Where to Focus: Investigating Hierarchical Attention Relationship for Fine-Grained Visual Classification

Yang Liu[1⋆], Lei Zhou[1⋆], Pengcheng Zhang[1], Xiao Bai[1], Lin Gu[2,3], Xiaohan Yu[4], Jun Zhou[4], and Edwin R. Hancock[5,1]

[1] School of Computer Science and Engineering, State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University, China
[2] RIKEN AIP, Tokyo, Japan [3] The University of Tokyo, Japan
[4] Griffith University, Australia [5] University of York, U.K.

Abstract. Object categories are often grouped into a multi-granularity taxonomic hierarchy. Classifying objects at coarser-grained hierarchy requires global and common characteristics, while finer-grained hierarchy classification relies on local and discriminative features. Therefore, humans should also subconsciously focus on different object regions when classifying different hierarchies. This granularity-wise attention is confirmed by our collected human real-time gaze data on different hierarchy classifications. To leverage this mechanism, we propose a Cross-Hierarchical Region Feature (CHRF) learning framework. Specifically, we first design a region feature mining module that imitates humans to learn different granularity-wise attention regions with multi-grained classification tasks. To explore how human attention shifts from one hierarchy to another, we further present a cross-hierarchical orthogonal fusion module to enhance the region feature representation by blending the original feature and an orthogonal component extracted from adjacent hierarchies. Experiments on five hierarchical fine-grained datasets demonstrate the effectiveness of CHRF compared with the state-of-the-art methods. Ablation study and visualization results also consistently verify the advantages of our human attention-oriented modules. The code and dataset are available at https://github.com/visiondom/CHRF.

Keywords: Fine-grained visual classification, multi-granularity, human attention, orthogonal fusion

1 Introduction

Fine-grained visual classification (FGVC) is more challenging than traditional image classification due to the highly similar appearance among subordinate categories. In the past decade, various approaches have been presented [2, 43, 23, 36, 45] to learn the fine distinction between highly similar objects. Thanks

⋆ Equal contribution.
Corresponding author: Xiao Bai (baixiao@buaa.edu.cn).

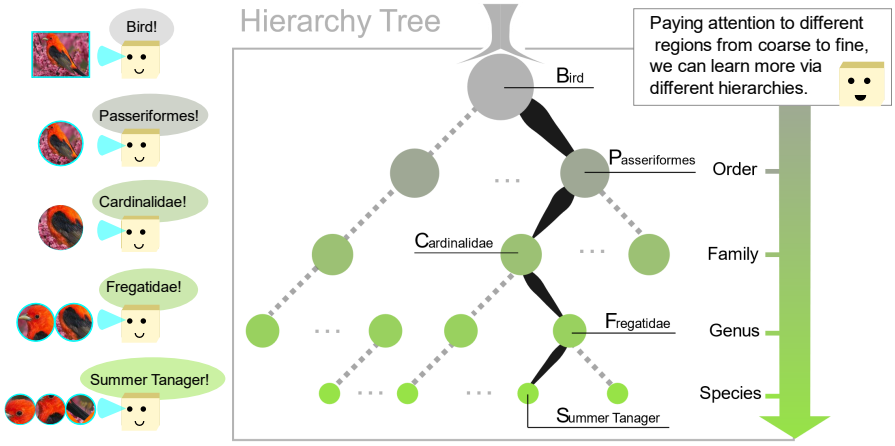


Fig. 1. Illustration of human attention behaviour for classification at different ranks. To recognize the *order* of a given bird image, humans generally glance at the entire bird, *i.e.*, large global attention that sufficiently discriminates categories of *orders*. When recognizing finer-grained categories, such as *genus* or *species*, humans ignore the shared characteristics in this *orders* and focus on smaller but significant local discriminative regions to find the minor inter-class differences between subordinate categories.

to the powerful capability of deep neural networks on discriminative representation learning, deep model-based fine-grained methods [21, 10, 44, 31] have achieved encouraging performance. However, most reported works ignore the multi-granularity relation among object categories, *e.g.*, different *orders* and *families* of birds, and directly train a classification model on one granularity or hierarchy.

The affinities of all the beings of the same class have sometimes been represented by a great tree[1]. Objects like animals, plants, cars, *etc.*, are often grouped into a taxon according to their shared morphological characteristics and given a taxonomic rank. Groups of a certain rank are aggregated to form a higher rank, thus creating a taxonomic hierarchy. Typically, closely related taxa under the same lower rank differ much less than more distantly related ones at higher levels. These hierarchical relationships are significant for designing computer vision models to solve the FGVC task. For example, to identify the *family* of a given bird, if its *order* is known, we then can focus on the differences between *families* that belonged to this *order* and ignore their common characteristics at the order rank, *i.e.*, more different discriminative regions should be paid attention to from coarse to fine levels. As illustrated in Fig. 1, the summer tanager, might be first classified to "passeriformes" according to their common characteristic perching-like shape, then grouped to "cardinalidae" with red belly, and finally classified to "summer tanager" due to the red crown and nape. In light of the taxonomic

[1] Charles Darwin, On the Origin of Species.

