


Article

Learning Attention-Aware Interactive Features for Fine-Grained Vegetable and Fruit Classification

Yimin Wang ¹, Zhifeng Xiao ²  and Lingguo Meng ^{1,*}¹ School of Microelectronics, Shandong University, Jinan 250101, China; qingnian666666@gmail.com² School of Engineering, Penn State Erie, The Behrend College, Erie, PA 16563, USA; zux2@psu.edu

* Correspondence: myyanghai@163.com

Abstract: Vegetable and fruit recognition can be considered as a fine-grained visual categorization (FGVC) task, which is challenging due to the large intraclass variances and small interclass variances. A mainstream direction to address the challenge is to exploit fine-grained local/global features to enhance the feature extraction and representation in the learning pipeline. However, unlike the human visual system, most of the existing FGVC methods only extract features from individual images during training. In contrast, human beings can learn discriminative features by comparing two different images. Inspired by this intuition, a recent FGVC method, named Attentive Pairwise Interaction Network (API-Net), takes as input an image pair for pairwise feature interaction and demonstrates superior performance in several open FGVC data sets. However, the accuracy of API-Net on VegFru, a domain-specific FGVC data set, is lower than expected, potentially due to the lack of spatialwise attention. Following this direction, we propose an FGVC framework named Attention-aware Interactive Features Network (AIF-Net) that refines the API-Net by integrating an attentive feature extractor into the backbone network. Specifically, we employ a region proposal network (RPN) to generate a collection of informative regions and apply a biattention module to learn global and local attentive feature maps, which are fused and fed into an interactive feature learning subnetwork. The novel neural structure is verified through extensive experiments and shows consistent performance improvement in comparison with the SOTA on the VegFru data set, demonstrating its superiority in fine-grained vegetable and fruit recognition. We also discover that a concatenation fusion operation applied in the feature extractor, along with three top-scoring regions suggested by an RPN, can effectively boost the performance.



Citation: Wang, Y.; Xiao, Z.; Meng, L. Learning Attention-Aware Interactive Features for Fine-Grained Vegetable and Fruit Classification. *Appl. Sci.* **2021**, *11*, 6533. <https://doi.org/10.3390/app11146533>

Academic Editor: Theodore Tsiligiris

Received: 29 May 2021

Accepted: 9 July 2021

Published: 16 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: fine-grained visual categorization; image classification; attentive feature representation; feature interaction; vegetable and fruit recognition

1. Introduction

Despite the consistent improvement in the application of convolutional neural networks (CNNs) to various computer vision tasks, fine-grained visual categorization (FGVC) is still a challenging task due to the large intraclass variance, small interclass variance, and the difficulties in obtaining part annotations [1,2]. To address these challenges, prior studies have explored a wide spectrum of FGVC methods, which can be generally divided into two categories [3]. The first type of methods start by locating the critical regions through a localization subnetwork [4–7], and then fusing global features from the whole image and local features from the critical regions for the final recognition. The other type of methods attempt to learn discriminative features directly via an end-to-end feature encoding network [8–12]. A common goal of these methods is to enhance a model's capability to exploit distinguishable fine-grained features from global or local regions for performance boosting. Their main difference is that the former focuses on certain informative regions of an image, while the latter aims to find critical patterns from the whole image.

As an FGVC task, the recognition of vegetables and fruits has high practical significance in the implementation of fine-grained cooking and food management. One of

the most influential domain-specific data sets for FGVC is VegFru [13], which contains 200 vegetable categories and 92 fruit categories, with more than 160,000 images in total and at least 200 images for each subcategory. Data sets such as VegFru can be utilized to build automated food management systems that can recognize raw food materials and recommend suitable recipes for individuals with different dietary needs.

Most CNN-based FGVC models [4–12] only take a single image as input each time during training. However, in FGVC tasks, objects in different subcategories share many visual similarities, which increases the learning difficulty for models that learn from individual images. In contrast, humans often recognize fine-grained objects by comparing image pairs to extract subtle visual differences that can be used as distinguishable features. Inspired by this intuition, recent efforts have explored ways to learn interactive features from image pairs. A typical study, Attentive Pairwise Interaction Network (API-Net) [14] is one of the novel networks motivated by this capacity of human beings. API-Net feeds a pair of images into a backbone CNN to obtain two individual feature vectors, which are used to create a mutual clue feature vector. In addition, API-Net takes the mutual and individual feature vectors to generate gate vectors that can highlight semantic differences between the two input images. In a nutshell, an individual feature activated by its own gate vector is encouraged to be more discriminative than the one activated by the other gate vector. API-Net has demonstrated state-of-the-art (SOTA) performance in several open FGVC data sets, such as Stanford Cars (95.3%) [15], NABirds (88.1%) [16], and Aircraft (93.9%) [17]. However, as we apply API-Net to the VegFru data set, the performance is worse than ResNet50 by 0.761% in accuracy, potentially owing to the fact that API-Net only adopts channel attentions that highlight *what* is meaningful in an input image, while spatial attention plays a more crucial role to identify where is meaningful in the image [18]. The lack of spatial attention could make API-Net vulnerable to complex background noise. Attention mechanisms have been widely adopted in computer vision tasks. Wang et al. [19] propose a residual attention network that stacks many attention modules to generate attention-aware features. Hu et al. design a squeeze-and-excitation attention block [20] to fuse both spatialwise and channelwise information across feature maps at each layer. A similar study named Convolutional Block Attention Module (CBAM) [18] also explores both channel and attentive spatial features that allow a model to learn what and where to focus.

