Supplementary Materials: 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.

Appendix A: ARISTO Collection

We collected the human gaze dataset Attention Reinforced Images on Species TaxonOmy (ARISTO) on the Caltech-UCSD birds (CUB) dataset [32] by designing a bird classification game consisting of two steps. As illustrated in Fig. 1, firstly, two random images from different categories were given for the participant with 6 seconds to learn the differences between these two categories. Secondly, a new image of one of the two categories was shown, and the participant needed to decide which category this image belonged to in 6 seconds. We utilized the Tobii Pro Nano eye-tracker device to record the human gaze data during the second step. To obtain the human gaze for image classification at different granularities, we adopted the four-level category hierarchy (13 *orders*, 37 *families*, 122 *genera*, and 200 *species*) of the CUB dataset organized by [5]. We repeated the game for four rounds on these four levels, respectively. In each round, the compared images in the first step of the game were randomly chosen from categories of different granularities. For example, for the *family* level classification, the compared images were randomly chosen from two different *families*. Ten participants of different genders and ages contributed to the human gaze study. We divided the entire CUB dataset into ten subsets, and each participant played the game at four hierarchies on one subset. Fig. 1 also shows more samples of the ARISTO.

Appendix B: Fusion Operation Analysis

The addition is an effective fusion operation to enhance the region representations. We can take the region representation $b_{l,m}$ and its orthogonal component $b_{l,m}^{orth}$ as bases and obtain more discriminative region representation through their

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

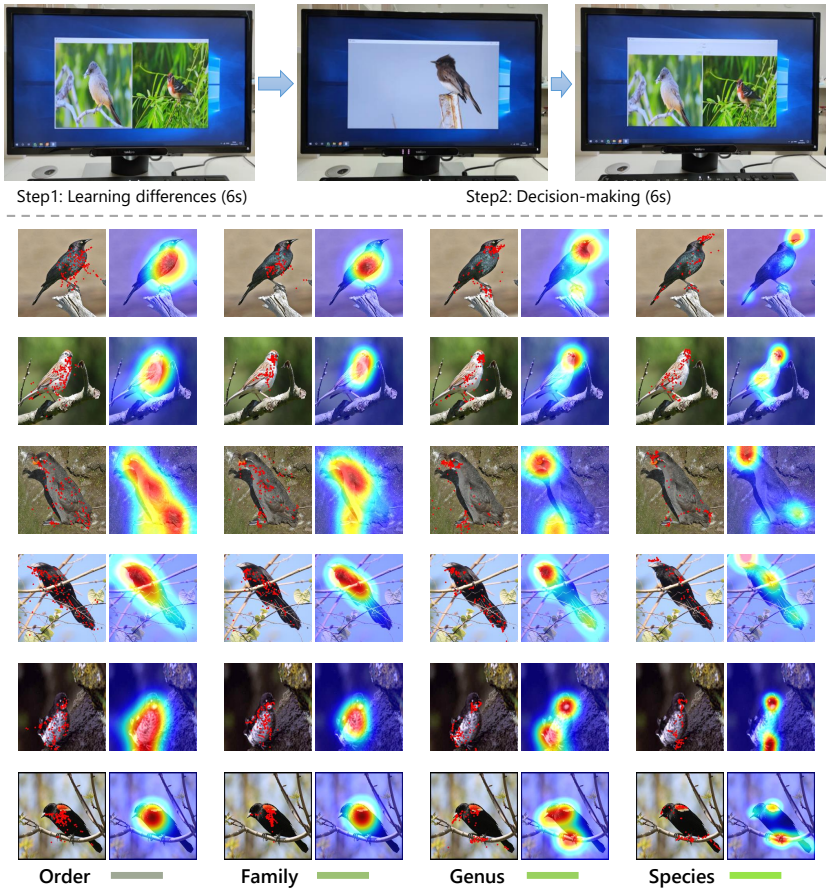


Fig. 1. The gaze data collection process and samples of the collected human gaze. We generate the heatmap of gaze points by a Gaussian blur for more obvious observation.

linear combination. A proper linear combination can can bring the region representation closer to the optimal region representation $o_{l,m}^{optimal}$.