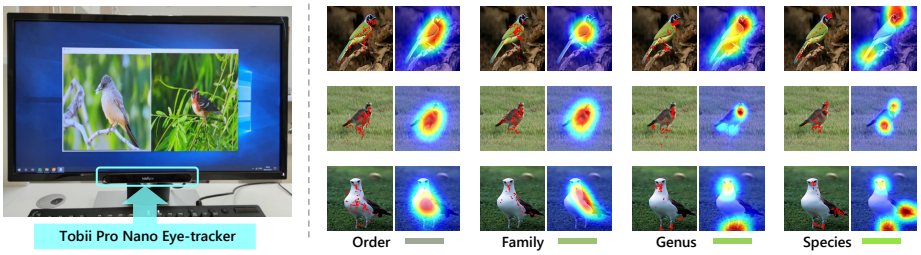


Fig. 2. Left: the eye-tracker device and classification game interface. Right: some samples of the collected human gaze on the CUB dataset [33]. From left to right are gaze data from the category hierarchy of *order*, *family*, *genus* and *species*, respectively.

hierarchy, in this paper, we study the relationship of human attention for image classification at different granularities.

To investigate the attention mechanism of the human visual system when handling multi-granularity image classification, we designed a bird classification game at each category hierarchy of the Caltech-UCSD birds (CUB) dataset [33] following [22] to collect human gaze data for human attention monitoring. An eye-tracker is used to record participants' gaze when they classify the birds under different category hierarchies. We name the collected human gaze dataset Attention Reinforced Images on Species TaxonOmy (ARISTO). The detailed collection process is introduced in Appendix A. Fig. 2 shows some samples of the ARISTO at different hierarchies. We can see that *at the coarser-level category hierarchy, humans prefer to glance at the entire bird, i.e., global attention. When classifying the finer-level categories, they attempt to find smaller local regions to distinguish the slight inter-class differences.* In addition, by observing the position of gaze points of the same image at different hierarchies, we can find the relationship between human attention at different category hierarchies: *the concerned regions at coarser-level classification tasks are usually different from the attention for finer-level classification.* This is because the attention of higher hierarchy often reflects the different attributes between the corresponding level category, while these attributes are common characteristics when classifying the sub-categories of one particular category. These results demonstrate that the human attention behavior on FGVC at different granularities coincides with the knowledge of taxonomic hierarchy.

Motivated by the adaptive human attention on different hierarchies, we propose a cross-hierarchical region feature (CHRF) learning framework to solve the FGVC problem at different granularities. There are two novel modules in the proposed framework: region feature mining (RFM) module and cross-hierarchical orthogonal fusion (COF) module. The RFM module mimics the human visual system that learns granularity-wise attention for individual category hierarchy. We extract granularity-wise semantic features to guide the learning of different region prototypes for each hierarchy. The COF module is designed to explore how human attention varies from higher hierarchy to lower one, further enhancing

the discriminability of finer-grained region representation. Specifically, we introduce a feature orthogonal fusion operation to implement interaction between region representations of different hierarchies. The finer-level region representation can be disentangled by vector orthogonal decomposition with coarser-level representation, which outputs more discriminative features (orthogonal component) for the current hierarchy. Finally, we apply a fusion operation on the region representation and its orthogonal component.

Our main contributions can be summarized as follows: (1) We propose a cross-hierarchical region feature (CHRF) learning framework with two novel modules, *i.e.*the region feature mining and cross-hierarchical orthogonal fusion modules, to mimic human attention behavior towards improved FGVC at different granularities. (2) We design an image classification game that collects a human gaze dataset on the CUB at different category hierarchies. From the collected ARISTO dataset, we learn hierarchical relationships of human attention at different granularities, which are significant for the FGVC research. (3) Extensive experiments on five hierarchical fine-grained datasets show that our proposed CHRF can learn more discriminative representation on all hierarchies. The performance of CHRF is superior compared with other hierarchy-based methods and is also competitive among the state-of-the-art FGVC methods.

2 Related Work

2.1 Fine-Grained Visual Classification

Recently, the development of deep learning has led to remarkable breakthroughs in FGVC [36, 40, 3, 44]. The primary stream methods of FGVC can be divided into two branches, *i.e.*, fine-grained feature learning [11, 6, 28, 16] and discriminative part learning [1, 15, 21, 10]. The former explores the invariant representation of images through end-to-end feature encoding. Methods with a bilinear structure [25, 13, 23] use high-order feature interactions to enhance the categorization and generalization abilities. However, the lack of spatial distributions of discriminative regions limits the performance of these feature learning methods when objects are severely deformed. On the other hand, methods based on part learning expect to locate the discriminative regions to help fine-grained recognition. Earlier researches in this direction [2, 43, 17] tend to improve classification performance by weak supervision of part or bounding box. However, the annotations for supervision are expensive to obtain. Therefore, some recent part-based works [38, 9, 20] use attention mechanisms to discover the distinguishable regions.

More recently, a few works [5, 4] attempt to promote fine-grained classification by exploiting the multi-granularity category hierarchy. Chen *et al.* [5] introduced a hierarchical semantic embedding framework that used the predicted category score of the coarse level as the prior knowledge to predict the finer level sequentially. Chang *et al.* [4] designed a multiple label prediction model that exploits the inherent coarse-to-fine hierarchical relationship to perform hierarchy-wise feature disentanglement. Although the prior hierarchy relationship is avail-

able in these works, the essential multi-granularity classification mechanism of the human visual system is still not well-modelled in computer vision. Different from these works, we study the relationship of human attention for image classification at different granularities.

