Method	Arch	CUB-200-2011				Stanford Cars 196				FGVC Aircraft			
		1	2	4	8	1	2	4	8	1	2	4	8
SCDA (**?**)	R50	57.3	70.2	81.0	88.4	48.3	60.2	71.8	81.8	56.5	67.7	77.6	85.7
PDDM (**?**)	R50	58.3	69.2	79.0	88.4	57.4	68.6	80.1	89.4	-	-	-	-
CRL (**?**)	R50	62.5	74.2	82.9	89.7	57.8	69.1	78.6	86.6	61.1	71.6	80.9	88.2
HDCL (**?**)	R50	69.5	79.6	86.8	92.4	84.4	90.1	94.1	96.5	71.1	81.0	88.3	93.3
DGCRL (**?**)	R50	67.9	79.1	86.2	91.8	75.9	83.9	89.7	94.0	70.1	79.6	88.0	93.0
DCML (**?**)	R50	68.4	77.9	86.1	91.7	85.2	91.8	96.0	98.0	-	-	-	-
DRML (**?**)	In3	68.7	78.6	86.3	91.6	86.9	92.1	95.2	97.4	-	-	-	-
CEP (**?**)	R50	69.2	79.2	86.9	91.6	89.3	93.9	96.6	98.1	-	-	-	-
MemVir (**?**)	R50	69.8	-	-	-	86.4	-	-	-	-	-	-	-
IBC (**?**)	R50	70.3	80.3	87.6	92.7	88.1	93.3	96.2	98.2	-	-	-	-
S2SD (**?**)	R50	70.1	79.7	-	-	89.5	93.9	-	-	-	-	-	-
ETLR (**?**)	In3	72.1	81.3	87.6	-	89.6	94.0	96.5	-	-	-	-	-
PNCA++ (**?**)	R50	72.2	82.0	89.2	93.5	90.1	94.5	97.0	98.4	-	-	-	-
Our FRPT	R50	**74.3**	**83.7**	**89.8**	**94.3**	**91.1**	**95.1**	**97.3**	**98.6**	**77.6**	**85.7**	**91.4**	**95.6**

Table 2: Comparison of different methods on CUB-200-2011, Stanford Cars 196 and FGVC Aircraft datasets. "Arch" denotes the architecture of using backbone network. "R50" and "In3" represent Resnet50 (**?**) and Inception V3 (**?**), respectively.

Method	Arch	Params	CUB	Cars	Air
Fine-tuning	R50	23.5M	69.5	84.2	70.1
Our FRPT	R50	2.9M	74.3	91.1	77.6
Fine-tuning	R101	43.6M	70.9	85.1	69.7
Our FRPT	R101	2.9M	75.6	90.4	76.2

Table 3: Comparison of fine-tuning strategy on CUB-200-2011 (CUB), Stanford Cars 196 (Cars) and FGVC Aircraft (Air) datasets. "R50" and "R101" represent Resnet50 and Resnet101(**?**), respectively.

Method	Arch	Params	10-shot	5-shot
Fine-tuning	R50	23.5M	63.1%	59.7%
Our FRPT	R50	2.9M	66.6%	62.8%
Fine-tuning	R101	43.6M	64.7%	61.5%
Our FRPT	R101	2.9M	68.9%	65.2%

Table 4: Recall@1 results on CUB-200-2011 about few-shot learning. 10-shot and 5-shot indicate that only 10 and 5 images per category are used during training, respectively.

Method	R@1	R@2	R@4	R@8
CAM	63.7%	74.3%	82.5%	89.7%
Bounding box	67.6%	79.3%	85.8%	91.6%
Our DPP	74.3%	83.7%	89.8%	94.3%

Table 5: Performance comparison with other prompts in terms of Recall@K on CUB-200-2011.

improvements on three object retrieval datasets, which validates stronger generalization ability of our sample adaptation prompts and feature adaptation head. Besides, when the pre-trained vision model is switched from Resnet50 to Resnet101, our FRPT does not introduce more learnable parameters and takes full advantage of the stronger representation power of larger models, thus resulting in the improvement of retrieval performance again. To better display the positive impact of our FRPT, we visualize the retrieval accuracy and training loss curves in Fig. 2. As can be observed from our FRPT curves, the increasing number of training epochs generally brings slow performance improvement and significantly increases the convergence speeds. One important reason of this phenomenon is that our FRPT only introduces fewer learnable parameters and thus attacks the issue concerning with convergence to the suboptimal solutions.

Effective few-shot learning. To deeper explore the effectiveness of FRPT, we conduct extensive experiments based on the few-shot setting with two different numbers of samples per subcategory: 10 and 5 on CUB-200-2011. Across the 5-shot and 10-shot experimental setting in Tab. 4, our FRPT consistently outperforms the fine-tuning strategy with different pre-trained vision models. Compared with fine-tuning pre-trained models using all images in CUB-200-2011, our FRPT only use 10 samples per subcategory but obtain the proximity performances. Since our FRPT only needs to learn a few parameters and attack the issue concerning with convergence to suboptimal solutions accordingly, our method further achieves better results than the fully fine-tuning strategy when facing only a few training samples. Therefore, the above results demonstrate the outperforming performance of FRPT owning to attaching few but effective parameters into the frozen backbone network.

