Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Original Article

Multi-Granularity Sequence Generation for Hierarchical Image Classification

Xinda Liu [1], **Lili Wang**[1,2](✉)

Abstract

Hierarchical multi-granularity image classification is a challenging task that aims to tag each given image with multiple granularity labels simultaneously. Existing methods tend to overlook that different image regions contribute differently to label prediction at different granularities, and also insufficiently consider relationships between the hierarchical multi-granularity labels. We introduce a sequence-to-sequence mechanism to overcome these two problems and propose a multi-granularity sequence generation approach (MGSG) for the hierarchical multi-granularity image classification task. Specifically, we introduce a transformer architecture to encode the image into visual representation sequences. Next, we traverse the taxonomic tree and organize the multi-granularity labels into sequences, vectorize them and add positional information. The proposed multi-granularity sequence generation method builds a decoder that takes visual representation sequences and semantic label embedding as inputs, and outputs the predicted multi-granularity label sequence. The decoder models dependencies and correlations between multi-granularity labels through a masked multi-head self-attention mechanism, and relates visual information to the semantic label information through a cross-modality attention mechanism. In this way, the proposed method preserves the relationships between labels at different granularity levels and takes into account the influence of different image regions on labels with different granularities. Evaluations on six public benchmarks qualitatively and quantitatively demonstrate the advantages of the proposed method. Our project is available at https://github.com/liuxindazz/mgsg.

Keywords Hierarchical multi-granularity classification; Vision and text transformer

1 Introduction

Recently, there has been increasing interest in applying and processing multi-granularity images [1–4] in the computer vision and multimedia communities. Research on hierarchical multi-granularity images plays a crucial role in bridging the gap between vision and semantics, since multi-granularity images naturally contain hierarchical label semantic information and visual feature representations at different granularity levels. Among the many related tasks, hierarchical multi-granularity image classification [5–7] is a fundamental and challenging task that simultaneously identifies each given image belonging to labels at different granularity levels.

Our investigations indicate that there are two significant difficulties in hierarchical multi-granularity image classification. Firstly, labels with different granularities have different effects on learning other granular features. As Chang *et al.* claim, coarse-level label prediction exacerbates fine-grained feature learning, yet fine-level features improve the learning of coarse-level classifiers. The essence of this problem lies in the loss of the relationships between the hierarchical multi-granularity labels. The network cannot rely on the relationships between the hierarchical multi-granularity labels for training and prediction. Secondly, hierarchical multi-granularity image classification involves the hard subtask of fine-grained image recognition. Fine-grained image recognition is challenging due to subtle inter-class differences and significant intra-class variance. Labels of different granularity correspond to different semantics, and the image regions they correspond to are likely to be different. Therefore, in the case of parallel learning of multi-granularity labels, there may be interactions between features and labels at different granularity levels. The essence of this problem is that the modalities of the label and the image are different, so it is impossible to directly find an appropriate regional expression in the image through the hierarchical multi-granularity label.

Existing methods usually use a particular mechanism to solve these problems: a common backbone extracts

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, 100191, China. Lili Wang is the corresponding author. E-mail: liuxinda@buaa.edu.cn, wanglily@buaa.edu.cn.

[2] Peng Cheng Laboratory, Shengzhen, 518000, China.

Manuscript received: 2022-01-01; accepted: 2022-01-01

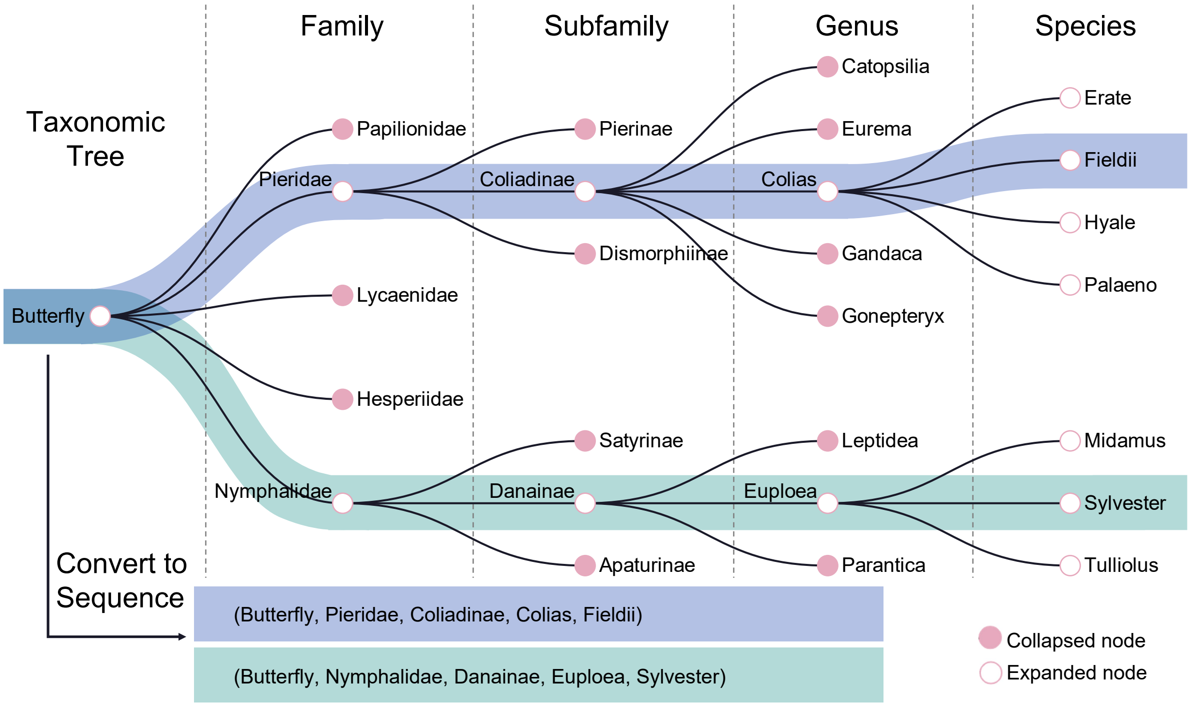

Fig. 1 An example to motivate our approach. The tree shows the hierarchical multi-granularity structure of butterfly taxonomic. Existing methods divide categories vertically, destroying the relationships between levels, but our method processes them horizontally to obtain sequences.

features and then different classification heads are used to predict labels at different granularity levels. As Figure 1 shows, existing methods usually divide different categories vertically into different granularity levels, in this case, family, subfamily, genus, and species, and then assign a classification header to each granularity level. These methods use different classification heads to separate labels at different granularity levels, thereby minimizing the impact of labels at different granularity levels on learning other granularity features. However, this mechanism often overlooks that different image regions contribute differently to label prediction at different granularities and fails to take into account relationships between the hierarchical multi-granularity labels.

Looking at Figure 1, we see that starting from the root node of the taxonomic tree and traversing to some leaf node will naturally generate a sequence. This sequence not only effectively preserves the relationships between the various categories but also maintains the granular level information of the categories by position. Based on this observation, we introduce a sequence-to-sequence mechanism to overcome the limitations of existing methods, and propose a multi-granularity sequence generation approach for the hierarchical multi-granularity image classification task.