2.2 Human Attention in Vision

Many researchers [7, 35, 12, 8] have exploited human attention behavior in different scenarios. Liu *et al.* [26] utilized human attention maps to guide the learning of attention maps for neural image caption. Huang *et al.* [18] proposed a hybrid model to predict human gaze by combining bottom-up visual saliency with task-dependent attention transition in egocentric videos. Liu *et al.* [27] tackled zero-shot recognition by learning discriminative attribute localization supervised by human attention when recognizing an unseen class. Human attention was also demonstrated to be able to enhance the medical application [19, 34]. Rong *et al.* [32] exploited human attention as a data augmentation step to improve the accuracy of fine-grained classification. Yu *et al.* [42] proposed vision Transformer by simulating the glance and gaze behavior of humans when identifying objects in the natural scenario. Partially motivated by these works, in this paper we design a CHRF framework by mimicking the human attention behavior to solve the FGVC task at different granularities.

3 Approach

To classify images at different granularities, humans will focus on different regions of objects. A global observation is helpful to distinguish coarse-grained objects. However, when humans classify finer-grained objects, they tend to explore more local discriminative regions which may be ignored during coarser-grained classification. Motivated by the relationship between the observed regions from coarse to fine by the human visual system, we investigate the interaction among interested regions at different hierarchies.

Problem Definition. Different from most existing FGVC tasks [40, 6, 10, 44], we follow a multi-grained classification setting [5, 4]. For a given image $\mathbf{x}$, the multi-grained hierarchical labels, $\{y^1, y^2, ..., y^l, ..., y^L\}$, are available from coarse to fine. The motivation of this setting is to simulate humans to study the interactions of hierarchies under different granularity views. In this section, we propose a cross-hierarchical region feature (CHRF) learning framework to simultaneously perform classification at different category hierarchies.

3.1 Overview

The overview of the proposed CHRF framework is depicted in Fig. 3. CHRF is a tree structure consisting of three parts, trunk, branches, and leaves. Given

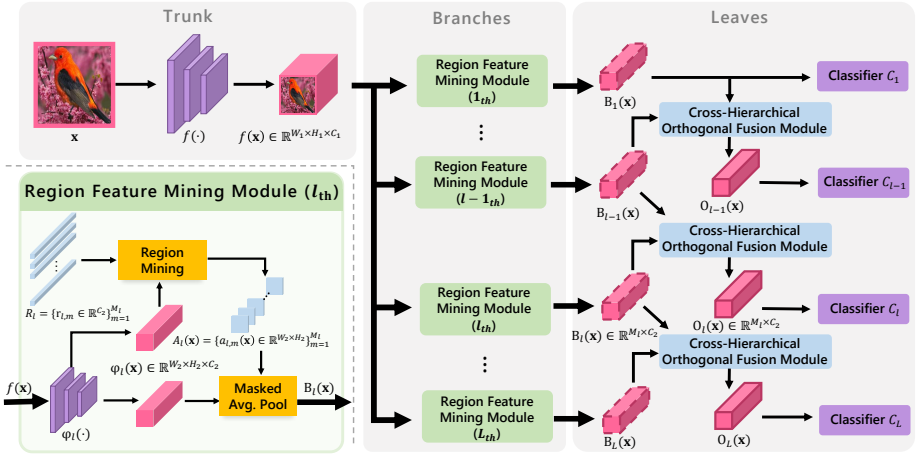


Fig. 3. An overview of the cross-hierarchical region feature learning (CHRF) framework, which consists of trunk, branches and leaves. The trunk is utilized to extract image features. Branches include L region feature mining (RFM) modules to mine different grained region representations. Leaves exploit the cross-hierarchical orthogonal fusion (COF) module to enhance the discriminability of the finer-grained representation by integrating the region representations of two adjacent hierarchies. The bottom left is the detailed RFM module. The COF module is shown in Fig. 4.

an image $\mathbf{x}$ with labels $\{y^1, y^2, ..., y^l, ..., y^L\}$, the trunk extracts image feature $f(\mathbf{x}) \in \mathbb{R}^{W_1 \times H_1 \times C_1}$ by a CNN $f(\cdot)$, where W_1, H_1 and C_1 denote the image feature's width, height and the number of channel, respectively.

Although $f(\mathbf{x})$ can describe the characteristics of $\mathbf{x}$, it lacks the insight in multi-granularity perspective. Therefore, we utilize the branches including L region feature mining (RFM) modules to mine the different grained region representations. Without loss of generality, we consider a multi-grained classification task with a category hierarchy of L levels. We use $1, 2, ..., l, ..., L$ to denote each level from coarse to fine. Each level contains one RFM module to simulate human cognitive behavior and find visual patterns corresponding to the level, *e.g.*, we will see the whole body of a bird or a butterfly in the order level, however, the head of the bird and the wing stripe of the butterfly will be focused on in the species level. Then the granularity-wise attention region representation $B_l(\mathbf{x}) \in \mathbb{R}^{M_l \times C_2}$ will be excavated, where M_l and C_2 denote the number of region at level l and the number of channels, respectively.

Leaves integrate the region representations of two adjacent levels through a cross-hierarchical orthogonal fusion (COF) module to enhance the finer-grained region representation. Our motivation of COF is to compare the difference between fine-grained observation and coarse-grained observation and improve the discriminability of the fine-grained representation. Specifically, for level l, COF takes as inputs $B_{l-1}(\mathbf{x})$ and $B_l(\mathbf{x})$ respectively produced by RFM_{l-1} and RFM_l and outputs the region orthogonal feature $O_l(\mathbf{x}) \in \mathbb{R}^{M_l \times C_2}$. For the most coarse-

grained hierarchy, *i.e.*, level 1, we use the region representation B_1 and y^1 to directly learn the classifier C_1. The classification objective of the first hierarchy can be formulated as:

$$\mathcal{L}_{cls,1} = \mathcal{L}_{CE}\left(C_1(B_1), y^1\right) \tag{1}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss. For level l among 2 to L, the discriminative region orthogonal feature $O_l(\mathbf{x})$ and y^l are taken as inputs to the classifier C_l. The classification objective of the l-th hierarchy can be formulated as:

$$\mathcal{L}_{cls,l} = \mathcal{L}_{CE}\left(C_l(O_l), y^l\right), l = 2, 3, ..., L \tag{2}$$

