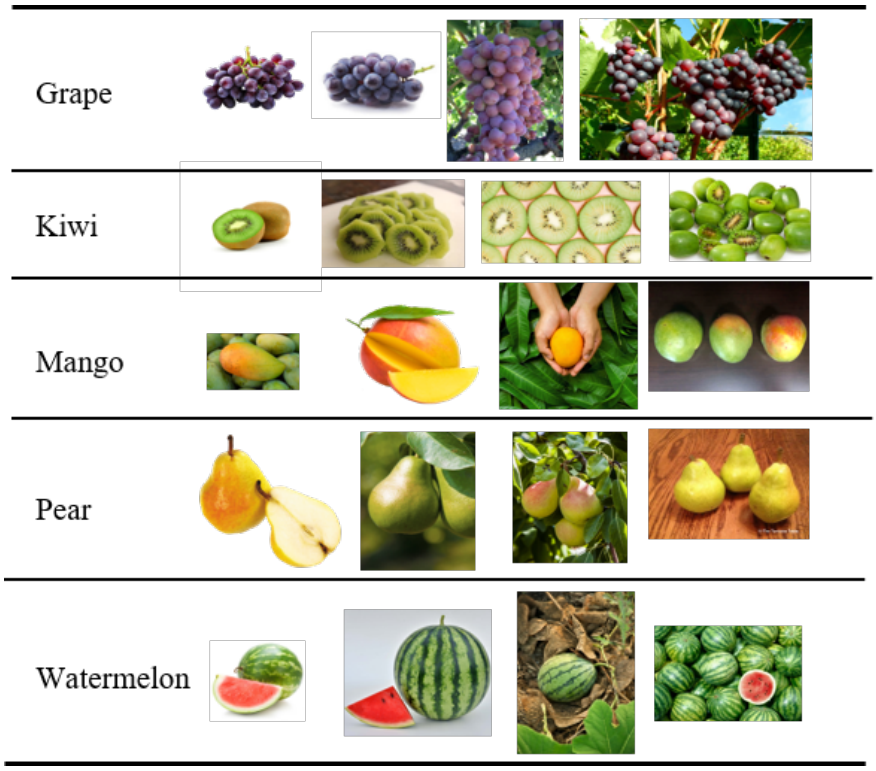

Figure 13. Examples of the FV dataset on six categories.

As for the FV dataset, each class only has 100 training samples and 20 testing samples, which is a relatively tiny dataset. More importantly, the sizes of samples are significantly different. We take the watermelon category as an example. The largest sample is 4416 by 3312 pixels and 9.2 MB in size, whereas the smallest sample is 267 by 260 pixels and

11.28 KB in size. The network is challenged in terms of training or evaluating under such disadvantage factors. Still, we trained ResNet-18, DenseNet-121, MobileNet-Tiny, and our proposed network on this public dataset. The relationship between training times and the average accuracy of the final prediction results are shown in Figure 14.

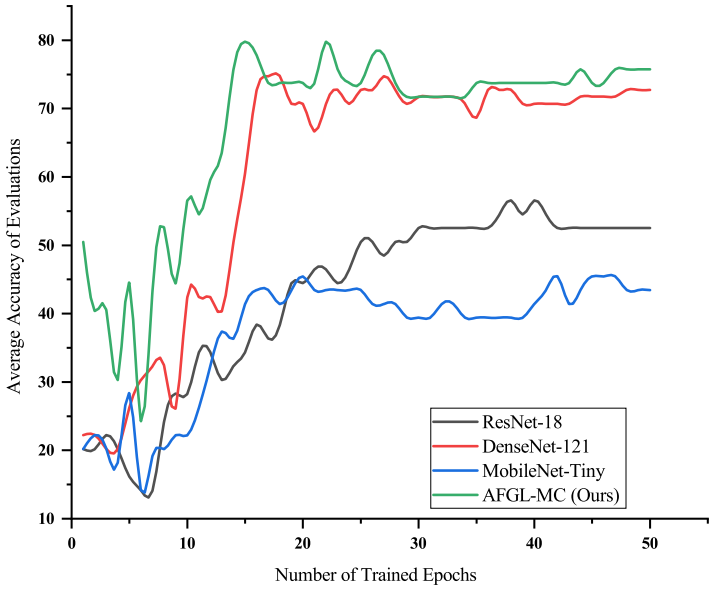

Figure 14. The evaluation results on FV benchmark.