Specifically, we first encode the image into a visual representation; the encoding methods may include different types of convolutional neural networks or vision transformer structures. Without loss of generality, the encoding process is introduced using a vision transformer as an example. The given image is first reshaped into a patch sequence without overlap and then linearly mapped to the sequential tokens. We build a stack of transformer encoder layers to encode given sequential image tokens into a visual representation sequence. Each transformer encoder layer contains a multi-head self-attention module, a multilayer perceptron, and a residual structure. We exploit pre-trained transformer-based vision models' excellent feature expression ability to obtain a more discriminative visual feature representation. Next, we traverse the taxonomic tree and organize the multi-granularity labels into sequences. These text label sequences are vectorized by initializing a series of label embeddings. Then we add location information to these label embeddings to maintain the granularity of labels. We build a stack of transformer decoder layers to decode the visual representation sequence to generate hierarchical multi-granularity label sequences. Each transformer decoder layer contains a masked multi-head self-attention module, a cross-modality attention module, a multilayer perceptron, and a residual structure. The proposed multi-granularity sequence generation method decoder takes visual representation sequences and semantic label embedding as input and outputs the predicted multi-granularity label sequence. The decoder preserves the dependencies and correlations between hierarchical multi-granularity labels by applying the masked multi-head self-attention mechanism to labels of different granularity levels. The decoder maps the visual information to the semantic information of labels by applying the cross-modal attention mechanism to the visual representation sequence and semantic label embedding. In this way, the proposed method preserves the relationships between labels at different granularity levels and considers the influence of different image regions on labels with different granularities.

To verify the effectiveness of our method, we have conducted extensive experiments on popular benchmarks for the hierarchical multi-granularity image classification task. They demonstrate that the proposed method achieves results competitive with state-of-the-art approaches. Qualitative experimental results also demonstrate the effectiveness of the method for modeling label relationships at different granularities and finding different image regions for different granularity labels.

In summary, we make two main contributions.

- We introduce a sequence-to-sequence mechanism and propose a multi-granularity sequence generation approach for hierarchical multi-granularity image classification. The proposed method effectively models the dependencies and correlations between multi-granularity labels and strengthens the contribution of different image regions to different granularity labels.
- Extensive quantitative and qualitative experiments demonstrate the effectiveness of the proposed method, which achieves performance competitive with state-of-the-art approaches on six public datasets. Visual results also confirm that the proposed method effectively models relationships between labels at different granularities and selects appropriate image regions to judge labels at different granularity levels.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed framework. Experimental results and analysis are reported in Section 4. Finally, we have a discussion in Section 5 and conclude the paper in Section 6.

2 Related Work

In this section, we review the most recent work on hierarchical multi-label classification, fine-grained image recognition, and the vision transformer architecture, especially as it relates to our own work. We also consider how our framework differs from previous ones.

2.1 Hierarchical Multi-granularity Image Classification

This section first discusses the relationship between hierarchical multi-granularity image classification and hierarchical multi-label classification, and then elaborates on the characteristics of hierarchical multi-granularity image classification . In hierarchical multi-label classification, samples are assigned one or multiple class labels organized in a structured label hierarchy [8]. Typical hierarchical multi-label classification problems are text classification [9, 10] and bioinformatics tasks such as protein function prediction [11] and gene function [12]. In the computer vision and multimedia, tasks such as image annotation [13], few-shot image recognition [14], and semantic segmentation [15] are also treated as multi-label classification problems.

Hierarchical multi-granularity image classification is a particular type of hierarchical multi-label classification. The general hierarchical multi-label classification task implies that objects contain different aspects of attributes at different levels, while the hierarchical multi-granularity image classification task emphasizes different levels of image perception. For example, in general hierarchical multi-label classification, document classification, a document containing the word *football* could be labeled both with *sport* and *outdoor activity* at the same time. In contrast, in the multi-granularity image recognition task, an image in the CUB-200-2011 dataset should first be recognized as a *bird* and then as a *flamingo* as bird knowledge increases. However, most existing works ignore this feature and only use different classification heads to process labels of different granularities simultaneously. Some work has explored solutions to this problem. For example, Chang *et al.* propose that the ability to recognize labels at different granularity levels can be increased by combining different classification heads. Wang *et al.* propose to use a hierarchy transition matrix to guide the classification head for training and prediction. Chen *et al.* propose to use an attention mechanism to integrate the output of the classification head, thereby improving the ability to model multi-granularity label relationships. Although these methods are successful, they still do not explicitly model multi-granularity label relationships. Many studies [16–19] on hierarchical classification tasks have extensively explored how to exploit the relationship between multi-granularity labels. Chen *et al.* propose a multi-granularity regularization method to extract hierarchical structure, Wang *et al.* [17] propose a deep fuzzy tree model to learn a better tree structure, and Wang *et al.* [18] use deep reinforcement multi-granularity learning to minimize the risk of hierarchical classification errors. Like these methods, we also used a tree structure to express the relationships between multi-granularity labels. Unlike these methods, our approach further transforms the tree structure into a collection of multiple sequences, and then models the relationships. In this paper, we change the way to approach hierarchical multi-granularity image classification, mimicking the natural process of cognition, generating the coarsest-grained labels first and then gradually generating fine-grained labels. We model and preserve hierarchical multi-granularity label relationships more efficiently by constraining multi-granularity label relationships using a sequence-to-sequence network structure.

2.2 Fine-grained Image Recognition

In the hierarchical multi-granularity image classification task, fine-grained image recognition [20–27] is more complicated than coarse-grained image recognition [28–30], due to subtle inter-class differences and significant intra-class variance.

There are two prevailing paradigms in current research into fine-grained image recognition: the local approach, and the global approach. Local approaches focus on locating discriminative semantic parts of fine-grained objects using supervised [31, 32] or weakly supervised [20–22] mechanisms to identify subtle differences between different object categories. They then build intermediate representations corresponding to these parts for final classification. Inspired by such local methods, we input patch-level features into the transformer decoder of the proposed method to predict fine-grained labels. Global approaches [23–27] typically learn discriminative representations with a specific distance metric so that samples of the same class are close while samples of different classes are separated.

Global and local approaches have different emphasis, and both can achieve satisfactory results on fine-grained image recognition tasks. However, in a hierarchical multi-granularity image classification, while fine-grained features lead to better learning of coarse-level classifiers, coarse-level label prediction makes fine-grained feature learning more difficult, as Chang *et al.* point out. Therefore, we propose to model labels with different granularities to reduce the adverse effects of coarse-grained labels on fine-grained feature learning.

2.3 Vision Transformers

The transformer is an attention-based [33, 34] encoder-decoder architecture, which was proposed to deal with sequences in the field of natural language processing (NLP) [35, 36]. Inspired by breakthroughs provided by transformer architectures in NLP, computer vision researchers have applied an additional attention layer in either spatial [37, 38] or channel domains [39, 40] to capture long-range dependencies. Inspired by these ideas, Dosovitskiy *et al.* [41] proposed a pure transformer by using image patches as input for image classification; it achieves state-of-the-art results on many image classification benchmarks. Subsequently, many recent works have applied transformers to computer vision tasks with comparable results [42]. These include image recognition [41, 43, 44], object detection [45, 46], segmentation [47], and image super-resolution [48].

Transformer-based methods [49–54] have also proved useful in fine-grained image recognition. Specifically, He *et al.* introduced a vision transformer as a backbone and proposed the TransFG approach to select discriminative image regions with the attention map. Chou *et al.* proposed a plug-in network that can effectively extract discriminative and uninformative areas in images, improving recognition accuracy. TransFG and related subsequent work FFVT [52], AFTrans [50], and RAMS-Trans [51] belong to the the local method paradigm. Liu *et al.* [53] exploit the transformer architecture using a peak suppression module and knowledge guidance module, in an approach belonging to the global method paradigm .

