Bi-Modal Progressive Mask Attention for Fine-Grained Recognition

Kaitao Song, *Student Member, IEEE,* Xiu-Shen Wei, *Member, IEEE,* Xiangbo Shu, *Member, IEEE,* Ren-Jie Song, and Jianfeng Lu, *Member, IEEE*

***Abstract*—Traditional fine-grained image recognition is required to distinguish different subordinate categories (*e.g.*, birds species) based on the visual cues beneath raw images. Due to both small inter-class variations and large intra-class variations, it is desirable to capture the subtle differences between these sub-categories, which is crucial but challenging for fine-grained recognition. Recently, language modality aggregation has been proved as a successful technique to improve visual recognition in the experience. In this paper, we introduce an end-to-end trainable *Progressive Mask Attention (PMA)* model for fine-grained recognition by leveraging both visual and language modalities. Our Bi-Modal PMA model can not only stage-by-stage capture the most discriminative part in the visual modality by our mask-based fashion, but also explore the out-of-visual-domain knowledge from the language modality in an interactional alignment paradigm. Specifically, at each stage, a self-attention module is proposed to attend to the key patch from images or text descriptions. Besides, a query-relational module is designed to seize the key words/phrases of texts and further bridge the connection between two modalities. Later, the learned representations of bi-modality from multiple stages are aggregated as the final features for recognition. Our Bi-Modal PMA model only needs raw images and raw text descriptions, without requiring bounding boxes/part annotations in images or key word annotations in texts. By conducting comprehensive experiments on fine-grained benchmark datasets, we demonstrate that the proposed method achieves superior performance over the competing baselines, on either vision and language bi-modality or single visual modality.**

***Index Terms*—Fine-Grained Visual Recognition; Multi-Modal Analysis; Deep Neural Networks; Language Modality.**

I. Introduction

The task of fine-grained image recognition is to identify the species of birds [1], flowers[2], cars [3] and aircrafts [4] by mining the visual cues beneath raw images. It has been applied in diverse real-world scenarios, *e.g.*, biological protection [5], [6], vehicle identification [7], product recognition [8] and so on. Since the subordinate categories are all similar to each other, different sub-categories can only be distinguished by slight and subtle differences, which makes fine-grained recognition a challenging problem. Compared to the general object recognition task, fine-grained recognition benefits more from learning critical parts of objects, which helps discriminate different sub-categories and align objects of the same sub-category [9], [10], [11].

In the literature, a number of effective fine-grained image recognition methods have been developed [12], [6], [11]. However, these methods focus on the vision technologies for improving classification accuracy, which might be merely restricted in the single visual modality. Recently, some works, *e.g.*, [13], [14], [15] of fine-grained recognition attempted to boost the recognition performance by leveraging bi-modality / cross-modality analysis. Specifically, except for the traditional visual modality, [14] simply combined the information of vision and language streams and obtained a good performance gain. [13], [15] introduced the knowledge-based information into fine-grained recognition to enhance fine-grained feature learning.

In this paper, we propose a novel fine-grained method, termed as *Progressive Mask Attention* (*PMA*), which explores bi-modal analysis for fine-grained recognition. PMA unifies a progressive mask strategy, which can be friendly applied in both visual and language modalities simultaneously, which reveals its flexibility and scalability. Additionally, compared with the strong supervisions of fine-grained images (*e.g.*, bounding boxes and part annotations), text descriptions (like sentences and phrases) are weak supervisions, and they can also provide semantics that the visual domain is unable to display. Moreover, text descriptions can be relatively accurately returned by ordinary people, rather than the domain experts.

In order to apply our progressive mask strategy to address fine-grained recognition, we also propose the Bi-Modal Progressive Mask Attention (Bi-Modal PMA) model. It can seize crucial information from both visual and text streams. More specifically, raw images and texts are firstly processed by convolutional neural networks and long-short term memory networks into the deep visual descriptors and noun phrases embeddings, respectively. Then, the processed bi-modal representations are fed into our Bi-Modal PMA to produce a joint-representation. In our Bi-Modal PMA, a self-attention module is designed to extract semantics from visual or language modality. By employing various mask templates, the semantics collected by self-attention mechanisms can be used to locate key parts in the visual modality, or capture the out-of-visual-domain knowledge in the language modality. In addition, a query-relational module is designed to bridge the connection from

• X.-S. Wei and J. Lu are the corresponding authors.
• K. Song, X. Shu and J. Lu are with School of Computer Science and Engineering, Nanjing University of Science and Technolody. X.-S. Wei is with the PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology. R.-J. Song is with Megvii Research Nanjing, Megvii Technolody. Part of this work was done when K. Song was an intern in Megvii Research Nanjing.
• This work is supported by the National Key Research and Development Program of China (2017YFB1300205), National Natural Science Foundation of China (Grant No. 61702265), and Natural Science Foundation of Jiangsu Province (Grant No. BK20170856)..
• Emails: {kt.song, shuxb, lujf}@njust.edu.cn, weixs.gm@gmail.com, songrenjie@megvii.com.

the key words/phrases of the language modality to the critical part of the visual modality. Moreover, thanks to the proposed attention approach, we can align the representations of two modalities and obtain more discriminative and abundant joint-features. After that, we can obtain global-level image features, part-level image representations, global-level text features and the aligned part-level text representations at the same time.

Beyond that, by introducing our proposed progressive mask strategy, we can stack multiple PMA modules in a stage-by-stage way. Our PMA is tailored for fine-grained recognition, which can attend on a set of distinct and non-overlap parts progressively. It will significantly boost the final fine-grained recognition accuracy. Concretely, after PMA outputs the attended parts of bi-modality at the previous stage, these part information is treated as the inputs to the following stage. We propose a mask-based strategy to discard the located part of objects in vision and filter out the relevant noun phrases of the located parts in languages simultaneously. Because the most important image regions and noun phrases are omitted, PMA at that stage can focus on the secondary important parts of both vision and language. Under this paradigm, the discarded regions/phrases will be progressively expanded. Therefore, our method can work in a multiple-stages fashion iteratively, which benefits to fine-grained recognition.

In addition, sometimes, we might meet some scenarios without any text-level data (*i.e.*, single-modality) in the downstream tasks. In order to utilize our model in such cases (*i.e.*, conducting model inference without textual information), we further develop a knowledge distillation approach to distill the generalization ability of our Bi-Modal PMA on bi-modality into a student model can only deal with image data as inputs. Thanks to our distiller, our student model is able to make accurate predictions merely with test images, and almost matches the accuracy of our model by using both visual and language modalities.

The major contributions of this paper are as follows:

- We introduce a unified framework, termed as Progressive Mask Attention, to incorporate discriminative cues of both visual and language modalities for dealing with the fine-grained recognition task.
- We specifically devise an attention-based method for each modality to capture the important object parts to form part-level representations. Moreover, a stage-wise mask-based strategy is developed to stack these attention units. Thus, the whole model can progressively locate a set of discriminative but different key parts, or utilize text descriptions to furnish the out-of-visual-domain knowledge.
- We further develop a knowledge distiller to compress the knowledge of visual and language modalities into the object-level model, which allows model to make predictions using only image data.
- We conduct comprehensive experiments on four fine-grained benchmark datasets, and our proposed model achieves superior performance over competing solutions on either bi-modality or single visual modality.

The rest of our paper is organized as follows: Section II reviews previous works in fine-grained recognition, multi-modality analysis and attention mechanism. Section III elaborates the detailed design of our method on visual and language modalities. Section IV presents our experimental settings, and Section V reports the results on four public datasets, as well as ablation studies. Finally, we conclude our work in Section VI.

II. Related work

In this section, we briefly review the related work about fine-grained recognition, multi-modality analysis and attention mechanism.