From the curves, we can see the final values for prediction accuracy of ResNet-18, DenseNet-121, MobileNet-Tiny, and our proposed FGAL-MC are 0.53, 0.48,0.43, and 0.74, respectively. Therefore, the average accuracy of the networks trained on the FV dataset is greatly decreased compared with the FD dataset. This implies that the size of the samples and other related properties play a decisive role in a network's final prediction results. However, on such kinds of tiny-scale datasets, our proposed network has a comparatively higher accuracy over other classical network models with a limited number of training epochs. Finally, we visualized the focused part using the Grad-CAM highlight method, as shown in Figure 15.

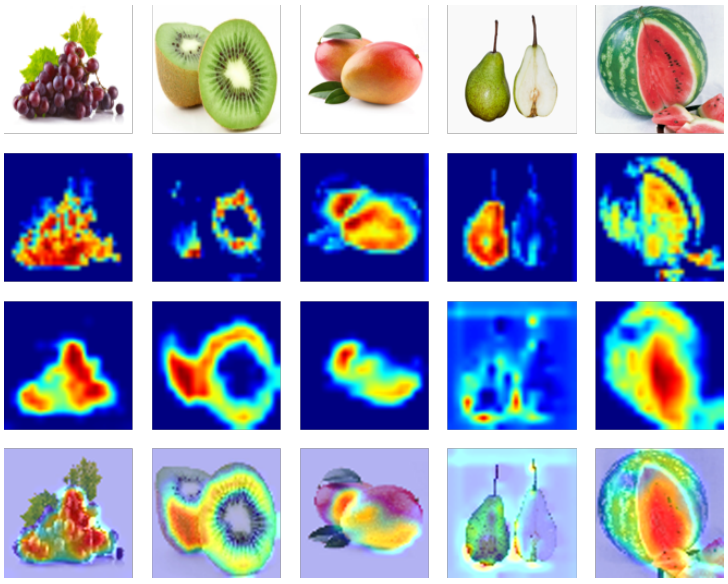

Figure 15. Visualizations of the feature map and class-specific discriminative regions for the predicted examples from the attention module of our proposed network.

Some feature maps lose the class-specific regions, e.g., kiwi and pear, whereas for mango and watermelon, the feature map did not cover all the regions of the objects. Therefore, compared to other datasets, the class-specific discriminative regions were not adequate in terms of the outputs of visualization from our attention module using the Gard-GAM method.

5.4. Parameters and Time Cost Evaluation

A model with the attributes of having lightweight parameters and low time cost is beneficial if it is to be conducted or promoted in practice. We took the number of parameters, the final pre-trained model size, and the time cost in terms of frames per second (FPS) as indicators, and the evaluation results are shown in Table 5.

Table 5. Comparison performance in parameters, memory and FPS.

Name	Number of Parameters (×105)	Memory (MB)	FPS (FQ)	FPS (FV)	FPS (Fuji Apple Maturity)
ResNet-18	111.79	44.8	472	27	386
DenseNet121	69.56	28.4	628	29	527
MobileNet-Tiny	5.40	2.3	857	30	836
AFGL-MC (Ours)	3.88	1.6	874	30	838

From the comparison results, our proposed AFGL-MC performs the best in terms of the number of parameters and memory. ResNet-18, DenseNet121, and MobileNet-Tiny are 20.7, 28.8, and 1.8 times that of AFGL-MC, respectively. The larger the number of parameters, the greater the number of parameters needed to be updated during the training process. The duration of each epoch is also increased. In terms of memory, which is proportional to the number of parameters, our proposed AFGL-MC achieved the best results. In terms of time cost, which is highly related to the resolution of the image being tested, we evaluated the performance of each comparison networks on the FQ, FV, and our proposed FM benchmark, respectively. The final results show that our proposed network achieved the best FPS values on FQ, FP, and our proposed FM dataset.