The total classification loss function can be then written as:

$$\mathcal{L}_{cls} = \textstyle\sum_{l=1}^{L} \mathcal{L}_{cls,l} \tag{3}$$

By minimizing the loss in Equation (3), CHRF is expected to achieve two goals: 1) the fine-grained level can enhance the discriminability of the regional observation by finding the difference compared with the coarse-grained level during the forward procedure. 2) the coarse-grained level feature can obtain extra supplementary details from the fine-grained level through the backward procedure. By the interaction of regions among different hierarchies, both coarse and fine levels can gain a performance improvement.

3.2 Region Feature Mining Module

In light of the insight that humans will focus on different regions and different extents of a region when classifying images at multiple granularities [41], we simulate this human attention mechanism to study the region representations at different hierarchies. The detailed network structure of the RFM module is shown in the bottom left of Fig. 3.

For the RFM of level l, we firstly extract granularity-wise semantic feature $\varphi_l(\mathbf{x}) \in \mathbb{R}^{W_2 \times H_2 \times C_2}$ by a CNN $\varphi_l(\cdot)$ from the image feature $f(\mathbf{x})$, where $\varphi_l(\cdot)$ is exclusive for the specific hierarchy. Specifically, a set of learnable region prototypes $R_l = \{r_{l,m} \in \mathbb{R}^{C_2}\}_{m=1}^{M_l}$ are introduced to discover M_l different regions of $\varphi_l(\mathbf{x})$, where $r_{l,m}$ denotes the m-th region prototype at level l. The feature vectors of $\varphi_l(\mathbf{x})$ are grouped into a series of related similarity map by calculating the dot product between the feature vector and region prototype. Then, we use a region mining operation implemented by batch normalization and ReLU activation to produce the region masks $A_l(\mathbf{x}) = \{a_{l,m}(\mathbf{x}) \in \mathbb{R}^{W_2 \times H_2}\}_{m=1}^{M_l}$. Finally, the vectors of semantic feature are weighted by the region masks and further aggregated to form region representation by global average pooling:

$$b_{l,m}(\mathbf{x}) = \frac{1}{W_2 H_2} \sum_{i=1}^{W_2} \sum_{j=1}^{H_2} a_{l,m}^{i,j}(\mathbf{x}) \varphi_l^{i,j}(\mathbf{x}) \tag{4}$$

where $b_{l,m}(\mathbf{x})$ denotes the m-th region representation and (i, j) denotes the spatial location. These region-level representations are futher concatenated to form

the observation $B_l(\mathbf{x}) = [b_{l,1}(\mathbf{x}), b_{l,2}(\mathbf{x}), ..., b_{l,M_l}(\mathbf{x})]$ of level l. The observation B_l including M_l regions can describe the image's patterns from different views that is helpful to investigate the relationship of the region observations among multiple granularity levels.

Then, L region observations $B_1, B_2, ..., B_L$ are obtained from different granularity levels by branches. These multi-grained attentions contain the spatial location information and the extent of the regions which are similar to the human attention mechanism, *i.e.*, the coarser-grained focuses on less different spatial locations with larger extent. In contrast, the finer-grained focuses on more different spatial locations with smaller extent. For finer-grained classification, smaller local observations are discriminative, which are not necessarily emphasized in coarse-grained observation.

3.3 Cross-Hierarchical Orthogonal Fusion Module

When classifying images at category level l, humans tend to overlook the common properties of the same category at level $l-1$ and pay more attention to the discriminative properties. We mimic this behavior to realize the interaction of the region representations of two adjacent levels. Inspired by [39, 37], the discriminative features are expected to be disentangled from the finer-grained region representation by feature vector decomposition. Specifically, we design a COF module to improve the discriminability of region representations at different hierarchies.

The structure of COF is shown in Fig. 4(a). Firstly, the global observation $G_{l-1}(\mathbf{x}) \in \mathbb{R}^{1\times C_2}$ of $B_{l-1}(\mathbf{x})$ is computed by average pooling operation:

$$G_{l-1}(\mathbf{x}) = \frac{1}{M_{l-1}} \sum_{m=1}^{M_{l-1}} b_{l-1,m}(\mathbf{x}) \tag{5}$$

Then, we calculate the projection $b_{l,m}^{proj}(\mathbf{x})$ of the m-th region representation $b_{l,m}(\mathbf{x})$ on the global observation. This operation can be written as:

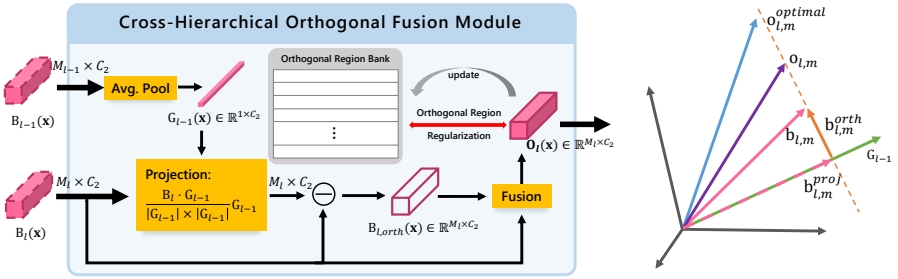
$$b_{l,m}^{proj}(\mathbf{x}) = \frac{b_{l,m}(\mathbf{x}) \cdot G_{l-1}(\mathbf{x})}{|G_{l-1}(\mathbf{x})|^2} G_{l-1}(\mathbf{x}) \tag{6}$$