A. Fine-grained Recognition

Fine-grained recognition [4], [2], [1] is a challenging problem in computer vision and has recently emerged as a hot topic [16], [10], [6], [9], [11], [17], [18]. To push a satisfactory fine-grained classification accuracy, researchers focus on how to locate the discriminative but subtle object parts, which is the crucial issue of fine-grained recognition. Specifically, Part-based R-CNN [19] applied R-CNN [16] for part localization and used CNNs to extract part representations for predictions. Pose-Normalized CNN [5] relied on the part annotations to accomplish the part alignments of the fine-grained objects. Part-Alignment CNN [10] proposed a concept of co-segmentation for part alignment. Although the promising results have been achieved with the help of the human annotations, it is still expensive and time-consuming to label bounding boxes or part annotations in large-scale data scenarios. Thus, recently, researchers attempted to analyze the convolutional response to explore discriminative representations only with image-level labels. For example, Bilinear CNN [6] computed the outer product of the outputs from two feature extractors to capture localized feature interactions. State-of-the-art methods, *e.g.*, [20], [21], [22], [23], learned part detectors via an unsupervised learning fashion. In addition, Region-Attention CNN [9] introduced an attention proposal network to extract multi-scale discriminative region features in a coarse-to-fine recurrent way. Multi-Attention CNN [11] leveraged channel grouping to attend on multiple discriminative parts and achieved good classification performance. However, these methods are restricted in the single visual modality and can not be applied to the language modality directly, let alone bi-modalities.

B. Multi-modality analysis

Multi-modality analysis, including bi-modality analysis, has attracted a lot of attentions with the rapid growth of multi-media data (*e.g.*, image, text, knowledge base, etc). It is studied and applied in many various computer vision applications [24], [25], [26], [27], such as image caption, visual question answering and so on. In fine-grained recognition, it is also used to take multi-modality information to establish joint-representation for better classification accuracy [28], [14], [13], [15]. In addition, compared with other methods using only image cues, incorporating multi-modality information is able to boost the recognition performance. Specifically, [28] first collected text descriptions and introduced a structured joint embedding for zero-shot image recognition by combining texts and images.

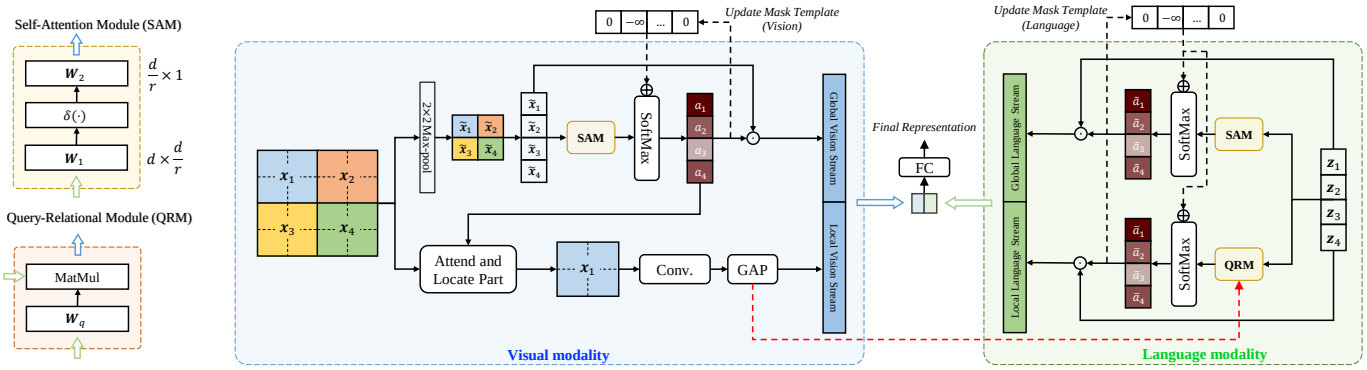


Figure 1: Detailed architecture of our Bi-Modal PMA module. Figures in the left is our defined components "SAM" and "QRM". "$\odot$" is the weighted sum operation, "$\oplus$" is the add operation, and $\delta(\cdot)$ is the non-linear activation function. The sub-module "Attend and locate part" means locating the most discriminative part of the feature map according to the maximum attention weight a_i. The red dashed means using the compressed output from the located parts to query language modality, and $\mathbf{z}_i$ of the language modality is the vector of phrase embedding. Here, we show four 2×2 blocks in that feature map as a simple and clear example.

[14], [29] combined the vision and language streams for fine-grained recognition, which achieved a good classification accuracy. Except for text data, some other work, *e.g.*, [13], [15], introduced the knowledge-based information into fine-grained recognition to enhance fine-grained feature learning. However, previous work merely focused on extracting joint-representations while ignored the correlations between different modalities to explore discriminative part features. Aiming at this point, our method is designed to capture the discriminative parts of bi-modalities (*i.e.*, key parts of fine-grained objects and key words of the corresponding text descriptions). More importantly, our model can also align them in an interactional way.

C. Attention mechanisms

Recently, attention mechanism is one kind of fundamental and effective strategy in deep learning, especially in natural language processing tasks [30], [31] and computer vision problems [25], [24], [32]. Attention is usually considered to excavate to what extent the input states will affect model accuracy. Generally, there existed many diverse attention forms in the previous work. Attention can be used to model the dependency among different domains (*e.g.*, source and target languages, images and texts) [30], [25] . Besides, there also existed some other work, *e.g.*, [31], abstracting semantic information from single domains. In this paper, we propose a unified framework, *i.e.*, bi-modal progressive mask attention, to perform attentions on both visual and language modalities. Specifically, a self-attention and a query-based attention method are developed for capturing crucial cues of images and texts, respectively. Furthermore, our bi-modal progressive mask attention can be able to stage-by-stage fuse the information of two modalities in an interactional way.

III. Approach

In this section, we introduce our Bi-Modal Progressive Mask Attention (Bi-Modal PMA) framework by elaborating its two key modules, *i.e.*, the visual-based PMA module and language-based PMA module. Specifically, Bi-Modal PMA is designed to fuse the information of visual and language modalities in an interactional way at each stage. While, we use PMA to iteratively locate the top-tier discriminative parts and the relevant phrases from bi-modality stage-by-stage. Figure 1 shows us the architecture of our Bi-Modal PMA model.

A. Notations

At the beginning of our introduction, we first give some pre-defined components which will be used in the following sections. Figure 1 also illustrates these components.

1) Self-Attention Module (SAM): SAM is a component used to gather semantic from the single modality. Assume $\boldsymbol{x} \in \mathbb{R}^d$ as the input, thus the formulation of SAM is presented by

$$\mathrm{SAM}(\boldsymbol{x}) = \mathbf{W}_2 \cdot \delta(\mathbf{W}_1 \cdot \boldsymbol{x}), \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times \frac{d}{r}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{d}{r} \times 1}$ are learnable matrices, and $r = 16$ is a reduction ratio. $\delta(\cdot)$ refers to the ReLU [33] activation function (in visual) or the tanh activation function (in language), respectively.

2) Query-Relational Module (QRM): QRM is used to establish the connection between visual and language modalities. It is able to guide visual feature to query the relevant keys in language modality. Let denote $\boldsymbol{x} \in \mathbb{R}^{d_1}$ as the key vector and $\boldsymbol{y} \in \mathbb{R}^{d_2}$ as the query vector, so the formulation of QRM is as:

$$\mathrm{QRM}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y} \odot (\mathbf{W}_q \cdot \boldsymbol{x}), \tag{2}$$

where $\odot$ represents the dot product and $\mathbf{W}_q \in \mathbb{R}^{d_2 \times d_1}$ is a learnable matrix.

3) Mask Template: We assume $\mathbf{M} = \{m_1, \ldots, m_n\}$ as the mask template, which will be adopted by our progressive mask strategy. Here n is consistent with the quantity of the input vectors and $m_i \in \{0, -\infty\}$, where $\boldsymbol{m}_i$ is 0 at initialization during each training step.[1] We define the mask

[1]Since the mask template is added before the softmax function, $m_i = 0$ will let attention model to keep the original weight while $m_i = -\infty$ will force the outputs of attention model as 0.

template $\mathbf{M}$ for visual modality and language modality as $\mathbf{M}^V = \{m_1^v, \ldots, m_n^v\}$ and $\mathbf{M}^T = \{m_1^t, \ldots, m_n^t\}$, respectively. Its element will be updated progressively by a stage-by-stage fashion, and will be elaborated in the following sub-section.

