

**Towards Efficient and Accurate Attention
Learning for Fine-grained Image
Classification**

Dichao Liu

Abstract

The studies on image classification can be divided into two Fine-grained image classification (FGIC) aims to recognize hard-to-distinguish object classes, such as different breeds of birds or models of cars. It is a very difficult task and capturing attention is key for solving the difficulty. The objective of this work is to explore how to efficiently capture attention information to improve the accuracy of FGIC. With this objective, we propose three novel attention-learning frameworks for FGIC. This paper has six chapters.

Chapter 1 gives the background of this research, discusses the motivation of this thesis as well as gives an overview of the proposed approaches.

Chapter 2 introduces the studies that are related to this research or the topic of fine-grained image classification.

Chapter 3 introduces a guided attention-learning framework, named as Attention-Guided Spatial Transformer Networks (AG-STNs), which focuses on capturing effective attention regions for FGIC. Traditional region-based attention learning approaches treat the localization and recognition of attention regions as two separate steps, during which the errors in each step can be accumulated. AG-STNs localize attention regions by deep neural networks, which can be optimized together with the recognition networks. Learning cropping attention regions is very difficult for deep neural networks, and AG-STNs solve the training difficulty by utilizing hard-coded attention as the guiding signal to initialize the localization network. Moreover, AG-STNs can generate multiple scales of attention regions, a fusion of whose predictions further improves the accuracy.

Chapter 4 introduces a multi-task attention-learning framework, named Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN). CRA-CNN is inspired by the human behavior of using the knowledge learned from one task to help learn another related task. During the training, CRA-CNN has two networks.

One is the major network used for the task of categorizing the given input image. The other is the subordinate network used for the task to make the deep features of the major network correspond more to the attention regions. After training, the subordinate network can be removed, and only the major network is kept for utilization. In this way, CRA-CNN does not require extra overhead for utilization and has no loss of information from input images.

Chapter 5 introduces a recursively-refined multi-scale attention framework, named Recursive Multi-scale Channel-spatial Attention Module (RMCSAM). Different from the approaches proposed in Chapters 3 and 4, RMCSAM is an insertable module that has small weights and can be embedded into various backbone networks. RMCSAM explores both channel-wise and spatial-wise attention from deep features, and recursively refines the learned attention information for more accurate attention. RMCSAM is lightweight and has strong versatility, and it can be combined with the Progressive Multi-Granularity Training (PMG), which is the state-of-the-art approach in the FGIC task, to further improve the accuracy. RMCSAM is also possible to combine with other training frameworks.

Chapter 6 gives the summary and prospect of this paper.

Contents

1	Introduction	1
1.1	Background and Motivation	1
1.2	Research overview and Thesis structure	3
1.2.1	Research Overview	3
1.2.2	Thesis Structure and Chapter Relationships	8
2	Related Studies	11
2.1	General Image Classification and Fine-grained Image Classification	11
2.2	Attention Learning	14
2.2.1	Region-based Attention Learning	14
2.2.2	Learning Attention from Deep Features	16
2.3	Machine Learning Techniques	19
2.3.1	Multi-task Learning	20
2.3.2	Contrastive Learning	20
2.4	Other Fine-grained Image Classification Approaches	21
3	Guided Attention Learning	23
3.1	Chapter Overview	23
3.2	Proposed Approach	27
3.2.1	STNs in attention region capturing	27
3.2.2	Attention-Guided STNs	28
3.2.3	Multi-stream AG-STNs	31
3.3	Experiments	33
3.3.1	Dataset and implementation details	33
3.3.2	Evaluation on detail-level attention learning	33
3.3.3	Evaluation on multi-stream AG-STNs	35

3.4	Summary of This Chapter	37
4	Multi-task Attention Learning	39
4.1	Chapter Overview	39
4.2	Proposed Approaches	42
4.2.1	Approach Overview	42
4.2.2	Attention-redundancy Transformer module	43
4.2.3	Multi-task Learning Pipeline	47
4.3	Experiments	47
4.3.1	Implementation details	47
4.3.2	Comparison with the Baselines	48
4.3.3	Ablation Study on different losses.	50
4.3.4	Comparison with Previous Studies	52
4.4	Summary of This Chapter	52
5	Recursively-refined Multi-scale Attention Learning	53
5.1	Chapter Overview	53
5.2	Proposed Approach	56
5.2.1	Multi-scale Channel-wise Attention Sub-modules	57
5.2.2	Multi-scale Spatial-wise Attention Sub-modules	60
5.2.3	Recursive Refinement	63
5.3	Experiments	64
5.3.1	Experimental Settings	64
5.3.2	Ablation Study	65
5.3.3	Comparison with the Baselines	67
5.3.4	Analysis of Attention Capturing	69
5.3.5	Comparison with the State-of-the-art Attention Modules in Fine-grained Image Classification Task	72
5.3.6	Comparison with the Previous Approaches in Fine-grained Image Classification Task	74
5.4	Summary of This Chapter	77
6	Conclusion	79

List of Figures

1.1	Inter-class similarity and intra-class variance	2
1.2	The importance of multi-scale attention regions	4
1.3	An overview of the approach proposed in Chapter 3.	5
1.4	A simplified illustration of the approach proposed in Chapter 4 . .	6
1.5	A simplified illustration of the approach proposed in Chapter 5. .	6
1.6	Relationships of the core chapters.	7
1.7	Changes of the deep-feature-based attention learning mechanism between Chapter 4 and 5	9
2.1	Illustration of generic and fine-grained image classification.	12
3.1	Hard-coded approaches	24
3.2	Spatial Transformer Networks and the proposed Attention-Guided Spatial Transformer Networks	26
3.3	The computation of guiding signal	30
3.4	Attention regions of different scales	32
3.5	Attention regions captured by different methods	34
3.6	Initial attention regions after the regressive guiding	35
3.7	The evolution of attention regions during the training	36
4.1	Motivation of Chapter 4	40
4.2	Contrastively-reinforced Attention Convolutional Neural Network	43
4.3	Visualization results of the baseline and CRA-CNN	49
4.4	Examples of transformed images obtained by CRA-CNN and STN	50
5.1	The main idea of the Chapter 5	55
5.2	Recursive Multi-scale Channel-spatial Attention Module	58

5.3	Visualization results of the network with/without RMCSAM . . .	70
5.4	Attention visualization of the images containing no target objects.	71
5.5	Attention precision with different thresholds	71

List of Tables

3.1	Size configuration for different levels	34
3.2	Comparison on the recognition performance between different approaches for exploiting detail-level attention	34
3.3	Results of multi-stream AG-STNs	36
3.4	Comparison with previous studies on CUB_200_2011	37
4.1	Comparison results with baselines.	49
4.2	Ablation study on different losses.	51
4.3	Comparison results on Stanford Cars.	51
4.4	Comparison results on CUB-200-2011.	51
5.1	Results of the ablation study	66
5.2	Comparison results with baselines	68
5.3	Comparison results with state-of-the-art attention modules in FGIC task	73
5.4	Comparison results with state-of-the-art approaches in FGIC task .	76

Chapter 1

Introduction

1.1 Background and Motivation

The studies on image classification can be divided into two sub-fields: generic image classification and fine-grained image classification (FGIC). Generic image classification aims to differentiate between distinctively different objects, such as birds and vehicles. In comparison, FGIC aims to differentiate between hard-to-distinguish object classes, such as different breeds of birds or models of cars.

As a fundamental, meaningful, and challenging subfield of image classification, fine-grained image classification (FGIC) has attracted much attention in recent years. However, FGIC is a very challenging task, and the challenges are principally related to two characteristics of its own: inter-class similarity and intra-class variance. As shown in Figure 1.1, inter-class similarity means the images of different categories of an FGIC task may look visually similar, and thus it is very difficult to learn discriminative features for distinguishing the images of different classes. Intra-class variance means the images of the same category of an FGIC task may look visually different because of different conditions, such as different poses, shooting angles, illuminations, etc. Thus it is very difficult to learn comprehensive and representative features for the images of the same class. Inter-class similarity and intra-class variance can easily confuse the classification models and heavily harm the classification accuracy.

Many previous studies have shown that accurately identifying visual attention (i.e., discriminative visual information) is the key to mitigate the adverse effect

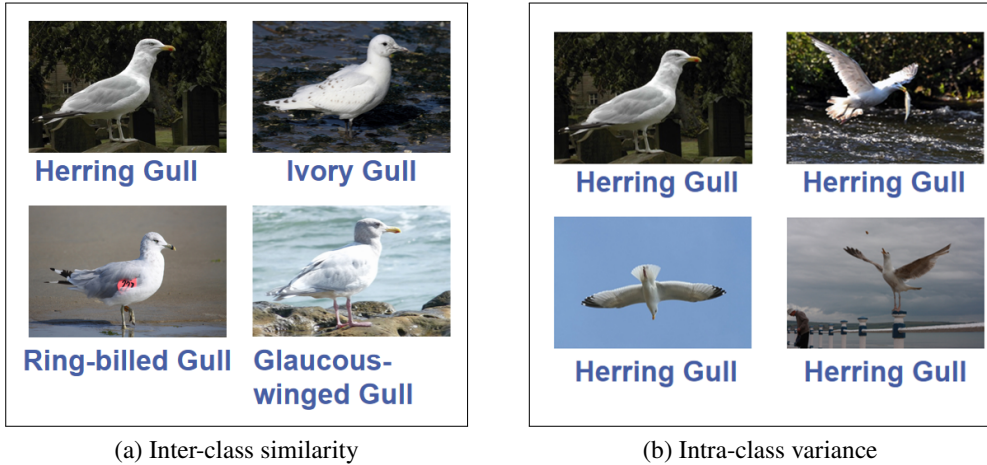


Figure 1.1: Example images showing inter-class similarity (a) and intra-class variance (b). (a) shows some images that are from different categories but look similar because of the same shooting angle. (b) shows some images that are from the same category but look different because of different poses, shooting angles, illuminations, etc.

caused by inter-class similarity and intra-class variance [33, 21, 96, 154, 42, 134, 28, 106, 130, 49, 69, 90, 139, 142] for the FGIC task. Visual attention is inspired by the fact that the human brain effectively filters out the majority of incoming visual information before it goes to the deeper levels of the brain [87]. For example, the human visual system intuitively focuses on certain important regions of the image, while ignoring other irrelevant regions [132]. Similarly, for the image classification task, attention learning implements this notion of importance by allowing the models to dynamically pay attention to only certain regions of the input image that help in improving the classification accuracy [7].

However, attention learning has heavy difficulties. First, attention learning is a hard-to-train task for deep neural networks. Second, attention learning always introduces heavy extra overheads. The above-mentioned problems result in the following research question that we try to address in this thesis: *how to efficiently capture and utilize accurate attention information to improve the classification accuracy in fine-grained image classification?* This thesis looks into this question by proposing novel deep-learning frameworks specializing in learning attention in efficient ways to improve the accuracy of FGIC. As discussed above, the key atten-

tion in FGIC is always subtle, and the position is difficult to predict. Consequently, the proposed frameworks should be able to knock over the difficulty to capture the key attention. Also, the framework should not introduce too much extra effort or overhead, which is unfavorable for applications.

1.2 Research overview and Thesis structure

1.2.1 Research Overview

Learning key attention has always been the leitmotif in FGIC and attracted great attention and effort from researchers. However, there is still plenty of room for research on how to capture key attention accurately and efficiently. As mentioned, the research question is *how to efficiently capture and utilize accurate attention information to improve the classification accuracy in fine-grained image classification*. Specifically, capturing attention information with deep learning approaches suffers from two serious difficulties. The first is hard to get accurate attentional information, and inaccurate attention information may harm the classification performance. The second is that learning attention information always introduces many extra overheads. In this thesis, we propose three frameworks, namely guided attention learning (Chapter 3), multi-task attention learning (Chapter 4) and recursively-refined multi-scale attention learning (Chapter 5) to address these difficulties. Thereinto, Chapter 3 mainly addresses the first difficulty. Chapter 4 and Chapter 5 mainly address the second difficulty.

Finding attention information is a difficult task for deep neural networks. A lot of existing studies capture key attention information relying on localizing and cropping attention regions [37]. Typically, the attention regions are firstly localized using manual regional annotations or conventional hand-crafted features. Then the localized attention regions are cropped and categorized [130, 135, 121]. However, such strategies are not only troublesome but also disconnect the localization and classification. That is, the errors during each step can be accumulated. The advent of deep neural networks [108, 40, 41, 111, 22] brings the possibility of connecting the steps of localization and classification and optimizing the two steps together. However, learning and cropping regions are very difficult for deep neural networks. It is because the early-stage noise causes huge errors, which is irreversible in later

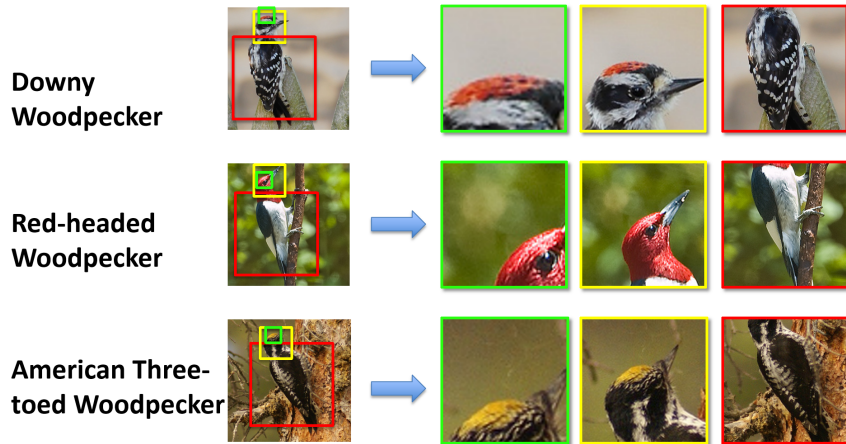


Figure 1.2: Examples of multi-scale attention regions for the images of different woodpeckers. Different scales of attention regions can capture different objects, such as nape, head, and body. All the information is important for distinguishing different woodpeckers. For example, Downy Woodpecker has a red nape. Red-headed Woodpecker has a bright-red head. American Three-toed Woodpecker has a black and white barred back and white breast.

stages [52, 71, 105]. For the first proposal, we propose a novel guided attention learning framework, named Attention-Guided Spatial Transformer Networks (AG-STNs), to solve the training difficulty of using deep neural networks to localize attention regions.

AG-STNs first use conventional hand-crafted features [141] to learn hard-coded attention regions and then use the hard-coded attention regions as the guidance to initialize deep neural networks for preventing the early-stage noise. Thereafter, the deep neural networks are trained to optimize the region localization by themselves. By doing so, the early-stage noise can be successfully avoided and the deep neural networks are able to find informative regions during the training. Furthermore, the scale of the hard-coded attention region can be set differently to guide the cropping of different scales of deep attention regions. As shown in Figure 1.2, attention regions of different scales are important. Especially, detailed regions can capture the subtle yet key attention in FGIC and play an important role. Moreover, multi-scale attention regions can provide complementary information. Thus, we learn multi-scale attention from the input image (as shown in Figure 1.3), and fusing the multi-scale attention information is proved to further improve the accuracy (see

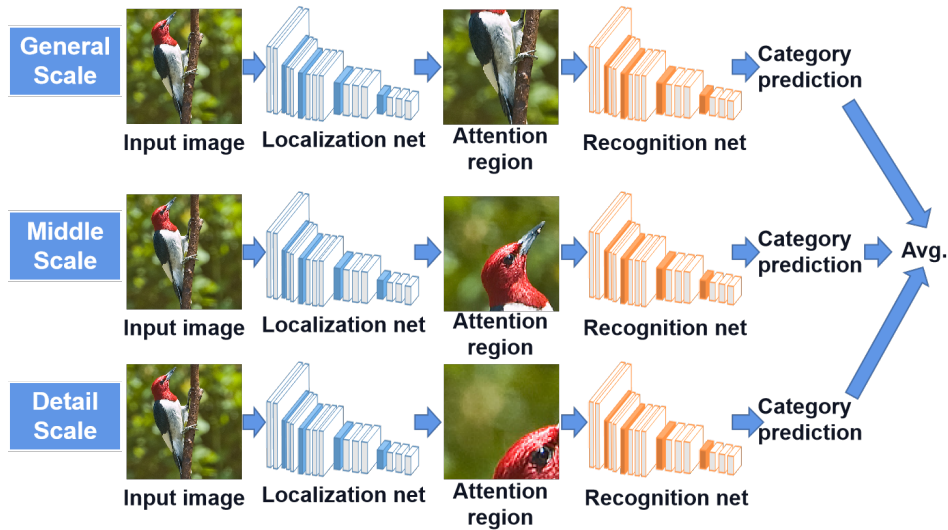


Figure 1.3: An overview of the approach proposed in Chapter 3.

more details about AG-STNs in Chapter 3).

Another problem for obtaining attention information with deep neural networks is heavy extra overheads. Region-based attention learning strategies have two inherent drawbacks: (a) cropping multi-scale attention regions introduces extremely heavy extra overhead; (b) the region-cropping strategy inevitably causes some loss of visual information (because the regions not included in the attention region are abandoned). Targeting these two drawbacks, we consider implementing attention in the feature level rather than the image level, for the second proposal. Namely, instead of directly cropping regions, we try to capture attention information from deep features and utilize the attention information to refine the deep features to better respond to attention regions. With this idea, we propose a novel multi-task attention learning framework, named Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN), based on the multi-task learning strategy. As shown in Figure 1.4, CRA-CNN treats the task of cropping attention region as an additional task and utilizes the additional task to improve the task of classification, which is the main task of CRA-CNN. We enable the additional task to automatically adjust the scale and location of the attention region since the key attention in FGIC is always subtle and has an uncertain position. CRA-CNN has to improve the awareness of attention in order to complete the additional task. The main task

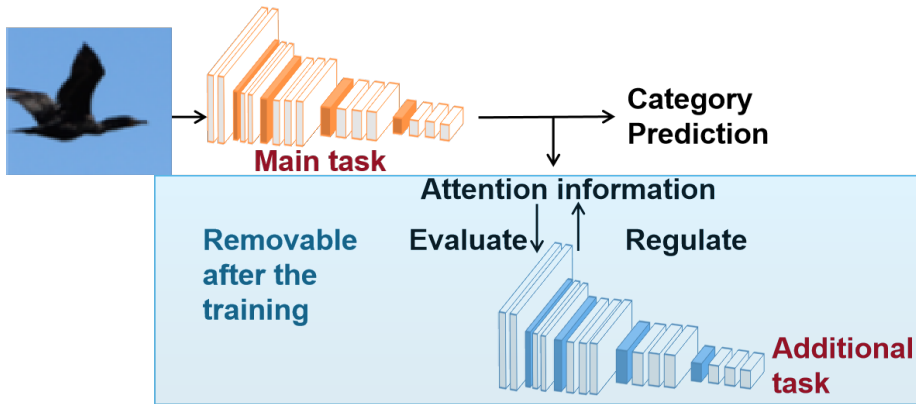


Figure 1.4: A simplified illustration of the approach proposed in Chapter 4. The shaded parts can be removed after training.

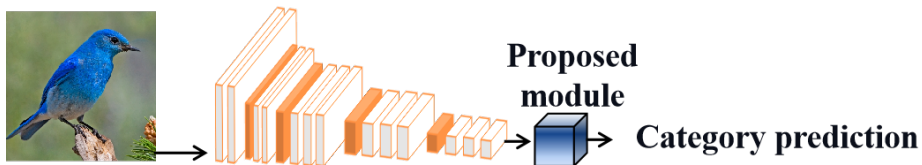


Figure 1.5: A simplified illustration of the approach proposed in Chapter 5.

of CRA-CNN takes the whole input image as input so that there is no loss of visual information. Moreover, after the training, the additional task can be removed, and thus CRA-CNN requires a small overhead for utilization (**see more details about CRA-CNN in Chapter 4**).

CRA-CNN mainly reduces the problem of heavy extra overhead for testing. To further reduce the extra overhead during the training procedure, we turn our eyes to the studies on attention modules, which are insertable deep neural modules that can explore attention information inside the networks and refine deep features according to the explored attention information. Existing attention modules are generally designed for generic image classification and do not have a good performance in the FGIC task. We suppose it is because existing attention modules only explore the suitable scale of attention information in the generic image classification task. In this thesis, we propose a novel attention module, named Recursive Multi-scale

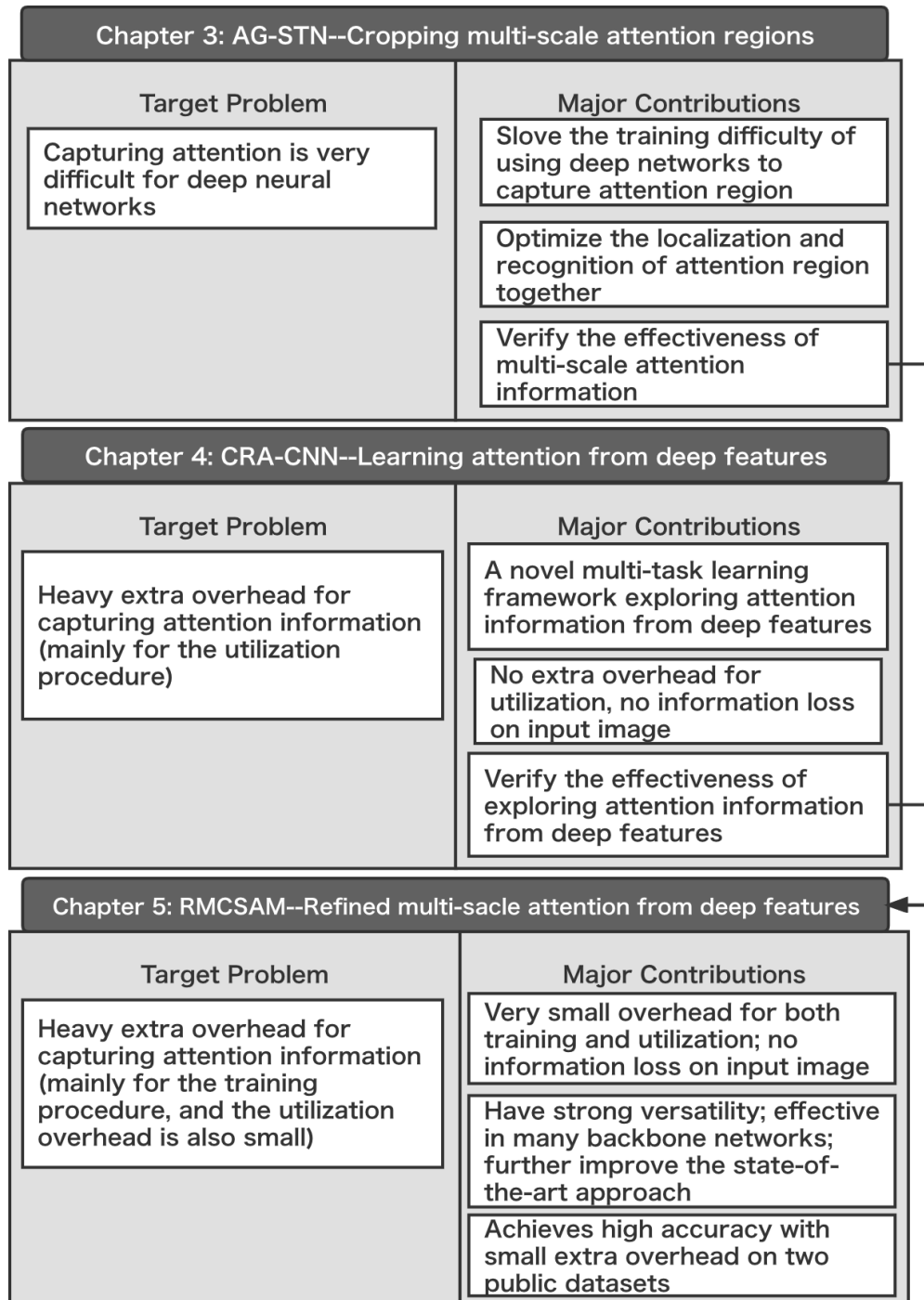


Figure 1.6: Relationships of the core chapters.

