Enhancing Fine-Grained Image Recognition with Multi-Channel Self-Attention Mechanisms: A Focus on Fruit Fly Species Classification

Yu Lu 1,†, Ke Yi 2,† and Yilu Xu 1,‡,

1 School of Software, Jiangxi Agricultural University, Nanchang 330045, China; w1973603232@163.com
2 School of Information Engineering, East China Jiaotong University, Nanchang 330013, China;
yk1223764339@163.com
‡ Correspondence: xu_yilu@jxau.edu.cn
† These authors contributed equally to this work.

Abstract: Fruit fly species classification is a fine-grained task as there is a small gap between species. In order to effectively identify and improve the recognition of fruit flies, a fine-grained image-recognition method based on a multi-channel self-attention mechanism was studied and a network framework for fine-grained image recognition based on deep learning was designed in this paper. In this framework, long-term and short-term memory networks are used to extract the underlying features in fruit fly fine-grained images. By inputting the underlying features in the multi-channel self-attention mechanism module, the global and local attention feature maps can be obtained. The weighted attention feature map can also be obtained by multiplying the weight of each channel and the attention feature map. The fine-grained image features of fruit flies were obtained by summing the weighted attention feature map. A softmax classifier was used to process the features and complete the recognition of the fruit fly fine-grained images. Two fine-grained image datasets of fruit flies were applied as experimental objects. Dataset 1 and Dataset 2 contain 11,778 images and 20,580 images from 46 different categories of fruit flies, respectively. The Kappa coefficient was used as the evaluation index to identify fruit fly images with different targets using the method proposed herein. The experimental results showed that, as the number of attention channels increased, the Kappa coefficient gradually increased, suggesting an improvement in the accuracy of fine-grained image recognition. The fine-grained image features extracted by introducing a multi-channel self-attention mechanism exhibited more distinct boundaries with a small amount of overlap, demonstrating strong feature extraction capabilities. When dealing with fine-grained images with either simple or complex backgrounds, the method proposed in this paper has good performance and generalization ability. Even if the target is small and varied in shape, it can still achieve highly accurate recognition.

Keywords: multi-channel self-attention mechanism; fine-grained image recognition; long-term and short-term memory network; softmax classification

1. Introduction

Fine-grained image-recognition technology can classify and recognize images with a high accuracy [1]; therefore, it has important applicative value in many practical applications, such as biological classification, medical diagnosis, and security monitoring [2]. In the agricultural field, it can be used to identify pests like fruit flies. The fruit fly is a common agricultural pest, which poses a serious threat to the growth and yield of crops [3]. By studying effective fine-grained image-recognition methods for fruit flies, we can help farmers to identify fruit fly hazards earlier and guide them to take scientific control measures in a timely manner and, therefore, improve their agricultural production. However, in general, fine-grained image recognition is a challenging task in the field of computer vision [4], as it requires classification and recognition of images with small differences. In the
case of the fruit fly, there are small differences between the species of Drosophila, which increase the difficulty in distinguishing them by traditional fine-grained image-recognition methods. More effective fine-grained image-recognition methods need to be developed to improve the recognition accuracy [5]. By studying effective fine-grained image-recognition methods, we can promote the control and monitoring of agricultural pests, help agricultural departments find fruit fly hazards in a timely manner, and guide agricultural departments to take scientific control measures to deal with fruit fly hazards, improve the control ability of agricultural pests, and improve the management level of agricultural production.

Fine-grained image recognition, especially in distinguishing closely related species like Drosophila (fruit flies), presents a significant challenge for computer vision. Traditional methods struggle with intricate details, prompting the exploration into advanced techniques such as convolutional neural networks (CNNs) and attention mechanisms. Successful fine-grained image recognition holds key implications for agriculture, enabling accurate pest identification like fruit flies. This facilitates targeted and efficient pest control, reducing pesticide use and the environmental impact. Future research should refine specialized models for agricultural pest challenges, fostering collaboration between computer vision and agriculture experts for practical solutions. In conclusion, addressing fine-grained image-recognition challenges in agriculture, particularly fruit fly identification, is crucial for sustainable pest management. Ongoing research and advanced technologies, like deep learning, enhance precision, benefiting global food production and ecosystem health.

In conclusion, our study introduces a novel fine-grained image-recognition method for the effective identification of fruit flies. Two primary innovations define our approach:

1. First, the incorporation of a multi-channel self-attention mechanism enhances feature extraction, allowing for the nuanced recognition of subtle differences between fruit fly species;

2. Second, the utilization of long-term and short-term memory networks as feature extractors contributes to the robustness of the framework, ensuring consistent and accurate recognition across diverse backgrounds. Together, these innovations mark a significant advancement in fine-grained image-recognition techniques tailored specifically for fruit fly identification.

