Trusted Fine-Grained Image Classification through Hierarchical Evidence Fusion

Zhikang Xu[1], Xiaodong Yue[1, 2, 3*], Ying Lv[1], Wei Liu[4], Zihao Li[1]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2] Artificial Intelligence Institute of Shanghai University, Shanghai, China
[3] VLN Lab, NAVI MedTech Co., Ltd. Shanghai, China
[4] College of Electronics and Information Engineering, Tongji University, Shanghai, China
{xuzhikangnba, yswantfly, lvying, zihao}@shu.edu.cn, ldachuan@outlook.com

Abstract

Fine-Grained Image Classification (FGIC) aims to classify images into specific subordinate classes of a superclass. Due to insufficient training data and confusing data samples, FGIC may produce uncertain classification results that are untrusted for data applications. In fact, FGIC can be viewed as a hierarchical classification process and the multilayer information facilitates to reduce uncertainty and improve the reliability of FGIC. In this paper, we adopt the evidence theory to measure uncertainty and confidence in hierarchical classification process and propose a trusted FGIC method through fusing multilayer classification evidence. Comparing with traditional approaches, the trusted FGIC method not only generates accurate classification results but also reduces the uncertainty of fine-grained classification. Specifically, we construct an evidence extractor at each classification layer and extract multilayer (multi-grained) evidence in hierarchical classification. To fuse the extracted multi-grained evidence from coarse to fine, we formulate evidence fusion with the Dirichlet hyper probability distribution and thereby hierarchically decompose the evidence of coarse-grained classes into fine-grained classes to enhance the performance of FGIC and reduce uncertainty. The ablation experiments validate that the hierarchical evidence fusion can improve the precision and also reduce the uncertainty of fine-grained classification. Comparison with the state-of-the-art FGIC methods shows that the proposed method achieves competitive performance.

Introduction

Fine-grained image classification (FGIC) aims to classify samples into correct categories with more specific concepts, e.g., the models of car and the species of bird. With the rapid development of deep neural networks, the performance of FGIC has made a significant progress (Zhao et al. 2017; Wei et al. 2021). In recent years, FGIC has been widely applied in some fields, such as image retrieval (Wei et al. 2017; Bhunia et al. 2021), personalized recommendation (Bai et al. 2020; Wei et al. 2019; Liu et al. 2016), and so on.

Discriminative information is exploited to improve the fine-grained classification, such as the discriminative subregions and features interactively extracted from images (Sun et al. 2018; Gao et al. 2020; Zhuang, Wang, and Qiao 2020).

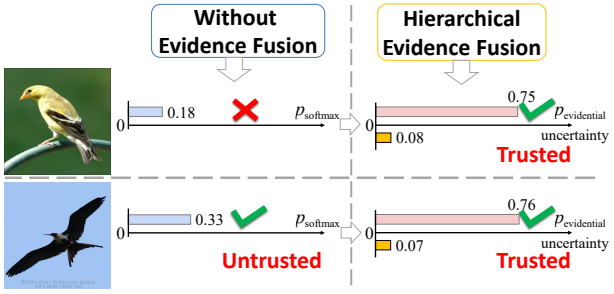

Figure 1: Comparison between the traditional FGIC method and the trusted FGIC method.

In addition, the relationship between classes and the correlation among class attributes are incorporated into FGIC to enhance the refined classification of similar images (Zhou and Lin 2016; Zhang et al. 2016). However, due to the insufficient training data and confusing samples (Van Horn et al. 2015), extant FGIC methods may produce uncertain classification results that are untrusted for data applications.

Trustworthy machine learning has received a lot of attention in recent years. Sensoy et al. (Sensoy, Kaplan, and Kandemir 2018) combined evidence theory and deep neural networks to implement the trusted deep learning method, where evidence theory is used to quantify the uncertainty of data samples and predictions with high uncertainty are seen as untrustworthy. Some researchers further introduced evidence theory into different fields and achieved trusted learning methods, e.g., trusted multi-view classification(Liu et al. 2022) and trusted long-tailed classification (Li et al. 2021).

In fact, FGIC can be viewed as a hierarchical classification process. In the class hierarchy (e.g., label tree), label nodes near the root node represent the coarser-grained concepts, while label nodes away from the root node represent the finer-grained concepts. Motivated by this, we adopt the evidence theory to measure the uncertainty and confidence in hierarchical classification process and propose a trusted FGIC method through fusing multilayer classification evidence. At each layer of classification granularity, we extract evidence that data samples belong to different classes. Evidence extracted from all layers is considered as multilayer

*Corresponding Author.

(or multi-grained) evidence. Multi-grained evidence facilitates to reduce uncertainty and improve the reliability of FGIC. To fuse the multi-grained evidence from coarse to fine, we formulate evidence fusion with the Dirichlet hyper probability distribution and thereby hierarchically decompose the evidence of coarse-grained classes into the corresponding fine-grained classes. Besides the algorithm implementation, we also theoretically demonstrate that the uncertainty of fine-grained classification can be reduced via the evidence fusion. Comparing with the traditional approaches, the proposed trusted FGIC method not only generates accurate classification results but also reduces the uncertainty of fine-grained classification. Figure 1 illustrates the comparison between the traditional FGIC method and our trusted method in the fine-grained classification of birds. In the first example, the traditional method without evidence fusion outputs the low probability for the ground-truth class via softmax and leads to the misclassification. Through evidence fusion, the trusted method produces the high probability and rectifies the prediction error. For the second case, the traditional method correctly recognizes the image but the low probability indicates the high uncertainty and the classification is untrusted. In contrast, our method produces the trusted classification result.