Inspired by prior efforts, we propose a framework named Attention-aware Interactive Features Network (AIF-Net) that enhances fine-grained feature learning through an integration of a biattention (spatialwise and channelwise attention) module [18] and a modified Attentive Pairwise Interaction module [14] into a backbone network. Specifically, the proposed AIF-Net consists of three components, including: (1) An attentive feature extractor that allows the network to identify and learn from critical areas in an image where distinguishable patterns may reside in. In addition, to exploit both global and local feature maps, we integrate a region proposal network (RPN) into the pipeline to generate a collection of informative regions; the top regions are selected to create attentive local features that are fused with the attentive global feature. The fusion output is utilized by the downstream components of the network to enhance mutual and individual feature learning. (2) An interactive feature learning module that learns to distinguish subtle pattern differences in an image pair through pairwise feature interaction, and (3) a softmax classifier with individual and pair regularization terms that can effectively utilize the attentive features to optimize the underlying neural network.

We conduct extensive experiments to evaluate the performance of the proposed AIF-Net on the VegFru data set. Two key design choices, including the fusion operation (concatenation vs. summation) and the number of top-scoring informative regions used for local feature extraction, are validated. We also report an overall performance comparison between AIF-Net and its peers, including ResNet [21], VGG [22], Compact Bilinear Pooling (CBP) [9], HybridNet [13], Destruction and Construction Learning (DCL) [23], Weakly Supervised Data Augmentation Network (19) (WS-DAN) [24], and API-Net [14], and the

latter four represent the SOTA. Results demonstrate a consistent performance improvement of our AIF-Net over other comparative models, validating the effectiveness of the proposed neural architecture in FGVC.

The rest of this paper is structured as follows. In Section 2, we review the related studies on FGVC. In Section 3, we provide the technical details of the proposed AIF-Net. In Section 4, we report the experimental results with analysis and insights. In Section 5, we provide a discussion and point out potential extensions.

2. Related Work

This section provides a summary of existing FGVC methods (1) with localization–classification subnetworks, (2) with end-to-end feature encoding, and (3) that use data augmentation.

2.1. Methods Based on Localization–Classification Subnetworks

Localization of critical regions can mitigate the challenge of intraclass variation. Early methods focusing on localization of critical regions rely on the manual part annotation [25], which is costly. Recent FGVC methods can achieve localization with image labels only. Jaderberg et al. [4] propose a spatial transformer network that is invariant to certain affine transformations. In [5], a long-short-term-memory (LSTM)-based neural architecture is applied to localize the subtle and discriminative regions in an iterative manner. Multiple regions with attention are localized in [26] by pooling the spatially related channels, and each group of channels corresponds to an attentive region. In [7], a Feature Pyramid Network (FPN) is used to localize multiple critical regions with different sizes or aspect ratios.

2.2. Methods Based on End-to-End Feature Encoding

Methods with end-to-end feature encoding can directly learn discriminative features from input images. One important work, Bilinear-CNN [11], represents the input image as a pooled outer product of features from a deep CNN, leading to remarkable performance improvement. However, due to the high dimension of features, the computational burden of Bilinear-CNN grows exponentially. To speed up computation, authors of studies such as [9,27] have explored ways of aggregating low-dimension embeddings through tensor sketching. Pairwise confusion, proposed in [28], reduces overfitting by intentionally introducing confusion in the activations and demonstrates efficacy in dealing with interclass similarity.

2.3. Methods Using Data Augmentation

Data augmentation methods can be used to enhance the data set diversity, which encourages models to learn more subtle and discriminative features. In [23], the spatial layout of an input image is destroyed to push the network to learn fine-grained features from the randomly shuffled inputs. API-Net [14] captures the pairwise interaction by learning mutual features from an image pair. In the Weakly Supervised Data Augmentation Network (WS-DAN) [24], high-quality features are kept and the useless features are dropped. Another direction to augment training set is through a Web-supervised network [29–31] that directly learns from the real-world Web images, which greatly increases the size of training set. A challenge with this approach is to eliminate irrelevant noisy images that are harmful to the training.

In summary, the three types of methods focus on different aspects of feature learning with their own merits and weaknesses. For the first type, focusing on the local features in critical regions may help discover informative patterns but could lose a global understanding of the whole image; on the other hand, the second line of efforts aims to mine sufficient patterns from the global image in an end-to-end framework, which may miss critical local information; the third type of method attempt to improve the quality of training data that could allow a learning algorithm to learn distinguishable features from more diverse input data. It is noted that the three methods are not mutually exclusive and can complement

each other to further boost the performance of FGVC tasks. Our work is driven by this idea. Compared to the prior efforts, the proposed AIF-Net aims to enhance the quality of extracted features via spatial and channelwise attention, an aggregation of local and global features, and interactive feature learning. The joint effects of these building blocks lead to a significant performance boost in the fine-grained vegetable and fruit recognition task.