B. Progressive Mask Attention in Visual Modality

Discriminative part localization is a common and core technique for fine-grained recognition in the visual domain. In this section, we design a self-attention mechanism to locate the most discriminative part from the original image. More importantly, we apply a progressive mask strategy into attention module to attend on a set of distinct and non-overlap parts stage-by-stage, while most of the existing attention-based methods for multiple discriminative part localization are only focusing on few important parts repeatedly. Specifically, for each attended stage, we use a mask to discard the located part in the previous stage. Thus, our PMA can locate discriminative but distinct parts in different stages. In order to boost fine-grained recognition performance, we also aggregate the global image semantic calculated by attention weights and discriminative part features as the final state of a single stage. In addition, the feature vector of the discriminative part in the current stage will be used in the language modality for textual guidance.

1) Input Preparation: Given an image, we use a conventional CNN to encode it and obtain the outputs from the last convolutional layer (*e.g.*, `conv5_3` in VGG-16). Let it be $\mathbf{X} \in \mathbb{R}^{h \times w \times d}$. For fine-grained recognition, large image resolutions could benefit to capture discriminative but subtle objects' parts, but will also increase computational burdens due to the increment of the number of deep descriptors. We hereby employ an additional 2×2 max-pooling operator to gather more compact information, which can also reduce the number of final descriptors without affecting subtle details captured by large resolutions. Meanwhile, the additional pooling could improve the receptive field of these pooled descriptors. Thus, we consider $\mathbf{X}$ as a set of 2×2 blocks, where $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and $n = \frac{h \times w}{4}$. $\boldsymbol{x}_i \in \mathbb{R}^{2 \times 2 \times d}$ represents the i-th 2×2 block containing four d-dimension deep descriptors. Then, we append a 2×2 max-pooling on $\mathbf{X}$ to make each feature map contain more compact information. Thus, the output is denoted as $\tilde{\mathbf{X}} = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n\}$, where $\tilde{\boldsymbol{x}}_i = f_{2\times2-\text{maxPool}}(\boldsymbol{x}_i) \in \mathbb{R}^{1 \times 1 \times d}$ is the aggregated local feature vector. After that, $\tilde{\mathbf{X}}$ is used as the input of our attention module in the visual modality to gather both global-level information and local-level information (*i.e.*, discriminative part cues) simultaneously.

2) Visual Representation: Given an image, after the aforementioned preparation, we can obtain $\tilde{\mathbf{X}} = \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_n\}$. Then, we introduce a self-attention and employ the visual mask template $\mathbf{M}^V$ as followed to evaluate the attention weight a_i^v corresponding to each local feature vector $\tilde{\boldsymbol{x}}_i$:

$$a_i^v = \frac{\exp(\text{SAM}(\tilde{\boldsymbol{x}}_i) + m_i^v)}{\sum_{j=1}^{n} \exp(\text{SAM}(\tilde{\boldsymbol{x}}_j) + m_j^v)}, \tag{3}$$

$\text{SAM}(\cdot)$ is equal to Eqn. (1) and m_i^v is the i-th element of $\mathbf{M}^V$. After that, we calculate the weighted sum of each local feature vector as the content vector $\mathbf{f}_{\text{global}}^v = \sum_{i=1}^{n} a_i^v \tilde{\boldsymbol{x}}_i$. Then, $\mathbf{f}_{\text{global}}^v$ can be regarded as the representation of the **global vision stream** to reflect image-level global visual information.

To further improve the fine-grained recognition accuracy, it is desirable to find and locate the key object parts which has the discriminative information to distinguish different sub-categories. Since the attention weight a_i^v can reflect the importance of the corresponding part for category predictions, we can locate key parts according to the scores of these attention weights. The largest a_i^{v} should be the most important one, and its corresponding 2×2 block $\boldsymbol{x}_i$ should be the key part we want to locate. We denote it as $\boldsymbol{x}_{\max}$. Based on that, we then employ 1×1-conv and global average-pooling operation over $\boldsymbol{x}_{\max}$ to get a compact part-level feature. We denote that part-level feature as $\mathbf{f}_{\text{local}}^v$ of the **local vision stream**. Please note that $\mathbf{f}_{\text{local}}^v$ gathers the most representative semantic in the current visual feature. Also, we will use $\mathbf{f}_{\text{local}}^v$ to bridge the connection to the language modality.

Finally, we concatenate the aforementioned global feature $\mathbf{f}_{\text{global}}^v$ and local feature $\mathbf{f}_{\text{local}}^v$ as $\mathbf{f}_{\text{vision}}$ to form the final representation in the visual modality:

$$\mathbf{f}_{\text{vision}} = [\mathbf{f}_{\text{global}}^v; \mathbf{f}_{\text{local}}^v], \tag{4}$$

where $[\cdot;\cdot]$ denotes the concatenate operation.

3) Progressive Mask Strategy: Based on Eqn. (3), we know a_i^v can reflect the degree which the block x_i should take attention. For the sake of multiple distinct parts localization, we design a mask strategy for visual mask template $\mathbf{M}^V$ to force the stacked modules to capture different discriminative visual parts in a stage-by-stage fashion. If we denote the largest attention weight returned in the current stage as $a_{\max}^v$. Then, at the end of each stage, we will update the elements of the mask template m_i^v as $-\infty$, if $a_i^v = a_{\max}^v$. At the following stage, our attention module will locate another important part on the basis of the updated mask $\mathbf{M}^V$. Benefit from our progressive mask attention strategy, we capture multiple discriminative parts without part overlapping.

C. Progressive Mask Attention in Language Modality

Learning to align the text information from the visual feature, is a frequently-used method in joint-representation learning across different modalities. To exert the advantages of language modality in our task, we employ a query-based attention to seize the relevant fine-grained noun phrases for each located part. These seized noun phrases can be considered as a textual representation for the visual domain. However, the language descriptions usually possess some information while the raw image can not express. Therefore, we also apply a self-attention mechanism with a progressive mask strategy to collect the out-of-visual-domain knowledge from language modality. Specifically, we discard some relevant noun phrases of the located part after each querying stage and gather the remaining phrases to form a global feature. Both the query-based (local) feature and the global feature will be fused as a final state at the current stage.

1) Input Preparation: Given the raw texts $\mathbf{T}$ describing the characteristics of the fine-grained objects in an image, we first adopt a sequence of commonly-used techniques in