2. Related Works

To date, some papers have conducted relevant research on fine-grained image-recognition methods. Palazzo, S. et al. [6] used prior knowledge to conduct a structural analysis and modeling of images to better capture the semantic and context information in images. By introducing structured advanced knowledge, features with discrimination can be extracted more quickly with more efficiency. In the classifier, the distinguishing features are the input, and the fine-grained image-recognition results are the output. This method is an effective image-recognition method with a high accuracy and precision. However, this method requires a large amount of annotated fine-grained image data for training. If the training data are insufficient or the sample distribution is uneven, it may not be able to fully learn the fine-grained features and semantic information. And, when the training data are insufficient or noisy, this method is prone to overfitting, which reduces the model’s generalization ability. Andriyanov, N. A. and others [7] made the neural network learn more diversified features and patterns in the training process by introducing antagonistic samples, and improved its generalization ability for unknown attacks. The antagonistic training of the neural network makes it more robust and resistant to common visual attacks, thus improving the accuracy and precision of fine-grained image recognition. In a neural network with a complete antagonism training, fine-grained image samples are input and fine-grained image-recognition results are output. However, the adversarial samples introduced during the training of this method may not fully represent the various types of attacks encountered in real application scenarios. Incorrect configuration may lead to overfitting problems in the neural network, so its performance may decrease in reality.
Ohri, K. et al. [8] applied an adversarial erasure strategy on the attention network to learn to locate more component regions that can identify known labels by erasing the most responsive part of a fine-grained image. After obtaining the most responsive component through the attention network, the component area is erased to drive the network to find another discriminant component. They used the similarity loss to limit the similarity of the two components, avoid feature overlap, increase feature diversity, and complete fine-grained image recognition. However, this method may overlook important components or contextual information in some fine-grained images, as erasing the most responsive part may result in the network being unable to accurately recognize the image. And, it has a strong dependence on existing data labels. If the labels contain noise or errors, it will have a negative impact on the learning of the model. Banerjee, A. et al. [9] established an anisotropic Weber operator by introducing angle parameters and scale parameters to solve the problem in the isotropic Weber operator related to the fact that the gray level change information is not fully reflected. According to the anisotropic Weber operator, the local features of the original fine-grained image are extracted and then input into the deep learning algorithm to obtain the fine-grained image-recognition results. Experiments showed that the method is effective for fine-grained image recognition under the conditions of changing lighting, posture, and local occlusion. But, this method mainly focuses on the local information of the image, ignoring global information. In some cases, global information is crucial for fine-grained image recognition, so the recognition performance of this method has certain limitations. Khan, A. et al. [10] used the powerful feature extraction ability of the deep CNN model to automatically learn the feature representation in fine-grained images and capture more semantic information and details. By fusing multi-level features, more abundant feature representation was obtained, and image discrimination was increased to complete fine-grained image recognition. Due to the excessive emphasis on features learned through deep CNN models, this method can easily lead to overfitting problems, which may result in a decrease in the model’s generalization ability on unseen data. Shuhan, L. U., and Si Jing, Y. E. [11] developed an efficient and accurate locust information acquisition technology to understand the correlation between the distribution density and community structure of locusts and the water, heat, and vegetation growth conditions of their habitats, in order to provide rapid and accurate warnings for locust invasion. This paper proposes a semi-automatic locust age information detection model based on locust image segmentation, locust feature variable extraction, and support vector machine classification by analyzing the differences in the morphological characteristics between East Asian locusts and East Asian locusts. Subsequently, its applicability and accuracy were tested based on the sample image data collected on site. However, the image segmentation process of this method is easily affected by factors such as lighting conditions and image quality, and the extracted features may have certain changes, which limits the applicability and accuracy of the method. Li, N., Gao, H., Ding, L. et al. [12] proposed an image segmentation method based on wheel imprint features, as segmenting wheel imprint areas from photos is an important prerequisite for feature extraction and parameter recognition. Compared with other commonly used graphic segmentation methods, by analyzing the mechanism of wheel terrain interaction, the morphological and frequency-domain characteristics of wheel imprints were identified, and feature vectors were constructed. In the wheel trajectory feature space, clustering algorithms are used to divide the image into trajectory and non-trajectory regions. The algorithm was evaluated using segmentation accuracy and processing speed. However, the applicability of the algorithm may be limited by factors such as scene complexity and image quality. In complex environments, segmentation results may be subject to interference or errors.

2.2. Introducing the Multi-Channel Self-Attention Mechanism

The multi-channel self-attention mechanism extracts more information by focusing on different channels of the image, such as color or texture. It can not only improve the accuracy of fine-grained image recognition, but also avoid the overfitting phenomenon [13]
and accelerate the efficiency of image recognition. In this way [14], we can extract and fuse the multi-channel features of the image and pay attention to the important areas of the image, thus improving the recognition accuracy and precision. Therefore, a fine-grained image-recognition method based on the multi-channel self-attention mechanism is studied herein to improve the fine-grained image-recognition effect. The introduction of input gates, forget gates, output gates, and internal memory units in long short-term memory networks solves the problems of vanishing and exploding gradients in current recurrent neural networks (RNNs). It takes account of the importance of both global and local features, which can extract more comprehensive and accurate fine-grained image features of fruit flies. By introducing the feature mean corresponding to each attention weight, higher-order features that are more relevant to the category can be extracted to improve the performance of feature fusion results. The softmax classifier is used to map weighted fruit fly features to categories and output a function of classification probability. Based on the center loss function, the similarity between fruit fly features and their class centers is increased, which reduces the distance between the two in the feature space and solves the problem of significant intra-class differences.

3. Fine-Grained Image-Recognition Method for Fruit Flies

Based on deep learning, this paper designs a network framework for the fine-grained image recognition of fruit flies, which combines fine-grained images \( X \) of fruit flies input into the feature extractor to obtain the underlying features \( F \), using long short-term memory networks (LSTM) as feature extractors. Through the multi-channel self-attention mechanism module, according to the underlying features, the global attention feature map and the local attention feature map are obtained. The weight of each channel is multiplied with the global attention feature map and the local attention feature map, respectively, to obtain the new attention feature representation after channel recalibration. The weighted attention feature map is calculated to obtain the fruit fly fine-grained image features \( T \), then the softmax classifier is used to process the features of the fruit fly fine-grained image and complete the recognition of the fruit fly fine-grained image.

3.1. Fruit Fly Fine-Grained Image Bottom Feature Extraction

Bottom level feature extraction is an important task in image processing, which can help transform the original image into a more meaningful and usable representation. In the fine-grained image processing of fruit flies, the function of low-level feature extraction is to extract relevant feature information from the original image for subsequent processing and analysis. In this section, by using LSTM to extract low-level features, the memory and adaptability of LSTM can be fully utilized to better capture feature information in fine-grained images of fruit flies. The results of such low-level feature extraction will serve as inputs for the subsequent multi-channel self-attention mechanism module, laying a solid foundation for the implementation of fine-grained image-recognition tasks. LSTM is an improved RNN by adding input gates to neurons \( h \), the forget gate \( f \), output gate \( o \), and internal memory unit \( c \). This makes it more advantageous in processing fruit fly fine-grained images, and alleviates the occurrence of gradient disappearance and explosion [15]. Compared with RNNs, it can extract the feature information of fruit fly fine-grained images more effectively [16]. The input gate \( h \) controls how much of the input \( X_t \), the current moment network is saved to the unit state \( Z_t \), the forget gate \( f \) determines how much of the unit state \( Z_{t-1} \) of the upper moment remains in the current moment \( Z_t \), and the output gate \( o \) controls how much of the unit state \( Z_t \) is output to the current output value \( F_t \) of LSTM.