Channel-spatial Attention Module (RMCSAM), for the third proposal. RMCSAM is able to explore both channel-wise and spatial-wise deep attention of multiple scales, which can meet the requirements of capturing the key attention in FGIC very well (**see more details about RMCSAM in Chapter 5**). RMCSAM can explore both spatial-wise and channel-wise attention information. The spatial-wise deep attention is the attention extended from the region-based attention strengthening strategy used in CRA-CNN (Chapter 4). The new mechanism can capture the attention corresponding to multi-scale regions with different weights if necessary. Moreover, as the effectiveness of multi-scale attention information is verified in Chapter 3, we make RMCSAM to capture multi-scale attention information with different scales of parameter kernel sizes in Chapter 5. As shown in Figure 1.5, the approach proposed in Chapter 5 is a lightweight and insertable module, which increases small overhead.

1.2.2 Thesis Structure and Chapter Relationships

This thesis contains six chapters. Chapter 1 gives an overview of the background of this research, discusses the motivation of this thesis as well as gives an overview of the proposed approaches. Chapter 2 introduces the studies that are related to this research or the topic of fine-grained image classification. Chapters 3~5 respectively introduce the three attention learning approaches proposed in this thesis in details, namely Attention-Guided Spatial Transformer Networks (AG-STNs), Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN), and Recursive Multi-scale Channel-spatial Attention Module (RMCSAM). Lastly, Chapter 6 concludes this thesis by reviewing the research contributions and results found through the thesis.

Chapters 3~5 are the core chapters of this thesis. The relationships of the core chapters are shown as Figure 1.6.

Chapter 3 introduces the AG-STN, which is a guided attention learning framework exploring attention from original input images. AG-STN solves the training difficulty of using deep neural networks to capture attention regions by initializing the networks with the guidance of traditional hard-coded approaches. By doing so, the localization can be performed by the initialized network, which can be concatenated with the network performing recognition. That is, the localization and

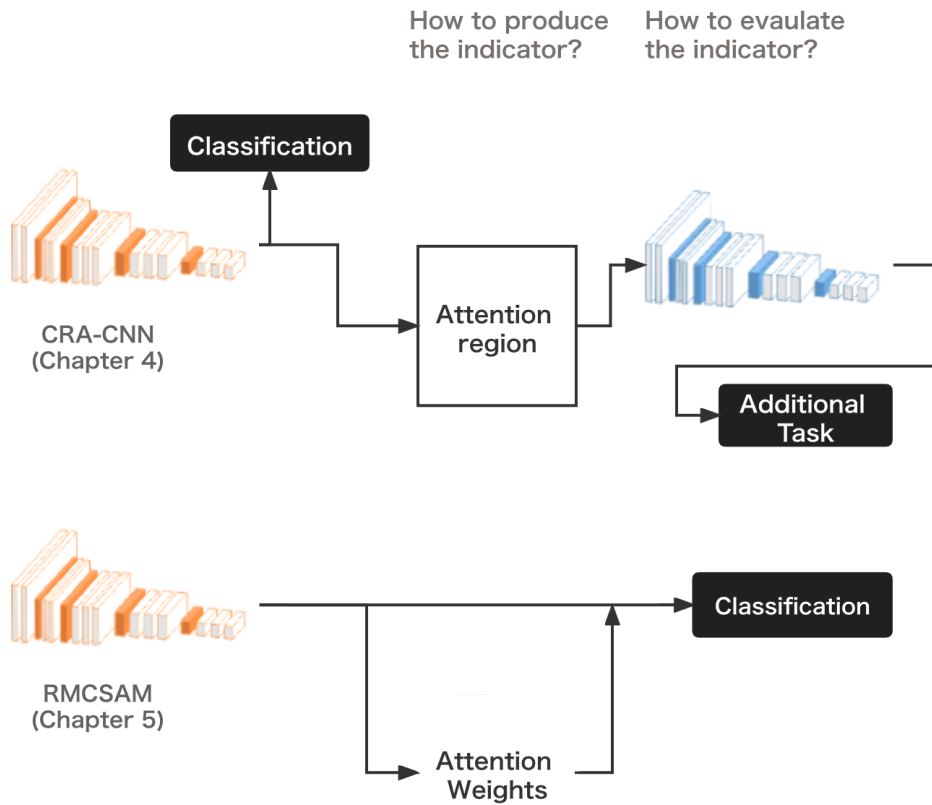


Figure 1.7: Illustration of the changes of deep-feature-based attention learning mechanism between the approaches proposed in Chapter 4 and Chapter 5.

recognition of attention regions can be optimized together. AG-STN can also be used to capture attention regions of multiple scales, which provide complementary information and further improve accuracy.

Chapter 4 introduces CRA-CNN, which is a multi-task attention learning framework targeting to solve the drawbacks of extra overhead during the testing procedure. In Chapter 4, instead of cropping regions on input images, we focus on exploring attention information from deep features, which means we try to increase the dependence on the features responding to attention information

and decrease the dependence on the features responding to non-attention information. In this way, there is no irreversible loss of information on the input image. Specifically, for training, CRA-CNN requires two backbone networks to perform the classification task and attention strengthening task, respectively. After the training, CRA-CNN only requires a single network backbone for utilization.

Chapter 5 introduces the RMCSAM, which is a novel attention module proposed by summing up the experience of Chapter 3 and Chapter 4 to solve the drawbacks of extra overhead during the training procedure. Chapter 3 verifies that multi-scale attention information is effective for FGIC. Chapter 4 verifies that exploring attention information from deep features is effective and very efficient. In Chapter 5, we try to explore multi-scale attention information from deep features with small training overheads.

In the approaches exploring attention from deep features, the training overhead is mainly influenced by two factors: (a) how to produce the indicator presenting the attention awareness of deep features; (b) how to evaluate the indicator and strengthen the attention awareness of deep features based on the evaluation.

As shown in Figure 1.7, in Chapter 5, the deep-feature-based attention learning mechanism is improved in terms of both factors. Instead of strengthening attention awareness by another network as the strategy in Chapter 4, we generate attention weights from the deep features and multiply the attention weights back to the deep features. Chapter 4 uses region prediction as the indicator to present the attention awareness of deep features. The indicator is evaluated by another network. Chapter 5 simply uses the attention weights as the indicator. The indicator does not only act as the indicator presenting the attention awareness of deep features but also involves forming the final prediction score. Thus, the indicator is simply evaluated by the classification loss. In this way, compared with the deep-feature-based attention learning strategy in Chapter 4, the deep-feature-based attention learning strategy in Chapter 5 reduces the training overhead a lot. This strategy can replace the multi-task learning strategy in Chapter 4 to achieve similar effects. Moreover, it can be easily extended to capture multi-scale attention information, following the experience in Chapter 3.

Chapter 2

Related Studies

2.1 General Image Classification and Fine-grained Image Classification

Image classification, which refers to the labeling of images into a fixed set of categories, is a core problem in computer vision. Image classification can be divided into two sub-fields, generic image classification [19] and fine-grained image classification (FGIC) [116, 63]. Generic image classification aims to classify distinctively different categories, such as birds, cars, etc., whereas FGIC aims to classify subordinate categories within an entry-level category, such as different species of birds, different models of cars, etc.

The research of FGIC has very important social significance. The most significant value of FGIC is that it aims to achieve a much stronger recognition ability than human beings. For example, it is impossible for an ordinary person to distinguish various bird species without long-time specialized training, whereas FGIC provides the capability to distinguish the species easily and quickly. With the recognition ability far beyond human brain's ability, FGIC provides the basic technology of a wide range of applications. For example, FGIC is the basic technology of biodiversity monitoring systems [101, 34], which are important for observing some global issues such as climate change [88, 38]. Also, FGIC can be applied for commercial use, such as counting the number of cars of a certain model on the highway [56, 68].

The two sub-fields of image classification, generic and fine-grained image clas-

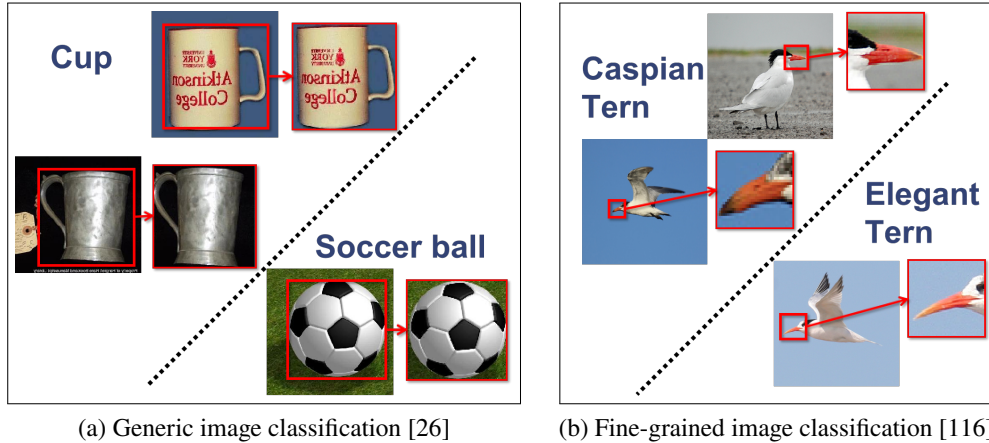


Figure 2.1: Some examples of a generic image classification task (a) [26] and an FGIC task (b) [116]. In the images of the generic image classification task, the key visual clues lie in the whole class-specific object. In the images of the FGIC task, the key visual clues lie in certain local regions of the class-specific object. For example, a long, drooping orange bill is the key feature of Elegant Tern [86]. Consequently, the bill is the key region to distinguish Elegant Tern from other similar birds, such as Caspian Tern, which has a thicker and redder bill than Elegant Tern. Also, the bill is the key region to estimate the Elegant Tern photos taken from different shooting angles to be in the same category.

sification, have different needs for attention learning. It is because that the classification clues for the generic image classification task lie in the whole class-specific object, but the key clues for the FGIC task always only lie in certain local regions of the class-specific object. As shown in Figure 2.1 (a), the class-specific objects of different categories of a generic image classification task are distinctively different. That is, the whole class-specific objects contain discriminative visual information. Consequently, the attention learning for generic image classification tasks usually means differentiating the whole class-specific object from the background, especially in early studies [84]. Typically, some earlier researchers, such as Gao et al. [29], Borji et al. [4], Ren et al. [98] and etc., tried to use visual saliency as a criterion to filter out the objects from the backgrounds and extract hand-crafted features (e.g., SIFT [74]) from the objects. Later, with the huge success of the convolutional neural network (CNNs), deep learning approaches became the mainstream tools for generic image classification [108, 40, 41, 111, 22, 22, 58, 136, 37]. Meanwhile, the

strategy that filters out the class-specific objects by using visual saliency became infrequent [84] because CNNs themselves have a degree of capability to localize class-specific objects [155]. With heavy extra overheads, filtering out the class-specific objects can hardly bring a good improvement over the powerful network architectures [108, 40, 41, 111] in terms of accuracy and become unnecessary. Nowadays, for generic image classification, most researchers implement attention inside the network architecture by assigning weights to different elements inside the network [22, 117, 85, 136]. For example, self-attention has attracted much interest because of its strong ability to capture long-distance information, meaning that it can easily derive global information [22]. As the global information of the class-specific object is the key to generic image classification, there is a wave of developing new network architectures based on self-attention [22, 58, 136, 37]. However, recently, there are also some researchers doubting whether attention is necessary for generic image classification. For example, Luke Melas-Kyriazi, a researcher from Oxford University, argued that attention is not the main reason responsible for the good performance of those self-attention-based architectures [78]. He replaced the attention layer of some self-attention-based architectures with simple feed-forward layers, and the accuracy for the generic image classification task did not decrease very much.

While the usefulness of attention in the sub-field of generic image classification has become to-some-extent controversial, attention learning is always the dominant theme in the FGIC sub-field [33, 21, 96, 154, 42, 134, 28, 147, 106, 114, 130, 49, 69, 90, 139, 142]. Moreover, the needs for attention learning in the FGIC sub-field are more complicated than the needs in the generic image classification sub-field. Generally, FGIC needs to explore attention in certain local regions of the class-specific object because not the whole object can provide useful information [157, 6, 33, 21, 96, 93, 154]. As shown in Figure 2.1 (b), due to inter-class similarity, many parts of the class-specific object might mislead the model while only certain key parts give the difference between the different categories. Also, due to intra-class variance, some same-category objects look different regarding most of the object parts because of different conditions, such as shooting angle. Only some key parts that are invariant to different conditions can help the model to recognize those objects.

Therefore, it is very important to capture the key attention for the FGIC task.

As the key regions always only occupy a small proportion of the whole scene and the position in the scene is uncertain, attention learning is a very difficult problem in the FGIC task. Moreover, in the FGIC task, the background objects can sometimes mislead the models, which further makes FGIC more difficult. For example, the bird called Palm Warbler generally hangs out in the understory of forests, and thus Palm Warbler always shows with understory plants in photos. The understory plants in the photos not only make the bird self-occluded but also can be wrongly regarded as the clues for recognizing Palm Warbler. Then, there will be classification mistakes when other species of birds happen to hang out in the understory.

Along with the hardness, another problem of learning attention for FGIC is it always introduces much extra effort and overhead. For capturing and utilizing attention information, many FGIC studies utilize extra manual bounding boxes or part annotations to localize attention regions, which improves the classification accuracy but is labor-intensive and limits the practicality of real-world applications [49, 69, 90, 139, 142]. Some other studies localize attention regions with weakly supervised localization schemes[52, 33, 21, 96, 93, 154, 42, 134, 28], which largely increase the computational overheads.

2.2 Attention Learning

2.2.1 Region-based Attention Learning

At a high level, FGIC can be regarded as a sub-field of the recognition task of computer vision (CV), and thus has been strongly influenced by the studies of visual recognition, including shallow approaches [64, 118, 119, 95, 92] and deep approaches [108, 40, 57, 65, 113, 123, 107, 120, 27]. Especially the deep ones have gained dramatical improvement in recent years. For example, in the generic image classification sub-field, convolutional neural networks (CNNs) [108, 40, 41, 111, 22, 22, 58, 136, 37] make a well-known success.

However, compared with other visual recognition tasks, such as generic image classification [19, 43], FGIC requires much effort to find and learn the key attention regions of the scenes rather than directly recognizing the entire scenes or just differentiating the class-specific objects from the background, as introduced in

Chapter 1. Exploring key attention is always the main theme in FGIC tasks, and many studies mainly rely on manual object bounding boxes or part annotations. For example, Xie *et al.* [130] propose to utilize the manual object bounding boxes to obtain image segmentation and give a descriptive image representation by building mid-level structures on the segmented regions. However, collecting manual annotations is time-consuming, labor-intensive and not feasible for real-world applications. To avoid this problem, hard-coded attention is proposed to localize the attention regions [84].

Hard-coded attention generally selects the attention regions before learning them. The selection is always implemented by solving a certain statistical problem, which strongly relies on human’s expert knowledges [3]. For example, with the experience that the regions with high visual saliency always contain key attention information, many researchers develop different computational saliency models to select the key attention regions [84, 83, 25, 140, 44, 45, 1, 94, 35]. The selected regions are then learned by recognition algorithms.

Hard-coded attention has some problems. The studies based on hard-coded attention divide attention region localization and attention region classification into two separate steps. The errors that occurred in each step can be accumulated finally. Moreover, the prior knowledge, on which the hard-coded attention is based, may sometimes be wrong and not lead the algorithms to find the most effective regions. The pre-defined region cropping strategy cannot be adjusted conditioned on specific datasets.

In this thesis, we propose novel Attention-Guided Spatial Transformer Networks (AG-STNs) to break the barrier between the attention region localization and attention region classification and optimize the two steps together (Chapter 3). AG-STN can be regarded as an attention-guided variant of Spatial Transformer Networks (STNs) [52]. The Spatial Transformers of the STNs can apply multiple transformations on the inputs. There is the theoretical possibility for the transformation to make the transformed images to be attention regions of the input images. However, outputting attention regions is a very difficult task for the Spatial Transformers. In this work, we add guidance signals, which are computed from hard-coded attention regions, to solve the training difficulty of Spatial Transformers. The hard-coded attention regions are obtained with the saliency map based on Minimum Barrier Distance (MBD) Transform [141]. By doing so, we successfully

solve the training difficulty and connect the attention region localization and classification steps. Both steps are optimized together for reducing the classification loss. That is, the localization of regions is adjusted for better accuracy.

The research [97] and [28] are both concerned with the learning of deep-learned attention on raw inputs. [97] proposed a non-uniformed sampling scheme on raw image inputs by using as guidance the saliency maps that are generated from a CNN. Conditioned on the saliency maps, [97] amplifies the regions that respond to more saliency. Compared with [97], our proposed work in Chapter 3 is different mainly in the following aspects: (i) our work uses saliency maps as an initial guidance signal and turns off such a guidance in the later stage. The purpose of doing so is to reduce the prior hypothesis bias caused by the saliency, in case the saliency learning scheme cannot bring the best recognition results. (ii) the sampling of [97] strongly relies on distorted zoom while our work is mainly a localization with slightly transformation for alignment. In our work, the spatial correlation of objects within same attention is still preserved. [28] recurrently locates the attention regions from coarse level to finer level. However, [28] only allows localization and scaling on input images while our work also allows other transformations (e.g., rotation, distortion, etc.) with Spatial Transformer module. Therefore, our work is able to align the captured regions.

2.2.2 Learning Attention from Deep Features

The approach proposed in Chapter 3 solves the training difficulty of cropping attention regions by deep neural networks and achieves the goal of learning and classifying attention simultaneously. However, there are drawbacks of heavy training and utilization overhead and loss of certain possibly useful information. Thus, we consider capturing attention from deep features inside the deep neural networks instead of directly cropping regions on input images.

Learning attention information from deep features has also been studied by previous researchers. Sharma *et al.* [103] takes the $7 \times 7 \times 1024$ - D feature cubes from CNNs as the inputs of their LSTM-based attention model. Jin *et al.* [55] propose to use a model containing two-stream networks, which respectively learn attention information with fast Matrix Power Normalized Covariance Pooling [66] and part feature matrix. Zhou *et al.* [156] propose a region selection model based

on the saliency constraint conditioned on the deep features from a CNN fed with the original input image. Then, another CNN is used to extract pyramidal features from the selected regions and the extracted pyramidal features are used to improve the attention capturing of the feature learned from the raw input.

Those approaches take full images as the input and explore attention information inside the deep neural networks, which has no information loss in the input procedure. However, those approaches still require the embedding of various backbone networks, such as CNNs+LSTMs [103, 33], multi-stage or multi-stream CNNs [156, 55].

To solve this problem, in this thesis, we propose a novel multi-task attention learning framework, named Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN), to improve the attention awareness of deep features (Chapter 4). CRA-CNN treats region capturing (including cropping, zooming and aligning) as an additional task to improve the main task, i.e., image classification. During the training procedure, our architecture looks somehow similar to some region-based attention learning architectures, such as [52]. However, the main idea of our work is to use the designed losses occurred by local regions to improve the attention awareness of the major network, rather than cropping local regions to replace the original images. Thus, our work explores the whole visual information of each image for classification. After the training, the network used for the additional task can be removed. Thus, CRA-CNN does not introduce extra overhead for utilization.

Recently, There emerge some studies learning attention information by self-supervised learning strategies for FGIC [5, 59, 127, 138]. Self-supervised learning strategies are a set of strategies that capture attention regions without using ground truth but generating some pseudo signals for supervising the training. The training of capturing attention regions is used as the pretext task, and the features learned from the pretext task are then transferred into the downstream task, namely classification. For example, Breiki *et al.* [5] propose to use three kinds of tasks as the pretext tasks: Jigsaw solving, adversarial learning [73], and SimCLR model [8]. In [5], Breiki *et al.* argue that the model trained with the pretext tasks will improve its attention, which benefits the downstream classification task. For example, in the Jigsaw solving pretext task, given the Jigsaw-transformed images, the model is trained to reconstruct the original image. During this process, the model can learn

the semantic meanings of different small regions.

Both self-supervised learning approaches and the approach proposed in Chapter 4 solves multiple tasks. However, self-supervised learning approaches have two separate steps. That is, they have to first complete the pretext task and then the downstream task [5, 59, 127, 138], which takes more work. The two steps are disconnected and cannot help to optimize each other. Moreover, some pretext tasks, such as Jigsaw solving, can introduce extra noises (e.g., the randomly shuffled local regions), which harm the classification performance [10]. The proposed approach in Chapter 4 is based on the multi-task learning strategy rather than the self-supervised learning strategy. In our work, the multiple tasks are solved simultaneously.

INSERTABLE ATTENTION MODULES. The proposed approach in Chapter 4 reduces the utilization overhead for attention learning. However, the capturing and utilizing of attention information still requires much overhead during the training procedure. Targeting this drawback, we turn our eyes to the recent studies of attention modules in the generic image classification sub-field, which can be inserted into backbone networks easily and introduce very small overhead.

Attention modules are designed to make CNNs learn to focus on the important information and ignore unuseful information by imitating the human visual attention mechanism [46, 126, 18]. Humans tend to process an image by regarding it as a sequence of partial glimpses and selectively concentrate on informative parts, instead of processing a whole scene at once. Inspired by this fact, there have been emerging efforts to incorporate attention modules into CNNs for improving classification accuracy in large-scale classification tasks, such as ImageNet [19].

Attention modules can explore two types of attention information, namely spatial-wise attention and channel-wise attention. The spatial-wise explores the attention information among the different spatial locations of the deep features. The channel-wise explores the attention information among the different channels of the deep features. The existing attention modules generally consist of some pooling layers, 2D convolutional layers, FC layers, and a sigmoid function at the end to generate a mask of the input feature map. For example, the SE module [46] squeezes global spatial information with 2D-pooling and excites the squeezed information into a set of channel weights to capture channel-wise dependencies.

The success of the SE module is succeeded by many studies. CBAM [126] uses a similar idea to the SE module to capture channel-wise attention and introduces spatial-wise attention encoding implemented by 2D-convolutional layers with large-size kernels. Dai *et al.* [18] propose channel-wise attention in multiple scales by varying the spatial pooling size.

The existing attention modules are designed for generic image classification and focus on exploring single-scale attention information [126, 46] or/and single-type attention information [18, 46], which is not enough for finding the subtle and location-unpredictable key attention for FGIC. In this thesis, we propose the Recursive Multi-scale Channel-spatial Attention (RMCSAM) for FGIC (Chapter 5). Different from the above-mentioned attention modules, our module can explore multi-scale attention of the input feature maps in both channel-wise and spatial wise. The multi-scale channel-wise attention in our work is implemented by using different numbers of the hidden units within the channel-wise sub-modules, which makes it different from the multi-scale channel-wise attention proposed in [18]. FC layers of different numbers of the hidden units can compress the features into different scales [133], the compressed features can then be used to generate multi-scale channel-wise dependencies. In this way, our work requires less overhead than [18] to explore multi-scale channel-wise attention. Besides, [18] only explores channel-wise attention and cannot explore the location information of attention. Moreover, unlike the above-mentioned attention modules, our module recurrently refines the features a predetermined number of times before outputting the final refined features. That is, our module can recurrently look into details to find the key attention for FGIC.

