Fine-grained Visual Classification with High-temperature Refinement and Background Suppression

Po-Yung Chou, Yu-Yung Kao, and Cheng-Hung Lin*
Department of Electrical Engineering, National Taiwan Normal University
{81075001H, 81075006h, brucelin}@ntnu.edu.tw

***Abstract*—Fine-grained visual classification is a challenging task due to the high similarity between categories and distinct differences among data within one single category. To address the challenges, previous strategies have focused on localizing subtle discrepancies between categories and enhencing the discriminative features in them. However, the background also provides important information that can tell the model which features are unnecessary or even harmful for classification, and models that rely too heavily on subtle features may overlook global features and contextual information. In this paper, we propose a novel network called "High-temperaturE Refinement and Background Suppression" (HERBS), which consists of two modules, namely, the high-temperature refinement module and the background suppression module, for extracting discriminative features and suppressing background noise, respectively. The high-temperature refinement module allows the model to learn the appropriate feature scales by refining the features map at different scales and improving the learning of diverse features. And, the background suppression module first splits the features map into foreground and background using classification confidence scores and suppresses feature values in low-confidence areas while enhancing discriminative features. The experimental results show that the proposed HERBS effectively fuses features of varying scales, suppresses background noise, discriminative features at appropriate scales for fine-grained visual classification.The proposed method achieves state-of-the-art performance on the CUB-200-2011 and NABirds benchmarks, surpassing 93% accuracy on both datasets. Thus, HERBS presents a promising solution for improving the performance of fine-grained visual classification tasks. code will be available: https://github.com/chou141253/FGVC-HERBS**

I. Introduction

FINE-GRAINED visual classification (FGVC) is a challenging task in computer vision that involves categorizing images into very specific and detailed categories, such as different species of birds[33], dogs[16], vehicle models[18], and medical images[50]. As shown in Fig.1, these four types of sparrows look almost identical, but from different perspectives, the same type of sparrow also looks very different. In contrast to coarse-grained classification, which involves identifying broad categories like "animals" or "vehicles," fine-grained classification requires the ability to recognize subtle differences in visual features, such as color, texture, shape, and pattern, which often exist in small regions. These regions are referred to as discriminative regions or foreground regions.

Fine-grained recognition can be achieved by dividing objects into parts, such as eyes, feet, etc., and comparing corresponding regions for easier identification[37], [8], [24], [43], [14], [3], [41]. However, these methods require manual annotation, which is costly and even requires expert annotation. To overcome this issue, the weakly supervised methods[42], [39], [7], [12], [6], [47], [36] are proposed to find discriminative regions through class activation mapping (CAM)[49], [27] were proposed, offering training the network through higher response areas in the feature map without labels. In addition, the attention-based methods[48], [51], [52], [44], [15] are proposed to locate discriminative regions by identifying common high-response areas among feature maps. Furthermore, the success of Vision Transformer (ViT) in image classification has led to its implementation in fine-grained visual recognition tasks. These methods[46], [29], [9], [35], [21], [13] use self-attention maps to get information on foreground regions. The main efforts were focused on enhancing the differentiation of discriminative regions, while neglecting unselected regions. However, in cases where the model cannot obtain strong enough discriminative regions, it is useful to first exclude unimportant regions, called background. Motivated by this concept, we propose the Background Suppression (BS) module.

The proposed BS module shows better performance in FGVC tasks. In the first step of the BS module, the output confidence scores are utilized to classify regions into foreground and background. The foreground represents the discriminative area, while the background refers to the unselected or noisy part. Subsequently, the BS module suppresses the feature values in low confidence regions and enhances the discriminative features, thus improving the details of the target object and reducing the noise. Therefore, the BS module can be helpful, especially in cases where it is difficult to distinguish between foreground and background areas.

The algorithm for extracting features from discriminative regions is important for the FGVG task. However, it can lead to the problem of losing contextual information due to the overuse of single or few specific categories of features. Therefore, we propose a high-temperature refinement module to enhance the learning of diverse features, including texture, shape, and appearance from various object categories. Specifically, the module initially uses higher temperatures to learn feature maps so that more global and contextual information can be captured. Subsequently, the feature maps were refined using lower temperatures to capture finer details. This approach allows to obtain richer features, to better classify similar objects, and to improve accuracy. It should be noted that the high-temperature refinement module can be considered as a form of knowledge distillation [11].

The high-temperature refinement module also maintains an appropriate size of the discriminative region, which is advantageous for FGVC tasks. If the feature size is too small,

*corresponding author

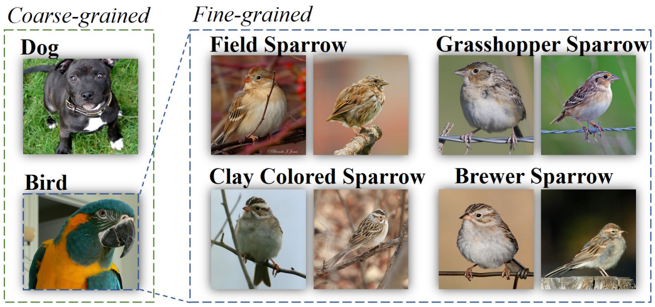


Fig. 1. Examples of visual classification for coarse-grained categories and fine-grained categories.

the algorithm may not be able to capture the overall features of the object, resulting in incorrect classification. Conversely, if the feature scale is too large, the accuracy of FGVC tasks may be reduced due to excessive noise and redundant information.