6. Conclusions

In this paper, we focused on the maturity classification task of Fuji apples. In order to achieve practical applications, we captured images from an open-world Fuji apple orchard and established a novel Fuji apple maturity benchmark that contains practical conditions. Then, we presented an attention-based fine-grained lightweight architecture for the Fuji apple maturity classification task, AFGL-MC. Our proposed AFGL-MC is advantageous in two ways. Firstly, the proposed AFGL-MC with an attention module guides the network to learn class-specific discrimination and regions for classification. Secondly, it has a slim structure, which has good performance in training and final prediction. Comprehensive experiments have shown that our proposed method has significant advantages in accuracy for maturity classification tasks, including many kinds of adverse factors from the orchard environment. Moreover, our proposed method is substantially superior to comparative models in terms of quantity of parameters and time cost. In addition, extensive experiments have shown that our proposed method has significant advantages in generalization ability, and the results implied our proposed method may provide a potential rethinking for more effective use of computer vision in the agriculture industry.

7. Future Work

In future work, we will attempt to deploy our model on a harvesting robot. Then, we will go one step further to focus on the fruit ripeness classification task under occlusion conditions, and try to realize the classification of fruit maturity under different occlusion conditions by complementing the multitude of information of nearby frames in the video.

Author Contributions: Conceptualization, L.Z., Q.H. and J.C.; methodology, L.Z.; software, L.Z.; validation, Q.H. and J.C.; formal analysis, L.Z.; investigation, J.C.; writing—original draft preparation, L.Z.; writing—review and editing, L.Z.; visualization, J.C.; supervision, Q.H.; project administration, J.C.; funding acquisition, J.C. and Q.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Beijing Nature Science Foundation of China (No. 4222017). Funding of Science And Technology Entry program under grant (KJFGS-QTZCHT-2022-008). National Natural Science Foundation of China (62275022).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