2.3 Machine Learning Techniques

In this section, we introduce the machine learning techniques related to CRA-CNN (the approach proposed in Chapter 4), including multi-task learning (Subsection 2.3.1) and contrastive learning (Subsection 2.3.2). CRA-CNN is based on the multi-task learning framework by treating attention region learning as an additional task. Contrastive learning is used to evaluate the region predicted in CRA-CNN.

2.3.1 Multi-task Learning

Multi-task Learning (MTL) is a branch of machine learning, where different training tasks are resolved at the same time for exploiting distinctness and commonness across the tasks. As pointed out in [146], the idea of MTL is inspired by the fact that human always uses the knowledge learned from one task to help learn another related task. For example, the knowledge of learning to ride a bike and a three-wheeled cycle helps each other. In machine learning practice, when compared to individually trained models, MTL can advance the prediction accuracy as well as training efficiency for the task-specific models.

MTL is widely used in different machine learning fields. For example, [151] proposes an MTL framework to effectively extract features for FGIC, which simultaneously solves ultra-fine-grained, fine-grained and coarse-grained image categorization tasks. [125] proposes an MTL CNN model for fingerprint image enhancement with the help of the ridge orientation, which reconstructs the fingerprint photos and the orientation field at the same time. [150] formulates choosing a business site to be an MTL problem and resolves it by an attention-based MTL framework, which specifies the shared features into separate tasks with relational attention for learning understandable features.

2.3.2 Contrastive Learning

Contrastive Learning (CL) is a family of training algorithms for deep neural networks, which acquires features by maximizing the similarity/dissimilarity between similar/dissimilar pairs of data samples. CL has an advantage in developing various learning approaches as it gives a unified framework. For example, [128] first builds independent embedding spaces, responsive to a particular augmentation by each (e.g., rotation, colour jittering, etc.) while insensitive to the others. Then, the proposed CL framework acquires visual representations by preserving the variance conditioned on each augmentation and capturing invariances to the augmentations. [9] first augments the images into different views and then learns visual features by maximizing the agreement among various views of the same image with a contrastive loss.

2.4 Other Fine-grained Image Classification Approaches

In this section, we introduce state-of-the-art FGIC approaches that are not directly related with our work. In this thesis, we compare these state-of-the-art approaches with RMCSAM (the approach proposed in Chapter 5) because RMCSAM is proposed by summing up the experience of two former approaches, and it is the most recommendable approach proposed in this thesis.

DECISION TREE. Decision tree refers to a process that selects the appropriate directions based on the characteristic of features [99]. The inherent interpretability of decision tree has attracted much interest in adapting it for the FGIC task. Nauta et al. [82] proposed the Neural Prototype Tree (ProtoTree) that consists of a CNN backbone followed by a binary tree structure. ProtoTree can be trained end-to-end and locally explain each prediction by describing a decision path. Ji et al. [53] proposed to combine convolutional operations along edges of the tree structure and determines the decision path using the routing functions in each node. The convolutional operations generate the representations of objects, and the tree structure provides a feature learning process to exploit the representations.

ELEMENT-WISE RELATION OF DEEP FEATURE. The intrinsic interrelationship between feature elements contains useful semantic information. Xu et al. [131] proposed a discrimination-aware mechanism (DAM) that improves the deep features conditioned to the analysis on the relation between deep feature elements. DAM can find the feature elements that are not well-learned and refine such elements for better FGIC performance. Zhao et al. [152] proposed a graph-based relation discovery (GaRD) approach to explore the high-order relationships among deep feature elements in the FGIC task. Given an input image, GaRD first generates a high-dimensional feature bank that is regularized with high-order constraints. Then GaRD utilizes a graph-based aggregating procedure to explore the relation between high-order elements of the feature bank and produce a low-dimensional feature representation.

PROGRESSIVE LEARNING. In the FGIC field, progressive learning approaches generally first divide a backbone CNN into several segments, and each segment progressively learns features and gives the prediction. Thereafter, the features learned by each segment are concatenated to give an overall prediction. Du et al. [23] proposed the Progressive Multi-Granularity (PMG), which uses a jigsaw

puzzle generator to produce the images with different levels of granularity and then learns cross-granularity information by progressive learning. Zhang et al. [143] proposed to explore the similarity between the images of the same category and the difference between the images of different categories.

TRANSFORMER. The latest success of transformer in some other fields [60, 100] has influenced the attention-based research in the FGIC field. A transformer is a deep learning model giving attention weights to each element of the input data. It was originally proposed for the natural language processing task [115] and has been adapted for computer vision tasks [22, 39]. He et al. [39] proposed a transformer-based multi-attention model specifically for FGIC use, which is called TransFG. TransFG first splits the input images into small regions, and the regions are projected into feature space by the transformer encoder. Thereafter, TransFG combines all raw attention weights of the transformer to be an attention map and uses the attention map as guidance for selecting discriminative regions. TransFG does not output the selected regions and then explore information from the selected regions. On the contrary, TransFG intuitively considers the attention link of the transformer as an indicator of attention. Specifically, before the last Transformer Layer, TransFG utilizes a part selection module (PSM) to select the tokens that correspond to the discriminative regions and only feed the selected tokens to the last transformer layer.

Though bringing a boost in terms of classification accuracy, these approaches have the problem of high overhead for memory, computation cost, etc. The huge computational expenses caused by their sophisticated architecture [82, 53, 131, 152, 39] or multi-stage framework [23, 143]. For example, TransFG [39] does not require directly localizing the attention regions by outputting the regions and achieves the best accuracy among the studies mentioned in this subsection. However, the backbone transformer, which itself has a extremely heavy computation overhead, together with the complicated part selection module (PSM) [39], makes TransFG require much more parameters, GFLOPs, and time than our approach. Different from these studies, our work provides an insertable, lightweight, and general module, which can be inserted into standard CNNs and only requires a little extra overhead. Moreover, as an insertable module, our approach is complementary to state-of-the-art framework [23], and further improve the accuracy.

Chapter 3

Guided Attention Learning

3.1 Chapter Overview

In this chapter, we focus on solving the problem of the heavy training difficulty of capturing attention with deep neural networks, which is a specific aspect of the research question. In FGIC, discriminative information is always contained in certain regions while the other regions contain much redundancy. Thus the intra-class variance is subtle, which makes FGIC an extremely difficult computer vision task. For solving this problem, many recent FGIC studies develop algorithms on the attention regions, rather than the whole scenes [12, 13]. In attention regions, much redundant visual information are discarded and the remainders are supposed to be discriminative. Mainstream attention-region-based FGIC approaches capture the attention regions by hard-coded methods, which generally locate attention regions by certain hand-crafted saliency features [84, 83, 25, 140, 44, 45, 1, 94, 35].

Hard-coded attention regions are generally not accurate enough. For example, as shown in Figure 3.1, most hard-coded methods assume the attention regions to have the maximum saliency values [140, 44, 1, 94] and consequently usually locate the body region as the attention region if we apply these approaches in bird image classification. However, the body may not always be the most informative region. As introduced in Chapter 1, the key attention for bird image classification may be very subtle and usually locates in very small regions, such as the head region.

Another problem of the approaches based on hard-coded attention regions is



Figure 3.1: Many hard-coded approaches localize the regions by saliency maps [84, 83, 25, 140, 44, 45, 1, 94, 35]. Typically, those approaches first generate saliency map from the input images and then use a value-heavy strategy to locate the regions that have the most saliency value.

that the localization and recognition of the attention regions are treated as two separate stages. The errors during each stage can be accumulated to be a huge final error, and the information learned during each stage cannot help optimize each other.

In this chapter, we try to solve this problem by using deep neural networks to complement the task of region localization. By doing so, the localization task can be optimized together with the task of recognition performed by other deep neural networks. The attention region learned by deep neural networks is named as deep-learned attention region. The problem of deep-learned methods is that they are extremely hard to train with only categorical information because they have to simultaneously complete two difficult tasks (i.e., region localization and classification).

In this chapter, we overcome this difficulty by proposing Attention-Guided Spatial Transformer Networks (AG-STNs), which is an attention-guided variant of the spatial transformer networks (STNs) [52]. AG-STNs have a mechanism named regressive guiding, which makes spatial transformers to capture the same regions as hard-coded attention regions in a certain scale by regression. With AG-STNs, we can first guide the localization network to capture the attention regions of the scales in an intended level (rather than performing meaningless transformations). Then we turn off regressive guiding and let the networks to adjust attention region capturing by themselves (with only categorical information). Finally, the deep-learned attention regions from AG-STNs will focus more meaningful parts. AG-STNs bring two benefits: (a) Additional guidance information makes the network easier to optimize; (b) With different scales of hard-coded attention regions, AG-STNs can be guided to capture different scales of attention information, which are complementary to each other. As mentioned in Chapter 1, we suppose multi-scale

attention information is effective for FGIC. To avoid the harmful influence of inter-class similarity and intra-class variance, it is very important to capture the subtle and discriminative attention. However, the attention regions of small scale are not the only informative regions, and the regions of other scales may sometimes also provide important clues.

In experiment, we explore attention regions of three levels, namely detailed level (224×224 attention regions from 896×896 input images), middle level (224×224 attention regions from 448×448 input images) and the general level (224×224 attention regions from 256×256 input images). Our experimental results show that AG-STNs are much easier to train and can capture meaningful regions. Moreover, multi-level attention information captured by the AG-STNs are complementary to each other, and a fusion can bring better results.

The contributions of the approach proposed in this chapter can be summarized as follows:

- We propose a framework that can capture attention regions by deep neural networks. In this way, the optimization of attention region localization and recognition can be optimized together.
- The proposed regressive guiding strategy successfully solves the training difficulty of using deep neural networks to perform the localization task. Besides, the proposed regressive guiding strategy can be used to guide the networks to capture the attention region of given scales.
- The experimental results show that the fusion of the prediction obtained from multi-scale attention regions can further improve the accuracy over the prediction obtained from single-scale attention regions and raw input images (i.e., without attention).

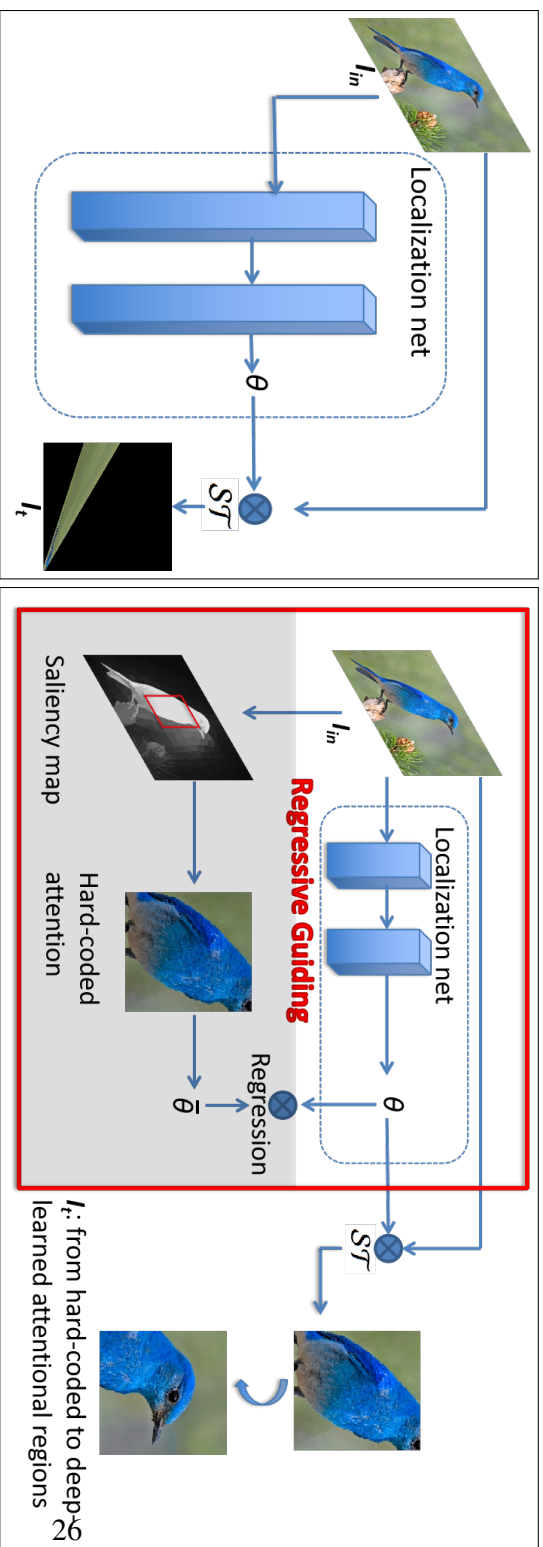


Figure 3.2: (a) illustrates the structure of the STNs. The localization network learns a set of transformation parameters θ from input image I_{in} . With θ , the Spatial Transformer module ST transfers I_{in} to I_t , which will be the input of a recognition network. The whole structure can be optimized together by standard back-propagation and I_t will be better recognized by the recognition network. However, the STNs severely suffer from “Distortion Effect” when learning detail-level attention information. Rather than intended attention regions, only distorted and weird images can be obtained. (b) illustrates the proposed AG-STNs. The parts in the red-frame box are what implements the regressive guiding. $\bar{\theta}$ is the set of transformation parameters, with which the Spatial Transformer module ST will output the same regions as hard-coded attention regions. Inside the red-frame box, except the shaded parts, the rest parts are actually existing parts of the STNs. During regressive guiding step, the parts outside of the red-frame box are turned off (truncated). During the regressive guiding step, we use $\bar{\theta}$ as the regression target and train the parts in the red-frame box by minimizing the L_1 -loss between the θ and $\bar{\theta}$. After the regressive guiding step, for further adjusting the attention regions, the shaded parts are turned off and all the other parts are turned on (linked up).

3.2 Proposed Approach

The proposed approach is an attention-guided variant of STNs, which addresses the training difficulty. With the guidance, AG-STNs can be restrained to capture certain regions of certain scales at first (*regressive guiding step*). Then such restraint is removed and the regions can be adjusted conditionally upon the classification loss in the same way of ordinary STNs (*joint training step*).

3.2.1 STNs in attention region capturing

The STNs contains two important parts: the localization network and the recognition network. As shown in Figure 3.2 (a), given an input image I_{in} , the localization network learns a set of transformation parameters $\theta = f_{loc}(I_{in})$, where f_{loc} denotes the function of the localization network. Thereafter a spatial transformation module \mathcal{ST} obtains the transformed image $I_t = \mathcal{ST}(\theta, I_{in})$, and I_t will be the input of the recognition network. The whole structure can be optimized together: thus, the transformation applied on I_{in} will make I_t better recognized by the recognition network. Rather than directly transforming I_{in} , \mathcal{ST} is in fact implemented by a transformation applied on a regular grid \mathbf{G} , on which I_t is defined. $\mathbf{G} = \{G_i\}$ and $G_i = (x_i, y_i)$ denotes the coordinates of \mathbf{G} . Let the transformation on \mathbf{G} be \mathcal{T} . $\mathcal{T}(\theta, \mathbf{G})$ denotes the transformed grid and it defines I_{in} . Thereafter, the sampler \mathcal{S} forms the pixel values of I_t by sampling the pixel values of I_{in} at particular locations defined by $\mathcal{T}(\theta, \mathbf{G})$. Thus, $I_t = \mathcal{ST}(\theta, I_{in})$ can be also written as $I_t = \mathcal{S}(\mathcal{T}(\theta, \mathbf{G}), I_{in})$. \mathcal{T} can be different transformations with different θ . For example, let $\mathcal{T}(\theta, \mathbf{G})$ be a 2D affine transformation on \mathbf{G} . Then $\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix}$ is the 6-dimensional affine transformation matrix and the transformation between $\mathcal{T}(\theta, \mathbf{G})$ and \mathbf{G} is implemented as

$$\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} \quad (3.1)$$

where (x'_i, y'_i) is a coordinate in $\mathcal{T}(\theta, \mathbf{G})$, and it defines the particular location of I_{in} , at which \mathcal{S} should samples the value for the location (x_i, y_i) in I_t . Those

coordinate are normalized by width/height. Thus, $-1 \leq x'_i, y'_i \leq 1, -1 \leq x_i, y_i \leq 1$. Bilinear sampling kernel is used for \mathcal{S} in [52], and our work uses the same sampling kernel.

In addition to affine transformation, STNs also allow other transformations by using different θ . In this work, we hope the Spatial Transformers to capture the attention regions, for which affine transformation is enough. Thus, we keep using affine transformation in this work, and thus the θ in this paper is always 6-dimensional. However, rather than capturing attention region, affine transformation can also result in other processing on I_{in} , such as rotation, warping, etc. It is hard to make the localization network automatically “*know*” our intention. The networks can hardly automatically perform the cropping of right scale, rather than other transformations, when only categorical information is provided. Furthermore, in some cases, especially when we try to capture attention regions in a very detailed level, the images obtained from the localization networks are severely distorted. Consequently, the recognition performance is extremely bad.

As shown in Figure 3.2 (a), the purpose of using the STNs is to learn the attention region from inputs. However, what we actually obtain from the STNs are distorted images that cannot be well recognized by the recognition network. This is because initial parameters of the localization network are not meaningful for the task. Therefore, at the beginning, the spatial transformation applied on images is meaningless. In many cases, especially when we want more detailed information, however the STNs are trained, they still obtain only distorted images. It is because in such cases, the STNs can hardly be optimized with only classification loss propagated from a recognition network. This phenomenon is referred as distortion effect.

3.2.2 Attention-Guided STNs

To solve the problem of “Distortion Effect”, we propose the *AG-STNs*. As shown in Figure 3.2 (b), the training of an AG-STN can be mainly divided into two steps. The first step is *regressive guiding*, in which the localization network is initialized with hard-coded attention regions. The second step is *joint training*, in which the regressive guiding is turned off and the localization network is jointly trained with the recognition network. The loss functions of these two steps are

defined as

$$\mathcal{L}_s = \begin{cases} \sum_{i=1}^2 \sum_{j=1}^3 |\theta_{ij} - \bar{\theta}_{ij}|, & \text{when } s = 1 \\ -\sum_{c=1}^k \lambda_c \log p(\rho = c), & \text{when } s = 2 \end{cases} \quad (3.2a)$$

$$(3.2b)$$

where s denotes the s_{th} step (e.g., $s = 1$ denotes the first step). \mathcal{L}_s is the loss function of the s_{th} step. As can be seen, \mathcal{L}_1 is in fact a $L1$ -loss and \mathcal{L}_2 a standard cross-entropy loss. More details about Equation (3.2a) and (3.2b) are given as followings:

REGRESSIVE GUIDING. Similar to the formulation in the Section 3.2.1, I_{in} is the input image and $\theta = f_{loc}(I_{in})$ is a set of transformation parameters directly outputted by the localization network. $\bar{\theta} = \begin{bmatrix} \bar{\theta}_{11} & \bar{\theta}_{12} & \bar{\theta}_{13} \\ \bar{\theta}_{21} & \bar{\theta}_{22} & \bar{\theta}_{23} \end{bmatrix}$ is also a set of transformation parameters. $\bar{\theta}$ is obtained as $\bar{\theta} = \mathcal{ST}^{-1}(R_h, I_{in})$, where R_h is the hard-coded attention region of I_{in} and \mathcal{ST}^{-1} is the inverse operation of \mathcal{ST} . Given I_{in} and R_h , \mathcal{ST}^{-1} outputs the transformation parameters $\bar{\theta}$ that makes $\mathcal{ST}(\bar{\theta}, I_{in}) = R_h$. Then in the step of regressive guiding, the localization network is optimized by reducing the loss defined by (3.2a). Obviously, after the training in this step, θ will get close to $\bar{\theta}$ and I_t will approximate R_h .

Then we introduce how we obtain the hard-coded attention regions and how we obtain $\bar{\theta}$ (namely $\mathcal{ST}^{-1}(\cdot)$).

HARD-CODED ATTENTION REGION GENERATION. The hard-coded attention regions are generated from saliency maps corresponding to each image. Let M be a saliency map and is obtained by utilizing the MB+ method in [141]. After obtaining the M of the images, we use a window of size $v \times v$ ($w > h > v$) to traverse M . Thereafter, we use the window which has the most saliency value to bound the hard-coded attention region R_h . The starting position of R_h in an image, whose size is $w \times h$ is defined as

$$(\alpha_{rh}, \beta_{rh}) = \underset{(\alpha, \beta)}{\operatorname{arg\,max}} \frac{1}{v^2} \sum_{i=\alpha}^{\alpha+v-1} \sum_{j=\beta}^{\beta+v-1} M_{(i,j)}^2 \quad (3.3)$$

$$(i \leq w - v + 1, j \leq h - v + 1)$$

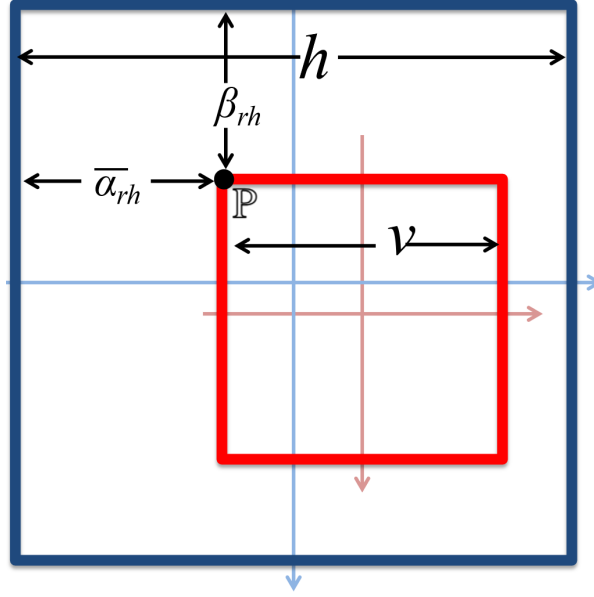


Figure 3.3: The dark-blue-frame box illustrates I_{in} and the light blue coordinate system illustrates $\mathcal{T}(\theta, \mathbf{G})$. The red-frame box illustrates intended I_t , which equals to R_h . The coordinate system in pale red illustrates \mathbf{G} . Note that in the coordinate systems used in the Spatial Transformers, the positive direction of vertical axis is downward.

THE COMPUTATION OF $\bar{\theta}$. For convenience of computation, we crop an $h \times h$ square part from the $w \times h$ images, and use the $h \times h$ part instead of the whole image as I_{in} . When cropping, we make sure R_h is included in the cropped part I_{in} .

Assume the starting position of R_h to be \mathbb{P} and the position of \mathbb{P} in I_{in} is $(\bar{\alpha}_{rh}, \bar{\beta}_{rh})$. Here we want I_t to be R_h . As shown in Figure 3.3, each point of I_t should be sampled from the same position of I_{in} . Take \mathbb{P} as an example, the pixel of \mathbb{P} in I_t should also be the pixel of \mathbb{P} in I_{in} . The position of \mathbb{P} in I_{in} is $(\bar{\alpha}_{rh}, \bar{\beta}_{rh})$, and thus the coordinate of \mathbb{P} in $\mathcal{T}(\theta, \mathbf{G})$ is $(-\frac{h-\bar{\alpha}_{rh}}{2}, -\frac{h-\bar{\beta}_{rh}}{2})$. Similarly, the position of \mathbb{P} in I_t is $(1, 1)$, and the coordinate of \mathbb{P} in \mathbf{G} is $(-1, -1)$.