In addition, the research works on trustworthy machine learning introduced above aim to deal with prediction uncertainties in different kinds of deep learning models. For FGIC tasks, to the best of our knowledge, our method is the first attempt to formulate and process uncertainty in coarse-to-fine classification and implement trusted fine-grained classification. Our contributions are summarized as follows.

- Propose a trusted fine-grained image classification method through fusing hierarchical classification evidence. We measure the uncertainty in the hierarchical refined classification based on evidence theory and improve the prediction reliability through fusing the evidence of multilayer classification.
- Formulate evidence fusion and decomposition in hierarchical fine-grained classification with the Dirichlet hyper probability distribution and prove that evidence fusion of multilayer classification can reduce uncertainty.

Related Work

FGIC Based on Hierarchical Classification

Due to the large intra-class variance and small inter-class variance, researchers have tried to incorporate more information except class labels to improve the performance of FGIC. Some researchers exploited discriminative information from images, e.g., to find subregions with parts of object that are discriminative (Zhang et al. 2019), or to perform feature interaction to improve feature discrimination (Zhao et al. 2021; Gao et al. 2020; Han et al. 2018).

In addition, fine-grained class labels can be merged into hierarchies based on taxonomies. Exploiting the relationship between class labels of different granularities is another way to improve performance. Existing methods in this paradigm can be divided into three types, (i) hierarchical classification learning based methods, (ii) multi-label learning based methods and (iii) metric learning based methods. Hierarchical classification based methods is to treat FGIC as the process of hierarchical classification (Xie et al. 2015; Wang et al. 2015; Chen et al. 2018; Chang et al. 2021). Wang et al. proposed Multiple Granularity Descriptors (MGD) for FGIC (Wang et al. 2015). MGD constructed the classifier at each classification layer separately. Features extracted from multiple classifiers are concatenated for fine-grained prediction. However, the relationship between classification layers is not exploited in MGD. Chen et al. (Chen et al. 2018) used the Kullback-Leibler divergence to measure the output of two classification layers to guide the training process of fine-grained classification. The method based on multi-label learning directly models the joint distribution of multiple layers of labels. Zhou and Lin (Zhou and Lin 2016) used class labels from the coarse classification layer to augment the last fully connected layer of neural networks, enabling the neural network to output multi-label predictions, and defined a loss objective to optimize the performance of multi-label classification. Metric learning based methods focus on using hierarchical information to optimize the feature representation. Zhang et al. (Zhang et al. 2016) defined the triplet loss on feature space to make the extracted features belonging to the same coarse-grained class closer than the features belonging to different coarse-grained classes. However, the objective is not specific to classification and the effectiveness of the resulting method may be suboptimal.

Existing hierarchical classification based FGIC methods mainly utilize hierarchical information in the training stage, but do not utilize it well in the testing stage. In addition, the uncertainty and trustworthiness of FGIC are not considered.

Classification with Evidence Theory

The Dempster-Shafer Evidence Theory (DST) is a generalization of the Bayesian theory to subjective probabilities (Dempster 1968b), and is widely applied in reasoning with uncertainty (Denoeux 1997; Liu, Pan, and Dezert 2013) and opinion fusion (Li et al. 2018; Si et al. 2014). DST represents the uncertainty in decision making by mass function. Combining DST with classification methods have implemented a variety of uncertainty inference methods, e.g., evidential k-nearest neighbors (Denoeux 2008), evidential logistic regression (Xu, Davoine, and Denœux 2015), and so on (Quost, Denoeux, and Li 2017). Sensoy et al. (Sensoy, Kaplan, and Kandemir 2018) extended DST by combining it with deep learning and proposed Evidential Neural Network (ENN) to measure the uncertainty of deep neural networks and perform uncertainty classification. ENN achieved end-to-end training without additional costs, which facilitates deep learning to perform uncertainty reasoning. In recent years, some researchers have introduced ENN into different fields to achieve trusted deep learning methods (Liu et al. 2022; Li et al. 2021; Han et al. 2021).

Method

In this section, we introduce the proposed trusted FGIC with hierarchical evidence fusion. Figure 2 presents the detailed framework. We first formulate FGIC as the process of hierarchical classification and extract multilayer (multi-grained)