In this paper, the proposed **H**igh temperatur**E** **R**efinement and **B**ackground **S**uppression (**HERBS**) can extract discriminative features and suppress background noise. This paper has two main contributions:

- The proposed HERBS can be integrated into various backbones, such as CNN-based networks and transformer-based networks. It also allows to perform end-to-end training.
- The proposed HERBS outperforms state-of-the-art approaches, improving the accuracy to 93.1% and 93.0% on CUB200-2011[33] and NABirds[31], respectively.

II. Related Work

A. Fine-grained visual classification

In the field of FGVC, there are two approaches for extracting discriminative features from subtle areas, broadly classified as object-part-based methods and attention-based methods.

Object-part based methods aim to find object local areas for recognition by using a model to generate candidate regions, then extracting discriminative features from them. MA-CNN[48] trains positioning and classification accuracy at the same time through clustering of feature maps into object parts. This unsupervised classification enhances feature learning by dividing patterns into object parts. The approach allows for simultaneous learning of discriminative features and positions. S3N[6] finds the local extremes of each category response on the feature map to enhance features. In addition, WS-DAN[12] augment the data by cutting out local extremes to discover other discriminative features.

Attention-based methods, on the other hand, use attention mechanisms to enhance feature learning and locate object details. MAMC[30] generates multiple sets of features enhanced by attention mechanisms, Cross-X[23] use attention maps from multi-excitation models to learn features from different caterories. API-Net[52] and PCA-Net[44] uses two images as input to calculate attention between feature maps to enhence discriminative representations. CAP[1] calculates the self-attention map of the output features to express the relationship between feature pixels, and SR-GNN[2] uses graph convolutional neural networks to describe the relationship between parts. CAL[25] adds a counterfactual intervention to the attention map to predict the category. With the development of Transformer[32] in the computer vision field, many improved Vision Transformer architectures have been proposed, such as FFVT[35], SIM-Trans[29], TransFG[9], and AFTrans[45], these methods utilize self-attention maps in transformer layers to enhance feature learning and locate object details.

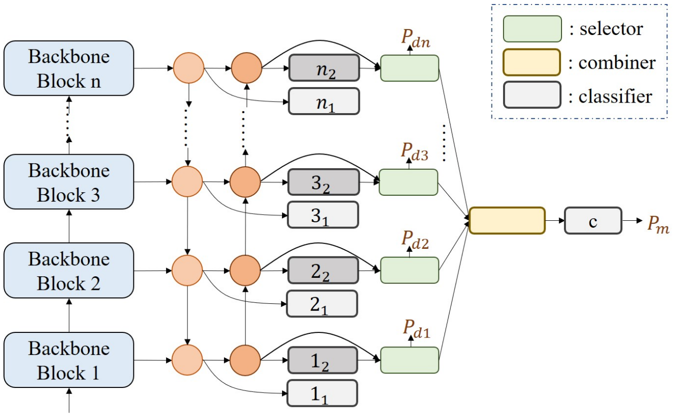


Fig. 2. The illustration of the model structure is shown, where the blue squares on the left represent the backbone blocks, which could be either Convolution-based or Transformer-based. The circles in the middle part denote the multi-scale feature fusion module, such as Feature Pyramid Network (FPN) or Path Aggregation (PA). The classifier, selector, and combiner on the right side depict the HERBS module.

B. Object detection

Supervised object detection methods have demonstrated significant results. The supervised YOLOv7[34] can achieve fast and high accuracy of detection. However, the manual labeling requirement for object positions limits its suitability for fine-grained visual recognition tasks.

Weakly supervised object detection (WSOD) has been introduced as an alternative to overcome the limitations. This method only requires classification labels and generates pseudo bounding box targets through algorithms. For instance, WCCN[5] generates class activation maps to identify regions of interest, which are then fed into the classifier and corrected through multiple instance learning. WSOD2[40] scores virtual candidate boxes through Top-Down and Bottom-Up approaches, with the virtual box with the highest score serving as the target output for the next layer. MIST[26] refines regions of interest through self-training, while WSCL[28] improves the features of regions of interest through data enhancement and contrastive learning. These methods gradually discover the whole object through refinement processes, utilizing the output of the previous stage as the virtual target.

III. Method

In Fig.2, the proposed High-temperatureE Refinement and Background Suppression (HERBS) network is composed of the backbones, the top-down features fusion module, the bottom-up features fusion module, and the HERBS. The backbone can be either a Transformer-based model (e.g., Swin Transformer) or a Convolution-based model (e.g., ResNet). The top-down and bottom-up features fusion module is similar to the path aggregation network (PA)[20], which can be treated as a feature pyramid network (FPN)[19] with an additional bottom-up path.

The proposed HERBS networks aims to learn diverse and discriminative features and improve the accuracy of several FGVC tasks. HERBS contains two modules: the background suppression (BS) module and the high-temperature refinement module. In the following sections, we refer to these two fusion modules as the top-down path and the bottom-up path. And the proposed HERBS, we will provide a comprehensive description of the design of the BS module and high-temperature module, including a detailed explanation of the use of the loss function and the integration of the HERBS module with various frameworks.