When the input vector matrix of the original fine-grained image of fruit flies is \( X = [x_1, x_2, \cdots, x_k] \), the update formula of LSTM is:

\[
h_t = \sigma(W_h \cdot [F_{t-1}, x_t] + b_h) \tag{1}
\]

\[
o_t = \sigma(W_o \cdot [F_{t-1}, x_t] + b_o) \tag{2}
\]
where $\sigma(\cdot)$ is the sigmoid activation function, $\tanh(\cdot)$ is a hyperbolic tangent function, $W$ is the corresponding weights, $b$ is bias, and $F_t$ is the bottom feature of the final output fruit fly fine-grained image. LSTM extracts low-level features as the input of the multi-channel self-attention mechanism module [17].

### 3.2. Multi-Channel Self-Attention Feature Fusion of Fine-Grained Images of Fruit Flies

By using the two branches of global attention representation and local attention representation, the underlying features are processed separately to obtain global and local attention feature maps. By adaptive max pooling, the spatial dimensions of the attention feature maps of the two branches are reduced to be consistent, and a correlation matrix is used to calculate the weights of each channel. By using matrix multiplication to multiply the attention feature map with weights, the attention feature representation after channel recalibration is obtained. By introducing global and local attention representations and combining them through matrix multiplication, important features in fine-grained images can be better captured, improving recognition accuracy.

In the multi-channel self-attention mechanism module, the idea of matrix multiplication is used to fuse the global and local attention representations, so as to reduce the information loss of fine-grained images of important fruit flies and improve the ability of extracting feature information of key fruit flies [18]. The multi-channel self-attention of the $N$ dimension corresponds to the selection process of $N$ pairs for inputting the characteristics of low-dimensional fruit flies, namely, $N$ is the number of attention feature maps.

The whole attention module is divided into $Q_1$, $Q_2$ branches:

1. $Q_1$ branch is the global attention representation. Note that the representation method utilizes a convolutional layer with $161 \times 1$ kernels, combined with the bottom features of Drosophila $F_t$, the global attention representation is obtained, and the important weights are more prominent at this time; finally, the attention feature map $F_1^2$ is obtained;

2. $Q_2$ branch is the local attention. This branch uses the CBAM (the lightweight attention module) spatial attention representation, and the calculation process of the spatial feature map $F_2^2$ is as follows:

$$F_2^2 = F_1 \cdot \sigma \left( \lambda^{1 \times 1} [\text{max Pool}(F_t); \text{avg Pool}(F_t)] \right)$$

where MaxPool and AvgPool are the maximum and average pooling layers, respectively; $\lambda^{1 \times 1}$ is convolution kernel with a size of $1 \times 1$.

The global fine-grained feature map is pooled to the maximum and average based on the channels, and the splicing and fusion operation is performed on multiple channels [19]. Then, the spatial feature map is obtained through the sigmoid activation function $F_2^2$. Because the maximum and average pooling will cause some information loss, it is called the local attention representation here [20].

In this paper, matrix multiplication is used to combine the global and local attention representation of fine-grained images of fruit flies. In order to meet the conditions of matrix multiplication, the size of $F_1^2$ and $F_2^2$ is compressed from two dimensions $\hat{H} \times \hat{W}$ into one dimension (the $N$ dimension), making the size $[Z \times 16 \times \hat{H} \times \hat{W}]$ become $[Z \times N \times 16]$ and $[Z \times 16 \times N]$, respectively, among which, $N = \hat{H} \times \hat{W}$, and $\hat{H}$ and $\hat{W}$ are the length and width of the feature. Multiplication of the $F_1^2$ and $F_2^2$ matrix give the $D$ matrix, the size of which is $[Z, N, N]$, which can be regarded as a correlation matrix, i.e., the correlation representation between the attention eigenvalues on $F_t$. The above process not only highlights
the important weight but also avoids the loss of important information. The calculation formula of matrix $C$ is as follows:

$$C_{ij} = \lambda_1 \left(F^1_{t,j}\right) \cdot \lambda_2 \left(F^2_{t,j}\right)$$

(7)

where $\lambda_1$ and $\lambda_2$ are convolution operations by branches $Q_1$ and $Q_2$, respectively; $C_{ij}$ can be understood as the element value in column $j$, row $i$, in matrix $C$, representing the influence of the $j$ th fine-grained local attention feature map in $F^2_t$ on the $i$ th fine-granular global attention feature map in $F^1_t$.

To prevent gradient explosion, the $C$ matrix is obtained line by line through the softmax function of the weight matrix $\omega$. The calculation formula of the matrix $\omega$ is as follows:

$$\omega_{i,j} = \frac{\exp(C_{ij})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \exp(C_{ij})}$$

(8)

where the sum of the elements $\omega_{i,j}$ in each row is 1; the influence of the $i$-row element in the $\omega$ matrix represents the attention feature graph of all the fine-grained images of all the fruit flies in feature $i$, which is the weight.

The steps of multi-channel self-attention feature fusion of fruit fly fine-grained images are as follows:

Step 1: Unify size of feature maps

Check the global attention feature map $F^1_t$ and local attention feature map $F^2_t$, conduct adaptive maximum pooling to reduce the spatial dimension to consistency, and maximum pooling can well preserve the texture information of the fruit fly images [21]; so, $F^1_t$ and $F^2_t$ are not only aggregates of a large amount of spatial information, but also effectively reduce the calculation amount of the deep learning model and improve the recognition accuracy of fruit fly fine-grained images [22].