We also analyze the other possible fusion operations. Table 1 shows the comparisons of different fusion operations on CUB. It is clear that the addition fusion with fixed λ achieves the best results. When λ is learnable, the blended region representations tend to be overfitting on the training set, resulting in poor performance. The concatenation operation denotes concatenating the region representation and its orthogonal component. Then, a fully connected layer is utilized to produce the region orthogonal feature. However, this operation will destroy the original structure of the region representation, resulting in performance degradation.

Table 1. Comparisons of different fusion operations on CUB.

Fusion Operation	Accracy(%)			
	Order	Family	Genus	Species
Addition (fixed $\lambda = 0.4$)	99.0	96.3	93.5	89.4
Addition (learnable λ)	99.0	95.9	93.2	88.8
concatenation	98.7	95.8	92.9	88.1

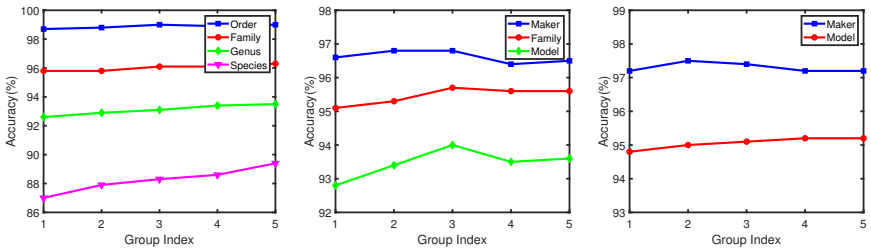
Table 2. Options of the number of region prototypes at different hierarchies on CUB, FGVC-Aircraft and Stanford Cars.

Group Index	CUB				FGVC-Aircraft			Stanford Cars	
	Order	Family	Genus	Species	Maker	Family	Model	Maker	Model
1	4	4	4	4	4	4	4	4	4
2	4	4	8	8	4	8	8	8	8
3	4	8	8	16	8	8	16	8	16
4	4	8	16	16	8	16	16	16	16
5	4	8	16	32	8	16	32	16	32

Appendix C: Supplementary Experiments

Implementation Details. Our CHRF is built on the widely used ResNet-50 pre-trained on ImageNet. Concretely, the ResNet-50 is divided into two CNNs including $f(\cdot)$ and $\varphi(\cdot)$. The former consists of the first three convolution groups (*i.e.*, conv1, conv2_x, and conv3_x) of the ResNet-50, whose parameters are fixed. And the latter consists of the rest components (*i.e.*, conv4_x and conv5_x), whose parameters can be learned. We set the number of region prototypes to 32 for the L-th RFM and those of the higher layer are divided by 2 in turn. λ is set to 0.4. Following the training and testing protocol of recent FGVC works [6,28,43], we use random cropping of 448×448 and horizontal flipping in training and center crop during inference. The data augmentation proposed by [15] is used to improve the attention. CHRF is trained for 160 epochs with a batch size of 8. We use the SGD optimizer with the initial learning rate of $1e$-3 with exponential decay of 0.9 after every two epochs, and momentum is set as 0.9. All experiments are conducted on a RTX 3090 GPU.

Region Prototypes in RFM. We further analyze the influence of the selected number of region prototypes. Based on the observation that the finer hierarchy should have more concerned regions, Table 2 shows five groups of candidate region numbers on CUB, FGVC-Aircraft, and Stanford Cars. Butterfly-200 and VegFru have a similar hierarchy structure to CUB and Stanford Cars, thus they share the same selections, respectively. We search for a comparable candidate as the hyper-parameter of all datasets. The results are shown in Fig. 2. We can see that CUB and Stanford Cars achieve the best result when the 5-th group



(a) CUB (b) FGVC-Aircraft (c) Stanford Cars

Fig. 2. The influence of the number of region prototype at different hierarchies on CUB, FGVC-Aircraft and Stanford Cars.

Table 3. Contribution of the orthogonal region regularization (ORR) on CUB. Fusion represents the addition operation.

Fusion	ORR	Accracy(%)			
		Order	Family	Genus	Species
		98.7	95.7	92.8	87.2
	✓	99.0	95.9	93.2	88.3
✓		98.9	95.9	93.0	87.8
✓	✓	99.0	96.3	93.5	89.4

candidate is chosen. The best result is achieved for FGVC-Aircraft when choosing the 3-th group candidate. Therefore, we finally choose the 5-th candidate as a trade-off hyper-parameter for all datasets in our experiments, *i.e.*we set the number of region prototypes to 32 for the L-th RFM and those of the higher layer are divided by 2 in turn.