Then, as introduced in Section 3.2.1, I_{in} and I_t are respectively defined on $\mathcal{T}(\theta, \mathbf{G})$ and \mathbf{G} . The transformation between $\mathcal{T}(\theta, \mathbf{G})$ and \mathbf{G} are defined as Equation (3.1). Thus,

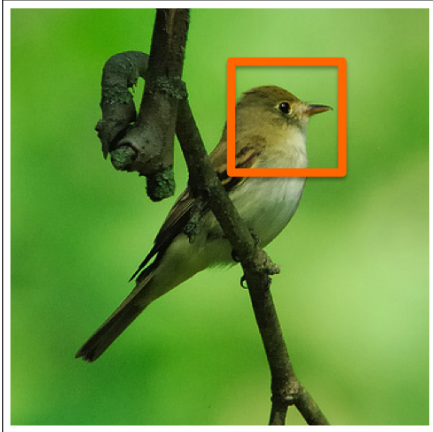
$$\begin{aligned}
\bar{\theta} = \begin{bmatrix} \bar{\theta}_{11} & \bar{\theta}_{12} & \bar{\theta}_{13} \\ \bar{\theta}_{21} & \bar{\theta}_{22} & \bar{\theta}_{23} \end{bmatrix} &= \begin{bmatrix} v/h & 0 & \frac{\frac{v}{2} + \alpha_r h - \frac{h}{2}}{\frac{h}{2}} \\ 0 & v/h & \frac{\frac{v}{2} + \beta_r h - \frac{h}{2}}{\frac{h}{2}} \end{bmatrix} \\
&= \begin{bmatrix} v/h & 0 & \frac{2}{h}\alpha_r h + \frac{v}{h} - 1 \\ 0 & v/h & \frac{2}{h}\beta_r h + \frac{v}{h} - 1 \end{bmatrix}
\end{aligned} \tag{3.4}$$

JOINT TRAINING. After the previous step, the localization network has been initialized to capture the region R_h . We then fuse the initialized localization network with the recognition network and train them together (joint training). Equation (3.2b) defines the loss function for a k -class classification problem. Let $\mathbf{l} = \{l_1, \dots, l_k\}$ be a k -dimensional vector of logits, which are outputted by the recognition network. In Equation (3.2b), ρ is the prediction of category for input instance and $p(\rho = c) = \frac{\exp(l_c)}{\sum_{i=1}^k \exp(l_i)}$. λ_c is a binary indicator (0 or 1), which equals to 1 if c is the true label of the input instance and 0 otherwise. During the step of joint training, all the networks are trained by Equation (3.2b). In other words, all the influencing factors are optimized toward the target of better classification performance. With the training, the networks will gradually locate from R_h to deep-learned attention region R_d , which will be more discriminative.

3.2.3 Multi-stream AG-STNs

When small attention regions can provide more details, sometimes general information is also crucial. For example, as shown in Figure 3.4, in some cases, more general information such as the body texture can be very important. Rather than exploiting only detailed attention information, AG-STNs can also be used for capturing multi-level attention information. To provide comprehensive information, we apply three levels of attention regions, namely the detail level, middle level and general level.

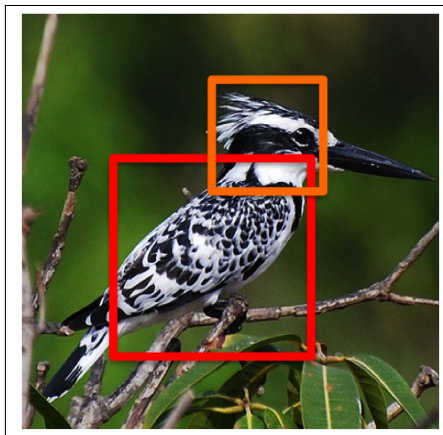
I_t is the input of the recognition network, which requires the inputs to be in a definite size (e.g., 224×224 for ResNet-101 [40]). I_t is in fact a part of I_{in} ($h \times h$), and thus if we resize original images to a larger size (i.e., larger $w \times h$), I_t will account for smaller proportion of I_{in} and otherwise larger proportion. Consequently, we can decide I_t to capture more detailed or general information by setting larger or



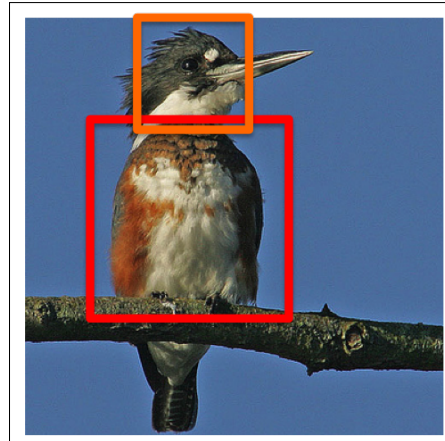
(a) Acadian Flycatcher



(b) Great-Crested Flycatcher



(c) Pied Kingfisher



(d) Belted Kingfisher

Figure 3.4: Examples of two pairs of hard-to-distinguish bird images: Acadian Flycatcher (a) and Great-Crested Flycatcher (b), Pied Kingfisher (c) and Belted Kingfisher (d). In most of the cases, such as the two pairs of images here, the detailed regions (e.g., the head region), provide very discriminative information. However, in some other cases, such as distinguishing between Pied Kingfisher and Belted Kingfisher, more general information, such as the body texture, is also very important.

smaller w and h .

As mentioned before, initial I_t depends on R_h . Therefore, we can compute R_h for multi-size I_{in} beforehand. Then we initialize multi-level localization networks

by regressive guiding with multi-level R_h . Thereafter, during the joint training step, the location and proportion of attention regions can be fine-tuned to a needed extend.

3.3 Experiments

3.3.1 Dataset and implementation details

DATASET. We use CUB-200-2011 [116], which is a bird image dataset across 200 species. There are totally 11788 images in this dataset, 5994 of which are training images and the left 5794 are testing images. The dataset also provides bounding boxes and detailed part annotations but we do not use them in this paper.

NETWORK ARCHITECTURE. We use the ResNet-101 model for localization networks and both ResNet-101 and DenseNet-161 model [48] for recognition network. We set the batch size as 8 for CUB-200-2011 dataset. We initially train the localization networks by regressive guiding and pre-train the recognition networks by random cropping. At this stage, we set the learning rate as 10^{-3} for localization networks and 10^{-4} for recognition networks. We then fuse localization and recognition networks together for joint training. At this stage, we set the learning rate as 10^{-6} for the localization networks and 10^{-5} for the recognition networks. When training CNNs directly on the hard-coded attention regions, the learning rate is set as 10^{-4} at first and then 10^{-5} when training status saturates.

INPUT SIZES. For all the datasets, we first resize all the images into a certain size $I_{r,s}$. Table 3.1 shows the size configurations for different levels. In order to feed I_{in} into the localization networks, we need to downscale I_{in} at first. Regarding the downscaling strategies, we utilize image resizing for the general level. For the middle and detail level, we respectively add $2\times$ and $4\times$ max pooling layers before the localization networks.

3.3.2 Evaluation on detail-level attention learning

In this section, we evaluate the performance of the AG-STNs for exploiting detail-level attention information. We compare AG-STNs with the CNNs trained on hard-coded attention regions and the original STNs. All the networks in this

Table 3.1: Size configuration for different levels

Level	I_{rs}	I_{in}	$I_t/R_h/R_d$
Detail	1189×896	896×896	224×224
Middle	594×448	448×448	224×224
General	340×256	256×256	224×224

Table 3.2: Comparison on the recognition performance between different approaches for exploiting detail-level attention

Methods	Accuracy
Hard-coded Attention Approaches	60.41%
Original STNs	31.90%
AG-STNs	80.15%

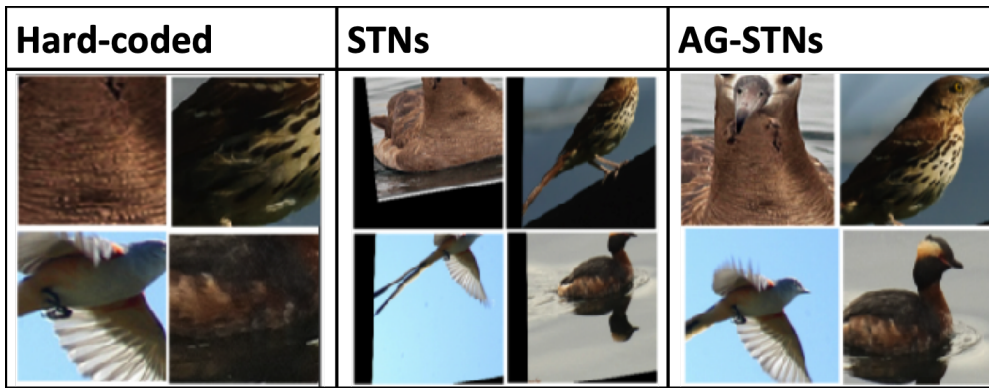


Figure 3.5: Illustration of the detail-level attention regions respectively captured by the hard-coded approach, STNs and AG-STNs. The STNs suffer from “Distortion Effect” so severely that the STNs can get very limited detail-level information. On the contrary, the AG-STNs can capture attention regions as intended without being affected by “Distortion Effect”. Guided by the hard-coded attention regions, the AG-STNs are able to gradually focus on more discriminative regions.

section are based on ResNet-101. Since the training of STNs involves randomness, especially regarding the initialization of network parameters, we run the experiments of STNs with different parameter initializations for 10 times and report the average results.

As shown in Table 3.2, the AG-STNs outperforms the STNs and hard-coded attention regions in all the aspects. Regarding the accuracy, the AG-STNs are



Figure 3.6: Examples of the detail-level attention regions extracted by the localization networks right after the regressive guiding stage (Subsection 3.2.2)

dramatically better than the other two.

Figure 3.5 illustrates some examples of different regions respectively captured by different approaches. As can be seen, the STNs can capture very limited attention information in detail level, which accounts for the low performance of STNs in this level. Whereas the AG-STNs guided by hard-coded attention regions can successfully capture the intended information. Though guided by hard-coded attention regions, the regions captured by the AG-STNs are far more discriminative. Therefore, the AG-STNs also outperform the hard-coded approach. Figure 3.6 shows the the initial attention regions captured by localization networks right after the guidance. Figure 3.7 visualizes how an AG-STN gradually moves its focus from the initial attention regions to more discriminative parts.

3.3.3 Evaluation on multi-stream AG-STNs

Beside the capability for exploiting detail-level attention information, the AG-STNs also have the ability to capture attention information in other levels. In this section, we present the results of multi-stream AG-STNs. All the localization networks are constructed from ResNet-101. We evaluate the performance by using both ResNet-101 and DenseNet-101 as recognition network. The performance of AG-STNs is also compared with no-attention baselines trained on whole images. The results are shown as Table 3.3. As can be observed, AG-STNs dramatically

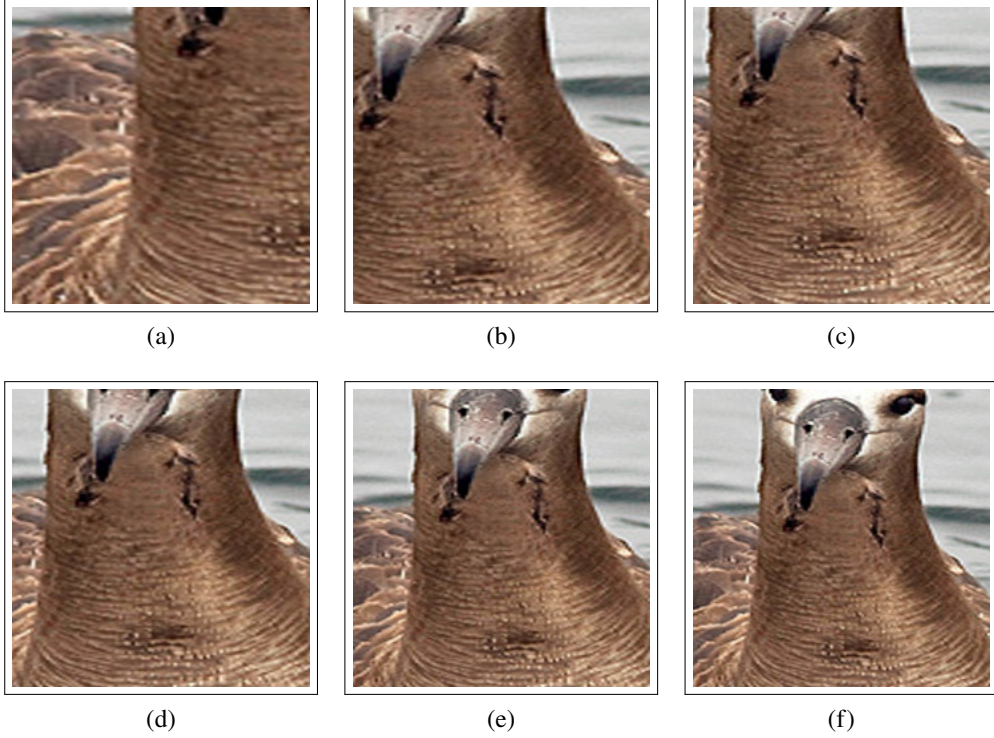


Figure 3.7: (a-f) show the detail-level attention regions captured by the AG-STNs during the joint training process. (a) is the attention region captured by the AG-STNs that are just initialized by regressive guiding. Thus, (a) can be regarded to be equal to the hard-coded attention regions. (f) is the attention regions captured by the AG-STNs when the joint training is finished. (b-e) are attention regions captured by AG-STNs in different stages between (a) and (f). It can be seen that from (a) to (f), AG-STNs gradually focus on more informative regions.

Table 3.3: Results of multi-stream AG-STNs

	Backbone Network	ResNet-101	DenseNet-161
With attention	Detail	80.15%	81.71%
	Middle	81.03%	82.93%
	General	78.29%	82.83%
	Fusion	83.36%	86.93%
Without attention		74.87%	81.29%

Table 3.4: Comparison with previous studies on CUB_200_2011

Part-based R-CNNs [142]	76.36%
PD+DCoP+flip+GT BBox+ft [62]	82.8%
Compact Bilinear Pooling [31]	84.0%
STNs (4×ST-CNN 448px) [52]	84.1%
LRBP [61]	84.21%
RN-50+SS [97]	84.5%
PD+FC+SWFV-CNN [144]	84.54%
RA-CNN (scale 1+2+3) [28]	85.3%
Improved Bilinear Pooling [70]	85.8%
BoostCNN [80]	86.2%
Kernel Pooling [16]	86.2%
MA-CNN ($L_{cls} + L_{cng}$) [153]	86.5%
ResNet-101+OSME+MAMC [110]	86.5%
PC-DenseNet-161 [24]	86.87%
Ours (multi-stream AG-STNs)	86.93%

outperforms the no-attention baselines. Besides, different streams of AG-STNs are complementary to each other. A fusion of them brings better results.

Table 3.4 shows a comparison with previous studies in CUB-200-2011. [52] is most related with our work. In [52], the best result is achieved by 4×ST-CNNs (224 px attention regions for 448 px inputs). In our work, we use the three levels of 1×AG-STNs. Compared with [52], our work is able to exploit more detailed attention information, which is proved to be complementary with attention information of other levels, such as the level used in [52] (i.e., 224 px attention regions for 448 px inputs). The result shows that our work is comparable with those previous studies on CUB-200-2011.

3.4 Summary of This Chapter

In this chapter, we focus on addressing the training difficulty of capturing attention with deep neural networks, which is a specific aspect of the research question that we try to figure out in this thesis. With this objective, we introduce a new extension model of STNs, the AG-STNs, for solving the problem of training difficulty. In the AG-STNs, at first, a mechanism named regressive guiding supervises

the Spatial Transformers with hard-coded attention regions. Then regressive guiding is turned off and the network is able to adjust the captured regions for obtaining more effective attention information. With the mechanism of regressive guiding, the Spatial Transformers are able to *understand their “mission”* and therefore capture the attention information in the intended level rather than implementing other transformations. Besides, with regressive guiding, the Spatial Transformers do not suffer from “Distortion Effect” any more. “Distortion Effect” is especially severe when capturing detail-level attention information and it is mainly caused by the deficiency of the only-provided categorical signals. Since the AG-STNs are provided with hard-coded attention information in addition to the categorical information, they successfully capture attention information in a very detailed level whereas the STNs fail to capture attention information in such level. Also, as attention region localization and recognition are optimized simultaneously, the AG-STNs can capture more discriminative regions than the hard-coded regions. Besides, regressive guiding can also be used to make the Spatial Transformers to capture multi-level attention regions by guiding with the relevant multi-level hard-coded attention regions. Our results show that AG-STNs outperform STNs and hard-coded approaches for capturing detail-level attention information. Moreover, the streams of multi-stream AG-STNs are complementary to each other. Therefore, the fusion of the streams brings better results. This chapter verifies that guiding the deep neural networks with traditional hard-coded attention regions helps solve the training difficulty and capture effective attention regions. This chapter also verifies the effectiveness of multi-scale attention.

Chapter 4

Multi-task Attention Learning

4.1 Chapter Overview

In this chapter, we mainly try to solve the problem of extra overhead for capturing attention information. Specifically, with the multi-task learning strategy, we propose a novel framework, where the extra attention-learning overhead can be removed after training, requiring no extra overhead during the testing.

Traditional region-based attention learning strategies suffer from two drawbacks. Firstly, region localization inevitably requires much extra overhead. Moreover, capturing three-scale attention regions further increase the overhead. For practical use, while the overhead in the training procedure can be to some extents avoided by training beforehand, the overhead in the testing (i.e., utilization) procedure is inevitably infeasible.

Secondly, region localization is a difficult task in itself. It is impossible to ensure to localize the imperfect region every time, and cropping wrong regions causes inevitable and irreversible information loss and introduces unfavourable noise to the training model. Sometimes, there might be several regions all contain useful information but the model has to abandon some of them. This problem largely limits the improvement of region-based attention learning. To avoid abandoning useful regions, some recent studies stack a lot of networks to capture dense attention regions, which, however, makes the overhead increase largely [42, 54, 96, 36, 52, 28, 129, 67, 122].

In this chapter, instead of following the typical region-based pipeline used in

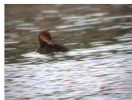
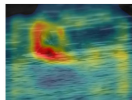



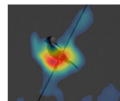



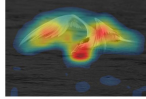


Input Image	CAM of the Input	Similar Image of the Wrongly Predicted Class	Dissimilar Image of the Ground Truth Class
			
			
			

Figure 4.1: Motivation of this chapter. This figure illustrates some classification results of a ResNet-50 [40] trained on CUB-200-2011 dataset [116]. The first column shows some input images that are wrongly classified by the trained ResNet-50. The second column shows the Class Activation Maps (CAMs) [155] generated by the ResNet-50 on the input images, which denotes what region the ResNet-50 uses as clues to identify the predicted category. The images in the third column are from a different class from the images in the first column, but the trained ResNet-50 predicts them to be in the same category. The images in the fourth column are from the same class as the images in the first column, but the trained ResNet-50 predicts them to be in different categories. The wrong classification is caused by that the visual clues found by the ResNet-50 are not robust enough.

Chapter 3, we try to explore attention information from deep features. On the one hand, the input image need not be cropped before going through recognition. Namely, there is no loss of information on the input image. On the other hand, the attention information can be obtained from the deep features. The obtained attention information can be used to refine the deep features to strengthen the features' responses towards attention regions and weaken the features' responses towards non-attention regions.

Following the above idea, in this chapter, we focus on strengthening the CNNs' awareness of the discriminative regions (i.e., we want to make CNN's correspondence stronger to the discriminative visual clues and weaker to redundant visual information). As shown in Figure 4.1, take bird image classification as an example. The classification difficulty mainly comes from two essential characteristics of the images. The first characteristic is the intra-class variance. The birds of the same species may look quite different in various poses, illumination, etc. The second

characteristic is the inter-class similarity. For foreground objects, the visual differences among bird species are subtle. For background objects, images of different birds may have the same habitats, such as trees or water. If the CNNs take the habitat background as an important clue, they tend to make mistakes.

For solving these problems, we suppose the classification networks requires: (a) mainly exploring clues on the discriminative object parts that are invariant in different poses, camera angles, etc. (e.g., bird heads); (b) referring to the other parts if necessary, otherwise not. For this objective, we propose the Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN) following the multi-task learning (MTL) strategy [104, 11, 30, 146, 151, 125, 150].

MTL is widely used for improving generalization by utilizing the domain knowledge in the training supervision of relevant tasks as an inductive bias. The proposed CRA-CNN consists of two network streams, namely a major network (\mathcal{N}_{maj}) and a subordinate network (\mathcal{N}_{sub}). \mathcal{N}_{maj} has two related tasks: (a) predict the correct category of a given image; (b) predict the attention information of the given image. For the second task, \mathcal{N}_{maj} is required to generate a set of attention parameters conditioned on the given input image. Thereafter the proposed Attention-redundancy Transformation module (ART module) takes as inputs the attention parameters and the given image and divides the input visual information into attention and redundancy. Then \mathcal{N}_{sub} evaluates the attention-redundancy proposal of \mathcal{N}_{maj} and regulates the \mathcal{N}_{maj} through standard backpropagation. We train \mathcal{N}_{maj} and \mathcal{N}_{sub} together during training while removing \mathcal{N}_{sub} and only using \mathcal{N}_{maj} during the testing process. Thus, CRA-CNN requires no more overhead than basic network backbones (e.g., the ResNets) in the testing procedure.

Generally, MTL requires multiple annotations for multiple tasks (e.g., category label for classification and bounding box for localization [30]). However, extra attention annotations besides necessary category labels, such as bounding boxes, require much extra manual effort. Thus we use category label as the only manual annotation for all tasks and design \mathcal{N}_{sub} to evaluate and improve the attention-redundancy proposal instead of extra attention annotations.

The evaluation is conducted from two aspects. Firstly, the proposed attention calls for contain discriminative visual information as much as possible. That means the proposed attention can be recognized as the correct category. Secondly, the proposed attention and redundancy are expected to be contrasted to each other,

which follows the contrastive learning (CL) strategy [128, 9, 17, 50] and is inspired by the fact that humans can learn discriminative clues by contrasting different images with categorical labels. For example, given an FGIC task of distinguishing bird species, a human instinctively contrasts images with the same/different labels and tries to find commonality/difference among them. Then, the human will find that the most discriminative regions are certain parts of the foreground objects (e.g., bird heads), rather than the background objects (e.g., tree branches), and the latter then become insignificant in the eyes of the human.

Corresponding to the two above aspects, we train the \mathcal{N}_{sub} by solving two tasks, which is based on a loss function by each. The first one is a softmax loss to teach \mathcal{N}_{sub} that the proposed attention is supposed to be recognized as the correct category. The second is a proposed contrastive learning loss to teach the \mathcal{N}_{sub} that, the attention-redundancy pairs of the same image should be pushed far apart, and the redundancies of different images should be pulled closer.

Our contributions in this chapter are:

- We propose a novel MTL framework that helps the networks to strengthen the awareness of condition-invariant attention.
- The proposed approach is easy to implement and computationally affordable, especially in the test procedure.
- Our approach clearly outperforms the baselines on CUB-200-2011 and Stanford Cars datasets. Experimental results show that exploring attention information from deep features is effective for improving the accuracy.