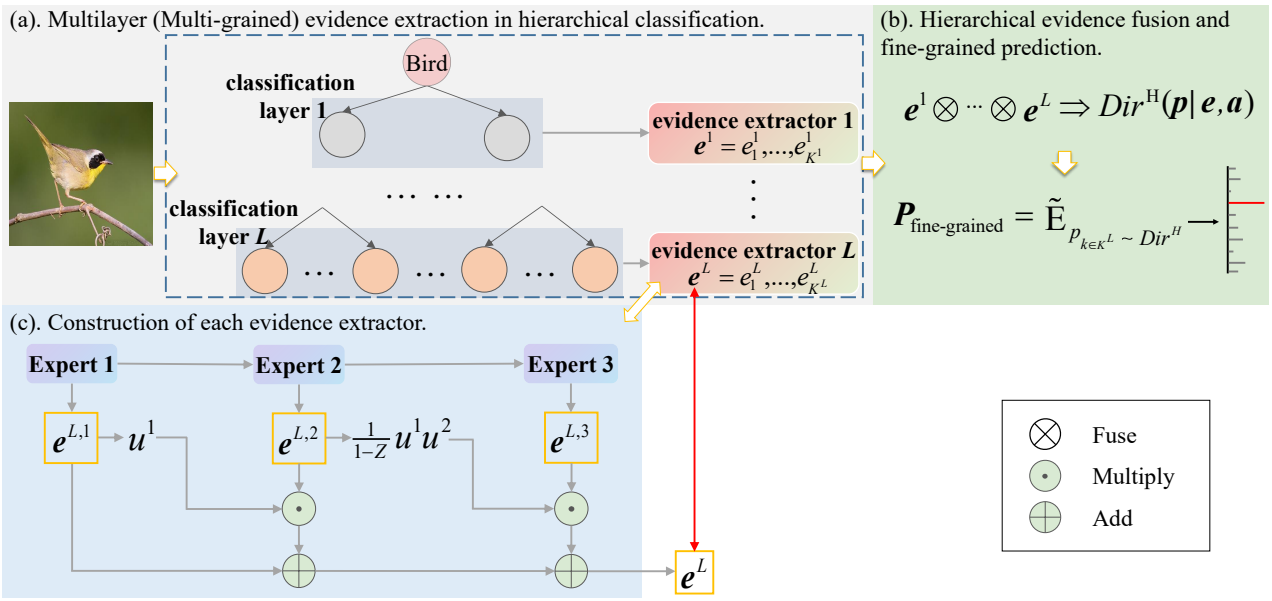

Figure 2: Overall framework of the trusted fine-grained image classification, which is formulated as a process of hierarchical classification and evidence fusion.

evidence in hierarchical classification (Figure 2(a)). At each classification layer, we construct an evidence extractor to extract evidence belonging to different classes at the corresponding classification granularity. Each evidence extractor is designed as three cascaded sub-extractors named expert 1, expert 2 and expert 3 (Figure 2(c)). Evidence extracted from the three sub-extractors is combined through weighted summation, and the weight of each sub-extractor is calculated based on the uncertainty of the previous sub-extractors. After that, in Figure 2(b), evidence extracted from all layers is fused by the Dirichlet hyper probability distribution (Dir HPDF). The probabilities of fine-grained classes are obtained by calculating the projected expectation of Dir HPDF, where evidence of coarse-grained classes is hierarchically decomposed into the corresponding fine-grained classes.

Multi-Grained Evidence Extraction in Hierarchical Classification

Dempster-Shafer Evidence Theory is an effective tool to measure the uncertainty of decision making which uses belief mass to represent the confidence and uncertainty about the opinion. Subjective Logic (SL) (Jsang 2016) extended the evidence theory to construct a mapping between belief mass and probability through the Dirichlet distribution

$$Dir(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{1}{\mathcal{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \tag{1}$$

where $p_k \in [0, 1]$ is the probability belonging to class k. α_k is the strength parameter, denoted as $\alpha_k = e_k + a_k W$. e_k represents the evidence supporting class k, derived from observations, while $a_k W$ is the prior preference for class k. (e.g., $W = K$ and a_k is set to $1/K$). By calculating the expectation of the Dirchlet distribution, we obtain the expected probability E_{p_k} belonging to class k

$$\mathrm{E}_{p_k} = \frac{\alpha_k}{S}, \tag{2}$$

where $S = \sum_{k=1}^{K} \alpha_k = \sum_{k=1}^{K} e_k + a_k W$. The belief mass belonging to class k and uncertainty are calculated by

$$b_k = \frac{e_k}{S}, u = \frac{W}{S}. \tag{3}$$

According to Eq. (3), we have $\mathrm{E}_{p_k} = b_k + a_k u$. The purpose of such design is to decouple the uncertainty u from probability so that uncertainty can be measured explicitly.

Based on the mapping between probability and belief mass, we construct an evidence extractor in each layer of class label hierarchy to extract evidence, and formalize it into Dirchlet distribution

$$Dir(\boldsymbol{p}^l|\boldsymbol{e}^l, \boldsymbol{a}^l) = \frac{1}{\mathcal{B}(\boldsymbol{e}^l, \boldsymbol{a}^l)} \prod_{k=1}^{K^l} p_k^{e_k + a_k W - 1}, \tag{4}$$

where K^l is the number of classes in layer l. Consequently, evidence extracted from multiple classification layers forms multi-grained evidence. Inspired by (Sensoy, Kaplan, and Kandemir 2018), we can use neural networks as evidence extractors. By adding an activation layer after the last fully connected layer, the non-negative outputs of neural networks are considered as the observed evidence $\boldsymbol{e}$. To improve the classification precision and reduce the uncertainty, the evidence extractor of each layer is designed as three cascaded sub-extractors (named expert 1,2 and 3). As shown in Figure 3, we input an image into expert 1 to extract evidence belonging to different classes. Then we crop a discriminative subregion by using AOLM (Zhang et al. 2021) and input it into expert 2. From expert 2 to expert 3, multiple subregions

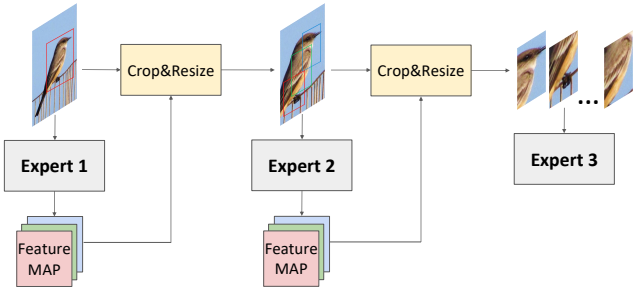

Figure 3: Structure of the evidence extractor.

with high response are further cropped out by using APPM (Zhang et al. 2021).