3. The Attention-Aware Interactive Features Network

In this section, we present the technical details of the proposed AIF-Net. Figure 1 depicts the system framework of AIF-Net, which consists of three components: an attentive feature extractor, an interactive feature learning module, and a softmax classifier with individual and pair regularization terms. The AIF-Net takes as input a pair of images from the same or different categories in the data set. Each image in the pair is processed by the attention-enhanced network with region proposal networks (RPNs). Global features extracted from the whole image and local features extracted from critical regions are fused through either concatenation or summation. The attentive fused features are then used to compute interactive features that are fed into the classification layer. Finally, the softmax classifier adopts regularization terms for both individual images and image pairs to construct the loss function for end-to-end training.

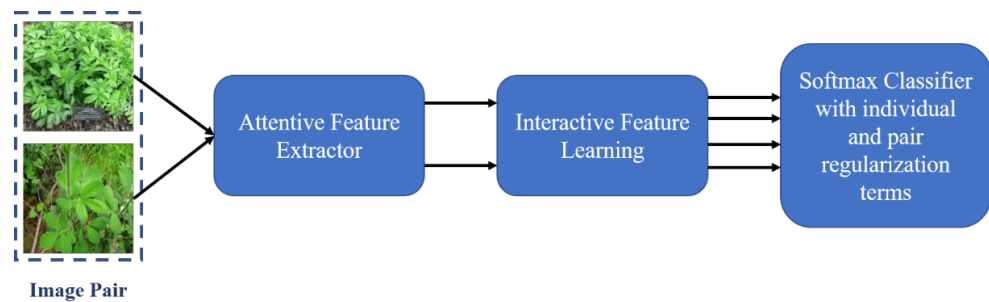


Figure 1. The system framework of AIF-Net.

3.1. Attentive Feature Extraction

Figure 2 shows the process of attentive feature extraction, which is broken down and described in the following two subsections.

3.1.1. Attention Modules in AIF-Net

The attention module employed in AIF-Net is similar to the one in CBAM [18]. Given an intermediate feature map $\mathbf{F} \in \mathbf{R}^{C \times H \times W}$ of a certain convolutional layer, attention module sequentially infers a 1D channel attention map $\mathbf{F}_C \in \mathbf{R}^{C \times 1 \times 1}$ and a 2D spatial attention map $\mathbf{F}_S \in \mathbf{R}^{1 \times H \times W}$. The overall process of the attention module can be summarized in Equations (1) and (2).

$$\mathbf{F}' = \mathbf{F}_C(\mathbf{F}) \otimes \mathbf{F} \quad (1)$$

$$\mathbf{F}'' = \mathbf{F}_S(\mathbf{F}') \otimes \mathbf{F}' \quad (2)$$

where \otimes denotes an elementwise multiplication. The intermediate feature map $\mathbf{F} \in \mathbf{R}^{C \times H \times W}$ is firstly processed by channel attention module. \mathbf{F} is pooled along the spatial dimension by maximum and average operations. The max-pooled features \mathbf{F}_c^{max} and average-pooled features \mathbf{F}_c^{avg} are processed by a shared network, which is composed of a multilayer perceptron (MLP) module with one hidden layer, to produce the channel attention vector $\mathbf{F}_C \in \mathbf{R}^{C \times 1 \times 1}$. The refined feature map \mathbf{F}' is obtained via an elementwise multiplication of \mathbf{F}_C and \mathbf{F} , as shown in Equation (1). The channel-attention-enhanced feature map \mathbf{F}' is then processed by a spatial attention module. Specifically, \mathbf{F}' is pooled along the channel dimension by both maximum and average operations. The max-pooled feature map \mathbf{F}_s^{max} and average-pooled feature map \mathbf{F}_s^{avg} are concatenated along the channel dimension to produce $\mathbf{F}_S \in \mathbf{R}^{2 \times H \times W}$. The \mathbf{F}_S is processed by a convolutional layer

with kernel size seven followed by a sigmoid function. In AIF-Net, the spatial-attention-enhanced feature map \mathbf{F}'' is then plugged into the network to obtain an attentive feature map $\mathbf{F}_{attention}$:

$$\mathbf{F}_{attention} = \mathbf{F} + \mathbf{F}'' \quad (3)$$

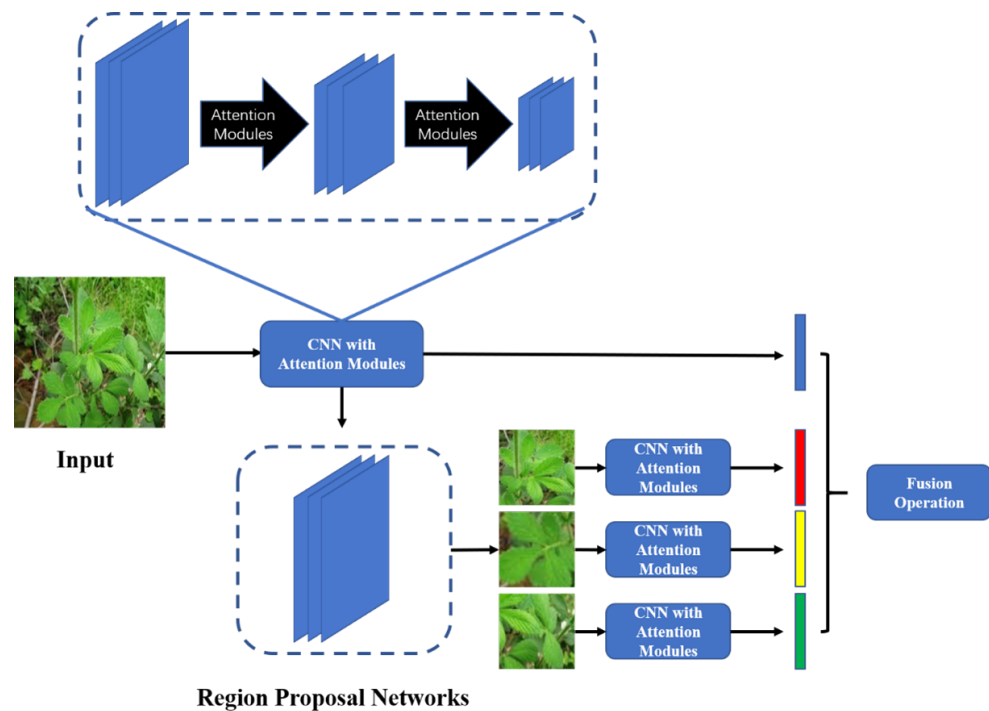


Figure 2. Attentive feature extractor.

3.1.2. Region Proposal Network

Attention modules in AIF-Net can reduce the adverse effects of background noises. Localization of critical regions can further avoid the effects of background noise. An RPN is plugged into the location that follows the last convolutional layer in the backbone network. The output of last convolutional layer is denoted as $\mathbf{F}_D \in \mathbf{R}^{C \times H \times W}$. We use convolutional layers to compute feature hierarchy layer by layer, followed by RELU and max pooling operations. Anchors in each convolutional layer of RPN correspond to regions with different size. For example, anchors in larger feature map correspond to smaller regions and anchors in smaller feature map correspond to larger regions. The convolutional activations are used as the informativeness of anchors. As shown in [7], each anchor is associated with sliding windows with different spatial positions, scales, and aspect ratios. A collection of regions $\{R_1, R_2, \dots, R_N\}$ are produced and each with a score denoting informativeness of the region, which is defined as $I(R_i)$, $i = 1, 2, \dots, N$. These regions are sorted as $I(R'_1) \geq I(R'_2) \geq \dots \geq I(R'_N)$, where N refers to the number of regions. In order to reduce the region redundancy, nonmaximum suppression (NMS) is applied on the regions based on their informativeness.

The top- M informative regions are taken from the sorted list and fed into the backbone network with an independent fully connected layer for complementary features extraction to get confidence as $\{C(R'_1), C(R'_2), \dots, C(R'_M)\}$. Parameters in both RPN and backbone network are optimized to ensure that the lists $\{I(R'_1), I(R'_2), \dots, I(R'_N)\}$ and $\{C(R'_1), C(R'_2), \dots, C(R'_N)\}$ have the same order. When informative regions are localized, the features extracted from the whole image and critical regions are fused through either summation or concatenation. Our empirical results show that concatenation is a better choice for this task, and more details are provided in Section 4.

3.2. Interactive Feature Learning

A pair of images is fed into the attentive feature extractor individually to produce two feature vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{R}^{D'}$. A concatenated vector $[\mathbf{x}_1, \mathbf{x}_2]$, as shown in Figure 3, is fed into a MLP with one hidden layer, which learns a mutual feature vector $\mathbf{x}_M \in \mathbf{R}^{D'}$:

$$\mathbf{x}_M = f_{mlp}([\mathbf{x}_1, \mathbf{x}_2]) \tag{4}$$

where f_{mlp} is the MLP mapping function. This mutual vector is then used to compare with the individual feature vectors \mathbf{x}_1 and \mathbf{x}_2 . Specifically, a gate vector $\mathbf{g}_i \in \mathbf{R}^{D'}$ is generated by

$$\mathbf{g}_i = \sigma(\mathbf{x}_M \odot \mathbf{x}_i), \quad i \in \{1, 2\} \tag{5}$$

where σ refers to a sigmoid function. Gate vectors \mathbf{g}_1 and \mathbf{g}_2 are used to activate the channels of features extracted from each individual image. Then the interactive features are obtained via residual attention as follows.

$$\mathbf{x}_1^{self} = \mathbf{x}_1 + \mathbf{x}_1 \odot \mathbf{g}_1 \tag{6}$$

$$\mathbf{x}_1^{other} = \mathbf{x}_1 + \mathbf{x}_1 \odot \mathbf{g}_2 \tag{7}$$

$$\mathbf{x}_2^{self} = \mathbf{x}_2 + \mathbf{x}_2 \odot \mathbf{g}_2 \tag{8}$$

$$\mathbf{x}_2^{other} = \mathbf{x}_2 + \mathbf{x}_2 \odot \mathbf{g}_1 \tag{9}$$

The outputs of the interactive feature learning module are $\mathbf{x}_1^{self}, \mathbf{x}_1^{other}, \mathbf{x}_2^{self}$, and \mathbf{x}_2^{other} , which are passed to the classification layer for the final prediction.