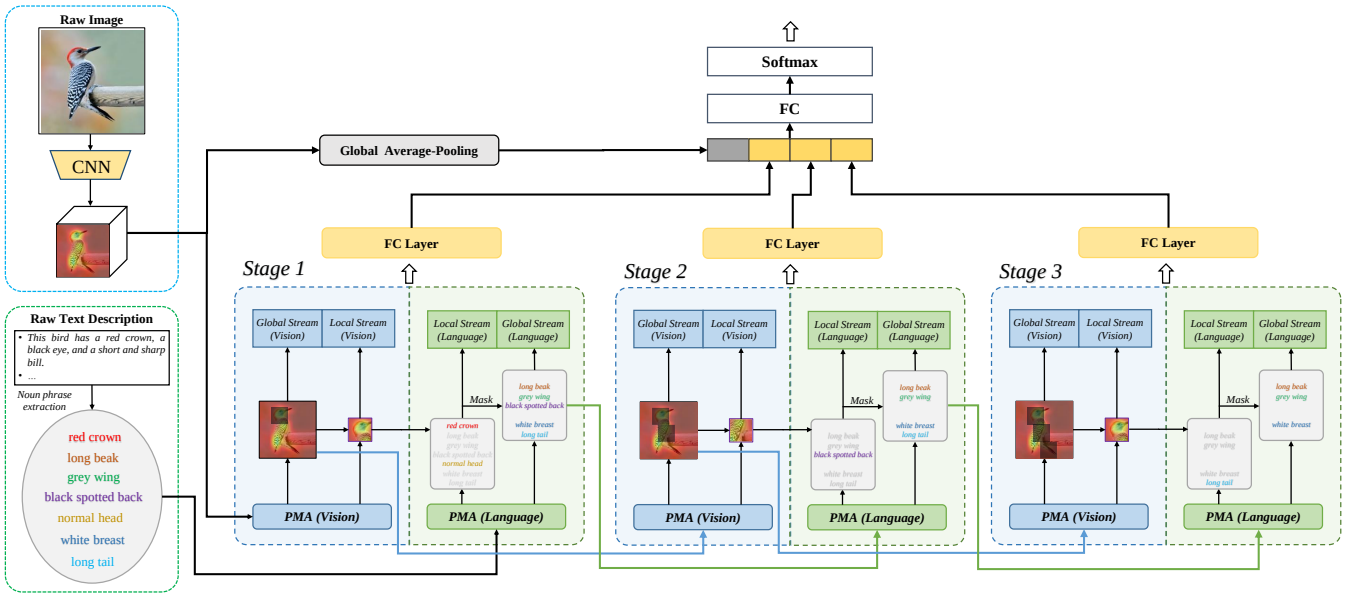


Figure 2: Overall framework of the proposed Bi-Modal Progressive Mask Attention model (with three stages) for fine-grained recognition. Raw images and texts are the original inputs. After CNN feature extraction of visual modality and noun phrase extraction of language modality, the processed bi-modal representations are fed into our Bi-Modal PMA module at the first stage. In visual-based PMA, it can locate the most discriminative part of images in that stage. Furthermore, a query-based attention approach is developed to align the information from both vision and language. Later, by employing the progressive mask strategy, we can stack multiple PMA modules corresponding to multiple processing stages. The mask-based approach in PMA is able to iteratively capture *different* discriminative parts and out-of-visual-domain knowledge in a stage-by-stage fashion.

natural language processing [34] (*i.e.*, word tokenization, part-of-speech tagging and noun phrase chunking) to extract noun phrases from $\mathbf{T}$. For each noun phrase, we use word embeddings and long-short term memory (LSTM) [35] to extract phrase-level embeddings. We denote these phrase embeddings as $\mathbf{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_p\}$, where p is the number of noun phrases and $\boldsymbol{z}_i \in \mathbb{R}^d$ is the vector of phrase embeddings. Therefore, $\mathbf{Z}$ is adopted to generate local-level and global-level textual semantic from our attention module in the language modality.

2) Language Representation: For the located part-level feature $\mathbf{f}^v_{\text{local}}$ and the transformed phrase embedding $\mathbf{Z}$, we design a query-based attention mechanism with text mask template $\mathbf{M}^T$ for language modality to generate its corresponding text representation. The attention weight a^t_i for each noun phrase is calculated as:

$$a^t_i = \frac{\exp(\text{QRM}(\mathbf{f}^v_{\text{local}}, \boldsymbol{z}_i) + m^t_i)}{\sum_{j=1}^{m} \exp(\text{QRM}(\mathbf{f}^v_{\text{local}}, \boldsymbol{z}_i) + m^t_i)}, \tag{5}$$

where $\text{QRM}(\cdot,\cdot)$ is equal to Eqn. (2) and m^t_i is the i-th element of $\mathbf{M}^T$. We form the content vector as $\mathbf{f}^t_{\text{local}} = \sum_{i=1}^{p} a^t_i \boldsymbol{z}_i$. More importantly, $\mathbf{f}^t_{\text{local}}$ can be considered as a **local language stream**, which aggregates the most typical textual semantic for the current visual outputs.

Beside the text features guided by the located part feature, we also want to mine some textual knowledge beyond the visual domain. Therefore, we discard phrases which are highly relevant to the located part and employ a self-attention mechanism over the remaining phrases to generate a textual representation, which gathers the out-of-visual-domain feature. The attention weight $\tilde{a}^t_i$ for each phrase is:

$$\tilde{a}^t_i = \frac{\exp(\text{SAM}(\boldsymbol{z}_i) + m^t_i)}{\sum_{j=1}^{m} \exp(\text{SAM}(\boldsymbol{z}_i) + m^t_i)}, \tag{6}$$

where $\text{SAM}(\cdot)$ is equal to Eqn. (1). We calculate the weighted sum of each noun phrase as the content vector $\mathbf{f}^t_{\text{global}} = \sum_{i=1}^{p} \tilde{a}^t_i \boldsymbol{z}_i$. Thus, $\mathbf{f}^t_{\text{global}}$ is regarded as the **global language stream** to collect the information which is out of visual domain.

Finally, we concatenate the global feature $\mathbf{f}^t_{\text{global}}$ and local feature $\mathbf{f}^t_{\text{local}}$ as $\mathbf{f}_{\text{text}}$ to form the final representation in the language modality:

$$\mathbf{f}_{\text{text}} = [\mathbf{f}^t_{\text{global}}; \mathbf{f}^t_{\text{local}}]. \tag{7}$$

3) Progressive Mask Strategy: Based on Eqn. (5), a^t_i reflects the relevance of the phrase $\boldsymbol{z}_i$ to the located part feature. As we expect to seize more textual descriptors which may not exist in the image, a mask strategy is designed for text mask template $\mathbf{M}^T$ to enforce the subsequent stage to explore more out-of-visual-domain features. After the query-based attention in the local stream (language modality), we update the element $\boldsymbol{m}^t_i$ as $-\infty$ as its weight a^t_i is ranked in top-three and higher than $1/p$. Please note, this mask strategy will be operated once the local stream is done at the current stage. Benefit from our progressive mask strategy, we can collect the text representation which in visual domain and out of visual domain simultaneously.

D. Stage-wise feature aggregation

As aforementioned, we respectively design two Progressive Mask Attention modules for visual and language modalities, where one is for key part localization and another is to seize key words of text. For each stage, we concatenate visual and text representation as the final output. After that, we append a shared fully connected layer after each stage outputs for dimensionality reduction:

$$\mathbf{f}^i_{\text{final}} = \text{FC}\left([\mathbf{f}^i_{\text{visual}}; \mathbf{f}^i_{\text{text}}]\right), \tag{8}$$

where $[\cdot;\cdot]$ represents concatenation and $\text{FC}(\cdot)$ denotes a fully connected layer. To exert the advantages of multi-stages ensemble, we also aggregate these output states as a final representation for predictions. Here we restrict the number of stages as three in our method. First, we extract the object-level representation $\mathbf{f}_{\text{object}}$ by conducting global average-pooling over the feature map $\mathbf{X}$ to obtain image-level visual information. Then, we concatenate object-level representation and the output of multiple stages to form the final representation $\mathbf{F}$:

$$\mathbf{F} = [\mathbf{f}_{\text{object}}; \mathbf{f}^1_{\text{final}}; \mathbf{f}^2_{\text{final}}; \mathbf{f}^3_{\text{final}}]. \tag{9}$$

Here, we use three stages for an example. After that, a fully connected layer with the softmax function is appended upon the final representation $\mathbf{F}$ to conduct final classification. The traditional cross-entropy loss is used to drive the whole network training, and our model can be end-to-end trainable. Figure 2 shows the whole framework of our Bi-Modal Progressive Mask Attention model in multi-stages.

E. Knowledge Distillation for Bi-Modal PMA

To support our bi-modality model to predict in the single-modality environment (*e.g.*, only using image data), we further perform a knowledge distillation approach [36] to compress the knowledge of both visual and language modalities into the student model. Here, we use the Bi-Modal PMA model as the teacher model, and a standard network (*i.e.*, only adopting the original image as inputs) as the student model.