Inspired by the above methods, we introduce a transformer architecture into hierarchical multi-granularity image recognition and propose a transformer decoder to generate a multi-granularity label sequence, which provides a strong basis for hierarchical multi-granularity image recognition.

3 Method

This section introduces our proposed framework, a transformer architecture for hierarchical multi-granularity image classification. Section 3.1 gives an overview of the model, Section 3.2 details the image encoder and label sequence construction, and Section 3.3 describes the proposed multi-granularity sequence generation approach.

3.1 Overview

As Fig. 2 shows, our framework has a clear transformer encoder and decoder architecture. The transformer encoder takes images as input and outputs representations of all tokens to the transformer decoder. The transformer decoder takes these representations as input, initially generates the coarsest-grained labels, and then combines the already generated labels to successively generate finer-grained labels.

Before we give details, we must define some necessary notation. Given the label space of g-level granularity with L

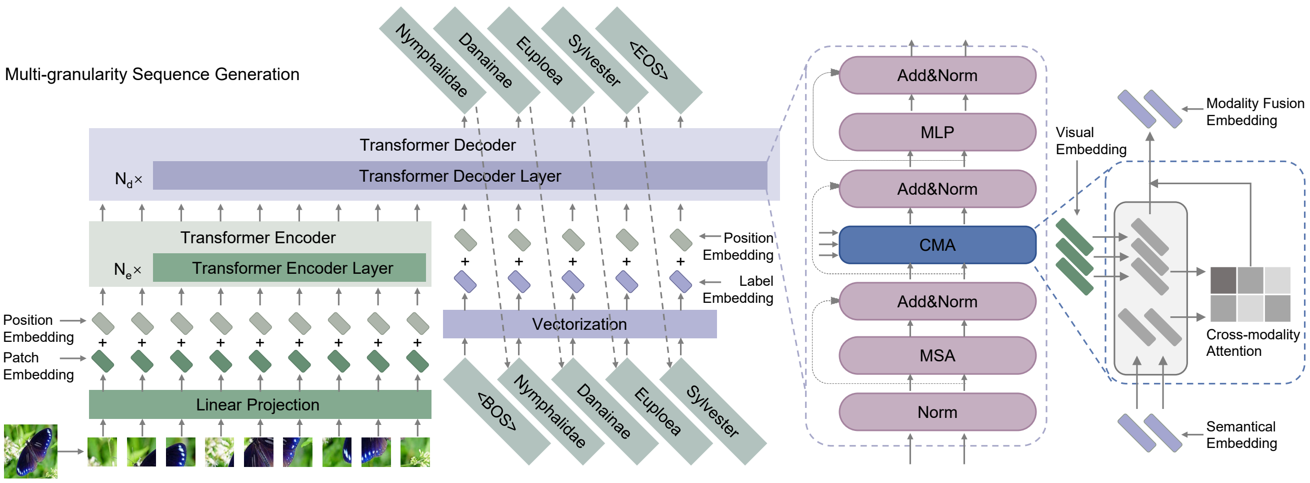

Fig. 2 Our framework. Here we visualize multi-granularity sequence generation given a training batch of butterfly images and corresponding hierarchical multi-granularity label sequences.

labels $\mathcal{L} = l_1, \ldots, l_L$ and an image x, the task is to assign a subset y containing g labels in the label space $\mathcal{L}$ to x. From the perspective of multi-granularity sequence generation, the hierarchical multi-granularity image classification task can be formalized as finding an optimal multi-granularity label sequence $y^\star$ that maximizes the conditional probability $p(y|x)$, calculated as follows:

$$p(y|x) = \prod_{i=1}^{g} p(y_i|y_1, \ldots, y_{i-1}, x). \tag{1}$$

Refer to Figure 2. The given image x is encoded into a series of feature representations by the transformer encoder. These representations serve as the global feature $\mathcal{F}$ for multi-granularity sequence generation. The transformer decoder takes the global feature $\mathcal{F}$ and the previous output state y_{t-1} of the decoder as the inputs to produce the output state y_t at time-step t. Finally, the loss is calculated by comparing the last output y_t with the ground truth, and the network parameters are updated by back-propagation.

3.2 Embedding Methods

We next describe in Sections 3.2.1 and 3.2.2 respectively how to construct the input and output required by the decoder of the multi-granularity sequence generation method. For the input image, we use the transformer encoder structure for encoding For the input label, we need to first convert the label into a label sequence, and then align it for vectorization.

3.2.1 Transformer Encoder

Let $x \in \mathbb{R}^{H \times W}$ denote a given training image of resolution (H, W). The image x is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{K \times P^2}$, where the resolution of each image patch is (P, P), and $K = HW/P^2$ is the resulting number of patches. These patches are converted to a D dimensional embedding $E_{\text{patch}} \in \mathbb{R}^{K \times D}$ as input tokens through a trainable linear projection. The learnable position embedding $E_{\text{pos}} \in \mathbb{R}^{K \times D}$ is added to the patch embedding to retain positional information, and the result is denoted $\mathcal{F}_0$. The transformer encoder takes this fused vector $\mathcal{F}_0$ as the initial input, and outputs a feature representation with the same dimension as the input. In detail, the transformer encoder is composed of a stack of N_e transformer encoder layers. Each encoder layer consists of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks. Layer normalization (LN) is applied before each block and residual connections are applied after each block. This process is shown in Equation 2.

$$\begin{aligned}
\mathcal{F}_0 &= E_{\text{patch}} + E_{\text{pos}}, \\
\mathcal{F}'_i &= \text{MSA}(\text{LN}(\mathcal{F}_{i-1})) + \mathcal{F}_{i-1}, \quad && i = 1, \ldots, N_e, \\
\mathcal{F}_i &= \text{MLP}(\text{LN}(\mathcal{F}'_i)) + \mathcal{F}'_i, && i = 1, \ldots, N_e.
\end{aligned} \tag{2}$$

It can be seen that the dimension of the final output $\mathcal{F}$ is $(K \times D)$.

3.2.2 Sequence Construction

For a common hierarchical multi-granularity image classification task, the multi-granularity labels are built as a tree structure. In order to convert the tree structure into sequences,for each image, we start from the root node and traverse the entire tree to generate a sequence corresponding to each leaf node. Therefore, each image corresponds to a sequence of length g from the coarsest-grained label to the finer-grained labels. To facilitate parallelization of the transformer decoder, we need to align the input sequence with the output sequence. Therefore, we add beginning and end of sequence markers ⟨BOS⟩ and ⟨EOS⟩ to the head and tail of the label sequence respectively, making the length of the input and output sequences $g + 1$. To bale to more accurately express and facilitate subsequent operations, we vectorize the elements in each sequence into a label embedding $E_{\text{label}} \in \mathbb{R}^{(g+1)\times D}$; each label embedding has the same dimension as the transformer encoder output. The learnable position embedding $E^{'}_{\text{pos}} \in \mathbb{R}^{(g+1)\times D}$ is added to the label embedding to retain the sequence context. The result is denoted $\mathcal{S}_0$.

3.3 Multi-granularity Sequence Generation

Section 3.3.1 describes how the decoder of the proposed method builds the relationships between the different granularity level labels, while Section 3.3.2 explains how we associate the visual embedding sequence with the semantic multi-granularity label embedding.