4.2 Proposed Approaches

4.2.1 Approach Overview

The outline of the proposed CRA-CNN is shown in Figure 4.2. The major network (\mathcal{N}_{maj}) is used to predict the category of the input image, and we use the subordinate network (\mathcal{N}_{sub}) to force the \mathcal{N}_{maj} to improve attention awareness. The two networks are linked by the proposed ART module, which consists of the attention transformation module (AT module) and the redundancy transformation module (RT module).

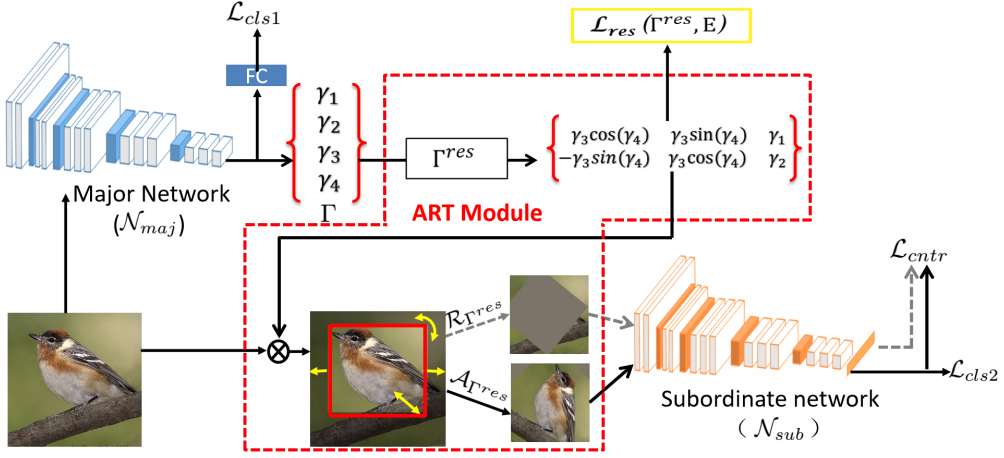


Figure 4.2: The pipeline of CRA-CNN. Given an input image, \mathcal{N}_{maj} is required to complete two tasks: predict the category (trained with \mathcal{L}_{cls1}) and output a set of transformation parameters Γ . The red dashed box illustrates the ART module. Γ is restricted to be in a reasonable range and becomes Γ^{res} , which is further limited by the proposed attention-restriction loss \mathcal{L}_{res} . $\mathcal{A}_{\Gamma^{res}}$ and $\mathcal{R}_{\Gamma^{res}}$ are parametrized transformations, which are implemented on the input image and form the attention and redundancy, respectively. The parametrized transformations allow localization, zooming, and rotation. Then the \mathcal{N}_{sub} evaluates the attention-redundancy proposal by recognizing the attention’s category (\mathcal{L}_{cls2}) and measuring the dissimilarity of the attention-redundancy pairs and similarity of different redundancies (\mathcal{L}_{ctr}).

In a practical sense, \mathcal{N}_{maj} has two tasks: (a) categorize a given image; (b) predict a transformation based on the features learned by \mathcal{N}_{maj} . \mathcal{N}_{sub} also has two tasks: (a) evaluate whether the proposed attention is discriminative; (b) evaluate whether the proposed redundancy is redundant.

4.2.2 Attention-redundancy Transformer module

ART module acts as an important bridge between \mathcal{N}_{maj} and \mathcal{N}_{sub} . That is, the ART module has to efficiently report the attention awareness of \mathcal{N}_{maj} to \mathcal{N}_{sub} , and then report \mathcal{N}_{sub} ’s evaluation to improve \mathcal{N}_{maj} ’s attention awareness.

Therefore, the ART module has to meet three requirements: (a) The ART module should be mathematically differentiable so that it can be embedded within CNNs. (b) The ART module should be able to transmit comprehensive information

of the attention between the networks, such as the locations, sizes, and alignment angles. (c) The design of the ART module should not be too complicated, and should be easy to optimize. Otherwise, the training difficulty would bring an accuracy decrease.

To meet the above requirements, we turn our eyes to the Spatial Transformer (ST) module, an essential component of the Spatial Transformer Networks (STNs)[52]. The ART module is adapted from the ST module by overcoming its shortcomings to meet the above requirements. The details of the STNs have been introduced in Chapter 3. Here, similar to the formulation in Chapter 3, given an input image I_{in} , the ST module outputs a transformed image I_t , and the transformation applied by the ST module is conditioned on I_{in} . Neglect the numbers of channels in I_t and let the 2-D size of I_t to be $H \times W$ (height, width), and $H \in [1, h]$, $w \in [1, w]$. $G = \{(y_1, x_1), (y_1, x_2), \dots, (y_2, x_1), (y_2, x_2), \dots, (y_w, x_{h-1}), (y_h, x_w)\}$ is a regular spatial grid that defines the I_t (i.e., G is the sampling grid for I_t). Similarly, let the 2-D size of I_{in} to be $H' \times W'$ (height, width), and $H' \in [1, h']$, $w' \in [1, w']$. Let $G' = \{(y'_1, x'_1), (y'_1, x'_2), \dots, (y'_2, x'_1), (y'_2, x'_2), \dots, (y'_w, x'_{h-1}), (y'_h, x'_w)\}$ to be the sampling grid that defines the I_{in} . The transformation is applied as Equation (3.1) in Chapter 3. Based on this formulation, we give the formulation of the proposed approach as below.

DETAILS OF ATTENTION-REDUNDANCY TRANSFORMER MODULE. The ST module is proposed to resolve spatial variations but faces some issues in practice.

The foremost common issue is that the localisation network has difficulty dealing with the early-stage noise, which consequently leads to irreversible loss of visual information and causes huge errors in classification (also indicated in [71]). Additionally, the ST module fails to resolve certain spatial variations (also indicated in [105]). Those issues make it rather difficult to optimize the localisation network and ST module. Due to the difficulty, [52] has to manually fix θ_{11} , θ_{12} , θ_{21} and θ_{22} and only optimizes θ_{13} and θ_{23} for learning attention region in the FGIC tasks (the details of θ_{11} , θ_{12} , θ_{13} , θ_{21} , θ_{22} and θ_{23} are given in Chapter 3), which however largely narrows the variety of possible transformations.

We propose the AT module to efficiently reveal the attention awareness of the \mathcal{N}_{maj} without heavy optimization difficulty. As mentioned before, the AT module should be able to reveal various attention information. In this paper, we design the

AT module to predict the the attention regions' locations, sizes, and angles. The angle information is used for aligning the attention region. As indicated in [36], the feature maps of strong visual semantics help to align objects. We assume the converse to be also true: completing alignment tasks promotes \mathcal{N}_{maj} to improve visual semantics.

Thus, along with the classifiers, \mathcal{N}_{maj} outputs $\Gamma = [\gamma_1 \ \gamma_2 \ \gamma_3 \ \gamma_4]$, which is a 4-dimensional vector and we refer it as attention parameters. γ_1 to γ_4 defines the horizontal location, vertical location, scale, and alignment angle, respectively. Here, we use I_{in} , I_t and G following the same formulation as given above. G^a and G^r are the grids defining the sampling destinations in I_{in} for the attention and redundancy, respectively. The transformation performed by the AT module is defined as

$$G^a = \mathcal{A}_\Gamma(G), \quad \text{where } \Gamma = f_{maj}^{att}(I_{in}). \quad (4.1)$$

In equation (4.1), f_{maj}^{att} denotes the function of predicting attention parameters with \mathcal{N}_{maj} . The transformation \mathcal{A} for g_{ij} is mathematically written as

$$\begin{pmatrix} x_i^a \\ y_j^a \end{pmatrix} = \mathcal{A}_\Gamma(g_{ij}) = \begin{bmatrix} \gamma_3 \cos(\gamma_4) & -\gamma_3 \sin(\gamma_4) & \gamma_1 \\ \gamma_3 \sin(\gamma_4) & \gamma_3 \cos(\gamma_4) & \gamma_2 \end{bmatrix} \begin{pmatrix} x_i \\ y_j \\ 1 \end{pmatrix}, \quad (4.2)$$

where (x_i^a, y_j^a) denotes a coordinate in G^a .

Compared with the ST module, the AT module has reduced the number of parameters, and each of the parameters only explores a determinate domain (i.e., locations, sizes or angle). Thus, the AT module is simpler to train than the ST module. Despite its simplicity, the AT module can define various attention transformations conditioned on different combinations of the attention parameters proposed by \mathcal{N}_{maj} . G^r is defined as

$$G^r = G - (G \cap G^a). \quad (4.3)$$

RESTRICTIONS ON THE PROPOSAL OF ATTENTION INFORMATION. In practice, the proposed transformation is possible to be meaningless (e.g., sampling largely

beyond the boundary of I_{in}). Thus, we give restrictions to the ART module to prevent this.

Concretely, since the coordinates of the sampling grids are all normalized to $[-1, 1]$, the AT module will sample outside I_{in} if the coordinates of G^a are beyond $[-1, 1]$. We constrain the location factors γ_1, γ_2 , and the scale factor γ_3 to be in reasonable range by:

$$\Gamma^{res} = \begin{bmatrix} \gamma_1^{res} & \gamma_2^{res} & \gamma_3^{res} & \gamma_4^{res} \end{bmatrix} = \begin{bmatrix} \alpha_p \tanh(\gamma_1) & \alpha_p \tanh(\gamma_2) & \alpha_s \tanh(\gamma_3) & \gamma_4 \end{bmatrix}, \quad (4.4)$$

where $\alpha_p \in [0, 1], \alpha_s \in [0, 1]$ ensure $\gamma_1^{res}, \gamma_2^{res} \in [-\alpha_p, \alpha_p]$ and $\gamma_3^{res} \in [-\alpha_s, \alpha_s]$. We keep $\gamma_4^{res} = \gamma_4$ because the angle factor γ_4 is actually restricted by trigonometric functions.

Instead of Γ , we use Γ^{res} to transmit attention information between the networks in practice. Γ^{res} helps to avoid irreversibly meaningless transformation, which inevitably misdirects the optimization of the networks.

Besides, we propose the attention-restriction loss to constrain the proposal of attention information further. Assume $E = [e_1, e_2, e_3, e_4]$ is the expectation of possible Γ . That is, the transformed image obtained with the $\Gamma = E$ likely represents the attention in the average situation. The attention-restriction loss is defined as:

$$\begin{aligned} \mathcal{L}_{res} = & \left(\frac{\max(0, |\gamma_1^{res} - e_1| - t_1)}{|\gamma_1^{res} - e_1| - t_1 + eps} (\gamma_1^{res} - e_1)^2 + \frac{\max(0, |\gamma_2^{res} - e_2| - t_2)}{|\gamma_2^{res} - e_2| - t_2 + eps} (\gamma_2^{res} - e_2)^2 \right. \\ & \left. + \frac{\max(0, |\gamma_3^{res} - e_3| - t_3)}{|\gamma_3^{res} - e_3| - t_3 + eps} (\gamma_3^{res} - e_3)^2 + \frac{\max(0, |\gamma_4 - e_4| - t_4)}{|\gamma_4 - e_4| - t_4 + eps} (\gamma_4 - e_4)^2 \right)^{\frac{1}{2}}, \end{aligned}$$

where eps denotes the epsilon.

(4.5)

In equation (4.5), $T = [t_1 \ t_2 \ t_3 \ t_4]$ is a set of thresholds. \mathcal{L}_{res} punishes \mathcal{N}_{maj} if the distances between E and the proposed Γ is larger than the thresholds. In experimental practice, E is manually set to be a half-width and half-length center cropping. While most prior studies only learn localization (i.e., cropping with a fixed size) [42, 52, 28], ART module proposes various transformations constrained with the thresholds.

4.2.3 Multi-task Learning Pipeline

Given an input image, our pipeline learns its category by minimizing four losses to solve multiple learning tasks. \mathcal{L}_{cls1} is the softmax loss between the ground-truth category labels and the category predictions that are mapped from the input image by \mathcal{N}_{maj} . \mathcal{L}_{res} is the restriction loss to guarantee proposals of attention/redundancy to be reasonable, as introduced in Subsection 4.2.2. \mathcal{L}_{cls2} is the softmax loss between the ground-truth category labels and the category predictions that are mapped by \mathcal{N}_{sub} from the attention proposed by \mathcal{N}_{maj} . \mathcal{L}_{ctr} is a contrastive learning loss that maximizes the discrepancy between the attention and redundancy while minimizing the discrepancy of different redundancies. Let (I_{in}^m, I_{in}^n) be a pair of input images and $(a^m, a^n), (r^m, r^n)$ are respectively the pairs of attention and redundancy corresponding to the image pairs. \mathcal{L}_{ctr} is mathematically defined as:

$$\begin{aligned} \mathcal{L}_{ctr} = & \max(d(f_{sub}(r^m), f_{sub}(r^n)) - d(f_{sub}(r^m), f_{sub}(a^m)) + margin, 0) + \\ & \max(d(f_{sub}(r^n), f_{sub}(r^m)) - d(f_{sub}(r^n), f_{sub}(a^n)) + margin, 0), \end{aligned} \quad (4.6)$$

where $d(\cdot)$ indicates the Euclidean distance and $f_{sub}(\cdot)$ denotes the deep representation learned by \mathcal{N}_{sub} , such as the output of the last fully connected layer in \mathcal{N}_{sub} . We minimize a multi-task loss function \mathcal{L} for training, which is defined as:

$$\mathcal{L} = \mathcal{L}_{cls1} + \mathcal{L}_{cls2} + \mathcal{L}_{res} + \mathcal{L}_{ctr}. \quad (4.7)$$

During the testing process, we remove \mathcal{N}_{sub} and the ART module, and only use the classifiers of \mathcal{N}_{maj} for predicting the categories. Thus, our pipeline has the same overhead as the basic backbone networks for testing.

4.3 Experiments

4.3.1 Implementation details

DATASETS. To evaluate the effectiveness of our approach, we carried out experiments on two widely-used and competitive datasets, namely CUB-200-2011 [116] and Stanford Cars [63]. CUB-200-2011 is also used in Chapter 3, which is a benchmark of 11,788 bird image across 200 different species. Stanford

Cars is a benchmark of car images across 196 car models with 16,185 images.

BASELINES. As our approach is actually a training strategy, we use as a baseline the VGG-16 [109], ResNet-50, ResNet-101 [40] and DenseNet-121 [48] pre-trained on ImageNet [20] and then fine-tuned on the three above-mentioned benchmarks. We compare the recognition performance between the baselines and the same network backbones trained by our approach. To simplify the setting, we always use the same network as the backbones of \mathcal{N}_{maj} and \mathcal{N}_{sub} . To obtain Γ , we add a fully-connected layer on the top of the ReLu7 layer of VGG-16 or the last pooling layer of the ResNets/DenseNets. As mentioned before, in the testing stage, \mathcal{N}_{sub} and the ART module are removed, and thus our approach has exactly the same structure as the baselines.

TRAINING AND TESTING DETAILS. We manually set $\alpha_p = 1$, $\alpha_s = 1$, $E = \begin{bmatrix} 0 & 0 & 0.5 & 0 \end{bmatrix}$, $T = \begin{bmatrix} 0.4 & 0.4 & 0.4 & \pi \end{bmatrix}$ and the threshold of Equation (4.6) as 0.7. For the training procedure, we resize the images to make the shorter side be 512, while keeping the aspect ratio being unchanged. Then we randomly crop a 448×448 part and feed the 448×448 images into \mathcal{N}_{maj} as the inputs for completing the tasks of category prediction and Γ generation. The output size of the ART module is 224×224. We train the models using standard Stochastic Gradient Descent (SGD) with the momentum of 0.9, batch size of 64, weight decay of 5×10^{-4} . We set the initial learning rate as 10^{-3} , and then reduce it to 10^{-4} after 50 epochs. Thereafter, the learning rate is reduced by 10^{-1} for every 45 epochs. Furthermore, after the first 50 epochs, \mathcal{L}_{res} is repeatedly turned off for 45 epochs and then turned on for 45 epochs.

For the testing procedure, initially, the images are resized in the same way as the training procedure. Then, we apply centre cropping on the the resized images (Subsection 4.3.2 and Subsection 4.3.3) or do not crop the resized image but average the final prediction scores outputted by the classifiers (Subsection 4.3.4).

4.3.2 Comparison with the Baselines

We compare our approach with the baselines on the two above-mentioned FGIC benchmarks, and the results are shown in Table 4.1. It is clear that our approach outperforms the baselines on all the datasets, whatever the backbone is. Note that our approach exactly has the same structure as baselines for testing,

Table 4.1: Comparison results with baselines.

		CUB Birds	Stanford Cars
VGG-16	Baseline	76.7%	85.2%
	CRA-CNN	80.0%	86.8%
ResNet-50	Baseline	84.2%	90.0%
	CRA-CNN	86.2%	92.6%
ResNet-101	Baseline	86.1%	91.8%
	CRA-CNN	87.6%	93.4%
DenseNet-121	Baseline	80.0%	89.1%
	CRA-CNN	84.2%	90.6%

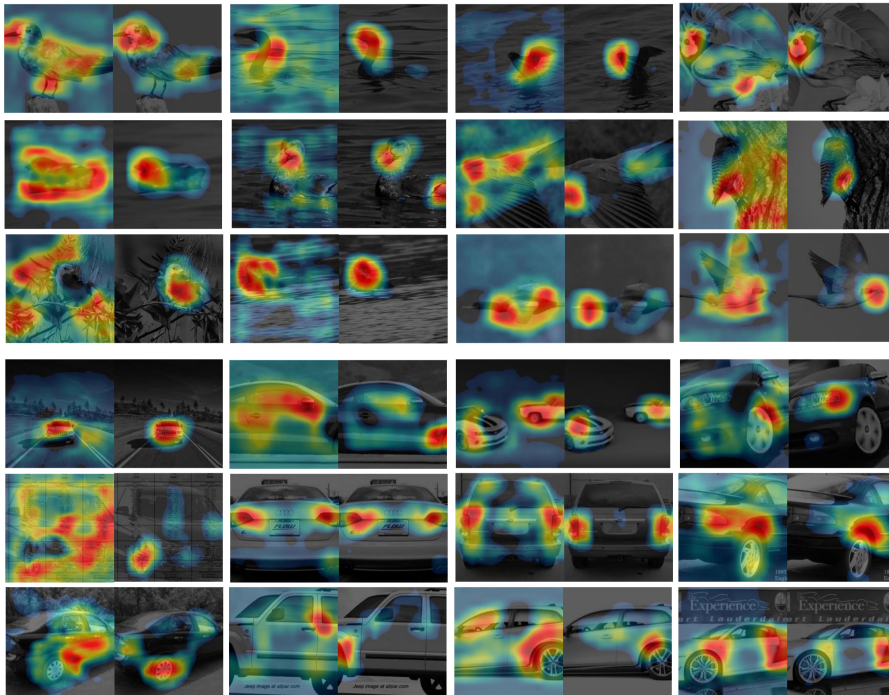


Figure 4.3: Examples of some CAMs respectively generated by the baseline ResNet-50 (the left image of each pair) and the ResNet-50 trained by CRA-CNN (the right image of each pair). CRA-CNN makes the network much more focused than the baseline network.

which shows the effectiveness of our proposed training strategy.

Figure 4.3 visualizes some examples of the CAMs respectively generated by the

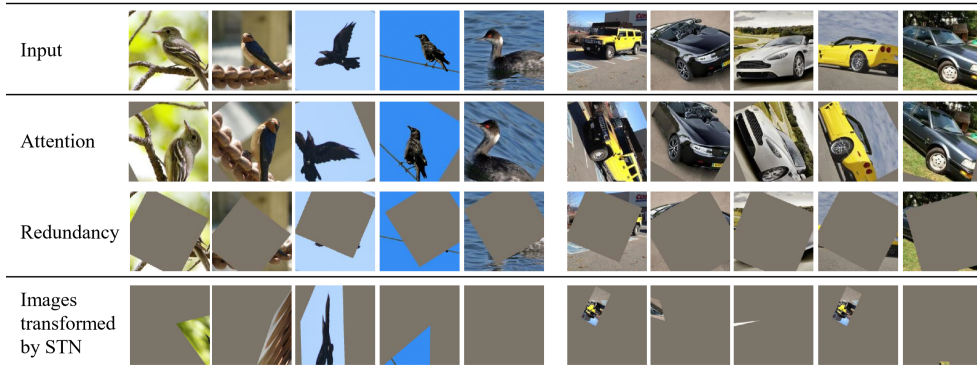


Figure 4.4: Examples of transformed images obtained by CRA-CNN and STN. It is very hard for the STN to capture meaningful attention due to the training difficulty. The CRA-CNN both captures and aligns the attention regions conditioned on certain objects, such as head, body contour, etc.

baseline ResNet-50 and the ResNet-50 trained by CRA-CNN. CRA-CNN forces the network to focus more on core attention, whereas the baseline network tends to be distracted. The core attention, which is forced to learn by CRA-CNN, helps to improve recognition accuracy.

Figure 4.4 shows the examples of some original input images, as well as the attention and redundancy images captured from the original input images. Moreover, since the ART module is adapted from the ST module, we also train an STN as a reference. We show the region learned by the STN also in Figure 4.4. For the sake of fairness, rather than fixing the first four transformation parameters of the ST module as is done in [52], we optimize all the six transformation parameters. It is obvious that the STN fails to capture any beneficial visual information in this setting, which is caused by the heavy training difficulty of the STN. Actually, in practice, we observe that the STN cannot converge at all unless we fix the first four transformation parameters. However, the networks trained by CRA-CNN captures useful visual semantics.

4.3.3 Ablation Study on different losses.

Compared with the baselines, our approach needs to be trained with three more additional losses (i.e., \mathcal{L}_{cls2} , \mathcal{L}_{res} and \mathcal{L}_{ctr}). Here we investigate the effectiveness of the three additional losses by respectively removing each of the

Table 4.2: Ablation study on different losses.

	w/o. \mathcal{L}_{cls2}	w/o. \mathcal{L}_{res}	w/o. \mathcal{L}_{cntr}	w/. all losses
CUB Birds	85.1%	85.7%	85.4%	86.2%
Stanford Cars	90.9%	91.8%	92.0%	92.6%

Table 4.3: Comparison results on Stanford Cars.

ResNet-101+OSME+MAMC [110]	90.3%	TASN [154]	93.8%
Subset B [15]	90.6%	SEF [76]	94.0%
Kernel Pooling [16]	92.0%	WS-DAN [47]	94.5%
RA-CNN (scale 1+2+3) [28]	92.5%	EfficientNet [112]	94.7%
MA-CNN ($L_{cls} + L_{cng}$) [153]	92.8%	AutoAugment [14]	94.8%
PC-DenseNet-161 [24]	92.9%	Ours(ResNet-50)	93.3%
MPN-COV [66]	93.3%	Ours(ResNet-101)	94.8%

Table 4.4: Comparison results on CUB-200-2011.

SPD representation [32]	72.4%	SEF [76]	87.3%
STNs (4×ST-CNN 448px) [52]	84.1%	TASN [154]	87.9%
RA-CNN (scale 1+2+3) [28]	85.3%	Subset B [15]	88.8%
Kernel Pooling [16]	86.2%	WS-DAN [47]	89.4%
MA-CNN ($L_{cls} + L_{cng}$) [153]	86.5%	Stacked LSTM [33]	90.4%
ResNet-101+OSME+MAMC [110]	86.5%	Ours(ResNet-50)	86.7%
PC-DenseNet-161 [24]	86.9%	Ours(ResNet-101)	88.3%

three losses and observing the change of classification accuracy. We adopt ResNet-50 as the backbone network for this ablation experiment and apply centre crop on the resized images. The results are shown in Table 4.2. On all three datasets, the classification accuracy decreases to some extent when any one of three additional losses is removed. In other words, each of the three additional losses contributes to the improvement of accuracy.