In addition, considering the differences in learning difficulty of data samples, we utilize uncertainty to measure the difficulty and follow (Li et al. 2021) to dynamically assign the experts for different samples during the training stage. Assuming that the uncertainty u^1, u^2 and u^3 are obtained from three experts respectively, if the uncertainty u^1 is smaller than a threshold ε, the expert 2 and 3 will not be activated. Otherwise, expert 2 will be activated, and then u^1 and u^2 will be fused by using Dempster's rule (Dempster 1968a). Given the belief masses $\boldsymbol{b}^1$, $\boldsymbol{b}^2$ of expert 1 and 2 for all classes in layer l calculated by Eq. (3), we have

$$u^{\text{fuse}} = \frac{1}{1-Z}\, u^1 \bullet u^2, \tag{5}$$

where $Z = \sum_{i \neq j} b_i^1 b_j^2$ is the conflict coefficient of expert 1 and expert 2. Based on Eq. (5), expert 3 will be activated only when the fused uncertainty u^{fuse} is greater than a threshold ε. Because the uncertainty indicates the learning difficulty, the activated expert 2 and 3 will focus more on discriminative subregions and features of the hard samples. In addition, according to the Dempster's rule, u^{fuse} is less than either of u^1 or u^2 and is related to the conflict between expert 1 and 2, i.e., the serious conflict will lead to high u^{fuse}.

In the following section, the uncertainties will also be used for evidence fusion during the test stage.

Hierarchical Evidence Fusion

Based on the extracted evidence and uncertainty introduced above, we formulate the evidence fusion in the hierarchical fine-grained classification. First, we define the evidence fusion in each classification layer below.

Evidence Fusion in Each Layer. The evidence extracted from all experts in the l-th classification layer is integrated through uncertainty weighted summation,

$$\boldsymbol{e}^l = \frac{\boldsymbol{e}^{l,1} + u^{l,1} \bullet \boldsymbol{e}^{l,2} + u^{l,\text{fuse}} \bullet \boldsymbol{e}^{l,3}}{1 + u^{l,1} + u^{l,\text{fuse}}}. \tag{6}$$

Because samples that are difficult to classify have high uncertainty, the expert 2 and 3 with large weights will contribute more in evidence fusion.

Next we extend the evidence fusion from single layer to multiple layers.

Evidence Fusion in Multiple Layers. Based on SL (Jsang 2016), assuming that the classes of different layers are independent of each other, we can map the multi-grained evidence extracted from different layers into a Dir HPDF.

$$Dir^{\text{H}}(\boldsymbol{p}|\boldsymbol{e},\boldsymbol{a}) = \frac{1}{\mathcal{B}(\boldsymbol{e},\boldsymbol{a})} \prod_{k=1}^{K^1 \cup \ldots \cup K^L} p_k^{e_k + a_k W - 1}, \tag{7}$$

where $\boldsymbol{p} = \boldsymbol{p}^1 \cup \boldsymbol{p}^2 \cup \ldots \cup \boldsymbol{p}^L$ and $\boldsymbol{e} = \boldsymbol{e}^1 \cup \boldsymbol{e}^2 \cup \ldots \cup \boldsymbol{e}^L$, and L is the total number of layers in the label hierarchy. Evidence fusion of multiple classification layers can reduce the uncertainty of hierarchical fine-grained classification and achieve the trusted FGIC.

Trusted FGIC with Hierarchical Evidence Fusion

Based on the evidence fusion in multiple layers, we use the generalized expectation formula of the Dirichlet distribution (Jsang 2016) to calculate the expected probabilities of each fine-grained class, where the evidence of coarse-grained classes is hierarchically decomposed into the corresponding fine-grained class.

$$\begin{aligned}
\widetilde{\mathrm{E}}_{p_k} &= \frac{\sum\limits_{k' \in K} \tilde{a}_{k|k'} e_{k'} + a_k W}{W + \sum\limits_{k' \in K} e_{k'}} \qquad (\forall k \in K^L) \\
&= \underbrace{\frac{e_k}{W + \sum\limits_{k' \in K} e_{k'}}}_{b_k^S} + \underbrace{\frac{\sum\limits_{k' \in K, k' \neq k} \tilde{a}_{k|k'} e_{k'}}{W + \sum\limits_{k' \in K} e_{k'}}}_{b_k^V} + \underbrace{\frac{a_k W}{W + \sum\limits_{k' \in K} e_{k'}}}_{u_k^F} \\
&= b_k^S + b_k^V + u_k^F,
\end{aligned} \tag{8}$$

where $K = K^1 \cup \ldots \cup K^L$. $\tilde{a}_{k|k'}$ is the relative base rate denoted as

$$\tilde{a}_{k|k'} = \begin{cases} \frac{1}{C_{k'}} & \text{if } k' \text{is the superclass of } k, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

$C_{k'}$ is the number of fine-grained class labels contained in k'. Eq. (8) satisfies $\sum_{k \in K^L} \widetilde{\mathrm{E}}_{p_k} = 1$. As can be seen in Eq. (8), $\widetilde{\mathrm{E}}_{p_k}$ is decomposed into three terms. b_k^S is the sharp belief mass representing the degree of direct support for the fine-grained class k. b_k^V is the vague belief mass where the evidence supporting class k is the weighted summation of the evidence of all its superclasses, by which we hierarchically decompose the evidence of coarse-grained classes into corresponding fine-grained classes to enhance the performance of fine-grained classification. u_k^F is the focal uncertainty transformed from overall uncertainty based on the ratio a_k. In addition, we can guarantee that the overall uncertainty of fine-grained classification can be reduce when the evidence of coarse-grained classification are fused.