3.3. Softmax Classifier with Individual and Pair Regularization Terms

In addition to $\mathbf{x}_1^{self}, \mathbf{x}_1^{other}, \mathbf{x}_2^{self}$, and \mathbf{x}_2^{other} , the part features extracted from the informative regions by the attentive features extractor, $\mathbf{f}_i^{part}, i = 1, 2, \dots, M$, are also fed into the softmax classifier. The loss function of AIF-Net is designed as follows:

$$L = L_{ce}^p + L_{ce}^i + \lambda_1 L_{rk}^p + \lambda_2 L_{rk}^i \tag{10}$$

where L_{ce}^p and L_{ce}^i denote the cross-entropy losses of pair and parts, and L_{rk}^p and L_{rk}^i denote the ranking regularization terms of pair and parts with coefficients λ_1 and λ_2 . In particular, L_{ce}^p is given as

$$L_{ce}^p = - \sum_{i \in \{1, 2\}} \sum_{j \in \{self, other\}} \mathbf{y}_i \log(\mathbf{p}_i^j) \tag{11}$$

where $\mathbf{p}_i^j = softmax(\mathbf{W}\mathbf{x}_i^j + \mathbf{b}), i \in \{1, 2\}, j \in \{self, other\}$. Cross-entropy loss of individual parts can be denoted as

$$L_{ce}^i = - \sum_{i \in \{1, 2, \dots, M\}} \mathbf{y}_i \log(\mathbf{p}_i) \tag{12}$$

where $\mathbf{p}_i = softmax(\mathbf{W}\mathbf{f}_i^{part} + \mathbf{b}), i \in \{1, 2, \dots, M\}$. The rank regularization of pair L_{rk}^p can be denoted as

$$L_{rk}^p = \sum_{i \in \{1, 2\}} \max(0, \mathbf{p}_i^{other}(C_i) - \mathbf{p}_i^{self}(C_i) + \epsilon) \tag{13}$$

It encourages \mathbf{x}_i^{self} be more discriminative than \mathbf{x}_i^{other} . The rank regularization term of individual L_{rk}^i can be denoted as

$$L_{rk}^i(I) = \sum_{(i,j): C_i < C_j} f(I_j - I_i) \tag{14}$$

where f is hinge loss function $f(x) = \max(1 - x, 0)$, C is the confidence function that maps the feature vector to its probability being ground-truth class and I is the informativeness of regions.