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

References

1. Bonomelli, C.; Mogollón, R.; Tonetto de Freitas, S.; Zoffoli, J.P.; Contreras, C. Nutritional relationships in bitter pit-affected fruit and the feasibility of vis-nir models to determine calcium concentration in 'fuji'apples. *Agronomy* **2020**, *10*, 1476. [CrossRef]
2. Menesatti, P.; Zanella, A.; D'Andrea, S.; Costa, C.; Paglia, G.; Pallottino, F. Supervised multivariate analysis of hyper-spectral NIR images to evaluate the starch index of apples. *Food Bioprocess Technol.* **2009**, *2*, 308–314. [CrossRef]
3. Zhang, B.; Zhang, M.; Shen, M.; Li, H.; Zhang, Z.; Zhang, H.; Zhou, Z.; Ren, X.; Ding, Y.; Xing, L.; et al. Quality monitoring method for apples of different maturity under long-term cold storage. *Infrared Phys. Technol.* **2021**, *112*, 103580. [CrossRef]
4. Muscato, G.; Prestifilippo, M.; Abbate, N.; Rizzuto, I. A prototype of an orange picking robot: Past history, the new robot and experimental results. *Ind. Robot. Int. J.* **2005**, *32*, 128–138. [CrossRef]
5. Baeten, J.; Donné, K.; Boedrij, S.; Beckers, W.; Claesen, E. Autonomous fruit picking machine: A robotic apple harvester. In Proceedings of the Field and Service Robotics, Chamonix, France, 9–12 July 2007; Springer: Berlin/Heidelberg, Germay, 2008; pp. 531–539.
6. Tu, S.; Xue, Y.; Zheng, C.; Qi, Y.; Wan, H.; Mao, L. Detection of passion fruits and maturity classification using Red-Green-Blue Depth images. *Biosyst. Eng.* **2018**, *175*, 156–167. [CrossRef]
7. Faisal, M.; Albogamy, F.; Elgibreen, H.; Algabri, M.; Alqershi, F.A. Deep learning and computer vision for estimating date fruits type, maturity level, and weight. *IEEE Access* **2020**, *8*, 206770–206782. [CrossRef]
8. Pathange, L.P.; Mallikarjunan, P.; Marini, R.P.; O'Keefe, S.; Vaughan, D. Non-destructive evaluation of apple maturity using an electronic nose system. *J. Food Eng.* **2006**, *77*, 1018–1023. [CrossRef]
9. Chagné, D.; Lin-Wang, K.; Espley, R.V.; Volz, R.K.; How, N.M.; Rouse, S.; Brendolise, C.; Carlisle, C.M.; Kumar, S.; De Silva, N.; et al. An ancient duplication of apple MYB transcription factors is responsible for novel red fruit-flesh phenotypes. *Plant Physiol.* **2013**, *161*, 225–239. [CrossRef]
10. Lunadei, L.; Galleguillos, P.; Diezma, B.; Lleó, L.; Ruiz-Garcia, L. A multispectral vision system to evaluate enzymatic browning in fresh-cut apple slices. *Postharvest Biol. Technol.* **2011**, *60*, 225–234. [CrossRef]
11. Gao, F.; Fang, W.; Sun, X.; Wu, Z.; Zhao, G.; Li, G.; Li, R.; Fu, L.; Zhang, Q. A novel apple fruit detection and counting methodology based on deep learning and trunk tracking in modern orchard. *Comput. Electron. Agric.* **2022**, *197*, 107000. [CrossRef]
12. Chen, S.; Xiong, J.; Jiao, J.; Xie, Z.; Huo, Z.; Hu, W. Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map. *Precis. Agric.* **2022**, *23*, 1515–1531. [CrossRef]
13. Huang, Y.P.; Wang, T.H.; Basanta, H. Using fuzzy mask R-CNN model to automatically identify tomato ripeness. *IEEE Access* **2020**, *8*, 207672–207682. [CrossRef]
14. Al-Mashhadani, Z.; Chandrasekaran, B. Autonomous Ripeness Detection Using Image Processing for an Agricultural Robotic System. In Proceedings of the 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 28–31 October 2020; pp. 743–748.
15. Liming, X.; Yanchao, Z. Automated strawberry grading system based on image processing. *Comput. Electron. Agric.* **2010**, *71*, S32–S39. [CrossRef]
16. Sadak, F. Strawberry Ripeness Assessment Via Camouflage-Based Data Augmentation for Automated Strawberry Picking Robot. *Düzce Üniversitesi Bilim Ve Teknol. Derg.* **2022**, *10*, 1589–1602. [CrossRef]
17. Razak, T.R.B.; Othman, M.B.; bin Abu Bakar, M.N.; bt Ahmad, K.A.; Mansor, A.R. Mango grading by using fuzzy image analysis. In Proceedings of the International Conference on Agricultural, Environment and Biological Sciences (ICAEBS'2012), Phuket, Thailand, 26–27 May 2012.
18. Jia, W.; Zhang, Z.; Shao, W.; Ji, Z.; Hou, S. RS-Net: Robust segmentation of green overlapped apples. *Precis. Agric.* **2022**, *23*, 492–513. [CrossRef]
19. Bramlage, W.; Autio, W. Determining apple maturity. *Pa. Fruit News* **1990**, *70*, 78–82.
20. Hossain, M.S.; Al-Hammadi, M.; Muhammad, G. Automatic fruit classification using deep learning for industrial applications. *IEEE Trans. Ind. Inform.* **2018**, *15*, 1027–1034. [CrossRef]