Step 2: Channel recalibration

The expression ability of attention features of the same scale in each channel is different. Calculate the weight of each channel by combining the correlation matrix of attention features with Equation (9) $\omega_i$. Multiply the weight matrix and the attention feature map to obtain a new attention feature representation after channel recalibration $F^1_t$ and $F^2_t$, enhancing the expression ability of the fine-grained feature map in the channel.

The method of the attention weight acting on fruit fly fine-grained image features can be seen as the process of encoding the input fruit fly features under an information selection mechanism [23]. For a single dimension soft attention weight map, the most common fusion method of attention and fruit fly features is to multiply the corresponding attention feature map using the attention weight in the form of a dot product:

$$F^1_t = \sum_{i=1}^{N} \omega_i \cdot F^1_{t,i}$$

$$F^2_t = \sum_{i=1}^{N} \omega_i \cdot F^2_{t,i}$$

(9)

In the deep learning model, the feature mean corresponding to each group of attention weights is introduced. The average attention is a network parameter that represents the average attention of each channel corresponding to the low-level features of all fine-grained fruit fly images. The introduction of the feature mean can extract higher-order features that are more relevant to the fruit fly category, improve the expression ability of the output feature fusion results, and improve the recognition accuracy of fruit fly fine-grained images.

For the fine-grained attention feature map and attention weight of each channel $\omega_i$, the mean value of attention can be expressed as $\bar{\omega}_i$. Then, the fusion method of the attention feature map and channel attention weight in Formula (9) can be rewritten as follows:

$$F^1_t = \frac{1}{N} \sum_{i=1}^{N} \omega_i \left( F^1_{t,i} - \bar{\omega}_i \right)$$

$$F^2_t = \frac{1}{N} \sum_{i=1}^{N} \omega_i \left( F^2_{t,i} - \bar{\omega}_i \right)$$

(10)
Step 3: Attention feature fusion. Sum the weighted attention feature map to obtain the fruit fly characteristics $T$. It effectively fuses low-level geometric information and high-level semantic information. The formula is as follows:

$$T = \left[ \sigma \left( \text{maxPool} \left( F_{i}^{1}, F_{i}^{2} \right) \right) \right] \times \left( \text{maxPool} \left( F_{i}^{1}, F_{i}^{2} \right) \right)^{T}$$  \hspace{1cm} (11)

where maxPool is maximum pooling; $T$ is the weighted fruit fly feature. By integrating global and local attention features combined with the attention mean, a comprehensive feature representation containing both low-level geometric information and high-level semantic information can be obtained, thereby improving the performance of fruit fly fine-grained image recognition.

3.3. Design of Fruit Fly Fine-Grained Image Classification and Recognition Device

Softmax is a commonly used classification recognition function in deep learning models, which can map the input weighted fruit fly features to various categories and return the probability results of classification. However, due to the poor performance of the softmax classification recognizer in tasks with significant intra-class differences, the A-softmax loss function was introduced as its improvement function. The A-softmax loss function adds a constraint variable between the parameter matrix and input features, resulting in greater angular separability between the learned fruit fly features. This loss function restricts the original softmax loss function, achieving better fine-grained image-recognition results.

3.3.1. Softmax Loss Function

Softmax is a commonly used classification and recognition function for deep learning models. Compared with other classification and recognition methods, it can directly output the probability of classification and recognition, which is simple and convenient. However, softmax does not require intra-class compactness and inter-class separation, which makes it difficult for classification and recognition tasks with large intra-class differences. The weighted fruit fly feature $T$ acquired in the input as described in Section 2.2 is mapped from the $M$-dimensional space to the category and returned to the classification identification results in a probability form. The recognition results of fruit fly fine-grained images output by the softmax classification recognizer are as follows:

$$p \left( y \mid T_{i}, T_{j} \right) = \text{softmax}(aT + \hat{b})$$

$$y = \arg \max_{y} p \left( y \mid T_{i}, T_{j} \right)$$  \hspace{1cm} (12)

where $a$ is the parameter matrix of softmax classification identifier; $\hat{b}$ is the offset parameter matrix; $T_{i}, T_{j}$ are the characteristics of fruit flies $i, j$; $p \left( y \mid T_{i}, T_{j} \right)$ is the probability of fine-grained image recognition; $y$ is the result of fine-grained image recognition of fruit fly. Softmax generally uses the cross-entropy loss function:

$$L_{1} = - \frac{\sum_{l} L_{1,l}}{M} = - \frac{\sum_{l} \log \left( \frac{\exp(y_{l})}{\sum_{l} \exp(y_{l})} \right)}{M}$$  \hspace{1cm} (13)

where $L_{1,l}$ represents the loss of the $l$ class; $M$ represents the number of fine-grained image categories of fruit flies; $y_{l}$ represents the output of the final softmax classification identifier for the $l$ components, namely, the result of fruit fly fine-grained image recognition. Because $y$ is determined by the parameter matrix $a$ and input into the characteristic vector
of fruit fly, \( y \) can be expressed as \( y = \arg \max \left[ \text{softmax}(aT + b) \right] \). The calculation formula of \( L_{1,l} \) is as follows:

\[
L_{1,l} = -\log \left( \frac{\exp \left( \| \arg \max \ p_l(y \mid T_i, T_j) \| \cdot \cos \xi_l \right)}{\sum_{\beta} \exp \left( \| \arg \max \ p_{\beta}(y \mid T_i, T_j) \| \cdot \cos \xi_{\beta} \right)} \right)
\]  

(14)

where \( \xi_l \) represents the angle between the fruit fly feature \( T_l \) and the parameter matrix \( a \) as input from the \( l \) fruit fly fine-grained image category.