A. Background suppression

Let hs_i denote the features map generated by the i^{th} backbone block, where $hs_i \in R^{C_i \times H_i \times W_i}$. Here, C_i represents the number of channels, H_i is the height, and W_i is the width of the features map. The first step of the background suppression (BS) module is to generate the classification maps from these features map, which can be expressed as:

$$Y_i = W_i hs_i + b_i \tag{1}$$

where W_i is the weight of the i^{th} layer classifier, b_i is its bias, and Y_i is the classification maps, with dimensions $R^{C_{gt} \times H_i \times W_i}$, where C_{gt} is the number of target categories. Then the maximum score map is calculated from classification map. The process can be expressed as:

$$P_{max,i} = \max(\text{Softmax}(Y_i)) \tag{2}$$

where $P_{max,i}$ represents the i-th layer's max score map. Next, the features with the top-K_i scores among all predictions are selected. The number of K_i is selected based on the principle that $K_i > K_j$ when $i < j$. Specifically, we set K_1 to 256, K_2 to 128, K_3 to 64, and K_4 to 32. We select this value based on the principle that earlier layers can limit the performance of subsequent layers, and our experiments show that the accuracy is relatively insensitive to variations in this parameter if this principle is followed.

A graph convolution module is then employed to merge the selected features and make a prediction based on the merged features. At this stage, the BS module has the non-selected classification maps, referred to as the dropped maps, denoted as Y_d, and the merged classification prediction, denoted as Y_m. This process is depicted through the selector and combiner components as shown in Fig.2.

The objective function of the merged classification prediction is a standard classification one, using cross-entropy to calculate the similarity between the prediction distribution P_m and the ground truth label y. The merged loss is calculated as follows:

$$P_m = \text{Softmax}(Y_m) \tag{3}$$

$$loss_m = -\sum_{ci=1}^{C_{gt}} y_{ci} \log(P_{m,ci}) \tag{4}$$

Here, y_{ci} is the ground truth of i^{th} class, and $P_{m,ci}$ is the predicted probability of the i^{th} class. The summation is performed over the number of target categories C_{gt}. This enhances the discriminative features in the selected area.

The other objective of the BS module is to suppress features in the dropped maps and increase the gap between the foreground and background. The hyperbolic tangent function, tanh, is applied to the dropped maps, Y_d, as shown in Eq.(5):

$$P_d = \tanh(Y_d) \tag{5}$$

Then, the dropped loss, $loss_d$, is calculated as the mean squared error between the prediction and a pseudo target of -1, as defined in Eq.(6):

$$loss_d = \sum_{i=1}^{C_{gt}} (P_{d,ci} + 1)^2 \tag{6}$$

Note that the hyperbolic tangent function in Eq. (5) maps the values of the prediction into a range that is not restricted to probabilities. This is because we really want to separate foreground and background features even if the background have some other classes appearances.

In order to prevent all blocks' feature maps from only having high responses in the same locations, we also incorporate the prediction of each layer into the training target as follows:

$$P_{li} = \text{Softmax}(W_i(\text{Avgpool}(hs_i)) + b_i) \tag{7}$$

$$loss_l = -\sum_{i=1}^{n} \sum_{ci=1}^{C_{gt}} y_{ci} \log(P_{li,ci}) \tag{8}$$

where Avgpool function aggregates all H_i, and W_i at each channel, and the number of blocks in the backbone is represented by n.

The total BS objective is given by the weighted sum of the merged loss ($loss_m$), dropped loss ($loss_d$), and average layer loss ($loss_l$), as shown in Eq.(9):

$$loss_{bs} = \lambda_m loss_m + \lambda_d loss_d + \lambda_l loss_l \tag{9}$$

where λ_m, λ_d, and λ_l are the weights for the merged loss, dropped loss, and average layer loss, respectively. Specifically, We set λ_m to 1, λ_d to 5, and λ_l to 0.3. These values were set to balance the foreground and background loss and were determined based on the training loss from the first three epochs.

B. High-temperature refinement

The classifier k_1 and classifier k_2 in the Fig.2 are followed by the k-th block features map, classifier k_1 is located within the top-down path while classifier k_2 is in the bottom-up one. The objective is to make classifier k_1 learn the output distribution of classifier k_2. We define the output of classifier k_1 as Y_{i1} and the output of classifier k_2 as Y_{i2}. The refinement objective function helps the model to learn more diverse and stronger representations in the earlier layers while allowing later layers to focus on finer details. In other words, the

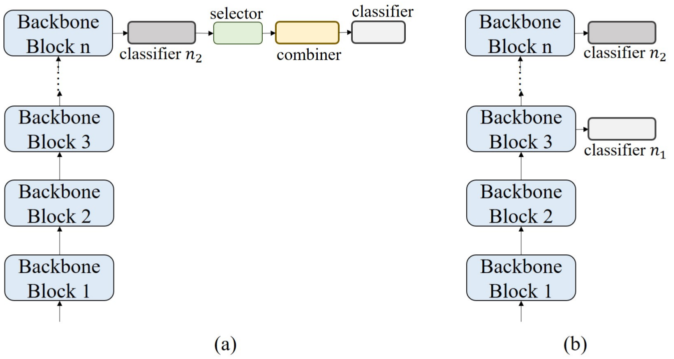


Fig. 3. Illustration of the structure of (a) basic background suppression module and (b) basic high-temperature refinement module.