3.3.1 Relationships between Labels

The transformer decoder treats $\mathcal{F}$ is $(K \times D)$, the output of the transformer encoder, as the global features of the images and takes the label sequence embeddings $\mathcal{S}_0$ as the input, finally, outputs the predicted values for labels at different granularity levels of the image.

The transformer decoder is composed of a stack of N_d transformer decoder layers. Each decoder layer consists of MSA, cross modality attention (CMA) and MLP blocks. We use residual connections and layer normalization to avoid over-fitting during the network training stage. To model the relationships between labels at different granularity levels, we first feed the label sequence embeddings into the multi-head self-attention layer. This process is similar to MSA in the encoder, but the self-attention layer in the decoder only allows attention to earlier positions in the output sequence. Therefore, we mask out the following sequences before the softmax step in the self-attention computation.

$$\begin{aligned} \mathcal{S}'_i &= \text{MaskedMSA}(\text{LN}(\mathcal{S}_{i-1})) + \mathcal{S}_{i-1}, && i = 1, \ldots, N_d, \\ \mathcal{S}_i &= \text{MLP}(\text{LN}(\mathcal{S}'_i)) + \mathcal{S}'_i, && i = 1, \ldots, N_d. \end{aligned} \tag{3}$$

3.3.2 Visual-Semantic Modality Fusion

The input to the decoder contains two parts: the visual embedding obtained from encoding the image patches and the semantic embedding obtained from vectorization of the multi-granularity labels. We now introduce how we relate the visual embedding sequences to the semantic embedding of multi-granularity labels. The fusion of different modality embeddings is a problem that has been widely studied. Liu *et al.* [55] suggest that each modality feature should be decomposed into a weighted sum of multiple low-rank features. Then, element-wise multiplication is performed to obtain fused multi-modality features. This approach inspired our design of the CMA module. However, rather than directly using element-wise multiplication, the CMA module uses attention between different modality features to fuse them to obtain multi-modality features. In order to solve the problem of different image regions corresponding to labels at different granularity levels, we apply the attention mechanism to all input image token embeddings and label sequence embeddings. However, there is a a semantic gap between multi-granularity label embedding space and visual feature space because of the modality difference. To solve this problem, we map the multi-granularity label embedding and image token embedding into a shared space through a set of learnable shared parameters, and then calculate their similarity and fuse them.

$$\begin{aligned} \mathcal{S}''_i &= \text{CMA}(\mathcal{S}_i, \mathcal{F}) + \mathcal{S}_i, && i = 1, \ldots, N_d, \\ \mathcal{S}_i &= \text{MLP}(\text{LN}(\mathcal{S}''_i)) + \mathcal{S}''_i, && i = 1, \ldots, N_d. \end{aligned} \tag{4}$$

After performing multiple attention-based operations, we output an embedding of the same size as the input, followed by an MLP layer that outputs a likelihood score for each category:

$$y = \text{softmax}(\text{MLP}(\text{out})), \tag{5}$$

where out denotes the class token vector of the output of the last transformer decoder layer. We guide network training by minimizing the cross-entropy loss between y and ground-truth labels.

In the training phase, the masked MSA allows the model to be trained in parallel to build the relationships between hierarchical multi-granularity labels. During inferencing, we can now automatically regress to generate the output from the initial ⟨BOS⟩ vector: in the process of continuous iteration, the predicted label is used to replace the real label for prediction, and finally a complete multi-granularity label sequence is generated.

The proposed multi-granularity sequence generation method builds a decoder that inputs a sequence of visual representations and semantic label embeddings, and outputs a predicted sequence of multi-granularity labels. The decoder maintains the dependencies and correlations between multi-granularity labels through the masked multi-head self-attention mechanism, solving the common label-category relationship loss problem in hierarchical multi-granularity image classification. The decoder also associates visual information with semantic information of hierarchical

Table 1 Multi-granularity datasets used to evaluate our proposed method.

Dataset	Level 1 labels	Level 2 labels	Level 3 labels	Level 4 labels	Training images	Testing images
Butterfly-200	5	23	116	200	5,135	15,009
CUB-200-2011	13	38	200	-	5,994	5,794
FGVC-Aircraft	30	70	100	-	6,667	3,333
Stanford Cars	9	196	-	-	8,144	8,041
ISIA Food-200	11	52	200	-	118,210	59,287
ISIA Food-500	11	60	500	-	239,379	120,143

Table 2 Hierarchical structure of the six experimental datasets.

Label Level	Butterfly-200	CUB-200-2011	FGVC-Aircraft	Stanford Cars	ISIA Food-200	ISIA Food-500
1	family	order	maker	maker	basic	basic
2	subfamily	family	family	model	ingredient	ingredient
3	genus	species	model	-	dish	dish
4	species	-	-	-	-	

multi-granularity labels through a cross-modal attention mechanism, solving the problem that cross-modal information cannot be effectively matched.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We conducted qualitative and quantitative experiments on six publicly available multi-granularity datasets: including Butterfly-200 [5], CUB-200-2011 [56], FGVC-Aircraft [57], Stanford Cars [58], ISIA Food-200 [59] and ISIA Food-500 [60] datasets. Statistical details of these datasets including the number of labels at each level and numbers of training and test images are summarized in Table 1. The label hierarchies for these datasets are shown in Table 2.

4.1.2 Implementation Details

We implemented the proposed method with Pytorch, using four Nvidia V100 GPUs. The input images were resized to 384 × 384. Following the setting used for Swin Transformer [44], we used data augmentation, including random cropping and horizontal flipping, during the training procedure. Only center cropping was performed during inferencing. The model was trained for 50 epochs with stochastic gradient descent. The batch size was set to 16 and momentum to 0.9 for all datasets. The learning rate was set to 5×10^{-4} initially, with a cosine decay schedule. We adopted Swin-Transformer pre-trained on ImageNet21k to initialize the image encoder parameters in all our experiments. We calculated the top-1 accuracy of different granularity levels as the evaluation metric.

4.2 Ablation and Related Analyses

We conducted a series of studies using the CUB-200-2011 dataset in order to understand better the working of the proposed multi-granularity sequence generation approach. Quantitative experiments were used to assess the influence of choice of backbone network on classification performance, and the influence of indiscriminately treating categories at different granularity levels on classification performance. Qualitative experiments were used to analyze how the proposed method affected the modeling of label relations.

4.2.1 Quantitative Experiments

In order to investigate the contribution of the CMA component in the proposed method, we omitted it, and used different backbones: ResNet-50 [61], Vision Transformer (Dosovitskiy *et al.*) and Swin Transformer [44]. We report the corresponding recognition accuracies in Table 3. We see that omitting the CMA module decreases average recognition accuracy of multi-granularity labels in each case, demonstrating the utility of CMA components for hierarchical multi-granularity image classification.

We also conducted experiments on feature learning using different prediction orders to explore the effect of multi-granularity labeling on feature learning at different granularities. Table 4 shows the experimental results: both forward order and reverse order sequential prediction are better than parallel prediction. This implies that coarse-grained label learning inhibits fine-grained feature learning in parallel learning of multi-granularity labels, as claimed by Chang *et al.*. One possible reason is the failure to model the relationship between multi-granularity labels in parallel prediction. Unlike parallel prediction methods, a sequential learning paradigm can better exploit correlation between multi-granularity labels, either using forward or reverse order prediction. Coarse-to-fine

Table 3 Results of ablating the CMA component, using the CUB-200-2011 dataset.