### 3.3.2. A-Softmax Loss Function

When dealing with fine-grained image recognition, the softmax classification recognizer is mainly based on target logic. When \( y_1 > y_2 \), it is determined as Category 1, otherwise it is determined as Category 2. There is no requirement for the distance between the two categories, and there is no prohibition within the restricted category. Therefore, when dealing with fine-grained image-recognition problems with large intra-class spacing or inter-class spacing, the recognition effect of the fruit fly fine-grained image will become worse. Due to the above defects of the softmax classification recognizer, an improved function for softmax has appeared in the field of fruit fly fine-grained image recognition, which is known as the A-softmax loss function. By adding a constraint variable \( \epsilon \) between parameters \( a \) and \( T_l \), the fruit fly fine-grained image-recognition conditions are more demanding, allowing for a greater angular separability between the learned fruit fly features. The A-softmax loss function limits the original softmax loss function, and sets \( \| a \| \) to 1 (that is, the parameter matrix is normalized), and sets the offset to 0. On the basis of A-softmax parameter matrix normalization and an offset of 0, the characteristics of fruit flies are also normalized. Under this restriction, the target logic of each category is only related to \( \cos \xi \).

The AM-softmax loss function can be written as:

\[
L'_1 = -\sum_i \log \left( \frac{\exp \gamma (\cos \xi_l - \epsilon)}{\exp \gamma (\cos \xi_l - \epsilon) + \sum_{\beta=1}^{M} \exp \gamma (\cos \xi_{\beta})} \right)
\]

(15)

where \( \gamma \) is a superparameter used to scale the cosine value; the hyperparameter \( \epsilon \) is used to control the inter-class interval of fine-grained images of fruit flies. According to Equation (16), the AM-softmax loss function only needs to be calculated during forward propagation \( \cos \xi - \epsilon \). If taking \( \cos \xi \) as an unknown quantity, the gradient in the backward propagation process is always 1, which greatly facilitates the forward and backward propagation calculation of deep learning network training.

However, because of the high inter-class confusion of fruit fly fine-grained images and the large intra-class differences, it is difficult to optimize the deep learning network to the ideal state by using only the AM-softmax loss function. Therefore, it is necessary to introduce a center loss function to increase the similarity between fruit fly features and their class centers, and to reduce the distance between the two in the feature space. The calculation formula of center loss is as follows:

\[
L_2 = 1/2 \| T_i - U_i \|^2
\]

(16)

where \( U_i \) is the corresponding fine-grained image category center of the fruit fly to \( T_i \), and the dimensions of \( U_i \) and \( T_i \) are the same.

In order to avoid the deviation of the decision boundary between fine-grained categories of fruit flies, the deep learning network should be as neutral as possible at the classification boundary of similar categories, and entropy \( y \) can be used to weight the loss function of the deep learning network, where the entropy of \( y \) is:

\[
\rho(y) = \sum_i -y_i \cdot \log(y_i)
\]

(17)
The entropy $y$ directly reflects the chaotic degree of the fine-grained image-recognition distribution of fruit flies in the deep learning network. For a sample, the lower the entropy of the recognition result, the fuzzier the judgment of the deep learning network. When the key features of fruit flies are weakened, the deep learning network should not have too high confidence in the recognition result, and should reason based on more valuable feature information of fruit flies. Although this information cannot directly determine the category, it is a common feature of a few categories, so it can be improved by increasing the entropy. Weighting entropy to the fine-grained image-recognition results of fruit flies and the loss of real labels, the loss function of attention impairment is obtained as follows:

$$L_3 = \left(1 - \frac{\sum_i -y_i \cdot \log(y_i)}{\log M}\right) \cdot \sum_i -y_i \cdot \log(y_i)$$  \hspace{1cm} (18)$$

where $\log M$ is the maximum value of entropy $y$, which is used to normalize entropy. In the task of fruit fly fine-grained image recognition, although the deep learning network can extract more valuable information from non-key fruit fly features, in the case of less fruit fly fine-grained image data and less fruit fly key features, the deep learning network should not be excessively bias towards a certain category when recognizing fruit fly fine-grained images. The prediction confidence between other similar categories should be consistent under the condition of ensuring accurate identification. The purpose of the AM-softmax loss function is only to make the probability of correct recognition higher, while the maximum entropy is to make all recognition results more balanced.

The data volume of the fruit fly fine-grained dataset is small and there are large differences within the data class. Many data are difficult to recognize due to a complex posture, poor angle, and other reasons, which is more likely to cause misjudgment in recognition. At the same time, the fine-grained features of some key data are relatively obvious, which plays a significant role in learning fine-grained features. Therefore, it is necessary to make an adaptive adjustment according to data characteristics. For this reason, the attention impairment loss function is improved, and the loss function weighted by the inverse sample weight is designed to protect the key fruit fly feature information. The formula is as follows:

$$L'_3 = (1 - D)T L_3 = (1 - D)^T \left(1 - \frac{\sum_i -y_i \cdot \log(y_i)}{\log M}\right) \cdot \sum_i -y_i \cdot \log(y_i)$$  \hspace{1cm} (19)$$

where $D$ indicates the correct confidence score of deep learning network recognition. For the fine-grained image samples of fruit flies with low confidence, the corresponding attention impairment loss is increased, and for the fine-grained image samples of fruit flies with high confidence, the learned characteristics of fruit flies are protected. $\tau$ is used to control the sensitivity of the loss function to confidence, and the greater the value, the lower the degree of weakening of the fine-grained image samples of fruit flies with high confidence.

Combining the AM-softmax loss, center loss, and inverse sample weight weighted loss gives the final loss function. The formula is as follows:

$$L = L'_1 + L_2 + L'_3$$  \hspace{1cm} (20)$$

The deep learning network was trained by $L$, and the network training was completed with the network parameters corresponding to the minimum $L$ as the best parameters. In the trained deep learning network, the original fine-grained image of the fruit fly is input, and the fine-grained image-recognition result of the fruit fly is output. This study solves the problems of vanishing and exploding gradients in current RNNs by introducing long short-term memory networks. In order to extract more comprehensive and accurate fine-grained image features of fruit flies, the feature mean corresponding to each attention weight is introduced to extract higher-order features that are more relevant to the category. Using a softmax classifier to map weighted fruit fly features to categories and output a function
of classification probability, based on the center loss function to increase the similarity between fruit fly features and their class centers, it solves the problem of significant intra-class differences and, therefore, achieves accurate image recognition.