high-temperature refinement module enables classifier k_1 to discover broader areas and classifier k_2 to focus on learning fine-grained and discriminative features. The refinement loss is calculated using the following equations:

$$P_{i1} = \text{LogSoftmax}(Y_{i1}/T_e) \tag{10}$$

$$P_{i2} = \text{Softmax}(Y_{i2}/T_e) \tag{11}$$

$$loss_r = P_{i2}\log(\frac{P_{i2}}{P_{i1}}) \tag{12}$$

where T_e represents the temperature at training epoch e. The value of T_e decreases as the training epoch increases, following a decay function defined as:

$$T_e = 0.5^{\left\lfloor \frac{e}{-\log_2(0.0625/T)} \right\rfloor} \tag{13}$$

We set the initial temperature T to a high value, such as 64 or 128, in comparison to the knowledge distillation approach[11]. The aim is to encourage the model to explore various features even if the initial predictions are inaccurate. Then, as training progresses, the temperature gradually decreases, allowing the model to focus more on the target class and learn more discriminative features. By using this decay policy, the model can obtain diverse and fine representations and make accurate predictions.

The total loss of HERBS can be formulated as:

$$loss_{herbs} = loss_{bs} + \lambda_r loss_r \tag{14}$$

where λ_r is the weight for refinement loss, which set to 1. And the HERBS network's final output is the softmax of the sum of nine classifier results, consisting of four from the top-down approach, four from the bottom-up approach, and one from the combiner.

Note that in the HERBS network, W_i and b_i belong to classifier k_2 when i equals to k. We separately describe them because the BS module and high-temperature refinement module can be applied to the backbone alone, which is very flexible. The experimental results show that both modules can improve accuracy. Of course, when using the entire HERBS network, the model's capability will result in even better performance.

Method	CUB-200-2011	NA-Birds
FFVT[35]	91.6	N/A
ViT-NeT[17]	91.7	N/A
TransFG[9]	91.7	90.8
IELT[38]	91.8	90.8
SIM-Trans[29]	91.8	N/A
SAC[7]	91.8	N/A
CAP[1]	91.9	91.0
SR-GNN[2]	91,9	91.2
DCAL[51]	92.0	N/A
MetaFormer[4]	92.4	92.7
HERBS	**93.1**	**93.0**

TABLE I
COMPARISON OF TOP-1 ACCURACY(%) WITH STATE-OF-THE-ART METHODS ON THE TWO BENCHMARKS, CUB-200-2011 AND NA-BIRDS.

Module			Backbone	
PA	Refinement	BS	Swin-Base	Swin-Large
			91.3	92.0
✓			91.9(+0.6)	92.5(+0.5)
	✓		91.5(+0.2)	92.3(+0.3)
		✓	91.8(+0.5)	92.4(+0.4)
✓	✓	✓	92.3(+1.0)	93.1(+1.1)

TABLE II
COMPARISON OF TOP-1 ACCURACY(%) ON CUB-200-2011 WITH DIFFERENT MODULE ADDED TO SWIN TRANSFORMER.

In this paper, we propose the HERBS module, composed of the first Background Suppression (BS) and the high-temperature refinement module. Both components can improve the backbone model's accuracy in FGVC tasks. We show the most basic BS and high-temperature refinement modules in Fig.3(a) and (b), respectively. The most basic BS module is added to the output of the final block and to achieve Eqs.(1)–(9). And the most basic High-temperature refinement module is applied to the last two blocks. The final classifier would be treated as classifier n_2 and another one as classifier n_1. Following Eq12, we calculate their KL-divergence as the objective function.

IV. EXPERIMENTS

A. Dataset and implement detail

The datasets used in this study are the CUB200-2011[33] and NA-Birds[31], two fine-grained bird classification datasets. The CUB200-2011 dataset has a total of 200 bird categories, including 5,994 training images and 5,794 testing data. Each category contains about 30 training and testing data. NA-Birds is larger than CUB200-2011, has 555 bird species, 23,929 training images and 24,633 test images. Both datasets provide image-level annotations and keypoint

Module			Backbone
PA	Refinement	BS	ResNet-50
			88.2
✓			88.6(+0.4)
	✓		88.7(+0.5)
		✓	88.4(+0.2)
✓	✓	✓	89.8(+1.6)

TABLE III
COMPARISON OF TOP-1 ACCURACY(%) ON CUB-200-2011 WITH DIFFERENT MODULE ADDED TO RESNET-50.

locations, but only image-level annotations will be used in this paper. When using ResNet-50[10] as the backbone network, the input image is a 448×448 color image, and when using Swin-Transformer[22], the input image is a 384×384 color image. The methods of data augmentation is as follows. If the input image size is 384×384, the first step is to scale the image to 510×510, and if the input image size is 448×448, it is scaled to 600×600. In training phrase, data augmentation is performed via Randon Crop, Random HorizontalFlip, Random GaussianBlur, and Normalizarion while in testing phrase, Center Crop and Normalizarion is used. During training, the learning rate is set to 0.0005, with cosine decay and weight decay set to 0.0005. The optimizer used is SGD, with a batch size of 8, gradient accumulation steps set to 4, and the model is trained for a total of 80 epochs. All experiments are completed on a single Nvidia GeForce RTX 3090, and the Pytorch toolbox is used as the main implementation substrate. It takes about 5 hours to complete the training on CUB200-2011, and about 16 hours for NA-Birds.