4.3.4 Comparison with Previous Studies

Table 4.3 shows comparison results with the previous studies on Stanford Cars, and our best result reaches the same result reported in [14]. [14] designs a search scheme for optimizing augmentation policy, which, however, introduces large computational expenses. Besides, as pointed out in [145], the proxy tasks' policies are sometimes not suitable for the target task. In comparison, the approach proposed in this thesis gives a computationally affordable and effective solution.

Even though our best result is a little behind [15, 47, 33] on CUB-200-2011, our work is still competitive because: (a) some of the methods, namely [15] and [75] (behind our results on Stanford Cars), are actually transfer learning approaches requiring larger extra data; (b) our approach is relatively easy to implement and quite light to utilize. For utilization, our approach achieves a high accuracy after removing all extra overhead and only using a single backbone CNN.

4.4 Summary of This Chapter

In this chapter, we focus on addressing the difficulty of extra overhead for attention learning. In this chapter, we propose the Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN) to enhance the attention awareness of deep neural networks. CRA-CNN is composed of two networks that are joined by the proposed attention-redundancy transformer (ART) module. The subordinate network helps the major network continuously explore core attention by evaluating the attention-redundancy proposal of the major network. Our approach is easy to implement and computationally affordable and largely reduces the extra attention-learning overhead. Our work is quite competitive with previous studies regarding its simplicity and categorization performance. This chapter verifies that exploring attention information from deep features is effective for FGIC, and the approach proposed in this chapter can solve the difficulty of extra overhead for the testing (utilization) procedure.

Chapter 5

Recursively-refined Multi-scale Attention Learning

5.1 Chapter Overview

In this chapter, we mainly focus on the difficulty of heavy overhead during the training procedure. By summing up the experience of previous chapters, we propose the recursive multi-scale channel-spatial attention module (RMCSAM) for addressing the above problems. RMCSAM provides a lightweight module that can be inserted into standard CNNs. Experimental results show that RMCSAM can improve the classification accuracy and attention capturing ability over baselines. Also, RMCSAM performs better than other state-of-the-art attention modules in fine-grained image classification, and is complementary to some state-of-the-art approaches for fine-grained image classification.

The proposed RMCSAM follows the success of the previous research on attention modules [149, 46, 18, 89, 126]. Attention modules refer to a set of insertable modules that enhance the feature representations generated by standard convolutional layers by giving weights among the channels or spatial locations of the feature. For example, the squeeze-and-excitation module (SE module) [46], which is one of the most prominent attention mechanisms, performs channel-wise attention by extracting global information from each channel and then generating a set of weights for each channel. By doing so, the SE module provides a boost of classification accuracy with a low additional overhead. The point-wise spatial at-

tention module (PSA module) [149] is another typical example. The PSA module uses self-adaptively predicted attention maps to aggregate long-range contextual information within images, which boosts the performance for the scene parsing task. These attention modules are generally insertable into different network architectures and able to improve the networks' focus on important information.

The RMCSAM is designed as an attention module that explores multi-scale attention information and uses the explored information to enhance the deep features learned in the FGIC task. As an attention module, RMCSAM can be easily placed inside various backbone CNNs, such as ResNet [40] or VGG models [108]. Trained together with the backbone CNNs, RMCSAM improves the correspondence to attention information for better classification accuracy. Clearly, our approach is different from previous FGIC approaches, which mainly design mechanisms placed as the output parts of the backbone CNNs yielding attention information (e.g., attention regions) [33, 21, 96, 154, 42, 134, 28, 147, 106, 114, 130].

Specifically, as shown in Figure 5.1, the main ideas of the proposed RMCSAM are summarized as follows:

- Rather than localization and categorization of attention regions, which is commonly used in previous FGIC approaches [33, 21, 96, 154, 42, 134, 28, 147, 106, 114, 130, 49, 69, 90, 139, 142], we focus on developing an insertable attention module for the FGIC task.
- We design the proposed attention module to explore both channel-wise and spatial-wise attention. For the channel-wise attention, we firstly spatially pool the given features and then use the pooled features to compute channel-wise weights with a set of fully connected (FC) layers. For the spatial-wise attention, we firstly pool the given features along the channel axis and then use the pooled features to compute spatial-wise weights with a set of convolutional layers. The features learned with the channel-wise and spatial-wise attention sub-module are aggregated by average.
- Following the prior experience that multi-scale attention is very important and effective for FGIC, we design the proposed attention module to perform three-scale channel-wise and spatial-wise attention. The different scales of the channel-wise sub-modules are defined with different numbers of the

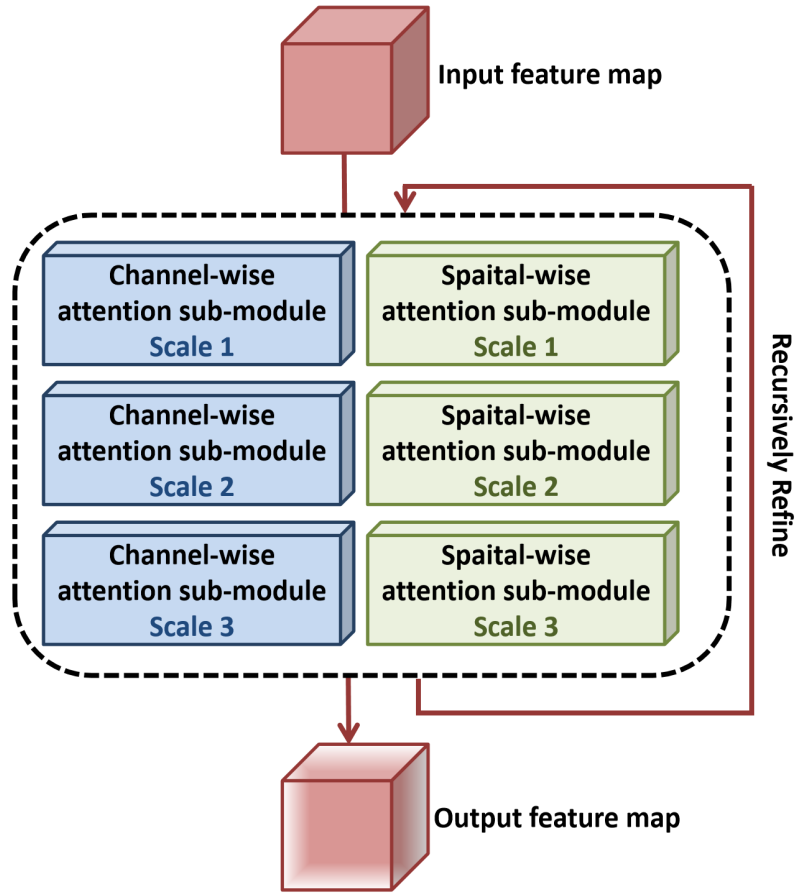


Figure 5.1: Illustration of the main ideas of our work. The proposed attention module has six sub-modules: three-scale channel-wise sub-modules and three-scale spatial-wise sub-modules. The input feature map is recursively refined through the six sub-modules for a predetermined number of times to output the finally refined feature map.

neurons in the FC layers within the sub-modules. The different scales of the spatial-wise sub-modules are defined with different kernel sizes in the convolutional layers within the sub-modules. The features refined by different scales of sub-modules are aggregated by average. Even though the proposed module is designed to perform three-scale channel-wise and spatial-wise attention, the whole module is still very lightweight because

each sub-module only requires a small number of parameters.

- We design the proposed attention module to progressively refine the learned attention. Starting from the feature map outputted by a standard convolutional layer, we design a cyclically learning scheduler to generate more effective features by iteratively treating the output of the former learned attention module as the input of the current attention module. The attention modules in the different stages share the same parameters.

The contributions of the approach proposed in this chapter can be summarized as follows:

- We propose a simple yet effective attention module that can explore multi-scale attention with negligible overhead for FGIC tasks.
- The proposed module can be easily inserted into standard CNNs and improve the classification accuracy for FGIC.
- We evaluate the proposed module on two benchmarks: CUB-200-2011 [116] and Stanford Cars [63]. We have validated the effectiveness of the design of the proposed attention module through extensive ablation studies. Experimental results show that RMCSAM can improve the classification accuracy and attention capturing ability over baselines. Also, RMCSAM outperforms other state-of-the-art attention modules [46, 18, 89, 126] in FGIC tasks.
- As an insertable attention module, our approach have very strong versatility. It can be combined with the previous approach achieving state-of-the-art accuracy in the FGIC task [23]. By combining our approach with the PMG framework [23], we achieve the best accuracy on the Stanford Cars and surpass the previous best accuracy obtained with the Resnet50 backbone on the CUB-200-2011.

5.2 Proposed Approach

In this section, we introduce the proposed RMCSAM in detail. As shown in Figure 5.2, given an input feature map, RMCSAM first processes it via six sub-modules: three channel-wise sub-modules in different scales and three spatial-wise

sub-modules in different scales. The processed feature maps are aggregated to be an output feature map. Thereafter, the output feature map is treated as the input feature map of the six sub-modules and processed again by the six sub-modules. This process is repeated a predetermined number of times to obtain the final refined feature map.

5.2.1 Multi-scale Channel-wise Attention Sub-modules

The multi-scale channel-wise attention sub-modules are used to exploit inter-dependencies among the channels of a given feature map. In CNNs, each channel of a feature map acts as an object detector [137]. Consequently, channel-wise attention tells what objects are discriminative or unimportant for distinguishing a given image [126]. For example, bird head and bird claw are generally discriminative objects for distinguishing different bird species, and some other objects, such as tree branches, are not important for classification. We describe the detailed operation of the multi-scale channel-wise attention sub-modules below.

Firstly, consider a single-scale channel-wise attention sub-module, which is implemented similarly to the SE module [46]. Let $X \in \mathbb{R}^{H \times W \times C}$ be an input feature map generated by the former layer within a CNN. H , W and C respectively represents the spatial height, width and number of channels. Let $\Omega_r^{chl}(\cdot)$ denote the function of the single-scale channel-wise attention sub-module. Note that r is a manual hyper parameter controlling the scale of the attention module, and it will be introduced in detail later in this subsection. An overview of the function of the single-scale channel-wise attention sub-module can be summarized as: output a 1D channel-wise weighted mask $M_r^{chl} \in \mathbb{R}^{1 \times 1 \times C}$ and then put M_r^{chl} on X for emphasizing the discriminative channels and de-emphasizing the unimportant channels. A mathematical definition of $\Omega_r^{chl}(\cdot)$ can be given as:

$$X_r^{chl} = \Omega_r^{chl}(X) = X \otimes M_r^{chl}, \quad (5.1)$$

where X_r^{chl} denotes the refined feature map outputted by the single-scale channel-wise sub-attention module, and \otimes denotes element-wise production. During \otimes , the values of M_r^{chl} are broadcasted along the spatial dimension to make M_r^{chl} have the same size as X .

M_r^{chl} is obtained from X with a set of pooling, fully connected (FC), and

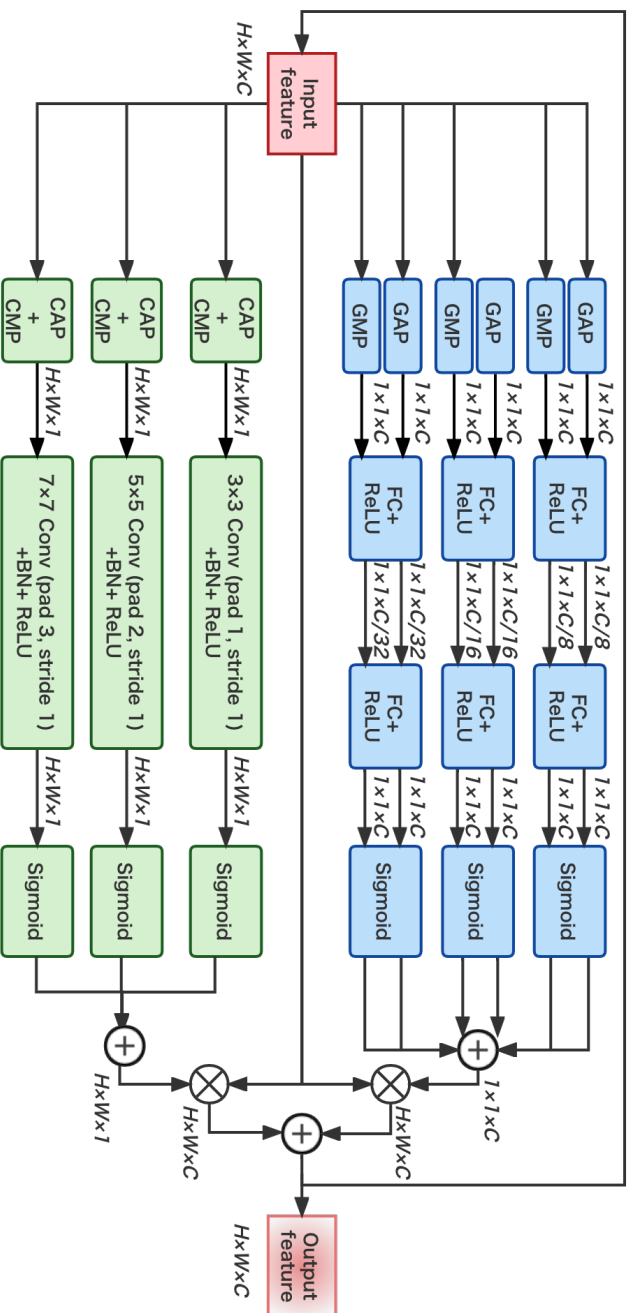


Figure 5.2: Illustration of the proposed recursive multi-scale channel-spatial attention module (RMCSAM). “GAP” and “GMP” respectively represent the global average and max pooling. “CAP” and “CMP” respectively represent the channel-wise average and max pooling. “FC” and “Conv” respectively represent fully-connected layer and convolutional layer. “BN” and “ReLU” respectively represent batch normalization [51] layer and ReLU layer. “ \oplus ” represents element-wise sum. “ \otimes ” denotes broadcast element-wise multiplication. The feature maps are denoted as feature dimensions, e.g. “ $H \times W \times C$ ” denotes a feature map with height H , width W and channel number C .

sigmoid operations. As average-pooled and max-pooled features provide complementary information [126], we first use both global average pooling and global max pooling to spatially shrink X to generate 1D channel-wise descriptors $D^{avg} \in \mathbb{R}^{1 \times 1 \times C}$ and $D^{max} \in \mathbb{R}^{1 \times 1 \times C}$ as:

$$d_c^{avg} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_{i,j,c}, \quad (5.2)$$

$$d_c^{max} = \max_{i=1}^H \max_{j=1}^W x_{i,j,c}, \quad (5.3)$$

where d_c^{avg} and d_c^{max} are respectively the values in the channel c ($c \in \{1, 2, 3, \dots, C\}$) of D^{avg} and D^{max} . $x_{i,j,c}$ denotes the value at the spatial location (i, j) in the channel c of X . Then both D^{avg} and D^{max} are processed by two successive FC layers as:

$$\begin{aligned} D^{avg'} &= \Phi^{avg}(D^{avg}) \\ &= f^{ReLU}(\phi_C^{avg}(f^{ReLU}(\phi_{\frac{C}{r}}^{avg}(D^{avg})))) \end{aligned} \quad (5.4)$$

$$\begin{aligned} D^{max'} &= \Phi^{max}(D^{max}) \\ &= f^{ReLU}(\phi_C^{max}(f^{ReLU}(\phi_{\frac{C}{r}}^{max}(D^{max})))) \end{aligned} \quad (5.5)$$

where $\Phi^{avg}(\cdot)$ denotes the layers processing D^{avg} , and $\Phi^{max}(\cdot)$ denotes the layers processing D^{max} . $f^{ReLU}(\cdot)$ denotes ReLU operation. $\Phi^{avg}(\cdot)$ and $\Phi^{max}(\cdot)$ share the same parameters in order to reduce overhead. For both $\Phi^{avg}(\cdot)$ and $\Phi^{max}(\cdot)$, the output size of the first FC layer (i.e., $\phi_{\frac{C}{r}}^{avg}(\cdot)$ or $\phi_{\frac{C}{r}}^{max}(\cdot)$) is set as $1 \times 1 \times \frac{C}{r}$, and this FC layer is used to compress the channel-wise information of D^{avg} or D^{max} into a certain scale. The output size of the second FC layer (i.e., $\phi_C^{avg}(\cdot)$ or $\phi_C^{max}(\cdot)$) is set as C , and this FC layer makes the output descriptor have the same size of channels of X (so that the element-wise multiplication in Equation (5.1) can be implemented).

Thereafter, M_r^{chl} is obtained as:

$$M_r^{chl} = \sigma(D^{avg'}) + \sigma(D^{max'}) \quad (5.6)$$

where σ represents the sigmoid operation, which makes each value range from

0 to 1 and thus gives the importance of each channel of X . The refined feature map X_r^{chl} can be obtained by substituting the M_r^{chl} obtained in Equation (5.6) into Equation (5.1).

The multi-scale channel-wise attention is obtained with different r . r controls the output size of $\phi_{\frac{c}{r}}^{avg}(\cdot)$ and $\phi_{\frac{c}{r}}^{max}(\cdot)$. A smaller output size makes the output information more compressed and gives a more abstract representation of the input descriptor (i.e., D^{avg} or D^{max}). A larger output size makes the output keep more information and gives a more detailed and inclusive representation of D^{avg} or D^{max} . In order to obtain all-sided channel-wise attention information, we build up multi-scale channel-wise attention sub-modules by using three different r : 8, 16 and 32. The refined feature map outputted by the multi-scale channel-wise attention sub-modules is defined as:

$$\begin{aligned} X_{multi}^{chl} &= \Omega^{chl}(X) \\ &= \Omega_8^{chl}(X) + \Omega_{16}^{chl}(X) + \Omega_{32}^{chl}(X). \end{aligned} \quad (5.7)$$

5.2.2 Multi-scale Spatial-wise Attention Sub-modules

The multi-scale spatial-wise sub-attention module are used to exploit inter-dependencies among the spatial locations of a given feature map. In convolutional neural networks, the receptive field is used to measure the region of the input space that affects a particular unit of the network. The receptive field increases as the depth of networks increases. The units of very deep layers have a very large receptive field, which is sometimes even larger than the input image. That is, the features of deep layers are affected by a large region of the input image rather than a small local region. However, in practice, each spatial location of the deep features only mainly corresponds to certain small local regions of the input images even though the spatial location can be affected by a large region in theory. Zhou *et al.* [155] proved in experiments that the deep features learned by CNNs in the classification task have a strong localization ability, and the spatial locations of the deep features can present the CNN's attention towards different spatial locations of the input image. Later, Luo *et al.* [77] proved how only a small number of pixels in the receptive field visibly contribute to the final deep features in theory and proposed the idea of Effective Receptive Field (ERF). Such studies can be regarded as the basis of spatial-wise attention. Spatial-wise

attention tells the spatial location of the discriminative objects. As introduced in Subsection 5.2.1, channel-wise attention tells what objects are discriminative for classification. Thus, these two types of attention are complementary to each other. We describe the detailed operation of the multi-scale spatial-wise attention sub-modules below.

Firstly, consider a single-scale spatial-wise attention sub-module. Similar to the formulation in Subsection 5.2.1, $X \in \mathbb{R}^{H \times W \times C}$ denotes an input feature map, and $\Omega_k^{spat}(\cdot)$ denote the function of the single-scale spatial-wise attention sub-module. Note that k is a manual parameter controlling the scale of the attention sub-module, and it will be introduced in detail later in this subsection. An overview of the function of the single-scale spatial-wise attention sub-module can be summarized as: output a 2D spatial-wise weighted mask $M_k^{spat} \in \mathbb{R}^{H \times W \times 1}$ and then put M_k^{spat} on X for emphasizing the discriminative spatial locations and de-emphasizing the unimportant spatial locations. A mathematical definition of $\Omega_k^{spat}(\cdot)$ can be given as:

$$X_k^{spat} = \Omega_k^{spat}(X) = X \otimes M_k^{spat}, \quad (5.8)$$

where X_k^{spat} denotes the refined feature map outputted by the single-scale spatial-wise attention sub-module, and during \otimes , the values of M_k^{spat} are broadcasted along the channel dimension to make M_r^{spat} have the same size as X .

M_r^{spat} is obtained from X with a set of operations including channel-wise pooling, 2D convolution, and sigmoid. The first step for obtaining M_r^{spat} is to shrink X along the channel dimension to generate 2D spatial-wise score maps $S^{avg} \in \mathbb{R}^{H \times W \times 1}$ and $S^{max} \in \mathbb{R}^{H \times W \times 1}$ as:

$$s_{i,j}^{avg} = \frac{1}{C} \sum_{c=1}^C x_{i,j,c}, \quad (5.9)$$

$$s_{i,j}^{max} = \max_{c=1}^C x_{i,j,c}, \quad (5.10)$$

where $s_{i,j}^{avg}$ and $s_{i,j}^{max}$ are respectively the values at the location (i, j) of S^{avg} and S^{max} . $x_{i,j,c}$ denotes the value at the spatial location (i, j) in the channel c of X .

Then S^{avg} and S^{max} are processed as:

$$S' = \psi_{k \times k \times 2 \times 1}(f^{cat}(S^{avg}, S^{max})), \quad (5.11)$$

where f^{cat} denotes a channel-wise concatenation operation. $\psi_{k \times k \times 2 \times 1}(\cdot)$ denotes a 2D convolutional layer whose kernel size is $k \times k \times 2 \times 1$, and this layer is used to encode the spatial-wise information of each $k \times k$ -size region inside S^{avg} and S^{max} . The padding size of $\psi_{k \times k \times 2 \times 1}(\cdot)$ is set as $\frac{k-1}{2}$ and the stride is set as 1. Consequently, $\psi_{k \times k \times 2 \times 1}(\cdot)$ does not change the spatial size of the input feature map.

Thereafter, M_k^{spat} is obtained as:

$$M_k^{spat} = \sigma(f^{ReLU}(f^{BN}(S'))), \quad (5.12)$$

where $f^{BN}(\cdot)$ denotes batch normalization operation [51]. The refined feature map X_k^{spat} can be obtained by substituting the M_k^{spat} obtained in Equation (5.12) into Equation (5.8).

The multi-scale spatial-wise attention is obtained with different k . k controls the kernel size of $\psi_{k \times k \times 2 \times 1}(\cdot)$. That is, k decides each value of M_k^{spat} to be corresponding to how large a region in S^{avg} and S^{max} . A 2D convolutional layer of a smaller kernel size has smaller Effective Receptive Fields and thus can capture more local information and more detailed clues. A 2D convolutional layer of a bigger kernel size has bigger Effective Receptive Fields and thus can “see” more information at once and capture relatively more global information, such as the dependencies among some local patterns. In order to obtain comprehensive spatial-wise attention information, we build up multi-scale spatial-wise attention sub-modules by using three different k : 3, 5 and 7. The refined feature map outputted by the multi-scale spatial-wise attention sub-modules is defined as:

$$\begin{aligned} X_{multi}^{spat} &= \Omega^{spat}(X) \\ &= \Omega_3^{spat}(X) + \Omega_5^{spat}(X) + \Omega_7^{spat}(X). \end{aligned} \quad (5.13)$$

5.2.3 Recursive Refinement

Our module recursively refines the given feature maps to focus on the discriminative visual information more finely. Let T denote how many times we refine the feature maps, and let X_t^{ref} ($t \in \{0, 1, 2, 3, \dots, T\}$) denote the feature map outputted at time t . We recursively refine the feature map by treating the output at time $t - 1$ as the input of time t . A mathematical definition is given as:

$$\begin{aligned} X_0^{ref} &= X, \\ X_t^{ref} &= \mathbf{\Omega}^{chl}(X_{t-1}^{ref}) + \mathbf{\Omega}^{spat}(X_{t-1}^{ref}). \end{aligned} \tag{5.14}$$

5.3 Experiments

5.3.1 Experimental Settings

To evaluate the effectiveness of our approach, we carried out experiments on two widely-used, competitive and standard benchmarks, namely CUB-200-2011 [116] and Stanford Cars [63], which are same as the datasets used in Chapter 4.