Method	Backbone	CUB-200-2011			
		l1: order	*l2*: family	*l3*: species	average
Without CMA Full MGSG	ResNet-50 (He *et al.*)	97.43 97.43	92.56 92.82	79.92 80.13	89.97 90.13
Without CMA Full MGSG	ViT (Dosovitskiy *et al.*)	99.35 99.60	97.67 98.00	90.10 90.23	95.71 95.94
Without CMA Full MGSG	Swin-T (Liu *et al.*)	99.43 99.66	98.65 98.65	91.22 91.84	96.43 96.72

Table 4 Effect on accuracy (in %, over all levels) of changing prediction order, using the CUB-200-2011 dataset.

Prediction Order	CUB-200-2011			
	l1:order	l2:family	l3:species	average
Parallel	99.31	98.65	90.66	96.04
Reverse	99.13	98.49	91.78	96.47
Forward	99.66	98.65	91.84	96.72

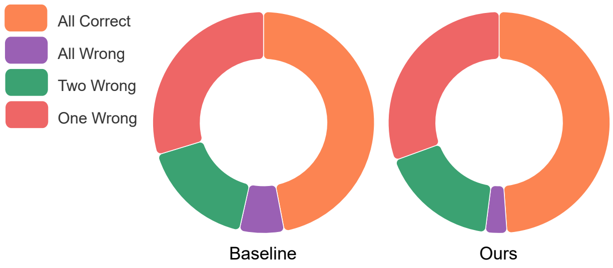

Fig. 3 Overall overview of the proportion of labels that were correctly and incorrectly predicted. Best viewed in color.

forward prediction is slightly better than reverse prediction, suggesting that in an asynchronous learning paradigm, coarse-grained labels may instead facilitate the learning of fine-grained features.

4.2.2 *Visual Assessment*

In addition to a quantitative analysis, we visualized proportional relationships between correct and incorrect predicted labels, to assess the effectiveness of the proposed method for hierarchical multi-granularity label modeling.

The CUB-200-2011 dataset has three hierarchical levels. Therefore, we divided label prediction results into four categories, as shown in Figure 3: all correct, one label wrong, two labels wrong, and all wrong. In order to more clearly see differences, we took the logarithms of all values. In this visualization, our method does not differ much from the baseline method in general.

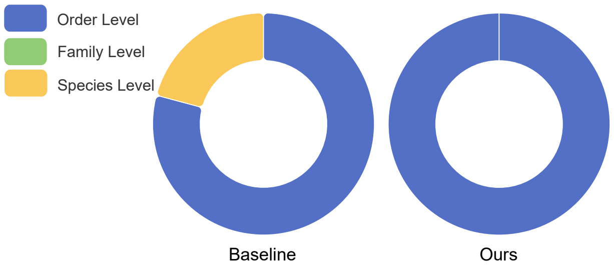

Fig. 4 Distribution of labels at each granularity level when only one label is correct.

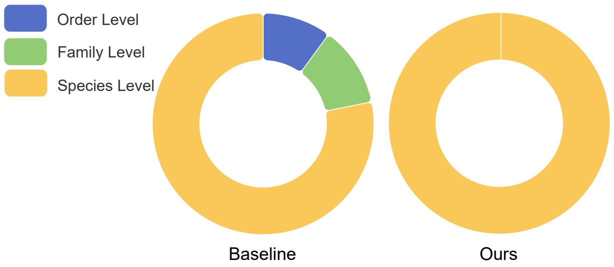

Fig. 5 Distribution of labels at each granularity level when only one label is wrong.

Ignoring the all-wrong and all-right cases, which are irrelevant to the experimental goal, we further analyzed the cases with one or two labelling errors. We first counted the number of correct labels at each level when only one label is correctly identified. The result is shown in Figure 4. If only one label is correct for the baseline method, it will most likely be the coarsest-grained label, the order level label, as one would expect. An interesting phenomenon is that in some cases, the only correct labels are the finest-grained labels, at species level; there are no cases where the only correct label is at an intermediate-level. This may mean that labels at different granularity levels guide the network to learn towards both ends, and is a topic requiring further exploration in future. However, it is unreasonable to correctly predict only the species level label but not the higher level labels, which indicates that the network ignores semantic relationships between labels at different granularity levels during training. Compared to the baseline method, in cases where only one label is correctly predicted, the proposed method is much more successful at predicting it at the coarsest-grained level. In the two-label error case, the proposed method outperforms the baseline method.

In the case of only one label in error, the results of the proposed method are very different from the baseline method. They are difficult to distinguish, since fine-grained images have significant intra-class differences and slight inter-class differences. Therefore, for the baseline method, if only one

Table 5 Accuracy (%) achieved at each level by various methods, for the CUB-200-2011 dataset.

Method	Backbone	CUB-200-2011			
		l1: order	*l2*: family	*l3*: species	average
LHT (Wang *et al.*)	ResNet-50 (He *et al.*)	98.19	92.92	79.29	90.13
HSE (Chen *et al.*)		98.80	95.70	88.10	94.20
FGN (Chang *et al.*)		96.37	90.39	77.95	88.24
MGSG (ours)		97.43	92.82	80.13	90.13
FGN (Chang *et al.*)	PMG (Du *et al.*)	97.98	93.50	82.26	91.25
MGSG (ours)		98.20	94.17	84.61	92.33
MHP	ViT (Dosovitskiy *et al.*)	99.19	97.42	89.40	95.34
MGSG (ours)		99.31	98.00	90.23	95.94
MHP	TransFG (He *et al.*)	99.24	97.98	89.72	95.64
MGSG (ours)		99.36	98.20	90.51	96.02
MHP	Swin-T (Liu *et al.*)	99.31	98.65	90.66	96.04
FGN (Chang *et al.*)		99.63	98.49	91.28	96.46
MGSG (ours)		**99.66**	**98.65**	**91.84**	**96.72**

label is mispredicted, this label is likely to at the finest-grained label, the species label. However, some of the only wrong labels are at the order level or family level, which means that the network model will incorrectly predict the family label when the order and species labels are correctly predicted.

In contrast, when only one label is wrongly predicted, the wrong labels predicted by the proposed method are all species-level labels, as Figure 5 shows: if the coarsest-grained label is wrong, the finer-grained label will also be wrongly predicted, and if the finest-grained label is correctly predicted, then coarser-grained labels are also correct.

The results of the above two experiments show that the baseline method loses the connections between labels during training, whereas the proposed method effectively models the relationships between labels, maintaining semantic consistency between labels at different granularity levels during training.

4.3 Comparison to Other Methods

We have also compared our method to state-of-the-art methods on the six publicly available datasets, and used a multiple head prediction (MHP) method as a strong baseline in order to demonstrate the effectiveness of MGSG. The MHP method uses the pre-trained transformer model as the encoder, followed by three fully connected layers as classification heads to classify labels at different granularity levels. As can be seen from our earlier description, the modules of our proposed method are loosely coupled, so the proposed method can easily be combined with pre-trained models for the hierarchical multi-granularity image classification task.