B. Ablation experiments

In Table I, we compare our proposed HERBS with state-of-the-art methods on CUB200-2011 and NA-Birds dataset. The middle column of Table I shows that the proposed HERBS can reach 93.1% in Top-1 accuracy, which is 0.7% higher than the previous best method. Table I last column shows that the proposed HERBS can reach 93.0% in Top-1 accuracy on NA-Birds dataset, beating the previous state-of-the-art approaches. These results demonstrate that the proposed HERBS can effectively filter out background noise and extract appropriately sized discriminative features, enabling the identification of fine-grained categories accurately.

To better understand the impact of each module proposed in HERBS, we separately added the PA, Refinement, and BS modules to the classification backbone. First, the Swin Transformer Base (Swin-Base) and Swin Transformer Large (Swin-Large) were used as the testing backbone. As shown in TableII, the original accuracies of Swin-Base and Swin-Large were 91.3% and 92.0%, respectively. After adding the PA, Refinement, or BS modules, there was a slight improvement in accuracy. The structure of only adding PA is shown in Fig.4(b), only adding Refinement is demonstrated in Fig.4(b), and only adding BS is shown in Fig.4(a). The last row in Table 2 shows that the HERBS module improves the backbone's accuracy by about 1%, demonstrating the module's effectiveness.

The HERBS module can be utilized not only with transformer structures but also with convolution-based methods. We chose ResNet-50 as the test backbone, and the results of adding different modules are presented in Table III. Interestingly, the HERBS module improves the accuracy of ResNet-50 (+1.6) more than the Swin Transformer (+1.1). This discrepancy may be attributed to the difference in input image resolutions, with ResNet-50 adapting to 448×448 while the Swin Transformer only adapts to 384×384. The resolution issue will be a topic for future discussion as it still needs to be addressed in this work. Generally speaking, the HERBS module demonstrates promising results on different types of backbones.

		Baseline		+HERBS	
Generic Class	Num.	Pr.(↑)	FP(↓)	Pr.(↑)	FP(↓)
Flycatcher	210	83.09	17	85.02	15
Gull	240	80.79	1	81.66	1
Kingfisher	150	93.33	1	94.67	0
Sparrow	629	91.05	4	92.36	2
Tern	209	75.12	0	83.73	0
Vireo	210	92.46	10	94.47	8
Warbler	750	94.73	13	95.54	14
Woodpecker	179	97.63	0	97.63	0
Wren	209	92.85	5	91.91	5
Average	310	89.01	5.67	**90.78**	**5.00**

TABLE IV
SHOW THE NUMBER (NUM.) OF GENERIC CLASSES, THE PRECISION (PR.) (%) AND THE NUMBER OF FALSE POSITIVES (FP) WITHIN THE FINE-CLASSES IN CUB-200-2011. THE SYMBOL ↑ DENOTES THAT A HIGHER VALUE IS BETTER, WHILE ↓ DENOTES THE OPPOSITE. WE PICK GENERIC CLASSES THAT CONTAIN MORE THAN SIX CATEGORIES.

About the capabilities of different model structures. We further investigate on the "growth" and "decline" of receptive fields and present the five stages of our experiments. First, we tested the original backbone, as shown in Fig.4(a), and the corresponding heat map is presented in Fig.5(b). It was observed that the model paid attention to a large amount of background area, indicating that the original backbone is not designed for detecting details in fine-grained data.

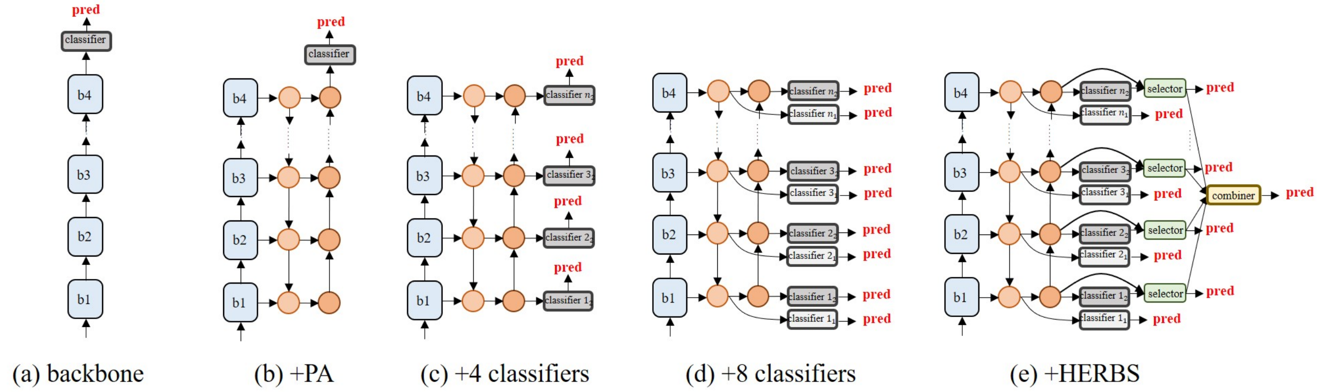
Second, we added the features fusion module PA to the backbone, and the corresponding heat map is presented in Fig.5(c), with the structure depicted in Fig.4(b). From the heat maps, we deduced that using only the last result, the response of the label could be focused on a small region. This improvement compared to the original model, however, still has a narrow focus. Next, we added classifiers to every block of the previous structure, as shown in Fig.4(c). This structure widened the attention area, as shown in Fig.5(d), and effectively utilized multi-scale features.