For the teacher model, we define the training corpus as $(s_i, y_i) \in \{\mathcal{S}, \mathcal{Y}\}$, where s_i means a pair of image and text data, and y_i is the ground truth. We use standard cross-entropy as the loss function for our model as:

$$\mathcal{L}_{teacher}(\mathcal{Y}|\mathcal{S};\theta_T) = \sum_{i=1}^{N}\sum_{j=1}^{C} 1\{y_i = j\} \log P(y_i|s_i;\theta_T), \tag{10}$$

where N and C is the number of training samples and classes, and θ_T is the parameter of our teacher model (*i.e.*, the Bi-Modal PMA model).

For the student model, we define the training corpus as $(t_i, y_i) \in \{\mathcal{T}, \mathcal{Y}\}$, where t_i is the image data. Instead of using the ground truth of images for prediction, our distiller enforces the student model to learn the output probability $P(y_i|s_i;\theta_T)$ of the teacher model. Therefore, the loss function for knowledge distillation can be formulated by

$$\mathcal{L}_{student}(\mathcal{Y}|\mathcal{T};\theta_S) = \sum_{i=1}^{N}\sum_{j=1}^{C} P(j|s_i;\theta_T) \cdot \log P(j|t_i;\theta_S), \tag{11}$$

Table I: Characteristics of datasets used in experiments.

Dataset	♯ category	♯ training	♯ test	Texts?
CUB-200-2011 [1]	200	5,994	5,794	✓
Oxford Flower [2]	102	2,000	6,149	✓
FGVC Aircraft [4]	100	6,667	3,333	
Stanford Car [3]	196	8,144	8,041	

where θ_S is the parameter of the student model. Based on Eqn. (11), we can distill the knowledge from two modalities into the visual modality, and thus allows the model be able to return predictions even without text data during inference.

IV. Experiments Setting

In this section, we will introduce the datasets, empirical setting and the comparison baseline methods.

A. Datasets

We conduct experiments on four widely used fine-grained benchmark datasets, *i.e.*, CUB-200-2011 [1], Oxford Flower [2], FGVC Aircraft [4] and Stanford Car [3]. The detailed information of each dataset is described in Table I. Among them, CUB-200-2011 and Oxford Flower provide text descriptions[2] of the language modality for each image. While, FGVC Aircraft and Stanford Car only contain raw images from the single visual modality without text descriptions.

B. Implementation details

To make a fair comparison, we choose VGG-16 [46] as the base model in the visual modality for obtaining the image-level feature. A pre-trained VGG-16 network on ImageNet [47] is used for parameter initializations. In addition, for final classification, we employ a dropout layer [48] with the 0.5 ratio before the fully connected layer. Stochastic gradient descent with a mini-batch size of 32 is performed as the optimizer. The models are trained in totally 100 epochs, and the learning rate starts from 0.005 and is divided by 10 at the 30^{th} and 60^{th} epoch. The weight decay is set to 10^{-4} and the momentum is set to 0.9. Following the previous strategy [6] in data augmentation, we conduct horizontal flips and random crop image patches as the 448×448 resolution from the original image. For text preprocessing, we extract noun phrases from text descriptions via the following stages [34]: word tokenization, part-of-speech tagging and noun-phrase chunking. We embed the word of each phrase and feed them into two-stacked 512-dimensional LSTM layers to get phrase embeddings. The word embedding is initialized as a 300-dimensions vector based on Glove 6B corpus [49]. Besides, to further explore recognition performance of our model, we conduct experiments with ResNet-50 [50] as the base model.

C. Comparison methods

To demonstrate the advantages of our model, we list the following baselines for comparisons:

[2]https://github.com/reedscot/cvpr2016

Table II: Comparison results on the CUB-200-2011 dataset. "Train/Test Anno." column means whether using bounding boxes or part annotations in the training or test phase. "One-stage" means whether the training stage can be accomplished in one-stage. "Modality" column represents whether using additional information.

Methods	Base model	Train Anno.	Test Anno.	One-stage	Modality	Accuracy
PA-CNN [10]	AlexNet	✓	✓	✓		82.8
MG-CNN [20]	VGG-19	✓				83.0
SPDA-CNN [37]	AlexNet	✓	✓	✓		85.1
PN-CNN [5]	AlexNet	✓				85.4
TLAN [38]	AlexNet					77.9
NAC [39]	VGG-19			✓		81.0
B-CNN [6]	VGG-16			✓		84.1
STN [12]	GoogLeNet					84.1
RA-CNN [9]	VGG-19					85.3
WARN [40]	Wide ResNet			✓		85.8
OPAM [41]	VGG-16					85.8
MoNet [42]	VGG-16			✓		86.4
MA-CNN [11]	VGG-19			✓		86.5
M2DRL [43]	VGG-16					87.2
DCL [44]	ResNet					87.8
TASN [45]	ResNet					87.9
CVL [14]	VGG-16			✓	+ Text	85.5
T-CNN [15]	ResNet			✓	+ Knowledge-Base	85.8
KERL [13]	VGG-16			✓	+ Knowledge-Base	86.3
TA-FGVC [29]	ResNet				+ Text	86.9
TA-FGVC [29]	ResNet	✓			+ Text	88.1
Baseline (VGG-16)	VGG-16			✓		78.9
Ours	VGG-16			✓		86.8
Ours	VGG-16			✓	+ Text	88.2
Baseline (ResNet-50)	ResNet-50			✓		84.5
Ours	ResNet-50			✓		87.5
Ours	ResNet-50			✓	+ Text	**88.7**

- **PA-CNN** [10]: a Part Alignment-based method which generates part via co-segmentation and alignment.
- **MG-CNN** [20]: learning multiple regions by Multiple-Granularity CNN for all the grain levels.
- **SPDA-CNN** [37]: a unified framework utilizes Semantic Part Detection and Abstraction.
- **PN-CNN** [5]: a Pose Normalized CNN to produce local feature from object's pose.
- **TLAN** [38]: Two-Level Attention Network on object and part domain for classification.
- **NAC** [39]: Neural Activation Constellation for unsupervised part discovery.
- **B-CNN** [6]: a Bilinear-CNN layer to model local pairwise feature interactions.
- **STN** [12]: a Spatial-Transformer Network to adaptively learn features in various transformed space.
- **RA-CNN** [9]: Recurrent-Attention Network to locate discriminative parts recurrently.
- **WARN** [40]: a gate attention mechanism for attend and rectify global and local features.
- **OPAM** [41]: an Object-Part Attention Model for fine-grained recognition.
- **MoNet** [42]: a framework unify G^2DeNet and bilinear pooling CNN from Moment Matrix to combine the compact representation.
- **MA-CNN** [11]: Multi-Attention CNN to extract multiple discriminative parts.
- **M2DRL** [43]: a Multi-scale and Multi-granularity Deep Reinforcement Learning approach.
- **DCL** [44]: a model with Destruction and Construction Learning.
- **TASN** [45]: a Trilinear Attention Sampling Network for fine-grained image recognition.
- **CVL** [14]: an approach Combines Visual and Language stream for performance boosting.
- **T-CNN** [15]: a CNN framework extract embedding from knowledge and text domain.
- **KERL** [13]: a method incorporates Knowledge-Embedded Representation Learning into fine-grained recognition.
- **TA-FGVC** [29]: a Text-Assisted Fine-Grained Visual Classification method.

V. Experimental Results

In the following, we show our classification results on four public datasets, as well as the ablation studies for analyzing and validating the effectiveness of our model designs.

A. Results

1) CUB-200-2011: The classification results of Caltech-UCSD birds are reported in Table II. We compare the accuracy of state-of-the-art fine-grained recognition methods with ours. As reported in that table, our model achieves 86.8% accuracy on VGG-16, which is slightly higher than the competing baselines. But, after equipped with the language modality, our VGG-based modal achieves 88.2% classification accuracy, which significantly outperforms other state-of-the-arts. By using a strong base model (ResNet-50), the accuracy obtains a further improvement and gets the 87.5% on single visual modality and 88.7% accuracy on bi-modality, which is the best classification accuracy on CUB-200-2011.