The projection contains redundant common properties of the m-th region representation of the finer-level. The discriminative region observation can be obtained by computing the orthogonal component:

$$b_{l,m}^{orth}(\mathbf{x}) = b_{l,m}(\mathbf{x}) - b_{l,m}^{proj}(\mathbf{x}) \tag{7}$$

A fusion operation is then used to enhance the discriminability of the region representation, which is demonstrated in Fig. 4(b), where we use $o_{l,m}^{optimal}$ to denote an optimal region representation for classification. To obtain a distinguishable feature closer to $o_{l,m}^{optimal}$, we add a component to $b_{l,m}(\mathbf{x})$ along the direction of $b_{l,m}^{orth}(\mathbf{x})$. Therefore, the m-th region orthogonal feature can be calculated by:

$$o_{l,m}(\mathbf{x}) = b_{l,m}(\mathbf{x}) + \lambda b_{l,m}^{orth}(\mathbf{x}) \tag{8}$$



(a) Framework of proposed cross-hierarchical orthogonal fusion module.

(b) Illustration of feature orthogonal fusion between l_{th} and $l-1_{th}$ level.

Fig. 4. An illustration of the proposed cross-hierarchical orthogonal fusion module.

where λ is the influence factor controlling the degree of blending orthogonal component. We further analysis the fusion operation in Appendix B. Then, all orthogonal features are concatenated to form the whole region orthogonal feature $O_l(\mathbf{x}) = [o_{l,1}(\mathbf{x}), o_{l,2}(\mathbf{x}), ..., o_{l,M_l}(\mathbf{x})]$ of level l.

The region orthogonal features will be used next for classification. Furthermore, we introduce an orthogonal region bank to store the center representation $c_m^{y^l}$ of the orthogonal region for every category at level l. Inspired by [30], we design an orthogonal region regularization to make each region more discriminative,, which can be written as:

$$\begin{aligned}\mathcal{L}_{orr,l} &= \frac{1}{M_l}\sum_{m=1}^{M_l}\left(1-\cos(o_{l,m}(\mathbf{x}), c_m^{y^l}) + \frac{1}{M_l-1}\sum_{\substack{j=1\\ j\neq m}}^{M_l}\left|\cos(o_{l,m}(\mathbf{x}), c_j^{y^l})\right|\right) \\ &= \frac{1}{M_l}\sum_{m=1}^{M_l}\left(1-\cos(o_{l,m}(\mathbf{x}), c_m^{y^l})\right) + \frac{1}{M_l(M_l-1)}\sum_{m=1}^{M_l}\sum_{\substack{j=1\\ j\neq m}}^{M_l}\left(\left|\cos(o_{l,m}(\mathbf{x}), c_j^{y^l})\right|\right)\end{aligned} \tag{9}$$

where $\cos(\cdot,\cdot)$ and $|\cdot|$ denote a cosine similarity and an absolute value operator, respectively. The first term makes the similarity between the region orthogonal feature and its center close to 1, which ensures clustering of the same region orthogonal feature. The second term makes the similarity close to 0, which ensures the orthogonality of different region orthogonal features to reduce their correlation, so that RFM can explore more different regions. The center representation $c_m^{y^l}$ of an orthogonal region is initialized from zero and optimized by momentum update:

$$c_m^{y^l} \leftarrow c_m^{y^l} + \beta\left(o_{l,m}(\mathbf{x}) - c_m^{y^l}\right) \tag{10}$$

where β is a momentum coefficient controlling the update rate of $c_m^{y^l}$. The total orthogonal region regularization can be written as:

$$\mathcal{L}_{orr} = \textstyle\sum_{l=2}^{L}\mathcal{L}_{orr,l} \tag{11}$$

Then, the whole objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{orr} \tag{12}$$

The joint training establishes the relationship between hierarchies, thus improving discriminability of region representation .

4 Experiments and Analysis

In this section, we report comprehensive experiments to verify the effectiveness of our method. We analyse the influence of hierarchy interaction and compare our CHRF with baselines on five hierarchical fine-grained datasets. CHRF is also compared with the state-of-the-art fine-grained methods on three widely used FGVC datasets. Finally, we present additional experiments to demonstrate the consistency of results and human vision system in handling the hierarchical data. The implementation details are provided in Appendix C.

4.1 Datasets

CUB [33] is the most widely used benchmark for FGVC task. It contains 11,877 images covering 200 species of birds. The dataset is divided into two sets including 5,994 training images and 5,794 test images. The 200 *species* of birds are grouped into 122 *genera*, 37 *families*, and 13 *orders* by a bird taxonomy hierarchy according to the ornithological systematics [5].
Butterfly-200 [5] is a newly released butterfly dataset, which has a hierarchical structure with 200 *species*, 116 *genera*, 23 *subfamilies*, and 5 *families* according to the insect taxonomy. The dataset contains 25,279 images, including a training set of 5,135 images, a validation set of 5,135 images and a test set of 15,009 images.
VegFru [14] is a dataset with fine-grained vegetables and fruits recognition covering 292 *subordinate* classes and 25 *upper-level* categories. VegFru dataset has 29,200 images for training, 14,600 for validation and 116,931 for testing.
FGVC-Aircraft [29] contains 100 fine-grained aircraft *models*, which are grouped into 70 *families* and 30 *makers* by tracing superclasses in Wikipedia pages [4]. The dataset has 10,000 images, 6,667 are for training and 3,333 for evaluation.
Stanford Cars [24] contains 196 car *models*, which can be re-organised into 9 *makers* by tracing superclasses in Wikipedia pages [4]. The dataset contains 16,185 images, including 8,144 images for training and 8,041 images for testing.