Orthogonal Region Regularization in COF. A group of region centers in the orthogonal region bank are used to gather the same orthogonal region features and distinguish different orthogonal region features. For simplicity, we name the orthogonal region regularization as ORR. As shown in Table 3, ORR can effectively enhance the accuracy whether the fusion operation is used or not. Furthermore, we demonstrate the cosine similarity of the centers in the orthogonal region bank at different granularities on CUB dataset, as shown in Fig. 3. We can see that ORR ensures the orthogonality of different region centers, which promotes the discriminability of different orthogonal region features.

More visualization results. We enumerate more attention maps of humans, Ours-RF and Ours-CHRF on CUB dataset in Fig. 4. Besides, comparisons of attention maps from Ours-RF and Ours-CHRF on Butterfly-200, FGVC-Aircraft, and Standford Cars are shown in Fig. 5 and Fig. 6. The visualization results consistently validate the effectiveness of our model.

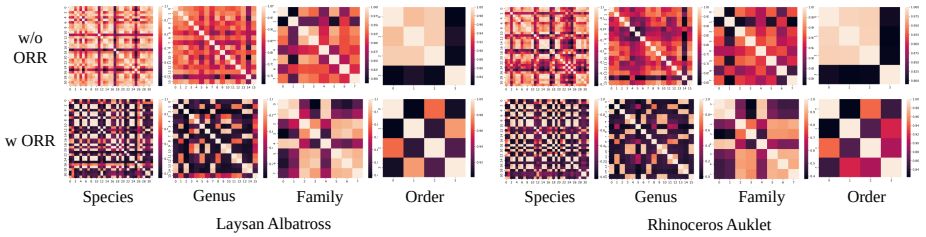


Fig. 3. Comparison of the cosine similarity of the centers in the orthogonal region bank. Four hierarchies of two classes ("Laysan Albatross" and "Rhinoceros Auklet") are demonstrated. The orthogonal region regularization (ORR) can improve the orthogonality of different region orthogonal features to explore more distinguishable regions.