Proposition 1. *The evidence fusion in hierarchical classification layers will reduce the uncertainty of fine-grained classification.*

Proof. According to Eq. (3), the uncertainty u can be obtained by

$$u = \frac{W}{S} = \frac{W}{\sum_{k \in K^L} e_k^L + W},$$

After fusing the coarse-grained evidence of classification layer l, the uncertainty becomes

$$u = \frac{W}{\sum_{k \in K^L} e_k^L + W + \sum_{k \in K^l} e_k^l}.$$

It is obvious that the uncertainty is reduced after fusing the evidence from coarse-grained layer l. □

For the model training, the loss function for optimizing the evidence extractor of each layer l is defined as the sum of the losses of all experts.

$$\mathcal{L}^l = \mathcal{L}^{l,1} + \mathcal{L}^{l,2} + \mathcal{L}^{l,3}, \tag{10}$$

and for each expert i, we have

$$\mathcal{L}^{l,i} = \mathcal{L}^l_{CE} + \lambda_t \mathcal{L}^l_{KL}. \tag{11}$$

Where $\lambda_t = \min\{1.0, t/50\}$ is the annealing coefficient and t denotes epoch. The first term $\mathcal{L}_{CE}$ is the Bayes risk of the cross-entropy loss with respect to Eq. (4) (Sensoy, Kaplan, and Kandemir 2018), denoted as

$$\begin{aligned} \mathcal{L}^l_{CE} &= \int \left[\sum_{k=1}^{K^l} -y_k \log p_k \right] \frac{1}{\mathcal{B}(\boldsymbol{\alpha}^l)} \prod_{k=1}^{K^l} p_k^{\alpha_k - 1} dp \\ &= \sum_{k=1}^{K^l} y_k \left[\psi(S) - \psi(\alpha_k) \right], \end{aligned} \tag{12}$$

where $\psi(\cdot)$ is the *digamma* function and $S = \sum_{k=1}^{K^l} \alpha_k$. In addition, we introduce Kullback-Leibler divergence term to reduce the evidence of false predictions.

$$\begin{aligned} \mathcal{L}^l_{KL} &= KL[Dir(\boldsymbol{p}^l | \tilde{\boldsymbol{\alpha}}^l) || Dir(\boldsymbol{p}^l | \mathbf{1})] \\ &= \log \frac{\Gamma(\sum_{k=1}^{K^l} \tilde{\alpha}_k)}{\Gamma(K^l) \prod_{k=1}^{K^l} \Gamma(\tilde{\alpha}_k)} + \sum_{k=1}^{K^l} (\tilde{\alpha}_k - 1)[\psi(\tilde{\alpha}_k) - \psi(\sum_{k=1}^{K^l} \tilde{\alpha}_k)], \end{aligned} \tag{13}$$

where $\Gamma(\cdot)$ is the *gamma* function, and $\tilde{\boldsymbol{\alpha}}^l = \boldsymbol{y} + (\mathbf{1} - \boldsymbol{y}) \odot \boldsymbol{\alpha}^l$ is the adjusted Dirichlet parameters which remove the predicted evidence belonging to the ground-truth class. Then the evidence extractor of each classification layer is trained individually by using Eq. (10).

Algorithm 1 presents the entire workflow of the trusted fine-grained image classification.

Algorithm 1: Trusted fine-grained image classification

Input: Depth L of label hierarchy, threshold ε.

Training :

1. **for each** $l \in L$ **do**
2. Initialize the neural network as evidence extractor of layer l;
3. Train evidence extractor of layer l using Eq. (10);
4. **end for**

Testing :

1. **for each** $l \in L$ **do**
2. Input test images to evidence extractor of layer l;
3. Extract evidence from multiple experts;
4. fuse evidence of all experts by Eq. (6);
5. **end for**
6. fuse evidence of all layers into a Dir HPDF, Eq. (7);
7. Calculate the projected probabilities of fine-grained classes by Eq. (8), and output the prediction.

Experiments

In this section, we perform experiments on three FGIC benchmark datasets, CUB-200-2011 (CUB), FGVC-Aircraft (AIR) and Stanford Cars (CAR). For CUB, we follow the setting in (Chen et al. 2018) to organize the label hierarchy of bird to 200 species, 122 genera, 37 families, and 13 orders. For AIR and CAR, we follow the setting in (Chang et al. 2021) to organize the label hierarchy of airplane to 100 models, 70 families and 30 makers, and organize the label hierarchy of car to 196 car makers and 9 car types. For each dataset, we use the bottom two layers in the class label hierarchy to implement the evidence fusion for fine-grained classification. The experiments consist of three parts. The first one is the ablation experiment to verify the improvements brought by the hierarchical evidence fusion for fine-grained classification. The second experiment is conducted to validate the superiority of the proposed method through comparing with other state-of-the-art FGIC methods. In the final experiment, we provide representative cases to interpret how the proposed method enhances the trustworthiness of fine-grained classification.

Implementation Details. We use ResNet50 (He et al. 2016) as the backbone network and add an activation layer on top of the last fully connected layer as evidence extractor to extract non-negative evidence. To reduce complexity, the three cascade experts of each evidence extractor share parameters. The hyperparameter ε is set to 0.5 for CUB and AIR. Since the CAR dataset has only 9 car types, we do not use ε to reduce experts when training the evidence extractor for car types. For fair comparison, we use only regular data augmentation, enabling the fine-tuned ResNet50 to achieve 84.6% accuracy on CUB. We use Stochastic Gradient Descent (SGD) with momentum=0.9 as optimizer. The learning rate is 0.001 and multiplied by 0.1 after 15, 30 and 45 epochs. The batch size is set to 6.

Ablation Study

The ablation experiments consist of two parts. **(I)** The First one is to verify the effectiveness of the hierarchical evidence fusion in improving the performance of FGIC. **(II)** The second ablation experiment is conducted to validate the effectiveness of uncertainty-based multi-expert combination in enhancing the performance of evidence extractors. In the following experimental results, we use **R** and **NR** to denote whether to reduce the number of experts in the training stage. In the testing stage, we denote the fusion of hierarchical classification evidence as **HE**, and denote the use of multiple experts in each evidence extractor as **ME**. We use **USUM** to denote the combination of experts by Eq. (6), and use **SUM** to denote the combination of experts by summing and averaging.