4.2 Hierarchy Interaction Analysis

Evaluation Metrics. Directly calculating the arithmetic mean of the accuracy across all hierarchies [4] cannot reasonably evaluate the overall performance of the model, since the classification difficulty at different hierarchies varies. Therefore, we propose a more convincing evaluation metric. First, we calculate the

Table 1. Comparison with different baselines on CUB, Butterfly-200 and VegFru under the multi-granularity setting. The best and the second best results are marked in red and blue.

Methods	CUB					Butterfly-200					VegFru		
	P_1	P_2	P_3	P_4	wAP	P_1	P_2	P_3	P_4	wAP	P_1	P_2	wAP
Baseline	98.5	95.4	91.6	85.4	88.9	98.9	97.4	94.4	84.3	88.8	90.6	88.5	88.7
Baseline++	98.6	95.5	91.4	85.3	88.8	98.9	97.3	94.2	84.4	88.8	90.8	88.8	89.0
HSE [5]	98.8	95.7	92.7	88.1	90.7	98.9	97.7	95.4	86.1	90.2	90.0	89.4	89.5
Ours-RF	98.7	95.7	92.8	87.2	90.3	98.9	97.8	95.3	86.5	90.4	92.0	90.6	90.7
Ours-CHRF	99.0	96.3	93.5	89.4	91.8	99.1	97.8	96.0	87.4	91.2	92.2	91.3	91.4

Table 2. Comparison with different baselines on CUB, FGVC-Aircraft and Stanford Cars under the multi-granularity setting. The best and the second best results are marked in red and blue.

Methods	CUB				FGVC-Aircraft				Stanford Cars		
	P_1	P_2	P_3	wAP	P_1	P_2	P_3	wAP	P_1	P_2	wAP
Baseline	98.5	95.7	85.4	87.6	95.9	93.8	91.5	93.0	96.7	93.5	93.6
Baseline++	98.6	95.5	85.3	87.5	96.0	94.1	91.9	93.3	96.9	93.4	93.6
FGN [4]	98.0	94.7	85.4	87.5	95.6	94.6	92.7	93.8	97.0	94.1	94.2
Ours-RF	98.7	96.0	87.2	89.1	96.4	95.2	92.5	94.0	97.2	94.1	94.2
Ours-CHRF	98.9	96.2	89.2	90.8	96.5	95.6	93.6	94.7	97.2	95.2	95.3

Top-1 precision of all hierarchies, respectively. Then, the hierarchical classification performance can be evaluated by the weighted average precision (wAP) of all hierarchies:

$$\text{wAP} = \sum_{l=1}^{L} \frac{class_num_l}{\sum_{k=1}^{L} class_num_k} P_l \tag{13}$$

where $class_num_l$ and P_l denote the number of categories and Top-1 classification accuracy at level l, respectively. The finer-grained hierarchy contains more categories, so the performance of the finer-grained hierarchy should account for a larger proportion.

Compared Methods. To verify the effectiveness of CHRF and different modules, we compare them with several baseline methods. **Baseline** contains fundamental structures including $f(\cdot)$ and $\varphi(\cdot)$, which is similar to CHRF. The shared former network is frozen and the latter is learnable to adapt different hierarchies. **Baseline++** has the same structure as Baseline, but the parameters of $f(\cdot)$ are freed. **HSE** [5] also adopts a hierarchical structure for multi-granularity setting, which focuses on the influence of the prediction score of coarse hierarchy on the classification of fine hierarchy. **FGN** [4] investigates the impact of transfer

between classification tasks at different granularities. **Ours-RF** is the Baseline model with the RFM module. **Ours-CHRF** is the full framework of our CHRF with hierarchy interaction. For fair comparisons, all methods were implemented with the same setting. For HSE, we show the results reported in [5]. For FGN, we re-produced the method and did experiments under the same setting as ours. We utilized the groups of CUB mentioned in [4].

Results and Ablation Study. The results are shown in Table 1 and Table 2. Baseline and Baseline++ exhibit similar performance under the metric of wAP on all datasets. We speculate that this is because the pre-trained $f(\cdot)$ can well extract visual feature representation. Thus, it is reasonable to fix the parameters of $f(\cdot)$ in our other models. When the RFM modules are added into Baseline, *i.e.* Ours-RF, the model extracts region representations to improve discrimination of different categories at all hierarchies and obtains 1.4%, 1.6%, 2.0%, 1.0%, and 0.6% improvement under wAP on all five datasets, respectively. Furthermore, compared with Ours-RF, Ours-CHRF can achieve 1.5%, 0.8%, 0.7%, 0.7%, and 1.1% wAP improvement on all five datasets, respectively. In the end, Ours-CHRF outperforms both HSE and FGN by a large margin.