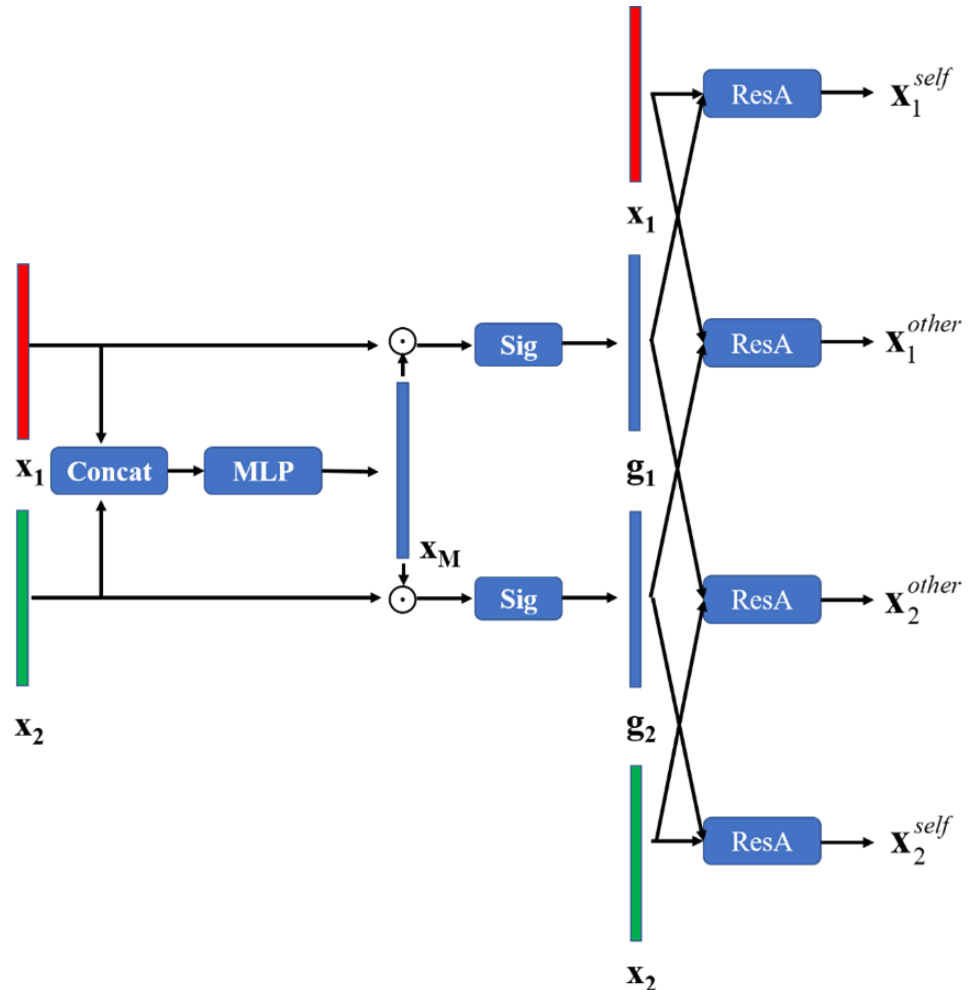


Figure 3. Interactive feature learning.

4. Experiments

4.1. Data Set

VegFru, which is a domain-specific data set, is utilized in the experiments. This data set contains vegetables and fruits of 25 upper-level categories and 292 subcategories. It consists of more than 160,000 images in total and at least 200 images for each subcategory. VegFru can be naturally divided into two subsets, i.e., Veg200 for vegetables and Fru92 for fruits.

4.2. Implementation Details

All of the experiments were implemented using PyTorch and conducted on a cluster of four GTX TITAN Xs. The input images in the training set were resized to 512×512 and randomly cropped to 448×448 . The input images in the test set were resized to 512×512 and center-cropped to 448×448 . We randomly sampled sixteen categories in each batch. For each category, we randomly sampled three images. For each image, we found its most similar image from its own class and from other classes. The backbone network of AIF-Net is chosen as ResNet50. The attention modules introduced in Section 3.1 were plugged in the last ResBlock of ResNet50. The RPN in AIF-Net had three convolutional layers. We tested different numbers of informative regions, including two, three, and four. A standard

SGD with a momentum of 0.9 for optimization and a weight decay of 0.0005 were used in the experiments. The initial learning rate was 0.001, and we adopted the cosine annealing strategy to adjust it. The model was trained in 200 epochs.

4.3. Performance Metric

We report the performance of top-1 and top-5 mean accuracy (Acc) for all experiments. The top-1 and top-5 mean Acc are commonly used in the literature for image classification tasks. Specifically, the top-1 Acc measures how many times the correct label has the highest score or confidence predicted by the classifier, while the top-5 Acc measures how many times the correct label is among the top five scoring classes. Obviously, the top-1 Acc is a more rigorous metric than the top-5 Acc.

4.4. Benchmarks

We choose a set of generic image classification models, including ResNet50, VGG, CBP, HybridNet, and a set of FGVC methods that represent the state of the art (SOTA), including DCL, API-Net and WS-DAN, as the benchmarks for our experiments. The configurations of these models are provided as follows.

- For ResNet50, the initial learning rate was set to 0.001, with an exponential decay of 0.1 after every 30 epochs.
- For VGG, CBP and HybridNet, the results are referenced from [13].
- For DCL, the division number for Region Confusion Mechanism (RCM) was set to 7.
- For API-Net, the experimental settings were the same as ours.
- For WS-DAN, the last convolutional layer was chosen as the feature map. The SGD with momentum of 0.9, a weight decay of 0.00001 were used. The initial learning rate was set to 0.001 with a exponential decay of 0.9 after every 2 epochs.

For DCL, API-Net, WS-DAN, and our AIF-Net, we chose ResNet50 as the backbone for a fair comparison.

4.5. Ablation Study

To investigate the properties of AIF-Net, we evaluate the key design choices on VegFru, Veg, and Fru, respectively.

4.5.1. Fusion Operation for Global and Local Feature Maps

We evaluate two choices of the fusion operation, including concatenation and summation, that can be applied on the global and local feature maps (see Figure 2). As shown in Table 1, the AIF-Net model with a concatenation fusion outperforms the model with a summation fusion in both top-1 and top-5 mean accuracy. A possible reason is that concatenation produces higher dimensional feature vectors with richer informative patterns that could increase the model's capacity, leading to a better generalization ability.

Table 1. An evaluation of different fusion operations.

Data Set	Fusion Operation	Top-1 Acc	Top-5 Acc
Veg200	Concatenation	89.154%	98.045%
	Summation	87.137%	96.724%
Fru92	Concatenation	91.058%	98.809%
	Summation	88.653%	97.913%
VegFru292	Concatenation	90.832%	98.619%
	Summation	89.317%	97.805%

4.5.2. Number of Local Informative Regions

We also evaluate the number of informative local regions selected from the outputs of the RPN. Three different values, including two, three, and four, are tested. Results

in Table 2 demonstrate that picking the top-three local regions to produce fine-grained feature representation achieved the best performance compared to the other two options. It is difficult to justify that why *three* is best. This empirical result can only suggest its superiority on the particular data set under the specific parameter setting.

Table 2. An evaluation of different numbers of informative regions.