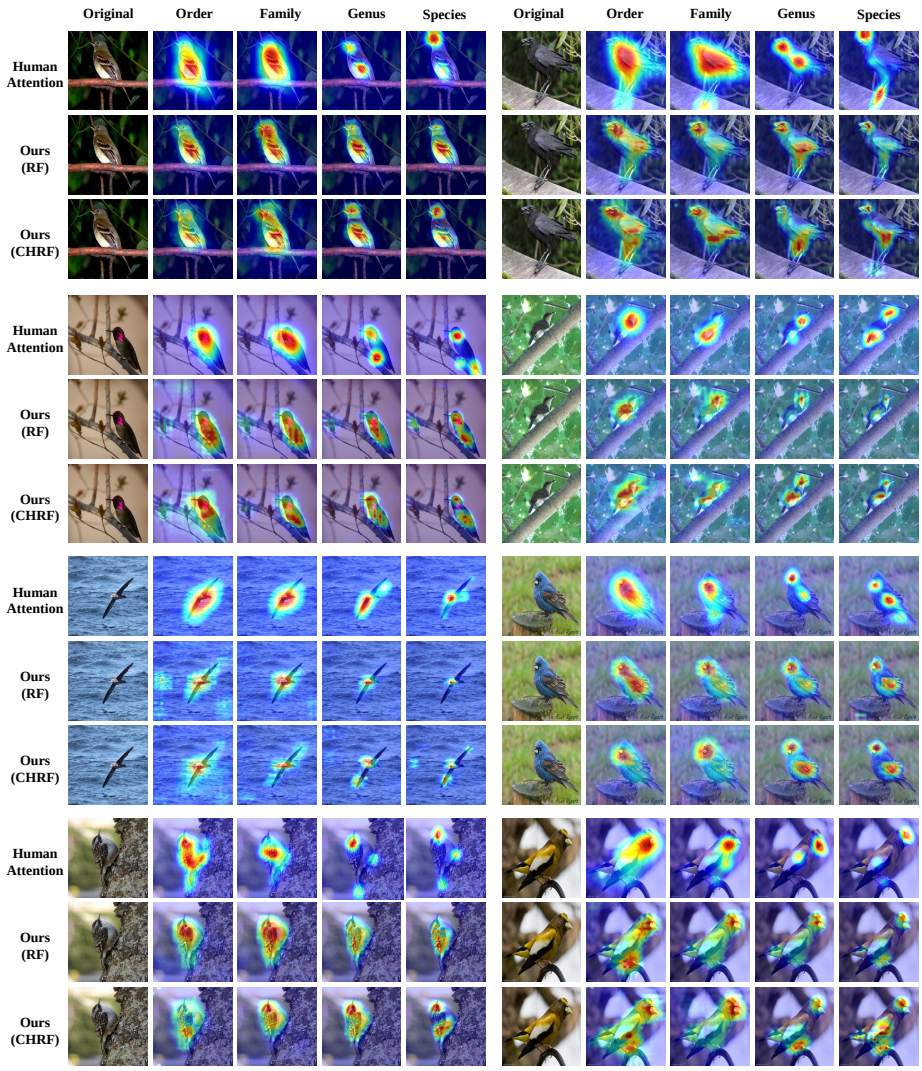


Fig. 4. Visualization of the attention maps from human, Ours-RF and Ours-CHRF on four hierarchies (*order*, *family*, *genus* and *species*) on CUB dataset.

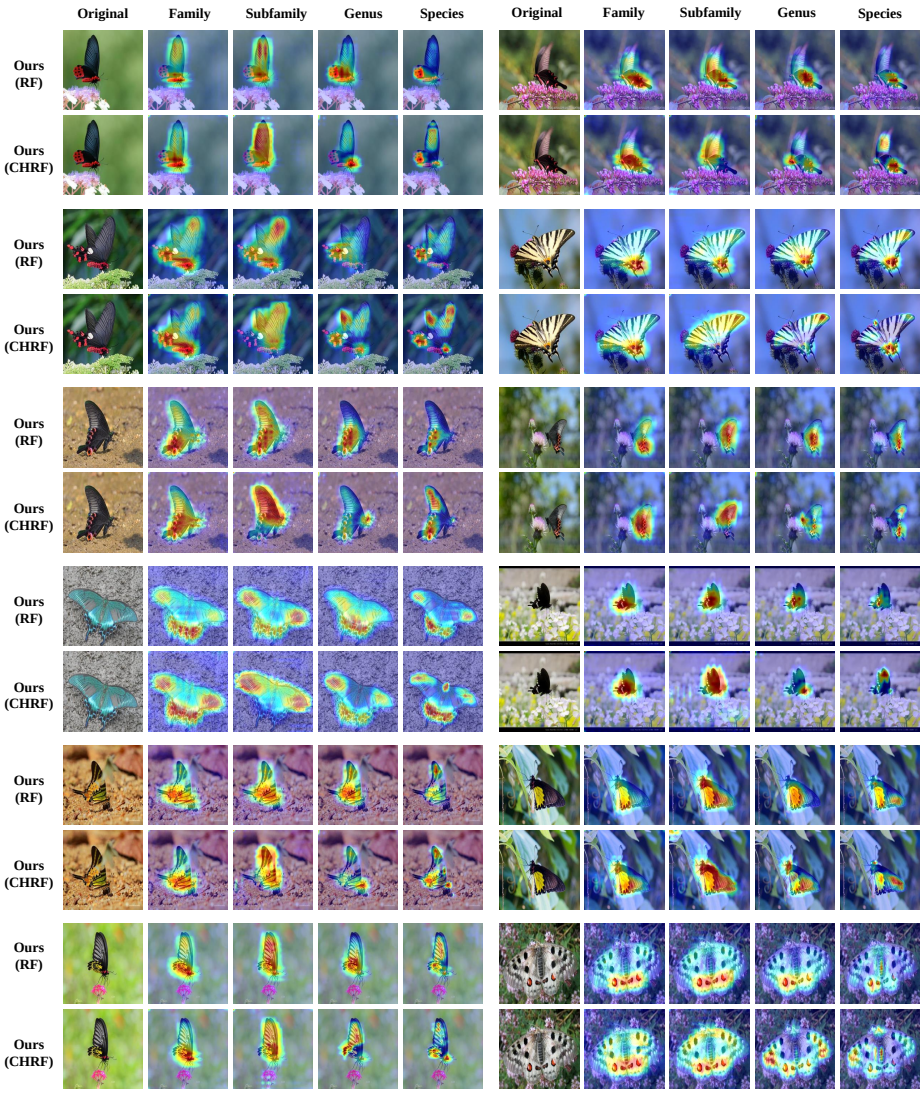


Fig. 5. Visualization of the attention maps from Ours-RF and Ours-CHRF on four hierarchies (*family*, *subfamily*, *genus* and *species*) on Butterfly-200 dataset.