Analysis. The experimental results show that our CHRF framework is effective in solving the FGVC task at different granularities. The proposed RFM and COF modules are the main technical contributions to ensure that the CHRF can mimic the human visual system, and the more discriminative regions are gradually focused on from coarse to fine. Different from other attention models [15, 21, 31], RFM explores granularity-wise attentions for different category hierarchies. On the coarser hierarchy, attention tends to be a global observation, so the improvement is not obvious. On finer hierarchy, attention tends to focus on more local regions, and more details are extracted, so Ours-RF has a significant improvement compared with the Baseline. Built on granularity-wise attention, COF investigates the interaction among attentions of different hierarchies. More discriminative regions of the finer hierarchy can be found by comparing coarse attention and fine attention. The interaction effectively boosts boost the classification accuracy of both coarse and fine granularity hierarchies.

Where to Focus? We visualize the attention maps of humans, Ours-RF, and Ours-CHRF in Fig. 5. By the region representations interaction through COF module, coarse-granularity and fine-granularity learn where to focus. Attention maps produced by Ours-RF and Ours-CHRF exhibit the consistent characteristic as human attentions, *i.e.*, global attentions are preferred to produce at coarser-granularity category hierarchy and more smaller discriminative local regions are attempted to explore when classifying the finer-granularity categories. From the visualization results, we can see that the attentions of finer-granularity are usually different from the concerned regions of coarser-granularity. The common properties of the coarser-granularity are overlooked and the discriminative

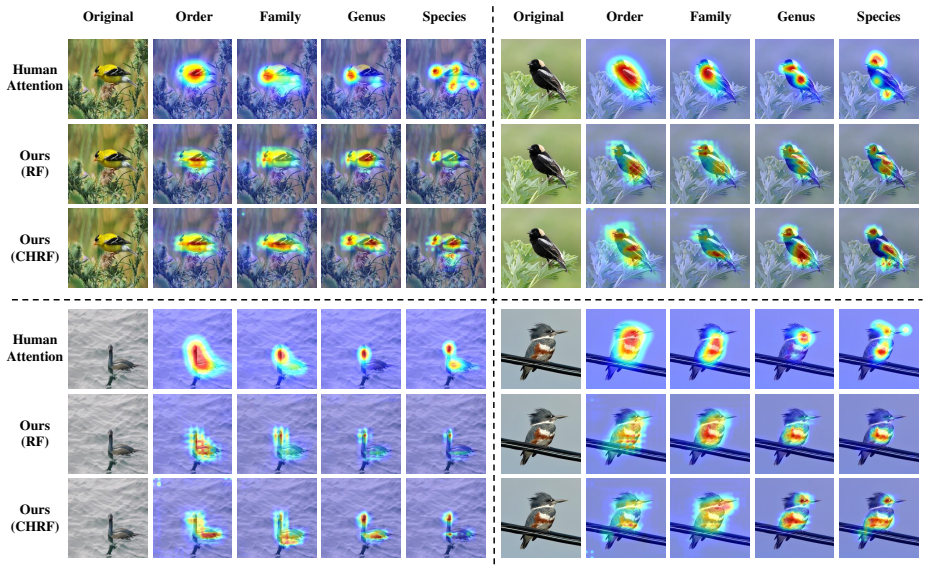


Fig. 5. Visualization of the attention maps from human, Ours-RF and Ours-CHRF on four hierarchies (*order*, *family*, *genus* and *species*) on the CUB dataset.

properties are concerned. In addition, attentions of Ours-CHRF are more distinct than Ours-RF for the discriminative regions mining at different hierarchies. This comparison validates the effectiveness of our COF module.

4.3 Evaluation on Traditional FGVC Setting

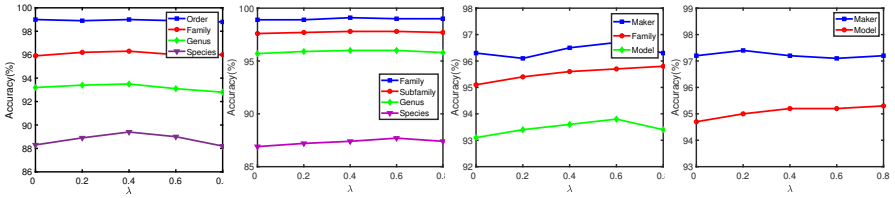
In this section, we also validate the effectiveness of our CHRF compared with recent state-of-the-art FGVC methods. We report the top-1 classification accuracy of CHRF at the bottom of the hierarchy. The results are shown in Table 3. For FGVC-Aircraft and Stanford Cars, CHRF outperforms all compared methods. For CUB, our CHRF can also achieve a competitive result which is only slightly lower than PMG. Notably, different from these methods, partial parameters of CHRF (*i.e.* $f(\cdot)$) are frozen. CHRF depends on the interaction among different hierarchies to improve the performance and achieve competitive results compared with state-of-the-art FGVC methods.

4.4 Further Analysis

Fig. 6 shows the effect of the influence factor λ. We vary λ among $\{0.0, 0.2, 0.4, 0.6, 0.8\}$ to observe the performance changes. The best results are achieved for CUB and Cars datasets when λ is 0.4. For Butterfly-200 and Aircraft datasets, the best results are achieved when λ is 0.6. Thus, we set λ to 0.4 to achieve a trade-off performance on all datasets. Due to the length limitation of the paper,

Table 3. Comparison of the proposed CHRF with the state-of-the-art methods on traditional FGVC setting. The best and the second best results are marked in red and blue, respectively.