Data Set	Number of Regions	Top-1 Acc	Top-5 Acc
Veg200	2	87.132%	97.513%
	3	89.154%	98.045%
	4	88.019%	97.861%
Fru92	2	90.629%	97.917%
	3	91.058%	98.809%
	4	90.681%	98.051%
VegFru292	2	86.795%	97.513%
	3	90.832%	98.619%
	4	88.019%	97.861%

4.6. Overall Performance Comparison

We report a performance comparison in Table 3 and provide our analysis as follows:

- AIF-Net presents the best performance in both top-1 and top-5 Acc on the VegFru292 set, outperforming the second-best, ResNet50, by 2.83% in top-1 and 0.627% in top-5. On Veg200, AIF-Net outperforms ResNet50 in top-1 Acc (89.154% vs. 88.195%) but slightly underperforms ResNet50 in top-5 Acc (98.045% vs. 98.187%). On Fru92, our AIF-Net outperforms the second-best API-Net by 1.14% in top-1 Acc, and also outperforms the second-best ResNet50 by 0.1% in top-5 Acc. It is observed that ResNet50, as a generic deep learning model, can achieve superior performance in this task, demonstrating its potential in FGVC. Additionally, the proposed AIF-Net presents its superior predictive power through interactive feature learning combined with a fusion of global and local attentive feature maps.
- Surprisingly, the SOTA methods (DCL, API-Net, and WS-DAN) that use ResNet50 as a backbone underperform ResNet50 in both top-1 and top-5 Acc on the VegFru292 set. Although API-Net demonstrated superior performance in other data sets [14], its performance in VegFru is slightly worse than its backbone network ResNet50, except on Fru92, where API-Net posts a top-1 Acc of 89.914%, with a 0.59% improvement over ResNet50. The results show that with interactive feature learning alone, the model does not present consistent performance improvement on the VegFru data set.
- The proposed AIF-Net, on the other hand, demonstrates a consistent improvement over both ResNet50 and API-Net, which means that a combination of an attentive feature aggregation and interactive feature learning can effectively push a model to learn subtle and fine-grained patterns from both local and global attentive feature maps, leading to consistent performance boost.

Table 3. A performance comparison of all models. The highest score of each metric is marked in bold-face.

Data Set	Method	Top-1 Acc	Top-5 Acc
Veg200	ResNet50	88.195%	98.187%
	VGG16	78.50%	-
	CBP	81.59%	-
	DCL	85.98%	97.53%
	API-Net	86.953%	97.210%
	WS-DAN	84.17%	96.71%
	AIF-Net	89.154%	98.045%
Fru92	ResNet50	89.323%	98.710%
	VGG16	79.80%	-
	DCL	85.07%	96.17%
	API-Net	89.914%	98.021%
	WS-DAN	87.32%	98.21%
	AIF-Net	91.058%	98.809%
VegFru292	ResNet50	88.002%	97.992%
	VGG16	77.12%	-
	CBP	82.21%	-
	HybridNet	83.51%	-
	DCL	87.13%	97.26%
	API-Net	87.241%	97.711%
	AIF-Net	90.832%	98.619%

5. Discussion

FGVC is becoming an increasingly significant computer vision task that has the potential to be applied in numerous scenarios. Human beings have the capability to quickly learn and accurately recognize fine-grained objects in different subcategories, because we can identify subtle distinguishable patterns in the course of learning. This intuition drives a wide spectrum of studies in the deep learning community. Deep CNNs enable automated feature extraction, while attention mechanisms allow CNNs to learn what and where to focus on, improving the quality of extracted features. Specific to FGVC, prior efforts have explored how to effectively represent fine-grained features through ways such as attentive feature learning, local feature map aggregation, multiscale feature extraction, and pairwise feature interaction. Our work also focuses on this core mission. The proposed AIF-Net aims to generate high-quality fine-grained features by fusing attentive local and global features and interactive feature learning. The novel neural structure is verified through extensive experiments and demonstrates consistent performance improvement in comparison with the SOTA. We also discover that a concatenation fusion operation applied in the feature extractor, along with three top-scoring regions suggested by an RPN, can effectively boost the performance.

Domain-specific FGVC models can be used to build recognition systems that fit various industrial needs. The proposed AIF-Net model can serve as a fine-grained vegetable and fruit classifier to automate applications in domestic cooking and food management. One interesting use case would be building a vegetable/fruit recognition software for educational or training purposes. To become a food/cooking/nutrition professional, one needs to be trained to recognize fine-grained vegetable/fruit subcategories. Such software can be used to generate exercises of different difficulty levels, asking trainees to distinguish vegetable/fruit types. A more intelligent classifier can even evolve with better recognition skills from food experts through active or reinforcement learning, and on the other hand, convey this novel knowledge to trainees or students, creating an efficient learning loop.