In the fourth step, we added four more classifiers to the model, as depicted in Fig.4(d). These eight classifiers constrain the attention area, but the accuracy decreased. Finally, we added the HERBS module onto the backbone, and the corresponding heat map is presented in Fig.5(f). This module maintained detail while capturing a wide range of information, with the attention area approximately between Fig.5(d) and Fig.5(e). The results demonstrate that HERBS achieved better accuracy.

In this example, only HERBS predicted the image correctly, while the other models predicted the image to the wrong but visually similar class. This demonstrates that fine-grained visual classification requires detailed features rather than features that are too narrow.

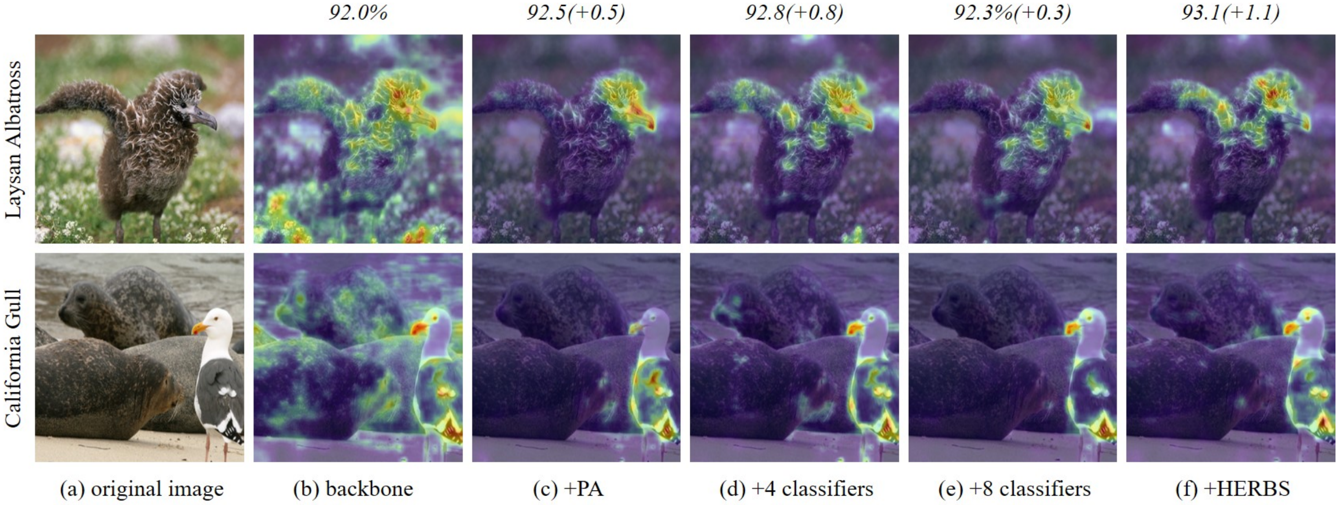
How the performance of HERBS on fine-classes? We evaluated the performance of HERBS on real fine classes in the CUB-200-2011 dataset, which contains about 70 generic classes with 1 to 25 subcategories each, including 9 generic classes with more than 6 categories. Table IV lists these 9 generic classes, and we evaluate the models' performance on them.

The results show that HERBS outperforms the Swin Transformer baseline regarding both precision and false positive



(a) backbone (b) +PA (c) +4 classifiers (d) +8 classifiers (e) +HERBS

Fig. 4. The structure of models, (a) original backbone, the blue box represent the backbone blokcs. (b) backbone + path aggregation module. (c) backbone + PA module with four classifiers on the last bottom-up path. (d) backbone + PA module with eight classifiers on the top-down and bottom-up path. (e) backbone + HERBS



(a) original image (b) backbone (c) +PA (d) +4 classifiers (e) +8 classifiers (f) +HERBS

Fig. 5. Visualization of heat maps generated from different model. (a) original color image, (b) Swin Transformer backbone, (c) backbone + PA, (d) backbone + PA with four classifier, (e) backbone + PA with eight classifiers. (f) backbone + HERBS. The number on the top of the images represents the accuracy of the corresponding model.

(FP) numbers. FP refers to cases where the model predicts a class that does not belong to the correct generic class. A lower FP number means that wrong predictions occur within the similar category, indicating that the model is not making serious mistakes. For example, doctors can trust the results from the fine-grained model and only need to check similar situations, reducing the effort required for double-checking.

How does the BS module suppress the background? We tested the suppression intensity λ_d from 0 to 9 and plotted their corresponding top-1 accuracy in Fig.6(a) (blue line). The corresponding heat maps for different λ_d values are shown in Fig.7. From the heat maps, we observed that when λ_d is set to 0, which means only the merged loss (selected areas) is used to constrain the feature map, the model still pays attention to some background areas. Comparing the heat maps of λ_d at 0, 5, and 9, we observed that the concentration level increases as the suppression intensity λ_d increases, demonstrating that the BS module can effectively suppress background values.

As mentioned before, the tanh function is used to map the classification results instead of the softmax function. In Fig.6(a), the blue line represents the tanh-based approach, while the green line represents the softmax-based approach. In the softmax-based method, the pseudo-target would be $\frac{1}{C_{gt}}$. However, this may lead to unstable training because even though we refer to the noisy or unselected area as "background", it is not necessarily the same as background elements such as the sky, trees, or ocean. Some unselected areas may still appear on the bird's body and could be present in other categories' appearances. Therefore, it is crucial to separate them by feature values rather than class probability. The impact of suppression intensity is shown in Fig. 7.