4.3.1 CUB-200-2011

Here, we compare the proposed method to state-of-the-art hierarchical multi-granularity image recognition models, with experimental results shown in Table 5. We conclude that:

(1) Overall, the proposed method performs better than the state-of-the-art fine-grained methods, including the attention-based approaches methods HSE (Chen *et al.*) and FGN (Chang *et al.*). It is worth noting that the proposed method with Swin Transformer is better by 1.61% (90.23% vs. 91.84%) on the sub-task of label classification at the finest-grained level compared to the method using Vision Transformer. This is understandable, as the sliding window mechanism of Swin Transformer is beneficial when extracting local information, which is crucial for fine-grained image recognition.

(2) The choice of backbone greatly influences the results; a strong backbone can significantly improve classification accuracy. We implemented FGN using Swin-T as the backbone to provide a fair comparison to the proposed method. The results show that the proposed method is still better than the FGN method. The average classification accuracy is improved by 0.78% (95.94% vs. 96.72%) when switching the backbone from the pre-trained Vision Transformer to the pre-trained Swin Transformer, which demonstrates that the proposed method can effectively exploit the expressive ability of the pre-trained model.

(3) For labels at increasingly finer levels, recognition accuracy of the network gradually decreases. For FGN using Swin-T as the backbone, accuracy drops by 1.14% from the order to the family level label and 7.21% from the family to the species level label. For our method, these values are 1.01% and 6.81%, respectively. This shows that although coarse-grained labels hurt the learning of fine-grained features, our method effectively mitigates this effect by modeling labels with different granularities.

(4) To verify the effectiveness of the proposed method, we tried different backbones, including ResNet-50, PMG,

Table 6 Accuracy (%) achieved at each level by various methods, for the Butterfly-200 dataset.

Method	Backbone	Butterfly-200				
		l1: family	*l2*: subfamily	*l3*: genus	*l4*: species	average
LHT (Wang *et al.*)	ResNet-50 (He *et al.*)	98.21	96.37	92.40	81.54	92.13
HSE (Chen *et al.*)		98.90	97.70	95.40	86.10	94.53
FGN (Chang *et al.*)		96.16	94.04	88.92	76.82	88.99
MGSG (ours)		97.28	95.81	91.56	82.30	91.74
FGN (Chang *et al.*)	PMG (Du *et al.*)	98.12	94.98	91.66	82.34	91.78
MGSG (ours)		98.37	95.60	94.13	84.56	93.17
MHP	ViT (Dosovitskiy *et al.*)	99.07	97.83	95.25	87.64	94.95
MGSG (ours)		99.12	98.27	95.71	88.27	95.34
MHP	TransFG (He *et al.*)	99.19	98.42	95.98	88.13	95.43
MGSG (ours)		99.22	98.95	96.21	88.38	95.69
MHP	Swin-T (Liu *et al.*)	99.24	98.62	95.12	88.44	95.36
FGN (Chang *et al.*)		99.54	98.62	95.38	88.62	95.54
MGSG (ours)		**99.66**	**99.06**	**96.78**	**89.24**	**96.19**

Table 7 Accuracy (%) achieved at each level by various methods, for the FGVC-Aircraft dataset.

Method	Backbone	FGVC-Aircraft			
		l1: maker	*l2*: family	*l3*: model	average
LHT (Wang *et al.*)	ResNet-50 (He *et al.*)	95.73	92.89	**88.56**	92.39
HSE (Chen *et al.*)		95.12	92.03	88.23	91.79
FGN (Chang *et al.*)		93.04	90.73	88.35	90.71
MGSG (ours)		94.09	92.17	88.41	91.56
FGN (Chang *et al.*)	PMG (Du *et al.*)	94.57	90.75	88.31	91.21
MGSG (ours)		95.31	91.87	88.47	91.88
MHP	ViT (Dosovitskiy *et al.*)	95.08	91.12	87.69	91.30
MGSG (ours)		95.31	91.80	88.01	91.71
MHP	TransFG (He *et al.*)	95.32	91.73	87.91	91.65
MGSG (ours)		95.76	92.20	88.11	92.02
MHP	Swin-T (Liu *et al.*)	95.57	92.15	88.23	91.98
FGN (Chang *et al.*)		95.83	92.53	88.46	92.27
MGSG (ours)		**96.67**	**93.21**	88.07	**92.65**

ViT, TransFG, and Swin Transformer. Note that we used the standard non-overlapping patch split when using transFG as the backbone while not using contrastive loss, to maintain consistency and fairness of the experiments. The results on CUB-200-2011 in Table 5 show that, when using ResNet-50 or PMG as the backbone, our proposed method outperforms the state-of-the-art FGN method by 1.89% and 1.08% in terms of average accuracy, respectively. With ViT or TransFG as backbone, the proposed method outperforms the baseline method MHP in terms of average accuracy.

4.3.2 *Butterfly-200*

The Butterfly-200 dataset is based on the hierarchical taxonomy used in biology, with items in 200 species, 116 genera, 23 subfamilies, and 5 families. We used this dataset to assess the proposed method when using longer label sequences. As Table 6 shows, our method still shows advantages on the butterfly-200 dataset, with accuracy 1.66% higher than the current state-of-the-art method HSE. As we consider labels at the genus level to labels at the species level, recognition accuracy of the HSE method drops by 9.30%, while for the proposed method, it drops by 7.54%, showing a clear advantage over the HSE method. For the Butterfly-200 dataset with more category levels, the proposed method also improves on the average accuracy of the state-of-the-art method, FGN, by 2.75% and 1.39%, respectively, when using ResNet-50 and PMG as the backbone.

4.3.3 *FGVC-Aircraft*

The FGVC-Aircraft dataset has 10,000 images covering 100 model variants. Table 7 reports the performance of several methods on this dataset. The third-level label of this dataset is model level (e.g., *767-200, 767-300*). Most pre-trained models based on Transformer perform poorly on this dataset, as is our method. It does not reach the state-of-the-art in this particular case, but the method's overall accuracy is still good, with average accuracy performance exceeding the state-of-the-art.

Table 8 Accuracy (%) achieved at each level by various methods, for the Stanford Cars dataset.

Method	Backbone	Stanford Cars		
		l1: maker	*l2*: model	average
LHT (Wang *et al.*)	ResNet-50 (He *et al.*)	96.74	89.67	93.21
HSE (Chen *et al.*)		96.89	91.32	94.11
FGN (Chang *et al.*)		95.58	89.66	92.62
MGSG (ours)		96.19	90.31	93.25
FGN (Chang *et al.*)	PMG (Du *et al.*)	96.42	91.05	93.74
MGSG (ours)		96.77	91.92	94.35
MHP	ViT (Dosovitskiy *et al.*)	96.50	91.19	93.85
MGSG (ours)		96.61	91.53	94.07
MHP	TransFG (He *et al.*)	96.64	91.60	94.12
MGSG (ours)		96.79	91.70	94.25
MHP	Swin-T (Liu *et al.*)	96.68	91.30	93.99
FGN (Chang *et al.*)		97.06	91.62	94.44
MGSG (ours)		**97.40**	**92.77**	**95.09**

Table 9 Accuracy (%) achieved at each level by various methods, for the ISIA Food-200 dataset.