This work has the following limitations that also point out our future research directions. First, the biattention module applied in the attentive feature extractor only employs a single MLP to learn the attentions, while a multichannel and multispatial attention mod-

ule [32] can be adopted for further improvement. Second, the attentive feature extractor only considers one image/feature scale, while a multiscale-based feature pyramid can be utilized to encourage the network to extract features in multiple granularities. Third, our study does not take advantage of data augmentation, which could be another performance booster. Since FGVC tasks usually involve many subcategories, and there are not sufficient images for each subcategory for training, adding a data augmentation module to enrich and diversify the training set could be an effective strategy to improve the accuracy of the network. Lastly, due to the introduction of several functional modules, such as the RPN and the attention module, AIF-Net is slower than the major baseline API-Net in both training and inference. Thus, an essential next step is to further optimize the network, making it more lightweight and efficient.

Author Contributions: Conceptualization and methodology, Y.W., Z.X. and L.M.; software, validation, and original draft preparation, Y.W.; review and editing, supervision, Z.X. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data Availability Statement: the VegFru292 data set supporting the conclusions of this article are available at <https://github.com/ustc-vim/vegfru> (accessed on 20 March 2021).

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

References

1. Pu, J.; Jiang, Y.G.; Wang, J.; Xue, X. Which looks like which: Exploring inter-class relationships in fine-grained visual categorization. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 425–440.
2. Zhang, L.; Huang, S.; Liu, W. Intra-class Part Swapping for Fine-Grained Image Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikola, HI, USA, 5–9 January 2021*; pp. 3209–3218.
3. Wei, X.S.; Wu, J.; Cui, Q. Deep learning for fine-grained image analysis: A survey. *arXiv* **2019**, arXiv:1907.03069.
4. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014*; MIT Press: Montreal, QC, Canada, 2015; Volume 2.
5. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*.
6. Sun, M.; Yuan, Y.; Zhou, F.; Ding, E. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. In *Proceedings of the Computer Vision—ECCV 2018, Salt Lake City, UT, USA, 18–23 June 2018*.
7. Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; Wang, L. Learning to Navigate for Fine-Grained Classification. In *Proceedings of the Computer Vision—ECCV 2018, Salt Lake City, UT, USA, 18–23 June 2018*.
8. Lin, T.; RoyChowdhury, A.; Maji, S. Bilinear CNN Models for Fine-Grained Visual Recognition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015*.
9. Gao, Y.; Beijbom, O.; Zhang, N.; Darrell, T. Compact Bilinear Pooling. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*.
10. Kong, S.; Fowlkes, C. Low-Rank Bilinear Pooling for Fine-Grained Classification. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*.
11. Lin, T.; RoyChowdhury, A.; Maji, S. Bilinear Convolutional Neural Networks for Fine-Grained Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1309–1322. [[CrossRef](#)] [[PubMed](#)]
12. Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; You, X. Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition. In *Proceedings of the Computer Vision—ECCV 2018, Salt Lake City, UT, USA, 18–23 June 2018*.
13. Hou, S.; Feng, Y.; Wang, Z. VegFru: A Domain-Specific Dataset for Fine-Grained Visual Categorization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*.
14. Zhuang, P.; Wang, Y.; Qiao, Y. Learning Attentive Pairwise Interaction for Fine-Grained Classification. *arXiv* **2020**, arXiv:2002.10191.
15. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 1–8 December 2013*; pp. 554–561.

16. Branson, S.; Van Horn, G.; Belongie, S.; Perona, P. Bird species categorization using pose normalized deep convolutional nets. *arXiv* **2014**, arXiv:1406.2952.
17. Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv* **2013**, arXiv:1306.5151.
18. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV 2018, Salt Lake City, UT, USA, 18–23 June 2018.
19. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
20. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [[CrossRef](#)] [[PubMed](#)]
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
23. Chen, Y.; Bai, Y.; Zhang, W.; Mei, T. Destruction and Construction Learning for Fine-Grained Image Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
24. Hu, T.; Qi, H. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification. *arXiv* **2019**, arXiv:1901.09891.
25. Gavves, E.; Fernando, B.; Snoek, C.G.; Smeulders, A.W.; Tuytelaars, T. Fine-grained categorization by alignments. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1713–1720.
26. Zheng, H.; Fu, J.; Mei, T.; Luo, J. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
27. Kim, J.-H.; On, K.W.; Lim, W.; Kim, J.; Ha, J.; Zhang, B.-T. Hadamard Product for Low-rank Bilinear Pooling. *arXiv* **2016**, arXiv:1610.04325
28. 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.
29. Zhang, C.; Yao, Y.; Liu, H.; Xie, G.S.; Shu, X.; Zhou, T.; Zhang, Z.; Shen, F.; Tang, Z. Web-supervised network with softly update-drop training for fine-grained visual classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12781–12788.
30. Zhang, C.; Yao, Y.; Zhang, J.; Chen, J.; Huang, P.; Zhang, J.; Tang, Z. Web-supervised network for fine-grained visual classification. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
31. Sun, X.; Chen, L.; Yang, J. Learning from web data using adversarial discriminative neural networks for fine-grained classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 273–280.
32. Yang, B.; Xiao, Z. A Multi-Channel and Multi-Spatial Attention Convolutional Neural Network for Prostate Cancer ISUP Grading. *Appl. Sci.* **2021**, *11*, 4321. [[CrossRef](#)]