To validate the effectiveness of hierarchical evidence fusion, we conduct experiments on CUB. From Table 1(b) to (d) and (c) to (e), after fusing hierarchical evidence, the classification accuracies are increased from 88.6% to 89.3% and

Method	Accuracy(%)
(a) FT-ResNet50	84.6
(b) ResNet + ME + NR + SUM	88.6
(c) ResNet + ME + R + USUM	89.3
(d) ResNet + ME + NR + SUM + HE	89.3
(e) ResNet + ME + R + USUM + HE	**89.9**

Table 1: Ablation experiments on CUB to validate the effectiveness of hierarchical evidence fusion, and uncertainty-based multi-expert ensemble of evid ence extractor.

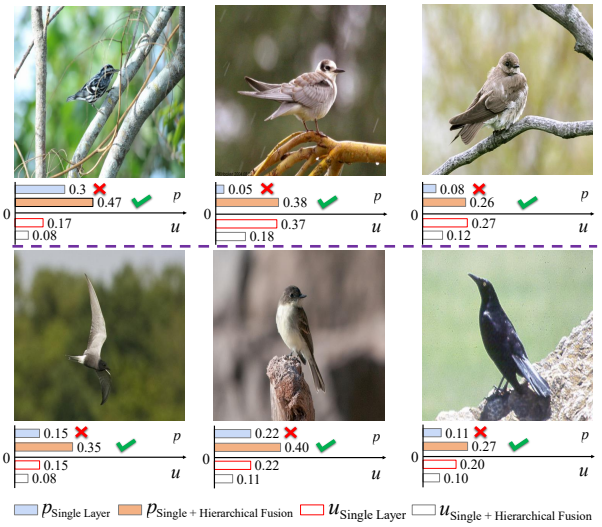

Figure 4: Six representative cases illustrate that fusing hierarchical classification evidence can help improve the performance of FGIC and reduce the uncertainty.

from 89.3% to 89.9% respectively. These results clearly validate that fusing hierarchical evidence can enhance the performance of FGIC. Moreover, we provide six representative cases to illustrate the effectiveness of hierarchical evidence fusion. In Figure 4, 'Single Layer' corresponds to Table 1(c) and 'Single + Hierarchical Fusion' corresponds to Table 1(e). As shown in Figure 4, fusing the hierarchical evidence corrects the wrong prediction of using single fine-grained classifier, and reduces the uncertainty. Similar conclusions can be observed from other examples.

Moreover, as shown in Table 1 from (b) to (c) and (d) to (e), using uncertainty-based multi-expert fusion improves the classification performance. In addition, we provide six representative cases to further illustrate how it works. In Figure 5, (1) to (3) are three different types of hard samples, i.e., confusing appearance, small object and background noise. (4) to (6) are three easy samples with clear background or significantly discriminative features. According to Eq. (5) and Eq. (6), the lower the uncertainties of experts 1 and 2, the smaller the weight values of experts 2 and 3 in the multi-expert fusion. For easy samples, expert 1 has high confidence and outputs low uncertainty values. Then the multi-expert fusion will be contributed mainly by expert 1. In contrast, for hard samples, both experts 1 and 2 have low confi-

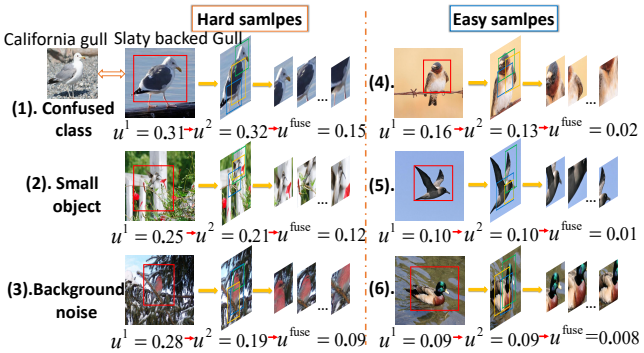

Figure 5: Two types of examples for illustrating the effectiveness of uncertainty-based multi-expert fusion. (1) to (3) are three subtypes of hard samples, and (4) to (6) are three easy samples. Compared with easy samples, experts 1 and 2 output higher uncertainty values for hard samples. The high uncertainties of experts 1 and 2 correspond to the large weight values of experts 2 and 3 in the fusion.

dence and output high uncertainty values. Then experts 2 and 3 will be assigned large weight values in the fusion. Thus the trustworthiness and accuracy for hard samples can be improved by the contribution of more experts.

Comparison with State-of-the-Art Methods

For a fair comparison, FGIC methods using ResNet50 as the backbone are compared, including two methods of incorporating class label hierarchy.

CUB-200-2011. As shown in Table 2, our proposed method achieves state-of-the-art performance on CUB that outperforms the GHORD and PMG by 0.3%, and outperforms other methods by at least 0.7%. In addition, our method significantly outperforms two state-of-the-art methods of combining label hierarchy by a large margin (1.8% for HSE and 3.1% for MGDR). These two methods use label hierarchy as additional supervised information to guide the learning of fine-grained classifier during the training stage. However, during the testing stage, the label hierarchy is not well exploited. As a result, the performance of FGIC is limited. In contrast, our method can fuse the label hierarchy during the testing stage and achieves better performance. These results validate that the hierarchical classification information should be better motivated during the testing stage.

FGVC-Aircraft. As shown in Table 2, our method achieves state-of-the-art performance on AIR that outperforms the second best method by 0.5%, and outperforms other methods by at least 1%. These results validate that fusing the evidence of classifying aircraft families (e.g., 'Boeing 737', 'A330') can significantly improve the performance of classifying aircraft models (e.g., 'Boeing 737-76J', 'A330-300'). Similar to CUB, our method outperforms MGDR by 2%, which verifies the effectiveness of fusing hierarchical evidence during the testing phase.