4. Experimental Analysis

The accurate recognition and analysis of cell images are crucial in fields such as cell biology, medicine, and drug development. In order to accurately extract and recognize cellular mechanical images, this paper proposes a method based on multi-channel self-attention mechanism. Assuming that the performance of fine-grained image recognition can be significantly improved by combining a multi-channel self-attention mechanism and long short-term memory network, the expected results are as follows:

1. The multi-channel self-attention mechanism can capture global and local features in an image, and strengthen task-related feature representations by assigning different weights to each channel;

2. As a feature extractor, LSTM can handle long-term dependencies in sequence data, thereby extracting low-level features in images, which are crucial for identifying fine-grained objects such as fruit flies.

By combining these two technologies, we expected to achieve efficient and accurate recognition of fine-grained images of fruit flies when processing images with complex backgrounds. Fine-grained image recognition is implemented through the following steps:

- Step 1: Use LSTM to extract the low-level features of the image and capture the feature information $F_i$ in the fine-grained image of fruit flies;
- Step 2: Multiply the attention feature map with weights, integrating low-level geometric information and high-level semantic information to obtain the fruit fly feature $T_i$;
- Step 3: The softmax classification recognizer outputs the image-recognition result $p(y \mid T_i, T_j)$;
- Step 4: Improve the softmax loss function to $A$-softmax $L_i'$;
- Step 5: Merge the AM-softmax loss, center loss, and inverse sample weighted loss values to obtain the final loss function $L$, achieving accurate image recognition.

4.1. Experimental Dataset

Taking two fine-grained image datasets of fruit flies as experimental objects, dataset 1 contains 11,778 images of 46 kinds of fruit flies, and the image background in this dataset is simple. The training set and the test set were divided according to the ratio of 1:1, including 5994 training pictures and 5794 test pictures. In addition, there are abundant annotation information in the dataset, including image category labels and 312 attributes, each of which is naturally visible. Dataset 2 also contains 20,580 images of 46 kinds of fruit flies, and the background of the images in this dataset is complicated, including 12,000 training pictures and 8560 test pictures. Dataset 1 mainly includes *Bactrocera dorsalis*, *Drosophila punctata*, *Drosophila hyde*, *Drosophila melanogaster*, and *Drosophila emii*. Dataset 2 mainly includes fruit flies in Queensland, fruit flies in guava, fruit flies in small berries, fruit flies in olives, and others.

4.2. Evaluation Criteria

The Kappa coefficient is an indicator of consistency test and can also be used to measure the effect of fine-grained image recognition. The Kappa coefficient reflects the consistency of fine-grained image-recognition results and actual classification results. When the Kappa coefficient value is 1, it denotes complete consistency; when the value is 0, it denotes accidental consistency. A value of $-1$ indicates complete inconsistency. Its mathematical expression is as follows:

$$Kappa = \frac{\Psi_0 - \Psi_1}{1 - \Psi_1}$$ (21)
where \( \Psi_0 \) represents the observation accuracy; \( \Psi_1 \) represents the expected accuracy. In the task of fine-grained image recognition, the confusion matrix between categories can directly reflect the recognition ability of the recognition method for confusing categories. The diagonal line in the confusion matrix indicates the recognition accuracy of each type of fine-grained image.

4.3. Parameter Analysis

For the multi-channel self-attention mechanism module described in this paper, the channel dimension \( N \) of the multi-channel self-attention feature map is a key parameter. When the number of self-attention weight channels is low, it may be difficult to provide enough attention information, affecting the results of fine-grained image recognition. When there are many self-attention weight channels, the network parameters increase and so does the complexity of the network. Simultaneously, the dimension of the image representation vector output after attention increases. It is difficult to obtain a compact image representation. We analyzed the Kappa coefficient of fine-grained image recognition obtained by training in dataset 1 when the number of channels in the self-attention weight map increased as follows: 4, 8, 16, 32 and 64. The analysis results are shown in Table 1.

<table>
<thead>
<tr>
<th>Number of Channels</th>
<th>Kappa Coefficient</th>
<th>Computational Cost/ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.75</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>0.95</td>
<td>32</td>
</tr>
<tr>
<td>16</td>
<td>0.96</td>
<td>58</td>
</tr>
<tr>
<td>32</td>
<td>0.97</td>
<td>77</td>
</tr>
<tr>
<td>64</td>
<td>0.98</td>
<td>103</td>
</tr>
</tbody>
</table>

It shows in Table 1 that there was a huge difference in the Kappa coefficient when the number of attention channels was 4 and 8. This suggests that the self-attention weight feature did not provide enough information, which had a great impact on the accuracy of fine-grained image recognition. When the number of channels of the self-attention weight map was greater than 16, the Kappa coefficients were relatively close, which means that, after the number of channels of the self-attention weight map exceeded 16, the multi-channel self-attention mechanism model in this method contained sufficient self-attention information; therefore, it can improve the precision of fine-grained image recognition. The experimental results also suggest that a good balance between the recognition accuracy and model complexity could be achieved when the number of attention weight map channels was 16 or 32.