What does high-temperature provide? We emphasize that the use of high-temperature is based on our experiment results, as shown in Fig.6(b), where we discovered that the best top-1 accuracy and top-3 accuracy occurred at a temperature of 64. A high temperature would cause the distribution to become

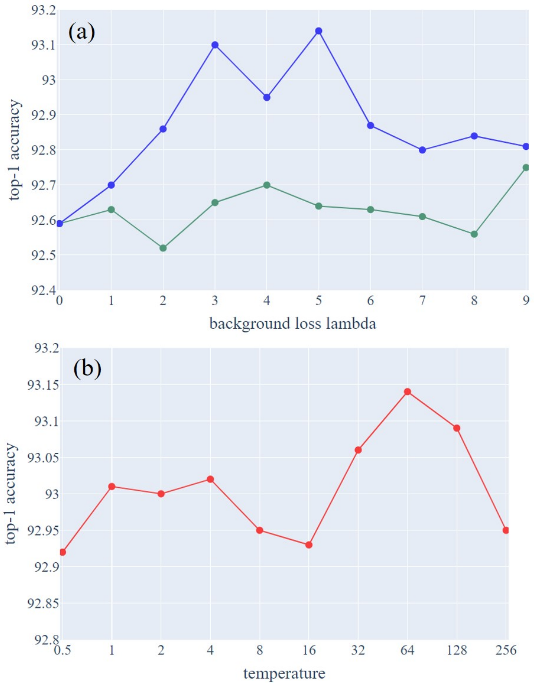


Fig. 6. Comparison of top-1 accuracy with different hyperparameters. (a) shows the top-1 accuracy for different λ_d values ranging from 0 to 9. The blue line represents the use of the tanh function, and the green line represents the use of the softmax function to map the classification results. (b) shows the top-1 accuracy for different temperatures ranging from 0.5 to 256.

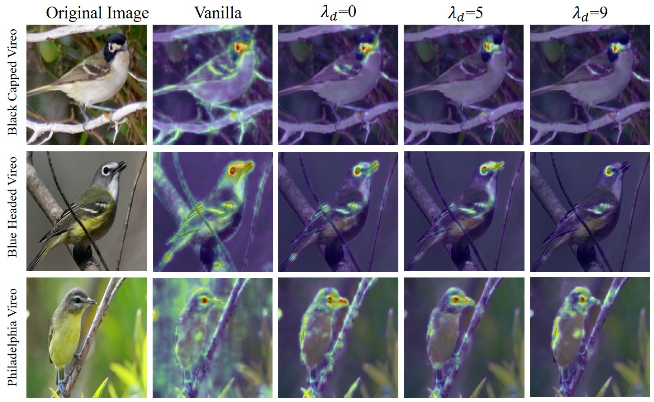


Fig. 7. Visualization of the heat maps for different λ_d values.

very flat, which means that the model has higher tolerance for misclassification. This tolerance allows the model to discover more diverse features and use multi-class features to enhance its capability.

To explain this further, we present the impactness of false-true rates in Fig.8. Here, false-true represents the number of wrong predictions in the top-bottom path but correct in the bottom-up path. The false-true rate is calculated using the following equation:

$$\text{false-true rate} = \frac{\text{false-true}}{(\text{false-true} + \text{false-false})} \tag{15}$$

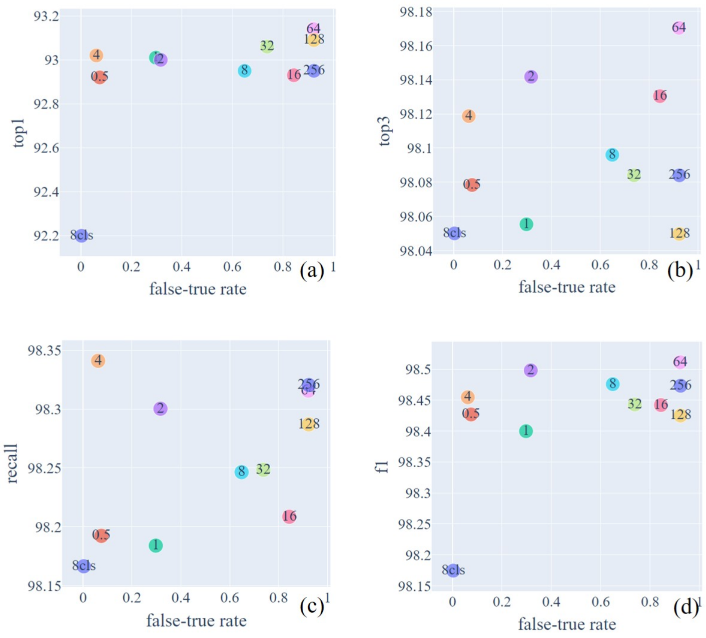


Fig. 8. The False-True rate and its relation with (a) CUB-200-2011 top-1 accuracy, (b) top-3 accuracy, (c) the recall of fine-classes, (d) the f1 score of fine-classes.

The higher false-rate means the model is not only focus on one target. We discovered that the top-1 and top-3 accuracy of the dataset, as well as the F1 score of the fine-classes, are slightly related to this. The structure used in the 8cls dot in Fig.8 is Fig.4(d), while the others are HERBS with different temperatures.