Methods	Accuracy(%)		
	CUB	FGVC-Aircraft	Stanford Cars
NTS-Net(ECCV'18) [40]	87.5	91.4	93.9
PC(ECCV'18) [11]	86.9	89.2	92.9
DCL(CVPR'19) [6]	87.8	93.0	94.5
S3N(ICCV'19) [9]	88.5	92.8	94.7
ACNet(CVPR'20) [21]	88.1	92.4	94.6
PMG(ECCV'20) [10]	89.6	93.4	95.1
SPS(ICCV'21) [16]	88.7	92.7	94.9
CHRF(Ours)	89.4	93.6	95.2



(a) CUB (b) Butterfly-200 (c) FGVC-Aircraft (d) Stanford Cars

Fig. 6. The effect of λ with different values on CUB, Butterfly-200, FGVC-Aircraft and Stanford Cars.

more analyses about region prototypes and orthogonal region regularization are given in Appendix C.

5 Conclusions

In this paper, we aim to solve the fine-grained visual classification task at different granularities. We study the relationship between hierarchical human attention by collecting human gaze data from a designed classification game. We designed a cross-hierarchical region feature learning framework to mimic human attention behavior that learns different discriminative representations for the corresponding category hierarchy. Extensive experiments on five hierarchical fine-grained datasets validate the superiority of the proposed human attention-oriented method. The code of our method and the collected human gaze dataset on four hierarchies of the CUB have been released. We believe there is tremendous potential for investigating the hierarchical human attention relationship for the multi-granularity image classification task.

References

1. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 955–962 (2013)
2. Chai, Y., Lempitsky, V., Zisserman, A.: Symbiotic segmentation and part localization for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 321–328 (2013)
3. Chang, D., Ding, Y., Xie, J., Bhunia, A.K., Li, X., Ma, Z., Wu, M., Guo, J., Song, Y.Z.: The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Transactions on Image Processing **29**, 4683–4695 (2020)
4. Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J.: Your" flamingo" is my" bird": Fine-grained, or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11476–11485 (2021)
5. Chen, T., Wu, W., Gao, Y., Dong, L., Luo, X., Lin, L.: Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 2023–2031 (2018)
6. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5157–5166 (2019)
7. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163**, 90–100 (2017)
8. Ding, S., Qu, S., Xi, Y., Wan, S.: Stimulus-driven and concept-driven analysis for image caption generation. Neurocomputing **398**, 520–530 (2020)
9. Ding, Y., Zhou, Y., Zhu, Y., Ye, Q., Jiao, J.: Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6599–6608 (2019)
10. Du, R., Chang, D., Bhunia, A.K., Xie, J., Ma, Z., Song, Y.Z., Guo, J.: Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In: European Conference on Computer Vision. pp. 153–168. Springer (2020)
11. 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. pp. 70–86 (2018)
12. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8554–8564 (2019)
13. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 317–326 (2016)
14. Hou, S., Feng, Y., Wang, Z.: Vegfru: A domain-specific dataset for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 541–549 (2017)
15. Hu, T., Qi, H., Huang, Q., Lu, Y.: See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891 (2019)
16. Huang, S., Wang, X., Tao, D.: Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 620–629 (2021)

17. Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked cnn for fine-grained visual categorization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1173–1182 (2016)
18. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Proceedings of the European conference on computer vision. pp. 754–769 (2018)
19. Huang, Y., Li, X., Yang, L., Gu, L., Zhu, Y., Seo, H., Meng, Q., Harada, T., Sato, Y.: Leveraging human selective attention for medical image analysis with limited training data. In: The British Machine Vision Conference (2021)
20. Huang, Z., Li, Y.: Interpretable and accurate fine-grained recognition via region grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8662–8672 (2020)
21. Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., Liu, X., Huang, F.: Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10468–10477 (2020)
22. Karessli, N., Akata, Z., Schiele, B., Bulling, A.: Gaze embeddings for zero-shot image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4525–4534 (2017)
23. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 365–374 (2017)
24. 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. pp. 554–561 (2013)
25. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)
26. Liu, C., Mao, J., Sha, F., Yuille, A.: Attention correctness in neural image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2017)
27. Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., Harada, T.: Goal-oriented gaze estimation for zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3794–3803 (2021)
28. Luo, W., Yang, X., Mo, X., Lu, Y., Davis, L.S., Li, J., Yang, J., Lim, S.N.: Cross-x learning for fine-grained visual categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8242–8251 (2019)
29. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
30. Ranasinghe, K., Naseer, M., Hayat, M., Khan, S., Khan, F.S.: Orthogonal projection loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12333–12343 (2021)
31. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1025–1034 (2021)
32. Rong, Y., Xu, W., Akata, Z., Kasneci, E.: Human attention in fine-grained classification. BMVC 2021 (2021)
33. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
34. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging (2022)

35. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4894–4903 (2018)
36. Wang, Y., Morariu, V.I., Davis, L.S.: Learning a discriminative filter bank within a cnn for fine-grained recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4148–4157 (2018)
37. Wu, A., Liu, R., Han, Y., Zhu, L., Yang, Y.: Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9342–9351 (2021)
38. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 842–850 (2015)
39. Yang, M., He, D., Fan, M., Shi, B., Xue, X., Li, F., Ding, E., Huang, J.: Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11772–11781 (2021)
40. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Proceedings of the European Conference on Computer Vision. pp. 420–435 (2018)
41. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 192–199 (2014)
42. Yu, Q., Xia, Y., Bai, Y., Lu, Y., Yuille, A.L., Shen, W.: Glance-and-gaze vision transformer. Advances in Neural Information Processing Systems **34** (2021)
43. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based r-cnns for fine-grained category detection. In: European conference on computer vision. pp. 834–849. Springer (2014)
44. Zhao, Y., Yan, K., Huang, F., Li, J.: Graph-based high-order relation discovery for fine-grained recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15079–15088 (2021)
45. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5012–5021 (2019)