When designing the AM-softmax loss function, this method includes two superparameters \( \gamma \) and \( \varepsilon \). In this paper, we conducted a superparameter sensitivity experiment on dataset 1, and used the Kappa coefficient as the evaluation criterion. The specific steps are described as follows:

1. Train softmax classifier \( L_1 \) on the training set using the extracted feature vectors and corresponding labels. (2) As needed, the center loss \( L_2 \) and attention weakening loss function \( L_3 \) can be added to the overall loss function \( L'_1 \) to increase the similarity between features and class centers. (3) Set the hyperparameter range: set the hyperparameter to an equidistant value between 0.1 and 1 in the experiment. (4) Set up training and validation process: Divide dataset 1 into training and validation sets. During the training process, the AM-softmax loss function is used as the optimization objective, and the values of hyperparameters are adjusted. In each training iteration, calculate the loss function based on the current hyperparameter settings and update the parameters. (5) Evaluation indicator selection: use the Kappa coefficient as the evaluation criterion to measure the performance of fine-grained image recognition. (6) Conduct hyperparameter sensitivity experiments: Train and validate each combination based on the set hyperparameter range, and record the
corresponding Kappa coefficients. Repeat multiple experiments to obtain multiple sets of Kappa coefficients with different combinations of hyperparameters. (7) Draw hyperparameter sensitivity analysis curve: Based on the experimental results, draw a hyperparameter sensitivity analysis graph (i.e., Figure 1). The horizontal axis represents the values of hyperparameters and the vertical axis represents the corresponding Kappa coefficients. By observing the trend and fluctuation of the curve, the sensitivity of hyperparameters to fine-grained image-recognition performance can be analyzed.

The superparameter sensitivity analysis results are shown in Figure 1.

![Figure 1](image1.png)

**Figure 1.** Sensitivity analysis results of hyperparameters. (a) Sensitivity test of hyperparameter $\gamma$. (b) Sensitivity test of hyperparameter $\epsilon$.

It shows in Figure 1a,b that when the hyperparameters $\gamma$ and $\epsilon$, and the feature selection rate changes, the Kappa coefficient of fine-grained image recognition did not fluctuate significantly. The experimental results suggest that the AM-softmax loss function in this method was insensitive to two superparameters.

4.4. Feature Extraction Effect Analysis

Taking dataset 1 as an example, we used the trained deep learning network to extract features from the original fruit fly fine-grained image, and then applied the t-SNE method to visualize the extracted features in two dimensions, so that the feature maps of different layers were obtained, as shown in Figure 2.

In Figure 2a, it can be seen that the distribution of features in the original fine-grained image of fruit flies is reduced to two-dimensions using the t-SNE method. The distribution of these features in two-dimensional space is relatively scattered, and there is a significant overlap of features between different categories without clear boundaries. This indicates that it was difficult to use the features of the original image to distinguish between different types of fruit flies without processing. Figure 2b,c show the feature maps of the LSTM layer and the last layer in the method proposed in this paper. From the graph changes in Figure 2b,c, it can be seen that, when the image passes through the LSTM layer in the deep learning network, the extracted low-level features are shown in Figure 2b. Compared with the original features, the distribution of these underlying features in two-dimensional space begins to exhibit a certain degree of structural and clustering characteristics. Although there is still some overlap between different categories, it can be seen that, compared to the original features, the LSTM layer was able to extract some useful information for distinguishing fruit fly species. After processing the deep structure of the deep learning network and introducing a multi-channel self-attention mechanism in the last layer, the final fine-grained image features extracted are shown in Figure 2c. This feature map shows a very obvious clustering effect, with clear boundaries formed by features of different categories in two-dimensional space, and the overlapping parts are greatly reduced. This indicates that, through the deep structure hierarchy extraction and the introduction of the
multi-channel self-attention mechanism in the method proposed in this paper, the model was able extract highly discriminative features, thus achieving good fine-grained image recognition of fruit flies.

![Figure 2](image.png)

**Figure 2.** Analysis results of feature extraction in fine-grained image of fruit flies. (a) Original fine-grained image feature distribution. (b) LSTM extracted low-level features. (c) Fine-grained image features finally extracted.

4.5. Analysis of Fine-Grained Image-Recognition Effect of Fruit Fly

A fine-grained image was randomly selected in dataset 1. We used the proposed method, the method in reference [6], the method in reference [7], and the method in reference [8] to identify fruit flies in this fine-grained image. The results of the fine-grained image identification of fruit flies are shown in Figure 3.

![Figure 3](image.png)

**Figure 3.** Recognition results of fine-grained images of fruit flies with a simple background [6–8].
In Figure 3, it shows that our proposed method could effectively identify the fine-grained image of fruit flies for the image with a simple background. According to the recognition results, there are three fruit flies in this fine-grained image. However, other methods could not accurately recognize fine-grained images of fruit flies, and the recognition effect still needs to be enhanced. This means that, after introducing the multi-channel self-attention mechanism into the deep learning network, this method could extract the local and global features of the fine-grained image, improve the comprehensiveness of the feature extraction of the fine-grained image, and thus improve the recognition accuracy of the fine-grained image. This method exhibited a strong adaptability to fine-grained images of fruit flies with different angles and postures, and could effectively overcome the influence of illumination changes, occlusion, and other factors on the recognition results.

Two fine-grained images were randomly selected from dataset 2, and the fruit flies were identified by this method. The results of the fine-grained image recognition of fruit flies are shown in Figure 3.

In Figure 4a, for fine-grained images with complex backgrounds and less obvious targets, the method in this paper could effectively recognize fine-grained images of fruit flies, which benefited from the multi-channel self-attention mechanism. By introducing the mechanism, our method could extract effective local and global features from images, improve the comprehensiveness of feature extraction, and thus improve the recognition accuracy of fine-grained images. The recognition results showed that there is a fruit fly hidden within a bunch of flowers. Although the color of the fruit fly is very close to the background color, the method in this paper could still accurately recognize the fruit fly, which indicates that the method proposed in this paper has a high robustness and generalization ability. It can be seen from Figure 4b that, for fine-grained images with complex backgrounds and small targets, this method could also effectively identify fine-grained images of fruit flies. Two fruit flies were recognized on the leaves. Although the size of fruit flies is small, this method could complete accurate recognition as usual, which suggest that this method has a good adaptability to process fruit flies of small size and different shapes. We adopted a multi-channel self-attention mechanism to extract local and global features in images. This mechanism can focus on important regions and features in the image while preserving image details, and fuse this information together for fine-grained image recognition. For images with complex backgrounds and less obvious targets, by integrating attention mechanisms from various channels, effective features could be extracted from the images, improving the recognition accuracy of fine-grained images. The fine-grained image-recognition performance of the method proposed in this paper was measured by the confusion matrix. Taking dataset 2 as an example, the results were analyzed under the self-attention weight map of 32 channels, as shown in Figure 5.