As our approach is actually a lightweight insertable module, we compare the FGIC performance of the standard networks without the proposed module, with the proposed module, with other state-of-the-art attention modules. Besides, we also compare our approach with the latest state-of-the-art FGIC approaches [152, 131, 143, 39, 53, 82, 23]. Following the experience in previous studies [126, 18], we insert the proposed module after the final convolutional block of each network. In order to perform apple-to-apple comparisons, we reproduced all the evaluated networks with the same training and testing configuration.

For the training procedure, we resize the images to make the shorter side be 512, while keeping the aspect ratio being unchanged. Then we randomly crop a 448×448 part augmented with random flipping as the input. Consequently, the GFLOPs in this paper are reported by computing with 448×448 input. For the testing procedure, we resize the images in the same way as the training procedure but use center cropping to obtain the 448×448 input images. For keeping the interference factors as few as possible and obtaining a stable result, we evaluate the time cost of the proposed approach as well as other approaches by handling a group of eight input images (unless otherwise specified), i.e., an $8 \times 3 \times 448 \times 448$ tensor, with a single Nvidia GTX 1080 Ti.

Regarding the parameter initialization, we use the network backbones pre-trained on the ImageNet [19] (provided by PyTorch [91]) and then fine-tune them on the fine-grained image classification datasets. The inserted RMCSAM, as well as other attention modules, are randomly initialized. However, in Subsection 5.3.6, to further improve the accuracy, we also implement the experiment of pre-training RMCSAM with the Resnet50 backbone on the ImageNet once before the fine-tuning (see more training details in Subsection 5.3.6). For all the other experiments, we use same experimental configuration:

- We reproduce all the experiments 10 times and report the average accuracy.

- We train all the networks using standard Stochastic Gradient Descent (SGD) with the momentum of 0.9, batch size of 32, weight decay of 5×10^{-4} , learning rate of 2×10^{-3} .
- All the experiments are implemented in the PyTorch framework [91] with 2×Nvidia GTX 1080 Ti (except for evaluating the time cost).

5.3.2 Ablation Study

In this subsection, we analyze whether and how multiple scales of the channel-wise, spatial-wise attention and recursive refinement are beneficial for FGIC tasks. We use a VGG11 network [108] with batch normalization [51] as the baseline, and evaluate the performance of: the baseline, the baseline + different single-scale channel-wise attention modules, the baseline + multi-scale channel-wise attention module, the baseline + different single-scale spatial-wise attention modules, the baseline + multi-scale spatial-wise attention module, the baseline + RMCSAM respectively refined 1~5 times.

The ablation study is conducted on both datasets, and the results are shown in Table 5.1.

Table 5.1: Results of the ablation study

	Accuracy		Parameters	GFLOPs
	CUB-200-2011	Stanford Cars		
Baseline	75.9%	89.3%	9.327M	30.031
$\Omega_8^{chl}(\cdot)$	81.2%	90.5%	9.393M	30.031
$\Omega_{16}^{chl}(\cdot)$	81.4%	90.4%	9.360M	30.031
$\Omega_{32}^{chl}(\cdot)$	81.6%	90.7%	9.343M	30.031
$\Omega_8^{chl}(\cdot)+\Omega_{16}^{chl}(\cdot)+\Omega_{32}^{chl}(\cdot)$	81.8%	90.8%	9.443M	30.031
$\Omega_3^{spat}(\cdot)$	81.7%	90.6%	9.327M	30.031
$\Omega_5^{spat}(\cdot)$	82.2%	90.6%	9.327M	30.031
$\Omega_7^{spat}(\cdot)$	81.6%	90.6%	9.327M	30.031
$\Omega_3^{spat}(\cdot)+\Omega_5^{spat}(\cdot)+\Omega_7^{spat}(\cdot)$	81.6%	90.8%	9.327M	30.031
$\Omega^{spat}+\Omega^{chl}, T = 1$	81.9%	91.3%	9.443M	30.031
$\Omega^{spat}+\Omega^{chl}, T = 2$	81.9%	91.5%	9.443M	30.032
$\Omega^{spat}+\Omega^{chl}, T = 3$	82.4%	92.1%	9.443M	30.032
$\Omega^{spat}+\Omega^{chl}, T = 4$	81.4%	91.9%	9.443M	30.032
$\Omega^{spat}+\Omega^{chl}, T = 5$	80.9%	91.6%	9.443M	30.032

Single-scale attention vs. multi-scale attention. On both datasets, the multi-scale channel-wise attention module performs better than all the single-scale channel-wise attention modules. Compared with the baseline, the multi-scale channel-wise attention module improves the accuracy by 5.9% on CUB-200-2011 and 1.5% on Stanford Cars. Multi-scale spatial-wise attention module performs better than all the single-scale spatial-wise attention modules. Compared with the baseline, the multi-scale spatial-wise attention module improves the accuracy by 5.7% on CUB-200-2011 and 1.5% on Stanford Cars.

The influence of refining times. Simply Aggregating both multi-scale channel-wise and spatial-wise attention (i.e., $\Omega^{pat}(\cdot) + \Omega^{chl}(\cdot)$ with $T = 1$) performs better than only using one of them, which suggests multi-scale channel-wise and spatial-wise attention are complementary to each other. Moreover, increasing refining times can further affect the accuracy. On both two datasets, the most suitable T is 3, because the accuracy tends to decrease with a T larger than 3. Compared with the baseline, by setting T as 3, RMCSAM improves the classification accuracy by 6.5% on CUB-200-2011 and 2.8% on Stanford Cars, while increasing only 0.116M parameters and 0.001 GFLOPs.

For all the rest experiments, the T for RMCSAM is set as 3.

5.3.3 Comparison with the Baselines

In this subsection, we empirically show how RMCSAM helps improve the classification accuracy over different baseline networks. We use as baselines six network models, namely VGG11 [108] with batch normalization, VGG16 [108] with batch normalization, Resnet18 [40], Resnet50 [40], Gluon_resnet18_v1b [41], and GoogLeNet [111]. We compare the networks with and without the proposed module, and the results are shown in Table 5.2. RMCSAM favorably improves the classification of all the baselines by 0.4%~6.5% on CUB-200-2011 and 0.4%~2.8% on Stanford Cars. In terms of the extra overhead, RMCSAM increases only 0.116M~1.841M parameters and 0.001~0.003 GFLOPs. In view of the negligible additional parameters and GFLOPs, our approach provides a good improvement in classification accuracy. Regarding the additional time cost, RMCSAM increases 1.065ms~6.100ms over different backbones for processing a group of eight input images, which is also a small overhead.

Table 5.2: Comparison results with baselines

	Accuracy		Parameters	GFLOPs	Time Cost
	CUB-200-2011	Stanford Cars			
VGG11_bn	75.9%	89.3%	9.327M	30.031	49.209ms
VGG11_bn+RMCSAM	82.4%	92.1%	9.443M	30.032	52.102ms
VGG16_bn	80.1%	91.9%	14.824M	61.549	97.108ms
VGG16_bn+RMCSAM	83.6%	93.0%	14.940M	61.550	98.173ms
Resnet18	79.9%	92.1%	11.277M	7.274	16.181ms
Resnet18+RMCSAM	80.5%	92.9%	11.394M	7.275	20.152ms
Resnet50	85.5%	93.2%	23.910M	16.438	48.000ms
Resnet50+RMCSAM	86.1%	94.2%	25.751M	16.449	54.100ms
Glunet_resnet18_v1b	81.9%	92.6%	11.277M	7.274	15.938ms
Glunet_resnet18_v1b+RMCSAM	82.7%	93.0%	11.394M	7.275	19.339ms
GoogLeNet	80.5%	93.4%	5.801M	6.016	25.126ms
GoogLeNet+RMCSAM	80.9%	93.8%	6.263M	6.018	29.023ms

5.3.4 Analysis of Attention Capturing

In this subsection, we evaluate whether the proposed RMCSAM actually helps a network focus on discriminative visual information by two methods, namely visualization and quantitative analysis. The experiments in this subsection are implemented with the VGG11 model with batch normalization.

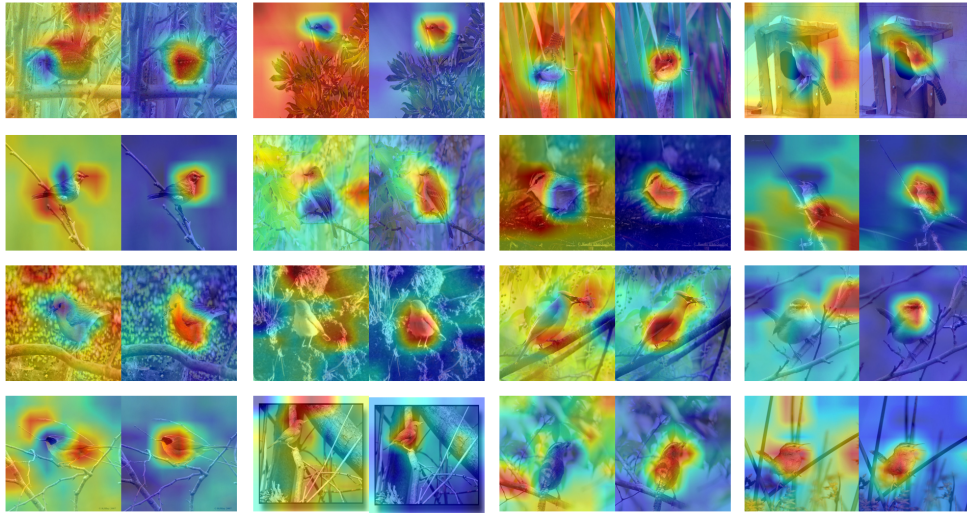
First, we use Grad-CAM [102] to visualize the focus of the networks. Grad-CAM uses the gradients of the predicted category, flowing into the final convolutional layer to generate a heatmap highlighting the important regions in the image for predicting the category. That is, the heatmap generated by Grad-CAM visualizes the “reason” why the network “thinks” a given image belongs to a certain category. The visualization results are shown in Figure 5.3. Compared with the baseline network, the network inserted with RMCSAM focuses more on discriminative regions and objects. Moreover, Figure 5.4. shows the visualization results of the attention generated with RMCSAM when there is no target object in the scene. As can be observed, the attention is not accurate but it still has some capability of telling foreground and background in certain cases. However, in some other cases, the attention makes no distinction between the major and the minor clues.

Second, we quantitatively analyze the attention capturing ability by attention precision. We first introduce the definition of attention precision. The computation of attention precision starts from generating a heatmap $Y \in \mathbb{R}^{H' \times W'}$ by Grad-CAM, which has the same spatial size as the input image ($\mathbb{R}^{H' \times W' \times 3}$). Regard Y as a set of pixels, namely $Y = \{y_{(1,1)}, y_{(1,2)}, \dots, y_{(\alpha,\beta)}, \dots, y_{(H',W')}\}$. Then Y is normalized as:

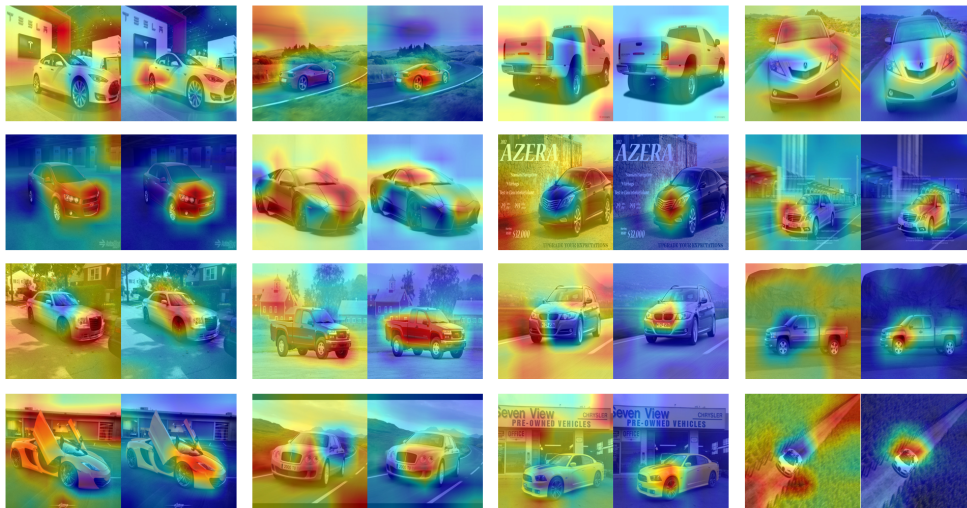
$$y'_{(\alpha,\beta)} = \frac{y_{(\alpha,\beta)} - \min(Y)}{\max(Y) - \min(Y)} \quad (5.15)$$

After the normalization, each value of the heat map ranges from 0 to 1. Then given a threshold λ ($0 < \lambda < 1$), all the values larger than λ are set as 1, and all the values no larger than λ are set as 0 as:

$$y''_{(\alpha,\beta)} = \begin{cases} 1, & \text{if } y'_{(\alpha,\beta)} - \lambda > 0 \\ 0, & \text{if } y'_{(\alpha,\beta)} - \lambda \leq 0. \end{cases} \quad (5.16)$$



(a) CUB-200-2011



(b) Stanford Cars

Figure 5.3: Visualization of Grad-CAM. In each pair of images, the left one is the visualization results using the baseline network. The right one is the visualization results using the network inserted with RMCSAM.

Thereafter, the attention precision AP is given as:

$$AP = \frac{N_{in}}{N_{in} + N_{out}}, \quad (5.17)$$

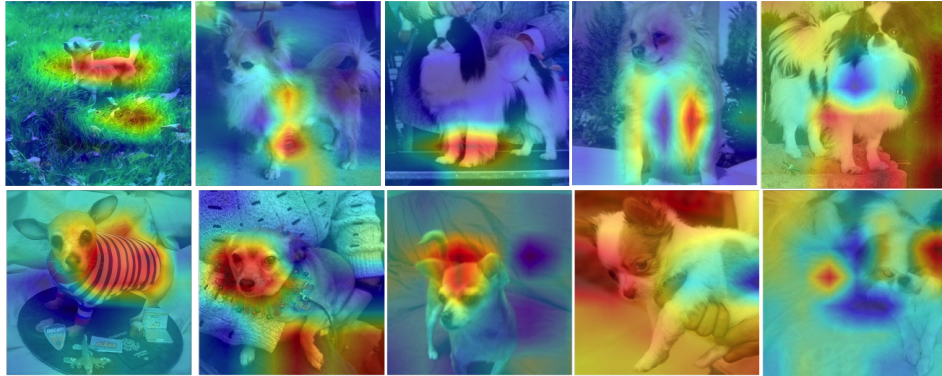


Figure 5.4: Visualization results of the attention of RMCSAM when there is no target object in the scene. In this figure, the network generating the Grad-CAMs is trained on Stanford Cars Dataset, but the input images are from the Stanford Dogs Dataset.

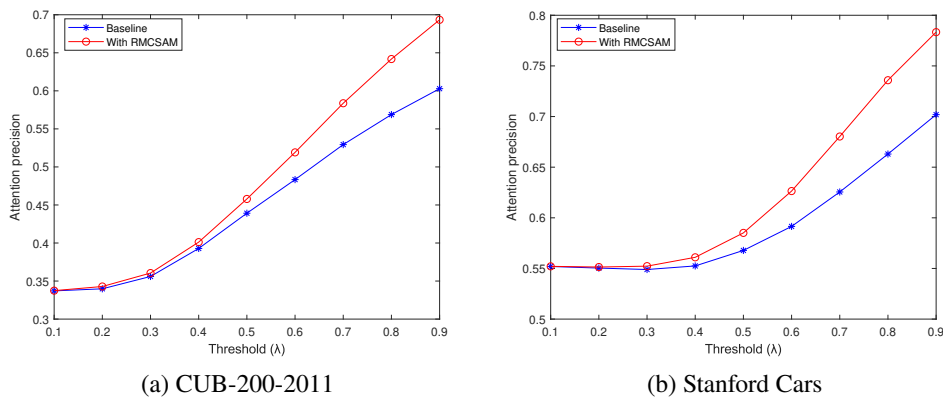


Figure 5.5: Attention precision with different thresholds

where N_{in} denotes the total number of pixels locating inside the manually labeled bounding box and having a value of 1. N_{out} denotes the total number of pixels locating outside the manually labeled bounding box and having a value of 1. The manually labeled bounding boxes are officially provided by the authors of the two datasets [116, 63]. The bounding boxes are widely used as the ground truth in fine-grained object detection or segmentation tasks [148, 2, 142].

The attention precision expresses the proportion of the pixels the networks “consider” to be discriminative actually are discriminative. We evaluate the at-

tention precision with different thresholds of 0.1~0.9. The results are shown in Figure 5.5. Overall, the network inserted with RMCSAM has much higher attention precision than the baseline. With the increase of λ , the gap of attention precision between them is getting wider and wider. A higher threshold selects the pixels that have more contribution to the final prediction. That is, the network inserted with RMCSAM tends to “consider” a higher proportion of pixels inside the bounding box as high-contribution pixels than the baseline.

5.3.5 Comparison with the State-of-the-art Attention Modules in Fine-grained Image Classification Task

In this subsection, we compare our proposed module with other state-of-the-art attention modules in FGIC tasks. We adopt Resnet50 as the backbone because it is the most commonly used network backbone for analyzing the performance of attention modules [46, 126, 89, 18]. The results are shown in Table 5.3.

Table 5.3: Comparison results with state-of-the-art attention modules in FGIC task

	Accuracy		Parameters	GFLOPs	Time Cost
	CUB-200-2011	Stanford Cars			
Resnet50+SE [46]	85.2%	93.9%	24.434M	16.439	52.487ms
Resnet50+AFF [18]	86.1%	94.1%	32.318M	17.676	54.294ms
Resnet50+iAFF [18]	85.6%	93.8%	34.423M	18.091	58.423ms
Resnet50+DAF [18]	85.8%	93.9%	28.108M	17.261	53.701ms
Resnet50+BAM [89]	85.7%	93.8%	24.998M	16.549	53.457ms
Resnet50+CBAM [126]	85.5%	93.4%	24.436M	16.440	53.322ms
Resnet50+RMCSAM (ours)	86.2%	94.2%	25.751M	16.449	54.100ms

The best accuracy and lowest overhead are highlighted in bold. Basically, the proposed attention module outperforms the other ones in terms of classification accuracy. The SE module [46], CBAM [126], and BAM [89] require lower overhead than our proposed module, but the accuracy of our proposed module is clearly higher than theirs on both datasets. AFF [18] has the closest classification accuracy to ours on both datasets but requires a little more time cost and much more GFLOPs and parameters.

5.3.6 Comparison with the Previous Approaches in Fine-grained Image Classification Task

In this subsection, we compare our proposed approach with the approaches achieving state-of-the-art accuracy in FGIC tasks. We use Resnet50 as the backbone because it is most widely used in those studies [82, 53, 131, 152, 23, 143]. As mentioned before, in previous subsections, we use the CNN backbones pre-trained on the ImageNet, but the parameters of the RMCSAM are initialized randomly. In this subsection, for better accuracy, we also present the experimental results by using the RMCSAM parameters pre-trained together with the Resnet50 on the ImageNet, which is marked as ★.

The pretraining is trained from scratch and conducted with the official Timm toolbox [124] on 2×Nvidia RTX 3080 Ti. We also train an original Resnet50 under the exact same configuration as a baseline. We turn on automatic mixed precision [79] and label smoothing [81]. We set the batch size as 256 and train the networks using standard Stochastic Gradient Descent (SGD) with the momentum of 0.9. We totally train the networks on the ImageNet for 180 epochs. Regarding the learning rate schedule, we divide the 180 epochs into 6×30 epochs. For the first 30 epochs, we train the Resnet50 with/without the RMCSAM by the constant learning rate of 0.1 for the quick decrease of training loss. From the second 30 epochs, we train the networks using cosine annealing [72], and the starting learning rate for the second 30 epochs is 0.05. Then, for every 30 epochs, we restart the cosine annealing schedule and decrease the starting learning rate by 0.7. The training of the baseline Resnet50 and the Resnet50 inserted with RMCSAM is conducted once. With RMCSAM, the average accuracy of the last 10 epochs on the validation set of the ImageNet is improved from 77.7% to 78.5%. The

best accuracy of the whole 180 epochs on the validation set of the ImageNet is improved from 78.1% to 78.9%.

All the other experiments in this subsection follow the general configuration of this paper. Namely, all the other experiments in this subsection are reproduced for 10 times, and we report the average accuracy. After the pre-training, we use the weights of the pre-trained RMCSAM to replace the randomly initialized RMCSAM weights for fine-tuning on the fine-grained image classification datasets.

Moreover, as an insertable module that can improve the accuracy of the backbone CNNs, our approach intuitively looks complementary to some state-of-the-art FGIC approaches. It is possible to combine our approach with other approaches for better accuracy. Specifically, we insert the RMCSAM pre-trained on the ImageNet into the Resnet50 backbone of PMG [23]. For a fair comparison, all the other parameters (including the parameters of the Resnet50 backbones) are initialized in the same way as the original PMG.

The comparison in this subsection is conducted in terms of both accuracy and computational costs. As many state-of-the-art FGIC approaches require extremely huge memory, such as [39], we test the time cost by processing one 448×448 image (i.e., a $1 \times 3 \times 448 \times 448$ tensor) to prevent the out-of-memory exception in this subsection. The comparison results are shown in Table 5.4. The best accuracy and lowest overhead are highlighted in bold. With the RMCSAM pre-trained on the ImageNet and Resnet50 backbone, the accuracy of our approach is very close to the state-of-the-art accuracy on the Stanford Cars and a little behind the state-of-the-art accuracy on the CUB-200-2011. TransFG achieves the best accuracy on the CUB-200-2011 but requires huge computational overhead regarding the parameters, GFLOPs, and time cost. In contrast, our approach requires much less overhead. Especially, our approach requires 13.694ms for processing a single image at once, which is the least time cost among the approaches and around 5.3% of the time cost of TransFG. Besides, our approach has the similar accuracy as TransFG on the Stanford Cars.