Table III: Comparison results on the Oxford flower dataset.

Methods	Accuracy
SDR [51]	90.5
Deep optimized [51]	91.3
RIIR [52]	94.0
NAC [39]	95.3
PBC [53]	96.1
Ours (VGG-16)	96.9
Ours (VGG-16 + Text)	**97.4**

Table IV: Comparison results on the FGVC Aircraft dataset. "Anno." column means whether using bounding boxes/part annotations.

Method	Anno.	Accuracy
MG-CNN [20]	✓	86.6
MDTP [54]	✓	88.4
B-CNN [6]		84.1
KP [55]		86.9
RA-CNN [9]		88.2
MA-CNN [11]		89.9
HBP [56]		90.3
Ours (VGG-16)		90.4
Ours (ResNet-50)		**90.8**

2) Oxford Flower: The results of Oxford Flower are reported in Table III. Our method achieves 97.4% accuracy on bi-modality, which has surpassed most of the reported results by 1.0%-6.6% accuracy. On the single visual modality, the obtained 96.9% classification accuracy is still significantly higher than the accuracy of previous work in the literature. These large marginal improvements validate the effectiveness of our attention approach in discriminative part learning and bi-modal joint-representation learning.

3) FGVC Aircraft: The results of FGVC Aircraft are reported in Table IV. Our method achieves 90.2% accuracy, which is also superior to other state-of-the-art systems. MDTP [54] are an advanced system which adopted human annotations. However, our method could brings a 2.0% relative improvement than MDTP [54] and does not use any annotations. MA-CNN [11] achieved 89.9% accuracy. Our PMA exceeds it by 0.5% accuracy. B-CNN [6] is a common method in fine-grained recognition with 84.1% accuracy, and KP [55] and HBP [56] are the improved methods based on B-CNN [6]. Nevertheless, our method is still superior to KP [55] and HBP [56]. In addition, when using ResNet-50 as the base model, our method can achieve 90.8% accuracy, which receives a 0.4% additional accuracy gain than VGG-based one. It further surpasses other previous fine-grained recognition models.

4) Stanford Car: The results of Stanford Car are reported in Table V. Our method achieves 93.0% accuracy which is also the best result over the competing baseline methods. The baseline of VGG-16 achieved 79.8% accuracy and our method surpasses the baseline by 13.2% accuracy gain. PA-CNN [10] and MA-CNN [11] are advanced fine-grained recognition systems which uses part annotations and disables annotations, respectively. Although promising results has been achieved by them, our method can still give a 0.3% accuracy gain comparing with PA-CNN [10] and MA-CNN [11]. Besides, our Bi-Modal PMA based on ResNet-50 can achieve 93.1% accuracy.

Table V: Comparison results on the Stanford Car dataset. "Anno." column means whether using bounding boxes/part annotations.

Method	Anno.	Accuracy
Part R-CNN [19]	✓	88.4
FCAN [57]	✓	91.3
PA-CNN [10]	✓	92.8
VGG-16		79.8
ResNet-50		84.7
WARN [40]		90.0
B-CNN [6]		91.3
RA-CNN [9]		92.5
MA-CNN [11]		92.8
Ours (VGG-16)		93.0
Ours (ResNet-50)		**93.1**

Table VI: Results of knowledge distillation on CUB-200-2011 (ResNet-50) and Oxford Flower (VGG-16). Please note that "Student model" means the network only deals with image data during inference.

Method	CUB-200-2011	Oxford flower
Teacher model (Bi-Modal PMA)	**88.7**	**97.4**
Student model	88.3	96.9

B. Knowledge Distillation

We conduct experiments on CUB-200-2011 and Oxford Flower datasets to evaluate the performance of knowledge distillation (*i.e.*, without using text information during inference) for our Bi-Modal PMA model. The results are displayed in Table VI. As can be seen, based on our distiller, our student model can achieve 88.3% and 96.9% accuracy on CUB-200-2011 and Oxford flower datasets respectively, which are slightly better than previous approaches even with only image-level data in the test phase. These observations show the potentiality to distill the knowledge of two modalities and also reveal the generalization ability of our Bi-Modal PMA in utilizing different modalities for feature learning.

C. Ablation studies

1) Effects of the number of learning stages: To further explore the effectiveness of our approach, we also conduct quantitative comparisons about our PMA model. We change the number of learning stages of our proposed model from 1 to 4 to perform comparisons. The results are reported in Table VII. As seen, we set the maximum number of the learning stage as 4, since our model does not receive further benefits with more stages. We argue this phenomenon might be possibly caused by overfitting. On CUB-200-2011, supposing that contains object-level information, our model achieves 85.9%, 86.5% and 86.8% accuracy when using 1, 2 and 3 stages, respectively. By increasing the number of learning stages, the classification performance of our model will receive consecutive gains by 0.6% and 0.3%. When disabling object-level information, our model gets 83.0%, 84.3% and 84.7% accuracy. When using the language information and object information, our approach receives 87.4%, 88.0% and 88.2% accuracy. We also further conduct experiments on Stanford Car. The results also indicates that our model achieves the best accuracy 93.0% with three learning stages, which is consistent with our conclusion on

Table VII: Ablation studies about the number of learning stages of our model on CUB-200-2011 (VGG-16) and Stanford Car (VGG-16). "+o/-o" in brackets means using/disabling object-level information which refers to the feature f_{object} extracted by conducting GAP on the feature maps. cf. Eqn. (9). "+t" means using text information of the language modality.

♯ stages	Acc. (-o)	Acc. (+o)	Acc. (+o/+t)
CUB-200-2011			
1	83.0	85.9	87.4
2	84.3	86.5	88.0
3	**84.7**	**86.8**	**88.2**
4	84.5	86.5	88.1
Stanford Car			
1	84.2	88.9	–
2	86.8	92.0	–
3	88.6	**93.0**	–
4	**88.7**	92.9	–

Table VIII: Ablation studies about the individual accuracy of each learning stage in our model on CUB-200-2011 (VGG-16) and Stanford Car (VGG-16).

♯ stage	Acc. (-o)	Acc. (+o)	Acc. (+o/+t)
CUB-200-2011			
1	83.0	85.9	87.4
2	81.9	85.4	86.8
3	79.5	84.8	85.9
Stanford Car			
1	84.2	88.9	–
2	83.1	88.2	–
3	81.0	86.9	–

CUB-200-2011. These quantitative comparisons with different conditions also satisfy the conclusion on experiments with object-level information. These incremental improvements also demonstrate the necessity of each learning stage for our model.

2) Classification accuracy of each learning stage: To better verify the contributions of each learning stage, we also test the accuracy by only using the outputs of each learning stage separately. The results are reported in Table VIII. We observe that the first learning stage achieves the best performance, and the accuracy will gradually decrease for the latter learning stages. We deem that the front stage always captures the most discriminative cues for classification, and the following stage will extract the less discriminative information which the previous stages do not attend. Meanwhile, by taking the results of Table VII into consideration, it can reveal that our model indeed captures discriminative but complementary information in the stage-by-stage manner. These observations are also consistent with the motivation of our progressive attention proposal.

3) Effects of the number of parts in each learning stage: In the ablation studies of this section, we attempt to predict multiple parts in a single stage for comparisons. We design two contrast experiments, where one is only using one stage and another uses three stages. Each stage in our ablation studies will predict multiple (three) parts. The results are shown in Table IX. We find that the model, which uses only one stage to predict three parts, will decrease 0.3%-0.5% accuracy on CUB-200-2011 and Stanford Car. It might caused by that, for a single stage, the attention model usually seizes the most discriminative part but ignores some inconspicuous parts. That means, except for the most discriminative part, the attention model disables to discover additional parts effectively in a single stage. Moreover, we find that if three learning stages all predict three parts, it will harm a lot accuracy by 0.9%-1.1%. We guess the reason might be that choosing too many parts will confuse the model to discover which part is critical. Overall, these comparisons prove that using three stages and selecting one part in each single stage are optimal for classification accuracy.