**Figure 4.** Recognition results of fine-grained images of fruit flies with a complex background. (a) Recognition results of fruit flies with less obvious targets. (b) Recognition results of fine-grained images of fruit flies with a complex background.
Figure 5. Confusion matrix analysis results of fine-grained image recognition of fruit flies [6–8].

As can be seen from Figure 5, for different kinds of fine-grained images of fruit flies, this method could effectively identify fine-grained images, suggesting the method has a good generalization ability and robustness. Among all the fruit flies tested, the accuracy of identification of Psidium guajava fruit flies was the lowest, which was 0.96. This is because the morphological characteristics of Psidium guajava fruit flies are different from other fruit flies, which made it difficult for this method to extract features and identify them. Although the recognition accuracy of Psidium guajava fruit flies was relatively low, it still remained at a high level, which shows that this method has a certain robustness and adaptability in dealing with challenging fine-grained images.

Using the F1 value as the test indicator, 200 images of the orange fruit fly were used as the test set for five experiments to analyze the recognition accuracy of different methods, as shown in Table 2.

Table 2. F1 value results from different methods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.98</td>
<td>0.88</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>0.96</td>
<td>0.86</td>
<td>0.93</td>
<td>0.83</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>0.90</td>
<td>0.94</td>
<td>0.79</td>
</tr>
<tr>
<td>4</td>
<td>0.95</td>
<td>0.88</td>
<td>0.92</td>
<td>0.81</td>
</tr>
<tr>
<td>5</td>
<td>0.97</td>
<td>0.89</td>
<td>0.95</td>
<td>0.82</td>
</tr>
</tbody>
</table>
According to Table 2, the F1 value of our method remained above 0.95 in all five experiments, indicating that our method has a high stability and accuracy in recognizing images of the fruit fly. Although the numerical results of the three comparison methods exhibited small fluctuations, they were always lower than the method proposed in this paper, which verifies the effectiveness of the method proposed in this paper. This is because the method proposed in this paper effectively integrates global and local attention features through a multi-channel self-attention feature fusion strategy, and introduces the concept of the attention mean to extract higher-order features that are more relevant to the category of fruit flies. This strategy enables the method in this paper to capture more fine-grained information when recognizing fruit fly images, thereby improving the accuracy and stability of recognition. In contrast, although the methods in references [6–8] may have certain advantages in certain aspects, their performance in processing the image-recognition task of the fruit fly was inferior to the method proposed in this paper. This was due to the shortcomings of these methods in feature extraction, attention mechanism design, or model optimization, which resulted in their inability to fully capture and utilize the fine-grained information of the fruit fly images.

5. Conclusions

The fine-grained image-recognition method of fruit flies based on the multi-channel self-attention mechanism is an effective image-recognition method. The multi-channel self-attention mechanism in this method can combine information from other modalities and explore multi-task learning methods, which have universality in a wide range of fields such as image processing, computer vision, and pattern recognition. This is especially interesting in scenarios that require capturing fine-grained information in images, fusing multimodal data, or performing multitasking learning.

Looking forward to the future, the fine-grained image-recognition method of fruit flies based on the multi-channel self-attention mechanism still has a lot of room for improvement. Here are some possible directions for improvement:

1. The introduction of more advanced deep learning models: with the continuous development of deep learning technology, we can introduce more advanced deep learning models into fruit fly fine-grained image recognition, such as the Transformer, CNN-RNN, etc., to improve the ability of feature extraction and classification;
2. Combining multimodal information: Fine-grained image recognition of fruit flies can combine other modal information, such as infrared images and ultraviolet images, which may make the method be more robust and accurate for recognition. Therefore, it would be worth trying to fuse the information of different modes into the model to improve the recognition accuracy;
3. Strengthening data enhancement technology: Data enhancement is an effective method to improve the generalization ability of the model. Image enhancement can be achieved by random transformation, cropping, and rotation, so that the model can better adapt to various scenarios and conditions. In the future, more effective data augmentation techniques will be explored to further improve the recognition accuracy of fine-grained images of fruit flies. In the future, more effective data augmentation techniques can be explored to further improve the recognition accuracy of fine-grained images of fruit flies.

Through this study, it is possible to further optimize the design and parameter settings of the multi-channel self-attention mechanism in the future to improve its performance. In addition to the fine-grained image recognition of fruit flies, multi-task learning methods involving other related tasks can also be explored. For example, the fine-grained image recognition of fruit flies can be combined with tasks such as key point detection and pose estimation to improve the overall performance and practicality of the system.
Author Contributions: Conceptualization, Y.L., K.Y. and Y.X.; Methodology, K.Y.; Software, Y.L.; Validation, K.Y. and Y.X.; Formal analysis, Y.L. and K.Y.; Investigation, Y.L. and K.Y.; Resources, Y.X.; Data curation, Y.L.; Writing—original draft, Y.L. and K.Y.; Writing—review & editing, Y.L. and K.Y.; Supervision, Y.X.; Project administration, Y.X.; Funding acquisition, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported by the Natural Science Foundation of Jiangxi Province (Grant no. 2022BAB206089).

Data Availability Statement: The raw data required for these findings cannot be shared at this time as they are part of ongoing research.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


18. Lourenco, V.; Paes, A. Learning attention-based representations from multiple patterns for relation prediction in knowledge graphs. Knowl.-Based Syst. 2022, 251, 109262.1–109262.12. [CrossRef]

19. retak, K.; Person, S. Should data ever be thrown away? pooling interval-censored data sets with different precision. Int. J. Approx. Reason. 2023, 156, 114–133. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.