Method	Backbone	ISIA Food-200			
		l1: basic	*l2*: ingredient	*l3*: dish	average
LHT (Wang *et al.*)	ResNet-50 (He *et al.*)	84.32	78.03	69.67	77.34
HSE (Chen *et al.*)		84.15	77.98	69.43	77.19
FGN (Chang *et al.*)		82.97	75.62	65.13	74.57
MGSG (ours)		83.50	77.17	67.61	76.09
FGN (Chang *et al.*)	PMG (Du *et al.*)	83.56	77.39	68.15	76.37
MGSG (ours)		84.43	78.78	69.10	77.44
MHP	ViT (Dosovitskiy *et al.*)	84.92	79.94	71.65	78.84
MGSG (ours)		85.22	81.80	73.11	80.04
MHP	TransFG (He *et al.*)	85.10	81.33	72.89	79.77
MGSG (ours)		85.38	82.00	73.90	80.43
MHP	Swin-T (Liu *et al.*)	85.37	80.19	73.30	79.62
FGN (Chang *et al.*)		85.97	80.81	74.22	80.33
MGSG (ours)		**86.54**	**81.29**	**75.12**	**80.98**

4.3.4 Stanford Cars

In order to further verify the effectiveness of the proposed method at fewer granularity levels, i.e., short label sequences, we conduct experiments on the Stanford Cars dataset, which has only two label levels. Table 8 reports the accuracy of several methods on the Stanford Cars dataset. Going from the maker level labels to the model level labels, FGN recognition accuracy drops by 5.37%, while the proposed method drops by 4.63%: the proposed method can still protect learning of fine-grained features when few label levels are used.

4.3.5 ISIA Food-200

In order to further explore the scope of application of our method, following [62], we explored the hierarchical multi-granularity image classification task on the ISIA Food-200 dataset. We re-organised this dataset into a three-level label hierarchy with 11 major food categories (e.g., *Cereals and cereal products* and *Meat and meat products*, 52 ingredient categories (e.g., *Bacon* and *Beef*) and 200 dish categories (e.g., *Bacon and eggs* and *Beef pie*). These food images lack fixed spatial structure and semantic patterns, and so it is challenging to capture semantic information at different granularities from these images. Our approach attempts to make correspondences between the semantics of different granularities and different image regions, which is more effective for non-rigid objects.

Table 9 reports the performance of several methods on the ISIA Food-200 dataset. In terms of average accuracy, our method achieved the best 80.98% accuracy, outperforming the state-of-the-art method FGN with Swin transformer by 0.65%. This result shows that our method provides more significant performance improvements in complex hierarchical multi-granularity image classification problems.

4.3.6 ISIA Food-500

ISIA Food-500 is a more comprehensive food dataset than ISIA Food-200, with more data, and higher diversity. We reorganized the ISIA Food-500 dataset as for ISIA Food-200, giving 11 major food categories, 60 ingredient categories, and 500 dish categories. We tested in the same way as for

Table 10 Accuracy (%) achieved at each level by various methods, for the ISIA Food-500 dataset.

Method	Backbone	ISIA Food-500			
		l1: basic	*l2*: ingredient	*l3*: dish	average
LHT (Wang *et al.*)	ResNet-50 (He *et al.*)	81.47	73.19	63.35	72.67
HSE (Chen *et al.*)		82.11	73.39	63.28	72.93
FGN (Chang *et al.*)		81.21	72.99	62.83	72.34
MGSG (ours)		82.30	74.01	64.39	73.57
FGN (Chang *et al.*)	PMG (Du *et al.*)	81.83	73.61	63.76	73.07
MGSG (ours)		82.44	74.91	65.75	74.37
MHP	ViT (Dosovitskiy *et al.*)	83.68	76.49	67.98	76.05
MGSG (ours)		84.13	77.52	69.02	76.89
MHP	TransFG (He *et al.*)	84.02	77.28	68.30	76.53
MGSG (ours)		84.48	77.93	69.27	77.23
MHP	Swin-T (Liu *et al.*)	85.12	78.14	69.58	77.61
FGN (Chang *et al.*)		85.26	78.36	70.12	77.91
MGSG (ours)		**85.33**	**78.84**	**70.94**	**78.37**

ISIA Food-200, with results given in Table 10.

Due to the greater amount of data and complexity of the ISIA Food-500 dataset compared to the ISIA Food-200 dataset, a significant decrease in accuracy was observed for all methods. The proposed method exceeds the accuracy of the original baseline method by 1.36% and the state-of-the-art method by 0.82% for the dish level, showing that our method is more effective at exploiting fine-grained semantic information in local regions of images.

For the Stanford Cars dataset with fewer category levels, the artificial product-centered FGVC-Aircraft dataset, and the irregularly shaped food-oriented datasets Food-200 and Food-500, our proposed MGSG method is effective using various backbones.

4.4 Qualitative Assessment

To more intuitively present the effectiveness of our method, we show cross-modality attention maps from our method's decoder, for sample images from several different datasets in Fig. 6. We can draw the following conclusions.

(1) For all datasets, our method shows a clear trend: as the label level becomes finer, the decoder combines the content of more images for classification and recognition. This trend demonstrates the ability of the proposed method to model multi-granularity label relationships from a visual perspective.

(2) For labels at different granularity levels, the location of critical areas for image classification may be different, as is particularly evident in the FGVC-aircraft and Stanford Cars datasets. Therefore, it is valuable to split images into sequences for processing and use cross-modality attention to fuse the semantic information of multi-granularity labels with the visual information of image sequences.

(3) For the coarsest-grained labels, the decoder can often determine the label from a small area. For finer-grained labels, the decoder typically needs more image information to assist judgment. In this way, the logic of our decoder is consistent with the cognitive logic of human beings.

(4) In [53], it is claimed that, for fine-grained level image recognition, the diversity of features is significant. Our method focuses on more image regions, which is also crucial for fine-grained classification.

5 Discussion

5.1 Is Patch Splitting Necessary?

The use of patch splitting in the encoder is effective, but whether it is necessary is a matter of debate. It can be easily seen that the encoder and decoder in the proposed method are loosely coupled, so the encoder module can be easily replaced, as verified by our experiments in Section 4.3. While splitting images into patches is not necessary, it can improve accuracy. We use this operation for three reasons: (i) several previous works have shown that splitting images into patches is an effective way of improving the accuracy of fine-grained image recognition, (ii) converting the image input into a sequence can be more readily extended to other multi-modal inputs, such as a list of food ingredients, and (iii) converting the image input to sequence form allows use of existing pre-trained vision transformer parameters, improving recognition performance. Nevertheless, the experiments using ResNet-50 and PMG as backbones show that splitting images into patches is unnecessary. We could also use other encodings, such as connecting several different fully connected networks as encoders after the feature maps. We intend to explore improvement of the encoder in further work.