V. Conclusion

In this paper, we proposed HERBS with the BS module and the high-temperature refinement module which can be easily applied to popular backbone networks. The method effectively filters out background noise and focuses on discriminative features while maintaining a proper attention area scale. Our experiments on fine-grained visual classification tasks show that HERBS significantly improves accuracy and outperforms state-of-the-art methods on the CUB-200-2011 and NA-Birds benchmark datasets. Future work can explore the use of adaptive strategies to choose the temperature or suppression intensity and investigate low computation cost methods based on this work. Overall, the proposed HERBS can achieve high accuracy up to 93% to provide a promising solution to improve the performance of fine-grained visual classification tasks.

References

[1] Ardhendu Behera, Zachary Wharton, Pradeep Hewage, and Asish Bera. Context-aware attentional pooling (cap) for fine-grained visual classification. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*. AAAI, 2021. 2, 4

[2] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. Sr-gnn: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, 31:6017–6031, 2022. 2, 4

[3] Steve Branson, Grant Van Horn, Pietro Perona, and Serge Belongie. Improved bird species recognition using pose normalized deep convolutional nets. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 1

[4] Qishuai Diao, Yi Jiang, Bin Wen, Jia Sun, and Zehuan Yuan. Metaformer: A unified meta framework for fine-grained recognition, 2022. 4

[5] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5131–5139, 2016. 2

[6] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6598–6607, 2019. 1, 2

[7] Tuong Do, Huy Tran, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Fine-grained visual classification using self assessment classifier, 2022. 1, 4

[8] E. Gavves, B. Fernando, C.G.M. Snoek, A.W.M. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In *2013 IEEE International Conference on Computer Vision*, pages 1713–1720, 2013. 1

[9] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):852–860, Jun. 2022. 1, 2, 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[11] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 1, 4

[12] Tao Hu and Honggang Qi. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *CoRR*, abs/1901.09891, 2019. 1, 2

[13] Yunqing Hu, Xuan Jin, Yin Zhang, Haiwen Hong, Jingfeng Zhang, Yuan He, and Hui Xue. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4239–4248, New York, NY, USA, 2021. Association for Computing Machinery. 1

[14] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1182, 2016. 1

[15] Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10465–10474, 2020. 1

[16] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei fei Li. L.: Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR (2011*, 2011. 1

[17] Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. ViT-NeT: Interpretable vision transformers with neural tree decoder. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 11162–11172. PMLR, 17–23 Jul 2022. 4

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 1

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 3

[20] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 3

[21] Xinda Liu, Lili Wang, and Xiaoguang Han. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. *Neurocomputing*, 492:137–149, 2022. 1

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 5

[23] Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S. Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8241–8250, 2019. 2

[24] Omkar M Parkhi, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. The truth about cats and dogs. In *2011 International Conference on Computer Vision*, pages 1427–1434, 2011. 1

[25] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1005–1014. IEEE, 2021. 2

[26] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 1

[28] Jinhwan Seo, Wonho Bae, Danica J. Sutherland, Junhyug Noh, and Daijin Kim. Object discovery via contrastive learning for weakly supervised object detection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 312–329, Cham, 2022. Springer Nature Switzerland. 2

[29] Hongbo Sun, Xiangteng He, and Yuxin Peng. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 5853–5861, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2, 4

[30] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 834–850, Cham, 2018. Springer International Publishing. 2

[31] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 595–604, 2015. 2, 4

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[33] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 2, 4

[34] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022. 2

[35] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 170. BMVA Press, 2021. 1, 2, 4

[36] Zhihui Wang, Shijie Wang, Shuhui Yang, Haojie Li, Jianjun Li, and Zezhou Li. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9746–9755, 2020. 1

[37] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *2013 IEEE International Conference on Computer Vision*, pages 1641–1648, 2013. 1

[38] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, pages 1–14, 2023. 4

[39] Shaokang Yang, Shuai Liu, Cheng Yang, and Changhu Wang. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. *CoRR*, abs/2102.09875, 2021. 1

[40] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8291–8299, 2019. 2

[41] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 1143–1152, 2016. 1

[42] Lianbo Zhang, Shaoli Huang, Wei Liu, and Dacheng Tao. Learning a mixture of granularity-specific experts for fine-grained categorization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8330–8339, 2019. 1

[43] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 834–849, Cham, 2014. Springer International Publishing. 1

[44] Tian Zhang, Dongliang Chang, Zhanyu Ma, and Jun Guo. Progressive co-attention network for fine-grained visual classification. *CoRR*, abs/2101.08527, 2021. 1, 2

[45] Yuan Zhang, Jian Cao, Ling Zhang, Xiangcheng Liu, Zhiyi Wang, Feng Ling, and Weiqian Chen. A free lunch from vit: adaptive attention multi-scale fusion transformer for fine-grained visual recognition. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3234–3238, 2021. 2

[46] Yuan Zhang, Jian Cao, Ling Zhang, Xiangcheng Liu, Zhiyi Wang, Feng Ling, and Weiqian Chen. A free lunch from vit: adaptive attention multi-scale fusion transformer for fine-grained visual recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3234–3238, 2022. 1

[47] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet-Anh Nguyen, and Minh N. Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, 25(4):1713–1725, 2016. 1

[48] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5219–5227, 2017. 1, 2

[49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1

[50] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, 2021. 1

[51] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4682–4692, 2022. 1, 4

[52] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13130–13137, Apr. 2020. 1, 2