Stanford Cars. As shown in Table 2, our method achieves competitive performance on CAR that performs inferiorly than GHORD and PMG by 0.2%. The reason is that evi-

Method	CUB	AIR	CAR
CIN (Gao et al. 2020)	87.5	92.8	94.5
Cross-X (Luo et al. 2019)	87.7	92.6	94.6
DCL (Chen et al. 2019)	87.8	93.0	94.5
API-Net (Zhuang et al. 2020)	87.7	93.0	94.8
ACNet (Ji et al. 2020)	88.1	92.4	94.6
GCL (Wang et al. 2020b)	88.3	93.2	94.0
S3N (Ding et al. 2019)	88.5	92.8	94.7
FDL (Liu et al. 2020)	88.6	93.4	94.3
DF-GMM (Wang et al. 2020c)	88.8	93.8	94.8
CSC-Net (Wang et al. 2020a)	89.2	93.8	94.9
PMG (Du et al. 2020)	89.6	93.4	**95.1**
GHORD (Zhao et al. 2021)	89.6	94.3	**95.1**
MGDR* (Chang et al. 2021)	86.8	92.8	94.3
HSE* (Chen et al. 2018)	88.1	-	-
Ours	**89.9**	**94.8**	94.9

Table 2: Comparison with other state-of-the-art methods on CUB, AIR and CAR. '$*$' denotes the methods with class label hierarchy.

dence extracted from coarse-grained classification layers is used to improve the confidence and reduce uncertainty of fine-grained classification. However, when the class hierarchy is too rough to reveal the correlation between classes from coarse to fine, fusing evidence from coarse-grained layers will provide very limited improvement to fine-grained classification. In Stanford Cars, the coarse-grained classes consist of only 9 car types, which is too rough to provide precise evidence to enhance the fine-grained classification of 196 car models. Therefore, the hierarchical evidence fusion on this data set does not work as well as on other data sets. In contrast, when the class hierarchy is moderate or specific, hierarchical evidence fusion can significantly improve the performance of FGIC.

In summary, these results well demonstrate that fusing hierarchical evidence can help improve the performance of FGIC.

Trusted FGIC vs. Traditional FGIC

In the final experiment, we further interpret the trustworthiness of the proposed method. Specifically, we compare our method with fine-tuned ResNet50 on CUB. Figure 6 gives the representative prediction results of two comparing methods. We find that our proposed method can achieve trusted FGIC from following three aspects. First, as shown in the first row of Figure 6, all samples have clear backgrounds and significantly discriminative features. However, baseline method classifies these samples as the wrong categories with high probabilities. In contrast, our method accurately predicts the correct class of the samples and outputs small uncertainty values to indicate that the prediction is trustworthy. Second, compared to the samples in the first row, the samples in the second row have a lot of noise. This allows the baseline method to classify these samples as the wrong category with high probabilities. In contrast, our method improves the probability of belonging to the correct class by hierarchical evidence fusion and identifies these untrusted

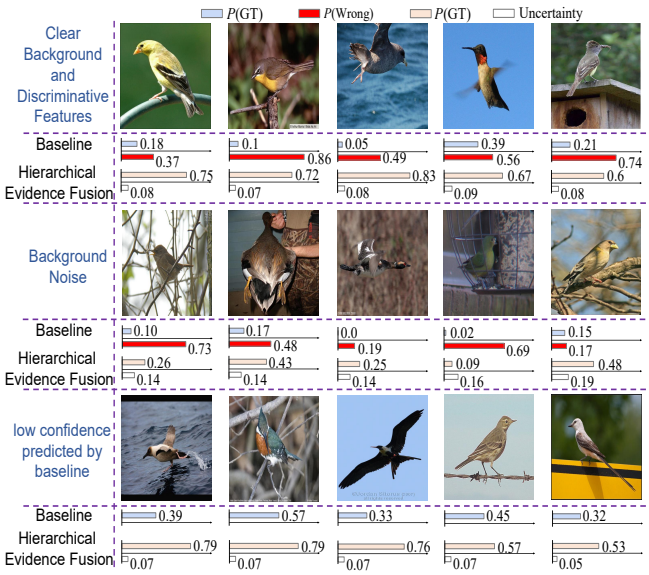

Figure 6: Three different types of representative cases illustrate how our method enhances the trustworthiness of FGIC.

samples by outputting large uncertainty values. In this way, we can further extend our model to perform uncertainty FGIC by rejecting the samples with large uncertainty values. Lastly, in the last row of Figure 6, baseline method can identify the correct class of samples, but the probability of belonging to the correct classes is low. This can lead that the prediction is untrusted. In contrast, our method improves the prediction probabilities by fusing hierarchical evidence.

In summary, by measuring uncertainty and fusing hierarchical evidence by evidence theory, our method achieves trusted FGIC from three aspects. 1) Our method can correct the wrong prediction and reduce uncertainty. 2) Our method can recognize noise samples. 3) Our method can improve the confidence of the prediction.

Conclusion

In this paper, we argue that in addition to classification accuracy, the trustworthiness of FGIC should also be concerned. Because insufficient training data and confusing samples can cause FGIC to produce untrusted classification results, we adopt evidence theory to explicitly represent uncertainty, and use uncertainty to measure trustworthiness. Furthermore, we view FGIC as a hierarchical classification and propose to fuse hierarchical evidence into fine-grained classification. In addition to improving the classification performance, our method can reduce uncertainty through theoretical proofs. The experimental results on three benchmark datasets validate that our method achieves competitive performance and trusted classification. In the future, we will attempt to extend our method to uncertainty FGIC and semi-supervised FGIC.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Serial Nos. 61991410, 61976134), Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province, China (No. CICIP2021001), and Natural Science Foundation of Shanghai (Serial No. 21ZR1423900).