Table 5.4: Comparison results with state-of-the-art approaches in FGIC task

	Backbone	Accuracy		Parameters	GFLOPs	Time Cost
		CUB-200-2011	Stanford Cars			
GARD [152]	Resnet50	89.6%	94.3%	23.871M	18.589	17.848ms
DAM [131]	Resnet50	87.5%	94.4%	23.508M	49.314	64.285ms
PCA-Net [143]	Resnet50	88.3%	94.3%	21.270M	184.317	61.202ms
TransFG [39]	ViT-B_16 [22]	91.7%	94.8%	85.762M	107.564	259.633ms
ACNet [53]	Resnet50	88.1%	94.6%	197.264M	155.497	184.287ms
ProtoTree [82]	Resnet50	87.2%	91.5%	24.032M	8.270	160.731ms
PMG [23]	Resnet50	89.6%	95.1%	45.132M	37.316	20.913ms
RMCSAM	Resnet50	86.2%	94.2%	25.751M	16.449	13.694ms
RMCSAM*	Resnet50	87.2%	94.7%	25.751M	16.449	13.694ms
RMCSAM*+PMG	Resnet50	89.9%	95.3%	46.973M	37.328	25.904ms

* illustrates the RMCSAM that is pre-trained on the ImageNet.

In this table, the time cost is evaluated with a single image (i.e., a $1 \times 3 \times 448 \times 448$ tensor).

Among the approaches, PCA-Net [143] has the fewest parameters, and ProtoTree [82] has the fewest GFLOPs. However, they require much more time cost than the proposed approach, which is caused by the complex feature extracting and aggregating framework (PCA-Net) or the tree architecture hardly parallelizable (ProtoTree). Besides, on the Stanford Cars, our approach has better accuracy than both PCA-Net and ProtoTree.

By combining with our approach, the accuracy is improved by on both datasets. Especially, RMCSAM^{*}+PMG achieves 95.3% accuracy on Stanford Cars, which surpasses the previous best accuracy on this dataset. It achieves 89.9% accuracy on the CUB-200-2011, which surpasses the previous best accuracy obtained with Resnet50 backbone on this dataset. Among the 10 times of repeated experiments of RMCSAM^{*}+PMG, the lowest accuracies are 89.7% (CUB-200-2011) and 95.3% (Stanford Cars), while the highest accuracies are 90.0% (CUB-200-2011) and 95.5% (Stanford Cars). On both datasets, the highest, lowest and average accuracies of RMCSAM^{*}+PMG are better than the best accuracies reported in [23], which shows our approach can bring stable improvement over the original PMG. Considering that the accuracy of PMG, the state-of-the-art approach, is already very high, it is interesting to see there is still room for improvement by our proposed module. The improvement over PMG might be caused by the capability of RMCSAM to capture multi-scale attention information. In [23], Du *et al.* force the model to learn multi-granularity information by randomly shuffling all the local regions of the input images before feeding the images into the model. During each stage of the progressive learning framework, the input images are partitioned by different granularity. Our proposed module can improve the attention awareness of PMG towards the partitioned regions of multiple granularities.

5.4 Summary of This Chapter

In this chapter, we focus on addressing the difficulty of extra overhead for attention learning in the training procedure. By summing up the experience of previous chapters, we propose the recursive multi-scale channel-spatial attention module (RMCSAM), a new approach for capturing attention information in fine-grained image classification (FGIC) tasks. RMCSAM is designed by following the previous experience that localizing multi-scale attention regions is very effective

for FGIC. However, instead of region localizing strategy, RMCSAM is designed as an insertable attention module, which can capture channel-wise and spatial-wise attention of multiple scales and accordingly refine the deep feature maps to better correspond to the visual attention. The feature maps are recursively refined a predetermined number of times to obtain the finer feature map. In this way, RMCSAM requires a very small additional overhead. The experimental results show that the multi-scale channel-wise and spatial-wise attention are complementary, and aggregation of them brings better performance. Besides, the recursive refinement can further improve the accuracy. The experimental results also show that RMCSAM can improve the classification accuracy of widely used network backbones and is able to improve the attention capturing ability. RMCSAM also outperforms other attention modules in FGIC tasks. Moreover, our approach have very strong versatility. The proposed approach can be combined with PMG framework, which is state-of-the-art approach in the FGIC task, to further improve the accuracy. Overall, the approach proposed in this chapter largely reduce the extra training overhead, and also the extra testing overhead is very small.

Chapter 6

Conclusion

The main objective of this work is to explore how to efficiently and accurately capture attention information for fine-grained image classification (FGIC). Fine-grained image classification (FGIC) is a very difficult task, and attention information is the key for this task. The research question of this thesis is how to efficiently capture and utilize accurate attention information to improve the classification accuracy in fine-grained image classification? The research question involves two specific difficulties. The first is that attention information is hard to capture, and the second is that learning attention information requires much extra overhead. In this thesis, we propose three novel frameworks to address the problems. Chapter 3 mainly focus on reducing the training difficulty of capturing attention information. Chapters 4 and 5 focus mainly on reducing the extra overheads.

Chapter 3 mainly aims to attention regions for FGIC. Based on the Spatial Transformers' capability of spatial manipulation within networks, we propose an extension model, the Attention-Guided Spatial Transformer Networks (AG-STNs). This model can guide the Spatial Transformers with hard-coded attention regions at first. Then such guidance can be turned off, and the network model will adjust the region learning in terms of the location and scale. Such adjustment is conditioned to the classification loss so that it is actually optimized for better recognition results. With this model, we are able to successfully capture detailed attention information. Also, the AG-STNs are able to capture attention information in multiple levels, and different levels of attention information are complementary to each other in

our experiments. A fusion of the information learned from them brings better results.

Chapter 4 mainly aims to reduce the extra overhead for the testing (utilization) procedure. Inspired by the human behavior of learning from experience to complete new tasks, we propose a multi-task learning framework, named Contrastively-reinforced Attention Convolutional Neural Network (CRA-CNN), which forces the major network to respond better to discriminative regions using a subordinate network. The major network is required to predict the categories and attention redundancy pairs of input images. The subordinate network evaluates the attention prediction by categorizing the attention and measuring the similarity and dissimilarity of attention/redundancy. CRA-CNN improves the attention awareness of the deep features in the major network by the tasks performed by the subordinate network and consequently improves the accuracy. The subordinate network can be removed after training, and CRA-CNN has no extra overhead for utilization.

Chapter 5 mainly aims to reduce the extra overhead for the training procedure, which also has very small testing (utilization) overheads. The approach in Chapter 5 is proposed by summing up the experience of Chapter 3 and Chapter 4. Chapter 3 verifies the effectiveness of multi-scale attention and Chapter 4 verifies the effectiveness of exploring attention from deep features. In Chapter 5, we propose the recursive multi-scale channel-spatial attention module (RMCSAM). Following the experience of previous research on fine-grained image classification, RMCSAM explores multi-scale attentional information. The attentional information is explored by recursively refining the deep feature maps of a convolutional neural network (CNN) to better correspond to multi-scale channel-wise and spatial-wise attention, instead of localizing attention regions. The spatial-wise attention of RMCSAM is extended from the region-based attention strengthening strategy used in Chapter 4. In Chapter 5, we use the attention module framework to replace the multi-task learning framework for capturing deep-feature-based attention. In this way, RMCSAM provides a lightweight module that can be inserted into standard CNNs. Experimental results show that RMCSAM can improve the classification accuracy and attention capturing ability over baselines. Also, RMCSAM performs better than other state-of-the-art attention modules in fine-grained image classification and is complementary to some state-of-the-art approaches for

fine-grained image classification.

DISCUSSION ON THE CAPTURED ATTENTION. Figure 3.5, 3.6, 4.3, 5.3, and 5.4 are visualization results of the approaches proposed in this thesis. From those visualization results, we can observe some phenomenons about the attention captured by the proposed approaches.

First, the ratio of foreground to background affects the captured attention. Especially in Figures 4.3 and 5.3, this phenomenon is obvious. For example, when the bird occupies a large portion of the whole scene, the important body parts, such as the head, are captured as attention. When the bird occupies a small portion of the whole scene, the attention mechanism mainly devotes effort to distinguishing the small bird from the complicated background.

Second, for the task of fine-grained image classification, some key parts are acting the most important roles. For example, for distinguishing different bird images, the head always acts the most important role. For distinguishing different car images, the brand logo and headlight always acts the most important role. However, only capturing those key parts is not enough for the FGIC task because of the strong intra-class variation. For example, in some cases, the bird head is occluded due to the complicated environment or unclear due to the bird being too small, and the model has to learn clues from more parts. Therefore, a good attention capturing strategy should not only be able to be sensitive to the most important parts but also adjust the capturing strategy condition on specific images. We suppose it is the main reason why the proposed deep-learned outperforms traditional hard-coded attention-learning approaches. Furthermore, our experimental results show that capturing multi-scale attention is also an effective way to deal with intra-class variation.

Third, as shown in Figure 5.4, when there is no target object in the scene, the attention mechanism still has some rough telling foreground and background in certain cases.

FUTURE PLAN AND PROSPECT. RMCSAM, the approach proposed in Chapter 5, is the most recommendable approach of this thesis due to its small extra overhead, good accuracy, and strong universality. However, compared with CRA-CNN, the approach proposed in Chapter 4, RMCSAM has a drawback. That is, in CRA-CNN, all the extra overhead required for attention learning can be removed after the training, and there is no extra overhead for utilization. Regarding RM-

CSAM, the extra training overhead is very small but still required for utilization. The future plan is to design a new framework, where the model is trained to capture multi-scale attention by small extra overhead, and the extra overhead can be removed after training, based on the experience of this thesis.

Another important thing that we need to verify in the future is whether capturing attention with the RMCSAM can be used as an additional task to strengthen the attention awareness of the networks. We plan to conduct new experiments by redesigning new additional tasks based on the RMCSAM. If we can realize using RMCSAM as an additional task we can further reduce the training overhead.

Moreover, the attention for FGIC tasks is very complicated due to various foreground/background ratios and occluded key parts, and it is important to adjust the attention capturing strategy conditioned on each specific image. Thus, it is worth expecting to further develop the attention capturing to automatically zoom in or out the captured attention for unification and complement the occluded key parts with generation models.

Acknowledgment

The author would like to express my sincere gratitude to my advisor Professor Jien Kato and Professor Kenji Mase for their continuous support of my Ph.D. study and other related research. Without their patience, kind guidance, and immense knowledge, this thesis would not have been actualized. I want to express my gratitude to Assistant Professor Yu Wang, as a co-author, and Ms Longjiao Zhao, as a colleague, for their support and insightful comments during my Ph.D. study. I would like to thank my thesis committee, Professor Yoshiharu Ishikawa, Associate Professor Deguchi Daisuke, and Lecturer Yu Enokibori for their precious time and insightful comments. Those comments are not only valuable and helpful for revising and improving my thesis, but also have great guiding significance to my research as well as the future plan.

Finally, I appreciate my family members, Shixian Liu, Yilai Liu, Sihan Zhang, Ningzhi Liu, for their understanding and kind support during my Ph.D. study.

Dichao Liu

Bibliography

- [1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009.
- [2] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 811–818, 2013.
- [3] Jimmy Ba, Ruslan R Salakhutdinov, Roger B Grosse, and Brendan J Frey. Learning wake-sleep recurrent attention models. In *Advances in Neural Information Processing Systems*, pages 2593–2601, 2015.
- [4] Ali Borji, Simone Frintrop, Dicky N. Sihite, and Laurent Itti. Adaptive object tracking by learning background context. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 23–30, 2012.
- [5] Farha Al Breiki, Muhammad Ridzuan, and Rushali Grandhe. Self-supervised learning for fine-grained image classification. *CoRR*, abs/2107.13973, 2021.
- [6] Dongliang Chang, Kaiyue Pang, Yixiao Zheng, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Your "flamingo" is my "bird": Fine-grained, or not. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11476–11485, 2021.

- [7] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32, 2021.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.
- [10] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.
- [11] Zhuo Chen, Fei Yin, Xu-Yao Zhang, Qing Yang, and Cheng-Lin Liu. Multirenets: Multilingual text recognition networks for simultaneous script identification and handwriting recognition. *Pattern Recognition*, page 107555, 2020.
- [12] Anoop Cherian and Stephen Gould. Second-order temporal pooling for action recognition. *arXiv preprint arXiv:1704.06925*, 2017.
- [13] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015.
- [14] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123, 2019.
- [15] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018.

- [16] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge J Belongie. Kernel pooling for convolutional neural networks. In *CVPR*, volume 1, page 7, 2017.
- [17] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017.
- [18] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3560–3569, 2021.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [21] Yifeng Ding, Zhanyu Ma, Shaoguo Wen, Jiyang Xie, Dongliang Chang, Zhongwei Si, Ming Wu, and Haibin Ling. Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, 30:2826–2836, 2021.
- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [23] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, pages 153–168. Springer, 2020.
- [24] Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification.

- In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2018.
- [25] Erkut Erdem and Aykut Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4):11–11, 2013.
- [26] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.
- [27] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.
- [28] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.
- [29] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):989–1005, 2009.
- [30] Fei Gao, Hyunsoo Yoon, Teresa Wu, and Xianghua Chu. A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications*, 143:112957, 2020.
- [31] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.
- [32] Zhi Gao, Yuwei Wu, Xingyuan Bu, Tan Yu, Junsong Yuan, and Yunde Jia. Learning a robust representation via a deep network on symmetric positive definite manifolds. *Pattern Recognition*, 92:1–12, 2019.

- [33] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019.
- [34] Habiba Gitay, Avelino Suárez, Robert T Watson, and David Jon Dokken. *Climate change and biodiversity*. 2002.
- [35] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 34(10):1915–1926, 2011.
- [36] Pei Guo and Ryan Farrell. Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1876–1885. IEEE, 2019.
- [37] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.
- [38] Oskar LP Hansen, Jens-Christian Svenning, Kent Olsen, Steen Dupont, Beulah H Garner, Alexandros Iosifidis, Benjamin W Price, and Toke T Høye. Species-level image classification with convolutional neural network enables insect identification from habitus images. *Ecology and evolution*, 10(2):737–747, 2020.
- [39] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille. Transfg: A transformer architecture for fine-grained recognition. *arXiv preprint arXiv:2103.07976*, 2021.
- [40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [41] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural

- networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.
- [42] Xiangteng He, Yuxin Peng, and Junjie Zhao. Fast fine-grained image classification via weakly supervised discriminative localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1394–1407, 2018.
- [43] Geoffrey E Hinton. Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10):428–434, 2007.
- [44] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2007.
- [45] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. 2009.
- [46] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [47] Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019.
- [48] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [49] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1173–1182, 2016.
- [50] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868, 2019.

- [51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [52] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.
- [53] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10468–10477, 2020.
- [54] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems*, pages 5995–6004, 2018.
- [55] Xuebo Jin, Zhi Tao, and Jianlei Kong. Multi-stream aggregation network for fine-grained crop pests and diseases image recognition. *International Journal of Cybernetics and Cyber-Physical Systems*, 1(1):52–67, 2021.
- [56] Shiva Kamkar and Reza Safabakhsh. Vehicle detection, counting and classification in various conditions. *IET Intelligent Transport Systems*, 10(6):406–413, 2016.
- [57] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
- [58] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.
- [59] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chelappa. The devil is in the details: Self-supervised attention for vehicle

- re-identification. In *European Conference on Computer Vision*, pages 369–386. Springer, 2020.
- [60] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–122, 2018.
- [61] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7025–7034. IEEE, 2017.
- [62] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5546–5555, 2015.
- [63] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.
- [64] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- [65] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368. IEEE, 2011.
- [66] Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 947–955, 2018.

- [67] Yang Li, Kan Li, and Xinxin Wang. Recognizing actions in images by fusing multiple body structure cues. *Pattern Recognition*, page 107341, 2020.
- [68] Mingpei Liang, Xinyu Huang, Chung-Hao Chen, Xin Chen, and Alade Tokuta. Counting and classification of highway vehicles by regression analysis. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2878–2888, 2015.
- [69] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1666–1674, 2015.
- [70] Tsung-Yu Lin and Subhansu Maji. Improved bilinear pooling with cnns. *arXiv preprint arXiv:1707.06772*, 2017.
- [71] Pau Rodriguez Lopez, Diego Velazquez Dorta, Guillem Cucurull Preixens, Josep M Gonfaus Sitjes, Francesc Xavier Roca Marva, and Jordi Gonzalez. Pay attention to the activations: a modular attention mechanism for fine-grained image recognition. *IEEE Transactions on Multimedia*, 2019.
- [72] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations*, 2017.
- [73] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, 2005.
- [74] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.
- [75] Zhichao Lu, Gautam Sreekumar, Erik Goodman, Wolfgang Banzhaf, Kalyanmoy Deb, and Vishnu Naresh Boddeti. Neural architecture transfer. *arXiv preprint arXiv:2005.05859*, 2020.

- [76] Wei Luo, Hengmin Zhang, Jun Li, and Xiu-Shen Wei. Learning semantically enhanced feature for fine-grained image classification. *IEEE Signal Processing Letters*, 2020.
- [77] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4905–4913, 2016.
- [78] Luke Melas-Kyriazi. Do you even need attention? A stack of feed-forward layers does surprisingly well on imagenet. *CoRR*, abs/2105.02723, 2021.
- [79] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.
- [80] Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *BMVC*, 2016.
- [81] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.
- [82] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.
- [83] Tam V. Nguyen and Luoqi Liu. Salient object detection with semantic priors. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4499–4505, 2017.
- [84] Tam V Nguyen, Qi Zhao, and Shuicheng Yan. Attentive systems: A survey. *International Journal of Computer Vision*, 126(1):86–110, 2018.
- [85] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

- [86] Cornell Lab of Ornithology. Similar species to elegant tern. https://www.allaboutbirds.org/guide/Elegant_Tern/species-compare/.
- [87] Rajarshi Pal. Computational models of visual attention: a survey. In *Research developments in computer vision and image processing: methodologies and applications*, pages 54–76. IGI Global, 2014.
- [88] Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisín Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10583–10592, 2021.
- [89] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. A simple and light-weight attention module for convolutional neural networks. *International Journal of Computer Vision*, 128(4):783–798, 2020.
- [90] Omkar M Parkhi, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. The truth about cats and dogs. In *2011 International Conference on Computer Vision*, pages 1427–1434. IEEE, 2011.
- [91] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [92] Xiaojiang Peng, Limin Wang, Yu Qiao, and Qiang Peng. Boosting vlad with supervised dictionary learning and high-order statistics. In *European Conference on Computer Vision*, pages 660–674. Springer, 2014.
- [93] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, 2017.
- [94] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012.

- [95] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- [96] Tianrong Rao, Xiaoxu Li, Haimin Zhang, and Min Xu. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing*, 333:429–439, 2019.
- [97] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018.
- [98] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2014.
- [99] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.
- [100] Kara Marie Schatz, Erik Quintanilla, Shruti Vyas, and Yogesh S Rawat. A recurrent transformer network for novel view action synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 410–426. Springer, 2020.
- [101] Rainer Schliep, Ulrich Walz, Ulrich Sukopp, and Stefan Heiland. Indicators on the impacts of climate change on biodiversity in germany—data driven or meeting political needs? *Sustainability*, 10(11):3959, 2018.
- [102] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.
- [103] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015.

- [104] Yiqing Shi, Wenzhong Guo, Yuzhen Niu, and Jiamei Zhan. No-reference stereoscopic image quality assessment using a multi-task cnn and registered distortion representation. *Pattern Recognition*, 100:107168, 2020.
- [105] Chang Shu, Xi Chen, Chong Yu, and Hua Han. A refined spatial transformer network. In *International Conference on Neural Information Processing*, pages 151–161. Springer, 2018.
- [106] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1143–1151, 2015.
- [107] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
- [108] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [109] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [110] Ming Sun, Yuchen Yuan, Feng Zhou, and Errui Ding. Multi-attention multi-class constraint for fine-grained image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–821, 2018.
- [111] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [112] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

- [113] Graham W Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. Convolutional learning of spatio-temporal features. In *European conference on computer vision*, pages 140–153. Springer, 2010.
- [114] Xiao Tianjun, Xu Yichong, Yang Kuiyuan, Zhang Jiaying, Peng Yuxin, and Zhang Zheng. 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*, pages 842–850, 2015.
- [115] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [116] 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.
- [117] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.
- [118] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.
- [119] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.
- [120] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2015.

- [121] Xiaoyu Wang, Tianbao Yang, Guobin Chen, and Yuanqing Lin. Object-centric sampling for fine-grained image classification. *arXiv preprint arXiv:1412.3161*, 2014.
- [122] Yanfei Wang, Fei Huang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. *Pattern Recognition*, 100:107148, 2020.
- [123] Yifan Wang, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Two-stream sr-cnns for action recognition in videos. In *BMVC*, 2016.
- [124] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019.
- [125] Wei Jing Wong and Shang-Hong Lai. Multi-task cnn for restoring corrupted fingerprint images. *Pattern Recognition*, 101:107203, 2020.
- [126] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [127] Di Wu, Siyuan Li, Zelin Zang, Kai Wang, Lei Shang, Baigui Sun, Hao Li, and Stan Z. Li. Align yourself: Self-supervised pre-training for fine-grained recognition via saliency alignment. *CoRR*, abs/2106.15788, 2021.
- [128] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.
- [129] Guo-Sen Xie, Xu-Yao Zhang, Wenhan Yang, Mingliang Xu, Shuicheng Yan, and Cheng-Lin Liu. Lg-cnn: From local parts to global discrimination for fine-grained recognition. *Pattern Recognition*, 71:118–131, 2017.
- [130] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan, and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In *Proceedings of the IEEE international conference on computer vision*, pages 1641–1648, 2013.

- [131] Furong Xu, Meng Wang, Wei Zhang, Yuan Cheng, and Wei Chu. Discrimination-aware mechanism for fine-grained representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 813–822, 2021.
- [132] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.
- [133] Qingyang Xu and Li Zhang. The effect of different hidden unit number of sparse autoencoder. In *The 27th Chinese Control and Decision Conference (2015 CCDC)*, pages 2464–2467, 2015.
- [134] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision*, pages 420–435, 2018.
- [135] Bangpeng Yao, Aditya Khosla, and Li Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR 2011*, pages 1577–1584, 2011.
- [136] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.
- [137] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [138] Chuanyi Zhang, Yazhou Yao, Xing Xu, Jie Shao, Jingkuan Song, Zechao Li, and Zhenmin Tang. Extracting useful knowledge from noisy web images via data purification for fine-grained recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4063–4072, 2021.

- [139] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152, 2016.
- [140] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013.
- [141] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE international conference on computer vision*, pages 1404–1412, 2015.
- [142] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.
- [143] Tian Zhang, Dongliang Chang, Zhanyu Ma, and Jun Guo. Progressive co-attention network for fine-grained visual classification. *arXiv preprint arXiv:2101.08527*, 2021.
- [144] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1134–1142, 2016.
- [145] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. In *International Conference on Learning Representations*, 2020.
- [146] Y. Zhang and Qiang Yang. A survey on multi-task learning. *ArXiv*, abs/1707.08114, 2017.
- [147] 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.
- [148] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, 14(2):119–135, 2017.
- [149] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018.
- [150] Jiejie Zhao, Bowen Du, Leilei Sun, Weifeng Lv, Yanchi Liu, and Hui Xiong. Deep multi-task learning with relational attention for business success prediction. *Pattern Recognition*, page 107469, 2020.
- [151] Junjie Zhao, Yuxin Peng, and Xiangteng He. Attribute hierarchy based multi-task learning for fine-grained image classification. *Neurocomputing*, 395:150–159, 2020.
- [152] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15079–15088, 2021.
- [153] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.
- [154] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 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*, pages 5012–5021, 2019.
- [155] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016.

- [156] Shaoxiong Zhou, Shengrong Gong, Shan Zhong, Wei Pan, and Wenhao Ying. Region selection model with saliency constraint for fine-grained recognition. In *International Conference on Neural Information Processing*, pages 365–376. Springer, 2019.
- [157] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13130–13137, 2020.