Table IX: Ablation studies of predicting different number of parts in each learning stage. Results are reported on CUB-200-2011 (VGG-16) and Standford Car (VGG-16). "♯ stages" means how many stages are used for prediction and "♯ parts" means how many regions are predicted in a single stage. "♯ stages=3" with "♯ parts=1" is our proposal's setting.

♯ stages	♯ parts	CUB-200-2011	Stanford Car
3	1	**86.8**	**93.0**
1	3	86.3	92.7
3	3	85.7	92.1

Table X: Ablation studies about the effect of different streams for classification accuracy on CUB-200-2011 (VGG-16) and Stanford Car (VGG-16). The meaning of global/local vision or language streams can refer to Figure 1.

Model settings	CUB-200-2011	Stanford Car
Only global vision stream	85.6	91.5
Only local vision stream	49.8	57.5
Both global & local vision streams	**86.8**	**93.0**
Only local language stream	59.3	–
Only global language stream	55.7	–
Both global & local language streams	**63.2**	–

4) Effects of different streams: In order to survey the necessity of global/local vision and language streams in PMA, we conduct a series of ablation experiments to analyze the effect of different streams. The results are reported in Table X.

For vision streams, we find that the model without global vision stream works quite badly. While global vision stream is available, the model can obtain 85.6%/91.8% accuracy in Bird/Car tasks. Equipped with both local vision stream (part-level) and global vision stream (object-level) simultaneously, it receives 1.2%/1.5% additional accuracy gains respectively. These comparisons indicate that the global vision stream is essential and local vision stream evidently boost the classification accuracy.

For language streams, we find that the model only using local language stream outperforms using global language stream by 3.6% accuracy gains. When integrating both local language stream and global language stream, it achieves 63.2% accuracy, which is superior to any single language stream. These comparisons indicate that local language stream gathers more text representation from salient parts and global language stream provide additional textual features for boosting accuracy.

In summary, each global/local stream in the visual or language modality is indispensable for our model.

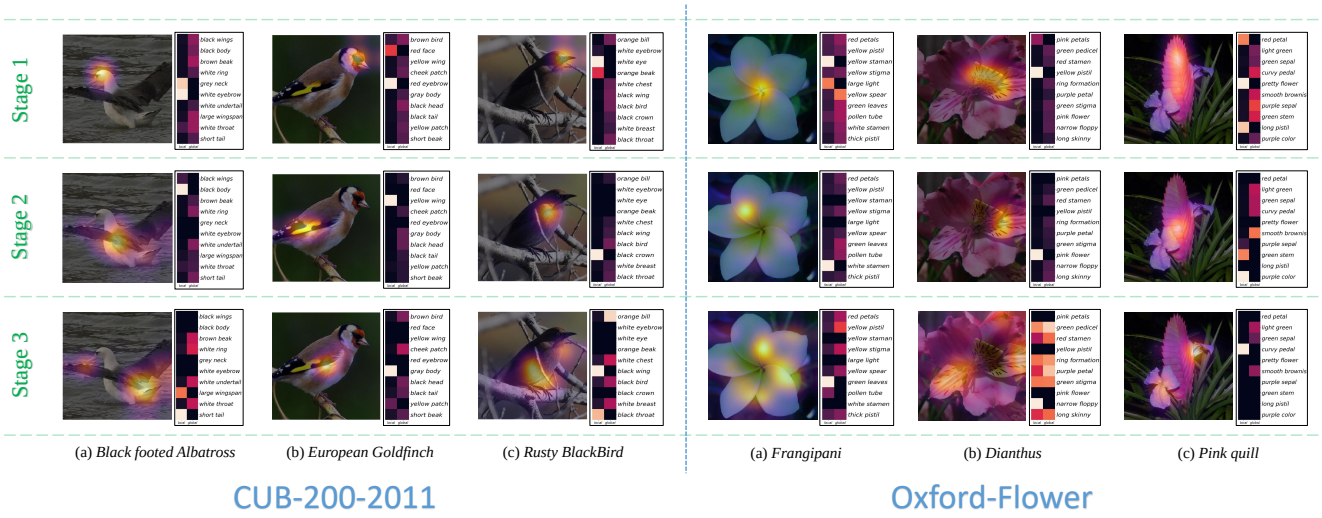
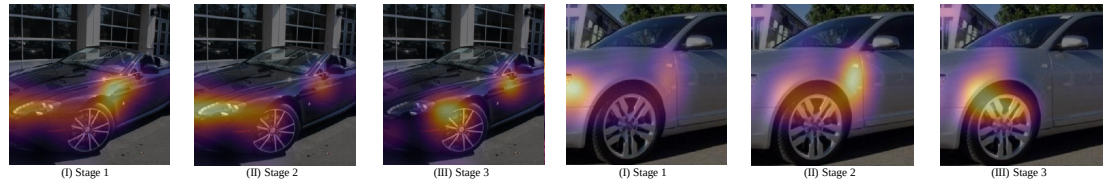


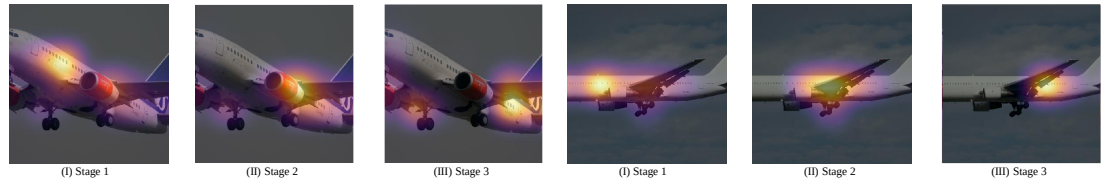
Figure 3: Visualization of the attended parts in images and the highlighted key words in texts of different learning stages. For visual modality, we visualize the attention weights on the original images. For the language modality, several key words are attended at each stage. Here, we highlight them with different background colors according to the attention weights from high (white) to low (dark) in both local (left column) and global (right column) language streams.



(a) *Aston Martin V8 Vantage Convertible 2012* (b) *Audi S6 Sedan 2011*

(c) *Cadillac Escalade EXT Crew Cab 2007* (d) *Dodge Dakota Club Cab 2007*

Figure 4: Visualization of the attended parts in images in different learning stages on Stanford Car.



(a) *Boeing 737-600* (b) *Boeing 767-200*

(c) *A340-300* (d) *Yak-42*

Figure 5: Visualization of the attended parts in images in different learning stages on FGVC Aircraft.

D. Visualization

Figure 3 shows the visualization of the attended part in raw images and texts (*i.e.*, noun phrases) in different learning stages. For the visual modality, we visualize the attention weights over the original images based on Eqn. (3). For the language modality, we visualize the attention weights of global/local streams on each key phrase based on Eqn. (5) and Eqn. (6). For the visual modality, it is clear to observe that at different stages, our model is able to: a) attend subtle but discriminative parts of fine-grained objects; b) locate different parts of objects with the help of our progressive mask attention strategy. For the language modality, we can find: a) the key words automatically seized by local language stream can be aligned with the corresponding image regions at each stage; b) the key words, which are low correlated with the image regions, will be activated in the global language streams, especially in the third stage. Besides, the visualization in image domain of Stanford Car and FGVC Aircraft are reported in Figure 4 and Figure 5. We find the attention visualization in these two datasets can also find different discriminative parts, which also satisfies our targets. The visualization results can validate the effectiveness of our proposed PMA in bi-modality from the qualitative perspective.

VI. Conclusion

In this paper, we proposed the Bi-Modal Progressive Mask Attention (Bi-Modal PMA) model for fine-grained recognition. Specifically, Bi-Modal PMA is a unified framework to incorporate information from both visual and language modalities. By our mask-based progressive strategy, our model can learn and capture a set of discriminative image regions of images and key words in texts in a stage-by-stage way. Furthermore, our Bi-Modal PMA can also extract the knowledge which is complementary to the visual modality. Experimental results of four benchmark fine-grained datasets validated the effectiveness of our proposed PMA model. In the future, we expect to explore the possibility of incorporating more modalities' information, like attributes, to further boost the fine-grained recognition accuracy.