21. Lal, S.; Behera, S.K.; Sethy, P.K.; Rath, A.K. Identification and counting of mature apple fruit based on BP feed forward neural network. In Proceedings of the 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), Chennai, India, 4–5 May 2017; pp. 361–368.
22. Hamza, R.; Chtourou, M. Apple ripeness estimation using artificial neural network. In Proceedings of the 2018 International Conference on High Performance Computing & Simulation (HPCS), Orleans, France, 16–20 July 2018; pp. 229–234.
23. Xiao, B.; Nguyen, M.; Yan, W.Q. Apple ripeness identification using deep learning. In Proceedings of the International Symposium on Geometry and Vision, Auckland, New Zealand, 28–29 January 2021; pp. 53–67.
24. Gunawan, K.C.; Lie, Z.S. Apple Ripeness Level Detection Based On Skin Color Features With Convolutional Neural Network Classification Method. In Proceedings of the 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang, Indonesia, 2 October 2021; pp. 1–6.
25. Mavridou, E.; Vrochidou, E.; Papakostas, G.A.; Pachidis, T.; Kaburlasos, V.G. Machine vision systems in precision agriculture for crop farming. *J. Imaging* **2019**, *5*, 89. [CrossRef]
26. Zhao, S.; Peng, Y.; Liu, J.; Wu, S. Tomato leaf disease diagnosis based on improved convolution neural network by attention module. *Agriculture* **2021**, *11*, 651. [CrossRef]
27. Lu, J.; Tan, L.; Jiang, H. Review on convolutional neural network (CNN) applied to plant leaf disease classification. *Agriculture* **2021**, *11*, 707. [CrossRef]
28. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
29. Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; Naik, N. Pairwise confusion for fine-grained visual classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 70–86.
30. Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. Multi-attention multi-class constraint for fine-grained image recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 805–821.
31. Wang, Y.; Morariu, V.I.; Davis, L.S. Learning a discriminative filter bank within a cnn for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4148–4157.
32. Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; Wang, L. Learning to navigate for fine-grained classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 420–435.
33. He, X.; Peng, Y. Fine-grained image classification via combining vision and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5994–6002.
34. Gebru, T.; Hoffman, J.; Fei-Fei, L. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1349–1358.
35. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 842–850.
36. Liu, X.; Milanova, M. Visual attention in deep learning: A review. *Int. Rob. Auto J.* **2018**, *4*, 154–155.
37. Luo, Y.; Jiang, M.; Zhao, Q. Visual attention in multi-label image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
38. Usda, U.; Ams, A. United States Standards for Grades of Apples. 2002. Available online: https://www.ams.usda.gov/sites/default/files/media/Apple_Standards.pdf (accessed on 15 January 2023).
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
40. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
41. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
42. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
44. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
47. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning. PMLR, Lile, France, 6–11 July 2015; pp. 448–456.
48. Hara, K.; Saito, D.; Shouno, H. Analysis of function of rectified linear unit used in deep learning. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–8.

49. Hahnloser, R.H.; Sarpeshkar, R.; Mahowald, M.A.; Douglas, R.J.; Seung, H.S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **2000**, *405*, 947–951. [CrossRef]
50. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef] [PubMed]
51. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
52. Yu, Y.; Xu, L.; Jia, W.; Zhu, W.; Fu, Y.; Lu, Q. CAM: A fine-grained vehicle model recognition method based on visual attention model. *Image Vis. Comput.* **2020**, *104*, 104027. [CrossRef]
53. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 618–626.
54. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
55. Sanjay, N.S.; Ahmadinia, A. MobileNet-Tiny: A deep neural network-based real-time object detection for rasberry Pi. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 647–652.
56. Meshram, V.; Patil, K. FruitNet: Indian fruits image dataset with quality for machine learning applications. *Data Brief* **2022**, *40*, 107686. [CrossRef]
57. Oltean, M. *Fruits 360 Dataset: A Dataset of Images Containing Fruits and Vegetables*; Kaggle: San Francisco, CA, USA, 2019.