References

Bai, Y.; Chen, Y.; Yu, W.; Wang, L.; and Zhang, W. 2020. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*.

Bhunia, A. K.; Chowdhury, P. N.; Sain, A.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2021. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4247–4256.

Chang, D.; Pang, K.; Zheng, Y.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2021. Your" Flamingo" is My" Bird": Fine-Grained, or Not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11476–11485.

Chen, T.; Wu, W.; Gao, Y.; Dong, L.; Luo, X.; and Lin, L. 2018. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM international conference on Multimedia*, 2023–2031.

Chen, Y.; Bai, Y.; Zhang, W.; and Mei, T. 2019. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5157–5166.

Dempster, A. P. 1968a. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2): 205–232.

Dempster, A. P. 1968b. Upper and lower probabilities generated by a random closed interval. *The Annals of Mathematical Statistics*, 957–966.

Denoeux, T. 1997. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern recognition*, 30(7): 1095–1107.

Denoeux, T. 2008. A k-nearest neighbor classification rule based on Dempster-Shafer theory. In *Classic works of the Dempster-Shafer theory of belief functions*, 737–760. Springer.

Ding, Y.; Zhou, Y.; Zhu, Y.; Ye, Q.; and Jiao, J. 2019. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6599–6608.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 153–168. Springer.

Gao, Y.; Han, X.; Wang, X.; Huang, W.; and Scott, M. 2020. Channel interaction networks for fine-grained image categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10818–10825.

Han, K.; Guo, J.; Zhang, C.; and Zhu, M. 2018. Attribute-aware attention model for fine-grained representation learning. In *Proceedings of the 26th ACM international conference on Multimedia*, 2040–2048.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ji, R.; Wen, L.; Zhang, L.; Du, D.; Wu, Y.; Zhao, C.; Liu, X.; and Huang, F. 2020. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10468–10477.

Jsang, A. 2016. Subjective Logic: A formalism for reasoning under uncertainty. *Springer Verlag*.

Li, B.; Han, Z.; Li, H.; Fu, H.; and Zhang, C. 2021. Trustworthy Long-Tailed Classification. *arXiv preprint arXiv:2111.09030*.

Li, S.; Yao, Y.; Hu, J.; Liu, G.; Yao, X.; and Hu, J. 2018. An ensemble stacked convolutional neural network model for environmental event sound recognition. *Applied Sciences*, 8(7): 1152.

Liu, C.; Xie, H.; Zha, Z.-J.; Ma, L.; Yu, L.; and Zhang, Y. 2020. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11555–11562.

Liu, W.; Yue, X.; Chen, Y.; and Denoeux, T. 2022. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7585–7593.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Liu, Z.; Pan, Q.; and Dezert, J. 2013. A new belief-based K-nearest neighbor classification method. *Pattern Recognition*, 46(3): 834–844.

Luo, W.; Yang, X.; Mo, X.; Lu, Y.; Davis, L. S.; Li, J.; Yang, J.; and Lim, S.-N. 2019. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8242–8251.

Quost, B.; Denoeux, T.; and Li, S. 2017. Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression. *Advances in Data Analysis and Classification*, 11(4): 659–690.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31.

Si, L.; Wang, Z.; Tan, C.; and Liu, X. 2014. A novel approach for coal seam terrain prediction through information fusion of improved D–S evidence theory and neural network. *Measurement*, 54: 140–151.

Sun, M.; Yuan, Y.; Zhou, F.; and Ding, E. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 805–821.

Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 595–604.

Wang, D.; Shen, Z.; Shao, J.; Zhang, W.; Xue, X.; and Zhang, Z. 2015. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision*, 2399–2406.

Wang, S.; Wang, Z.; Li, H.; and Ouyang, W. 2020a. Category-specific semantic coherency learning for fine-grained image recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 174–183.

Wang, Z.; Wang, S.; Li, H.; Dou, Z.; and Li, J. 2020b. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12289–12296.

Wang, Z.; Wang, S.; Yang, S.; Li, H.; Li, J.; and Li, Z. 2020c. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9749–9758.

Wei, X.-S.; Cui, Q.; Yang, L.; Wang, P.; and Liu, L. 2019. RPC: A large-scale retail product checkout dataset. *arXiv preprint arXiv:1901.07249*.

Wei, X.-S.; Luo, J.-H.; Wu, J.; and Zhou, Z.-H. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6): 2868–2881.

Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-Grained Image Analysis with Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xie, S.; Yang, T.; Wang, X.; and Lin, Y. 2015. Hyper-class augmented and regularized deep learning for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2645–2654.

Xu, P.; Davoine, F.; and Denœux, T. 2015. Evidential multinomial logistic regression for multiclass classifier calibration. In *2015 18th International Conference on Information Fusion (Fusion)*, 1106–1112. IEEE.

Zhang, F.; Li, M.; Zhai, G.; and Liu, Y. 2021. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *International Conference on Multimedia Modeling*, 136–147. Springer.

Zhang, L.; Huang, S.; Liu, W.; and Tao, D. 2019. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8331–8340.

Zhang, X.; Zhou, F.; Lin, Y.; and Zhang, S. 2016. Embedding label structures for fine-grained feature representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1114–1123.

Zhao, B.; Feng, J.; Wu, X.; and Yan, S. 2017. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2): 119–135.

Zhao, Y.; Yan, K.; Huang, F.; and Li, J. 2021. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15079–15088.

Zhou, F.; and Lin, Y. 2016. Fine-grained image classification by exploring bipartite-graph labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1124–1133.

Zhuang, P.; Wang, Y.; and Qiao, Y. 2020. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13130–13137.