Fixed prompts vs. Learnable prompts. More insight into the prompt scheme can be obtained by simple switching the processing manner of input images. As can be seen from Tab. 5, switching the processing method from the discriminative perturbation prompt (DPP) to the fixed prompt strategy, *i.e.* directly zooming objects, leads to a significant performance drop. Concretely, we use the class activation map (CAM) or the bounding boxes provided by the annotation information to localize the objects and then crop them

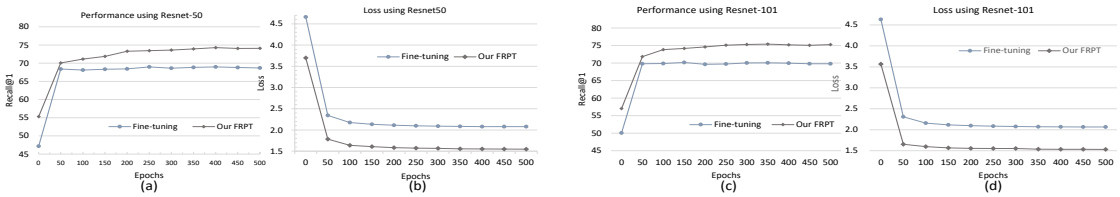

Figure 2: Curve visualization based on CUB-200-2011. (a) (c) denote the recall@1 curves about retrieval performance using Resnet-50 and Resnet-101, respectively. (b)(d) are the loss curves using Resnet-50 and Resnet-101, respectively.

Figure 3: Visualization of the object content perturbation. The first and second rows denote the original and modified images, respectively.

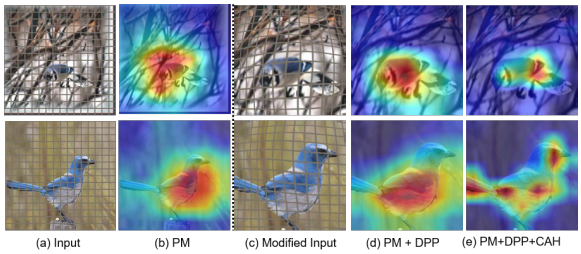

Figure 4: Illustration of class activation maps (CAM). (a)(c) are the original and modified images, respectively. (b)(d)(e) denote CAMs generated by diverse networks.

from the original images. Tab. 5 shows a significant performance improvement when we utilize more accurate localization manner to remove the background information and preserve more object regions as much as possible. However, our FRPT zooms and even exaggerates the idiosyncratic elements contributing to decision boundary rather than to simply amplify objects and remove background, thus making the FGOR task aided by the discriminative perturbation prompt close to the solved task during the original pre-training and forming a steady improvement.

What makes a network retrieve objects visually? With this question in our mind, we exhibit the visualization results of original and modified images in Fig. 3. These visualization images can interpret why and how our approach can correctly identify diverse subcategories. As shown in the second row, our sample prompting scheme can enhance the visual evidence of object parts via the dense sampling operation while suppressing the background and even non-discriminative parts, thus instructing the pre-trained model to pay more attention to discriminative details and improving the retrieval performance accordingly. It should be clarified that we manually put grid lines on the images to better display the pixel shift in the images after our prompt processing. In Fig. 4, in addition to showing the original and modified images, we present the discriminative activation maps of three representation models, i.e, pre-trained model (Fig. 4(b)) , our FRPT without CAH (Fig. 4(d)), and our FRPT (Fig. 4(e)). It is clear that using DPP module can make the network focus on the object rather than background information, thus improving the discriminative ability of feature representation. Compared to Fig. 4(d), the activation maps (e) can pay more attention to the category-specific details via introducing CAH module. Based on these visualizations, our model generates clearer object boundaries and emphasises the discriminative details, thus providing higher retrieve performance.

Conclusion

In this paper, we propose Fine-grained Retrieval Prompt Tuning (FRPT), which aims to solve the issue of convergence to sub-optimal solutions caused by fine-tuning the entire FGOR model. FRPT design the discriminative perturbation prompt (DPP) and category-specific awareness head (CAH) to steer frozen pre-trained vision model to perform fine-grained retrieval task. Technically, DPP zooms and exaggerates some pixels contributing to category prediction, which assists the frozen pre-trained model prompted with this content perturbation to focus on discriminative details. CAH optimizes the semantic features extracted by pre-trained model via removing the species discrepancies using category-guided instance normalization, which makes the optimized features sensitive to fine-grained objects within the same meta-category. Extensive experiments demonstrate that our FRPT with fewer learnable parameters achieves the state-of-the-art performance on three widely-used fine-grained datasets.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant NO. 61976038 and NO.61932020. Neque autem nulla incidunt ratione unde recusandae non, sit tenetur dolorum doloribus fuga consectetur dignissimos voluptates illo, expedita minus iste laudantium voluptatem magni aliquid qui quidem, quod tempore quisquam, iste unde aperiam at optio debitis sint