References

[1] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," no. CNS-TR-2010-001, 2010.

[2] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008, pp. 722–729.

[3] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *ICCV Workshop*, 2013, pp. 554–561.

[4] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Tech. Rep., 2013.

[5] S. Branson, G. V. Horn, S. J. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *BMVC*, 2014, pp. 1–14.

[6] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.

[7] X.-S. Wei, C.-L. Zhang, L. Liu, C. Shen, and J. Wu, "Coarse-to-fine: A RNN-based hierarchical attention model for vehicle re-identification," in *ACCV*, 2018, pp. 1–16.

[8] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, pp. 1–24, 2019.

[9] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017, pp. 4438–4446.

[10] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *CVPR*, 2015, pp. 5546–5555.

[11] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *ICCV*, 2017, pp. 5209–5217.

[12] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *NIPS*, 2015, pp. 2017–2025.

[13] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, "Knowledge-embedded representation learning for fine-grained image recognition," in *IJCAI*, 2018, pp. 627–634.

[14] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *CVPR*, 2017, pp. 5994–6002.

[15] H. Xu, G. Qi, J. Li, M. Wang, K. Xu, and H. Gao, "Fine-grained image classification by visual-semantic embedding," in *IJCAI*, 2018, pp. 1043–1049.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE TPAMI*, vol. 38, no. 1, pp. 142–158, 2016.

[17] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples," *IEEE TIP*, vol. 28, no. 12, pp. 6116–6125, 2019.

[18] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognition*, vol. 76, pp. 704–714, 2018.

[19] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based RCNN for fine-grained detection," in *ECCV*, 2014, pp. 834–849.

[20] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *ICCV*, 2015, pp. 2399–2406.

[21] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *CVPR*, 2016, pp. 1134–1142.

[22] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE TIP*, vol. 25, no. 4, pp. 1713–1725, 2016.

[23] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE TIP*, vol. 26, no. 6, pp. 2868–2881, 2017.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[25] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual question answering," in *ICCV*, 2015, pp. 2425–2433.

[26] X. Shu, J. Tang, Z. Li, H. Lai, L. Zhang, and S. Yan, "Personalized age progression with bi-level aging dictionary learning," *IEEE TPAMI*, vol. 40, no. 4, pp. 905–917, 2017.

[27] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *ACM MM*, 2015, pp. 35–44.

[28] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning deep representations of fine-grained visual descriptions," in *CVPR*, 2016, pp. 49–58.

[29] J. Li, L. Zhu, Z. Huang, K. Lu, and J. Zhao, "I read, I saw, I tell: Texts assisted fine-grained visual classification," in *ACM MM*, 2018, pp. 663–671.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2014, pp. 1–15.

[31] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *ICLR*, 2017, pp. 1–15.

[32] X. Shu, J. Tang, G.-J. Qi, W. Liu, and J. Yang, "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE TPAMI*, DOI: 10.1109/TPAMI.2019.2942030, 2019.

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[34] A. Handler, M. Denny, H. M. Wallach, and B. T. O'Connor, "Bag of what? simple noun phrase extraction for text analysis," in *EMNLP*, pp. 114–124.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

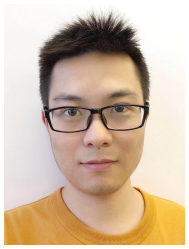
[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv*, 2015.

[37] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A. Elgammal, and D. Metaxas, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *CVPR*, 2016, pp. 1143–1152.
[38] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *CVPR*, 2015, pp. 842–850.
[39] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *ICCV*, 2015, pp. 1143–1151.
[40] P. Rodriguez, J. M. Gonfaus, G. Cucurull, F. XavierRoca, and J. Gonzalez, "Attend and rectify: a gated attention mechanism for fine-grained recovery," in *ECCV*, 2018, pp. 349–364.
[41] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE TIP*, vol. 27, no. 3, pp. 1487–1500, 2018.
[42] M. Gou, F. Xiong, O. Camps, and M. Sznaier, "MoNet: Moments embedding network," in *CVPR*, 2018, pp. 3175–3183.
[43] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *IJCV*, vol. 127, no. 9, pp. 1235–1255, 2019.
[44] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *CVPR*, 2019, pp. 5157–5166.
[45] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *CVPR*, 2019, pp. 5012–5021.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015, pp. 1409–1556.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, pp. 1929–1958, 2014.
[49] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *EMNLP*, 2014, pp. 1532–1543.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[51] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *CVPR Workshops*, 2015, pp. 36–45.
[52] L. Xie, J. Wang, W. Lin, B. Zhang, and Q. Tian, "Towards reversal-invariant image representation," *IJCV*, vol. 123, no. 2, pp. 226–250, 2017.
[53] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "PBC: Polygon-based classifier for fine-grained categorization," *IEEE TMM*, vol. 19, no. 4, pp. 673–684, 2017.
[54] Y. Wang, J. Choi, V. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *CVPR*, 2016, pp. 1163–1172.
[55] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *CVPR*, 2017, pp. 2921–2930.
[56] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *ECCV*, 2018, pp. 574–589.
[57] X. Liu, T. Xia, J. Wang, and Y. Lin, "Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition," in *ArXiv preprint: abs/1603.06765*, 2016.

Kaitao Song (S'19) currently is a PhD candidate in Nanjing University of Science and Technology, China. He received the B.S. degree in the computer science and engineering from the Nanjing University of Science and Technology in 2015. His current research interests include multimodal analysis, natural language processing, deep learning and machine learning.

Xiu-Shen Wei (M'18) received his BS degree in computer science, and received his Ph.D. degree in computer science and technology from Nanjing University. He is a Professor at Nanjing University of Science and Technology (NJUST). Before joining NJUST, he served as the Founding Director of Megvii Research Nanjing, Megvii Technology. He has published more than thirty academic papers on the top-tier international journals and conferences, such as IEEE TPAMI, IEEE TIP, IEEE TNNLS, IEEE TKDE, Machine Learning, CVPR, ICCV, ECCV, IJCAI, ICDM, ACCV, etc. He achieved the first place in the iWildCam competition (in association with CVPR 2020), the first place in the iNaturalist competition (in association with CVPR 2019), the first place in the Apparent Personality Analysis competition (in association with ECCV 2016) and the first runner-up in the Cultural Event Recognition competition (in association with ICCV 2015) as the team director. He also received the Presidential Special Scholarship (the highest honor for Ph.D. students) in Nanjing University, and received the Outstanding Reviewer Award in CVPR 2017. His research interests are computer vision and machine learning. He has served as a PC member of CVPR, ICCV, ECCV, NeurIPS, IJCAI, AAAI, etc. He is a member of the IEEE.

Xiangbo Shu (M'16) is currently an Associate Professor in School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received his Ph.D. degree in July 2016 from Nanjing University of Science and Technology. From 2014 to 2015, he worked as a visiting scholar in the Department of Electrical and Computer Engineering at National University of Singapore. His current research interests include computer vision, multimedia computing, and deep learning, and published He has authored over 35 journal and conference papers in these areas, including IEEE TPAMI, IEEE TIP, CVPR, ICCV, and ACM MM, etc. He has received the Best Student Paper Award in MMM 2016, the Best Paper Runner-up in ACM MM 2015, the Excellent Doctoral Dissertation of CAAI, and the Excellent Doctoral Dissertation of Jiangsu Province.



Ren-Jie Song is currently the researcher in Megvii Research Nanjing, Megvii Technology, China. He graduated from Nanjing University of Science and Technology in July 2015, and then received the Master degree from Nanjing University in July 2018. His current research interests include deep learning, computer vision and their applications.

Jianfeng Lu is currently a Professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include image processing, machine learning, data mining and intelligent robot.