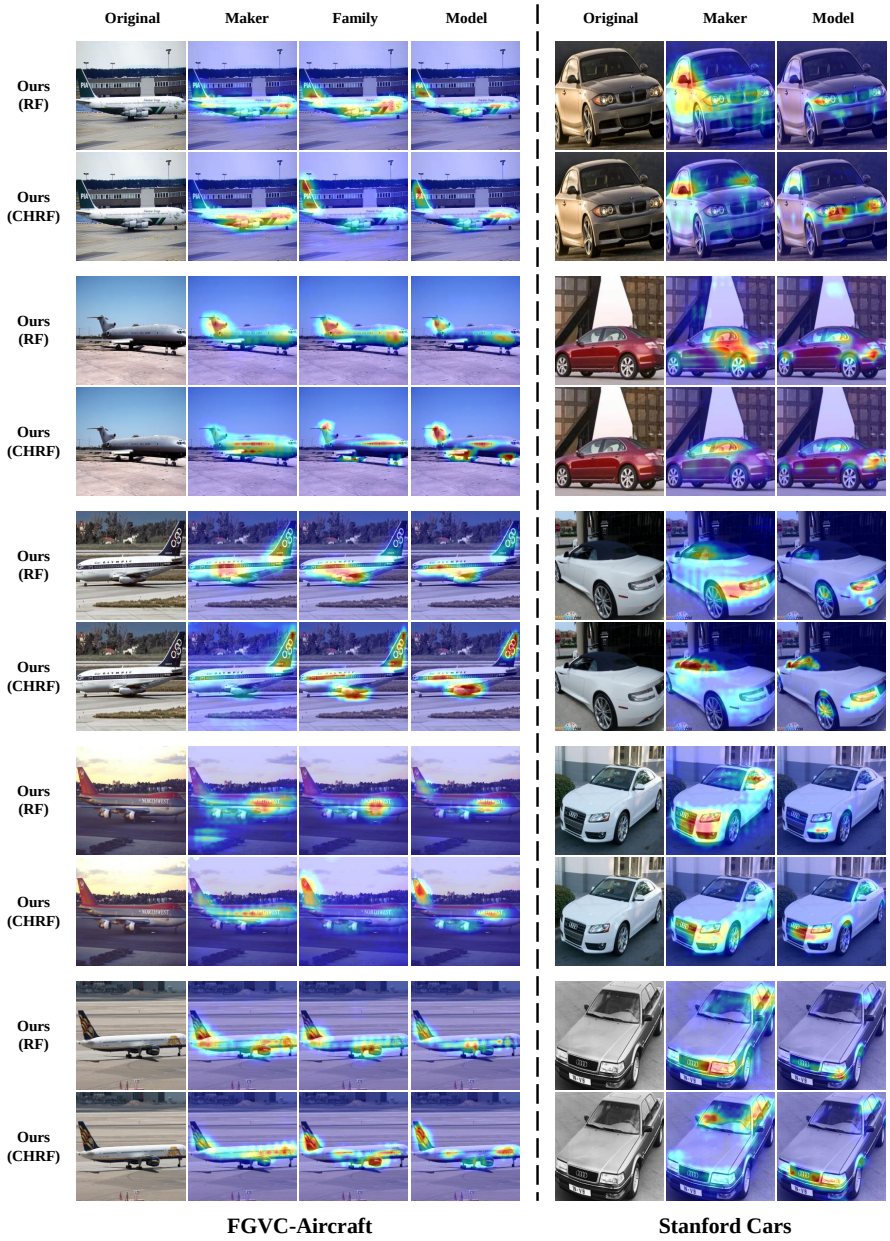


Fig. 6. Visualization of the attention maps from Ours-RF and Ours-CHRF on FGVC-Aircraft (three hierarchies including *maker*, *family* and *model*) and Stanford Cars (two hierarchies including *maker* and *model*).