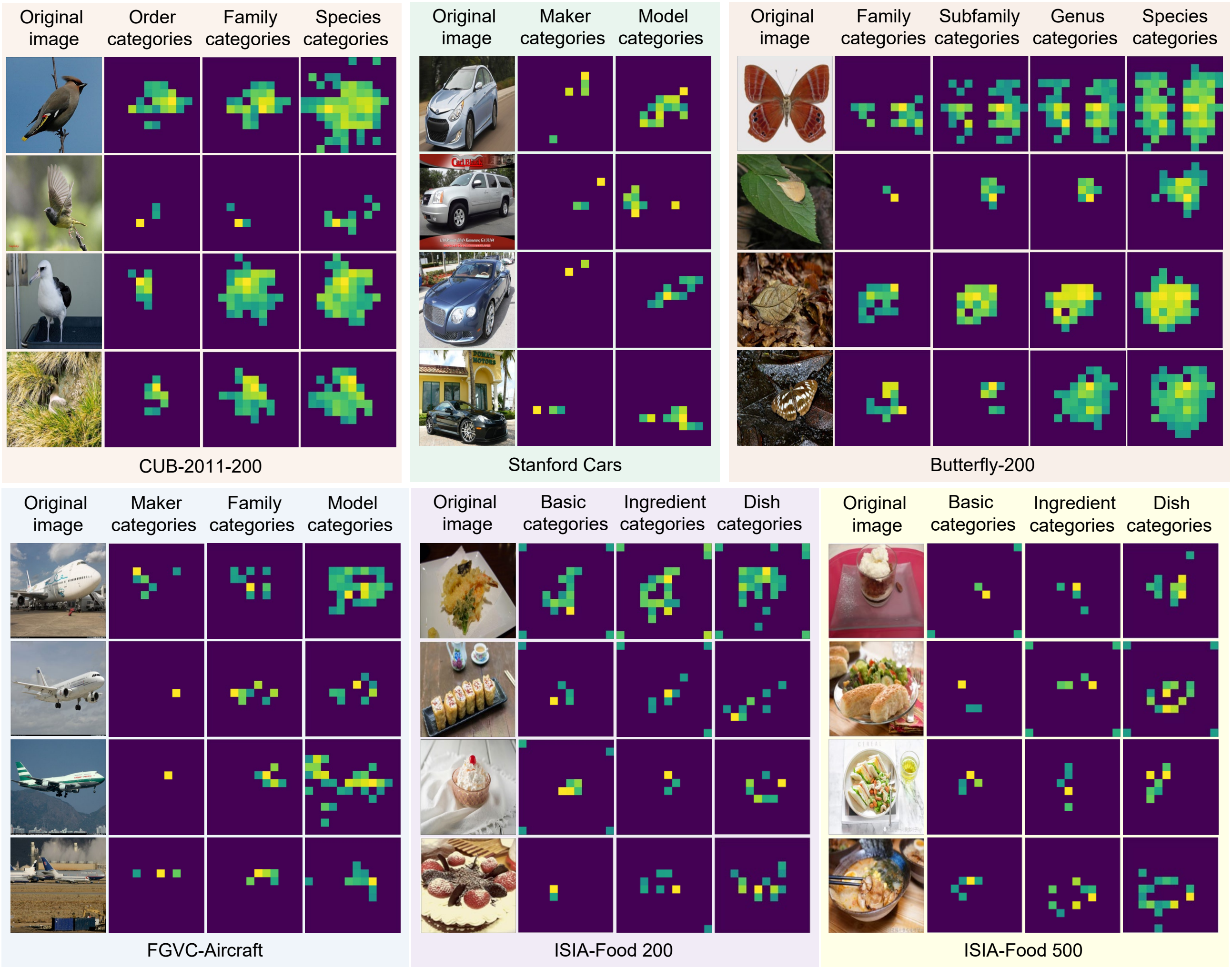

Fig. 6 Cross modality attention weights of different granularity level labels. The yellow means high weights and the green means relatively low weights. We adaptively remove the part smaller than the average value for a better display effect.

5.2 Are Coarse-grained Labels Beneficial or Detrimental to the Learning of Fine-grained Features?

The impact of coarse-grained labels on fine-grained feature learning is worth discussing. Chang *et al.* claim that coarse-level label prediction is detrimental to fine-grained feature learning. Other literature [63–65] concludes that coarse-level information can be beneficial for fine-level learning. Zhao *et al.* and Fan *et al.* propose using a tree classifier instead of the traditional N-way flat softmax classifier. Wang *et al.* propose a coarse-to-fine diagnosis framework to use the knowledge structure. Compared to these works, our problem of multi-granularity labeling is different, and therefore different conclusions are drawn. Our experiments also found that the average accuracy in parallel prediction without modeling relationships is significantly lower than when using sequential forward or reverse order prediction, modeling relationships. Therefore, we conclude that using relationships between different granularity labels is critical in multi-granularity feature learning.

6 Conclusion

In this paper, we have investigated hierarchical multi-granularity image classification and analyzed its particular problems. The first is that the relationships between hierarchical multi-granularity image labels are challenging to construct, and the second is that labels and visual content are difficult to match. We introduce a sequence-to-sequence mechanism to address these two issues, and propose a multi-granularity sequence generation method for hierarchical multi-granularity image classification tasks. The proposed multi-granularity sequence generation method builds a decoder that inputs a sequence of visual representations and semantic label embeddings and outputs a predicted sequence of multi-granularity labels. The decoder solves the first problem above by maintaining the dependencies and correlations between multi-granularity labels through a masked multi-head self-attention mechanism. The decoder also addresses the second problem above by associating visual information with semantic information from hierarchical multi-granularity labels through a cross-modal attention mechanism. Quantitative experiments show that the proposed method can provide results superior to those from state-of-the-art methods. Qualitative experiments show that the method effectively models label relationships at different granularities and finds distinct image regions for labels targeting different levels. Phenomena of interest were found during the experiment,e.g. the network may ignore labels at intermediate levels, which deserve further study.

Acknowledgments

This work was supported by National Key R&D plan 2019YFC1521102, by the National Natural Science Foundation of China through Project 61932003, and by Beijing Science and Technology Plan Project Z221100007722004.

Declaration of Competing Interest

The authors have no competing interests to declare relevant to the content of this article.

References

[1] Niu K, Huang Y, Ouyang W, Wang L. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 2020, 29: 5542–5556.

[2] Du R, Chang D, Bhunia AK, Xie J, Ma Z, Song YZ, Guo J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 2020, 153–168.

[3] Liu D, Wu L, Zheng F, Liu L, Wang M. Verbal-Person Nets: Pose-Guided Multi-Granularity Language-to-Person Generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[4] Ren Y, Wu J, Xiao X, Yang J. Online multi-granularity distillation for gan compression. In *International Conference on Computer Vision*, 2021, 6793–6803.

[5] Chen T, Wu W, Gao Y, Dong L, Luo X, Lin L. Fine-Grained Representation Learning and Recognition by Exploiting Hierarchical Semantic Embedding. In *Multimedia Conference on Multimedia Conference*, 2018, 2023–2031.

[6] Chang D, Pang K, Zheng Y, Ma Z, Song Y, Guo J. Your "Flamingo" is My "Bird": Fine-Grained, or Not. In *Conference on Computer Vision and Pattern Recognition*, 2021, 11476–11485.

[7] Wang R, Xiao K, Jia X, Han X, Meng D, et al.. Label Hierarchy Transition: Modeling Class Hierarchies to Enhance Deep Classifiers. *CoRR*, 2021.

[8] Silla CN, Freitas AA. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 2011, 22(1): 31–72.

[9] Rousu J, Saunders C, Szedmák S, Shawe-Taylor J. Kernel-Based Learning of Hierarchical Multilabel Classification Models. *J. Mach. Learn. Res.*, 2006, 7: 1601–1626.

[10] Cesa-Bianchi N, Gentile C, Zaniboni L. Incremental Algorithms for Hierarchical Classification. *J. Mach. Learn. Res.*, 2006, 7: 31–54.

[11] Triguero I, Vens C. Labelling strategies for hierarchical multi-label classification techniques. *Pattern Recognition*, 2016, 56: 170–183.

[12] Barutçuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. *Bioinform.*, 2006, 22(7): 830–836.

[13] Dimitrovski I, Kocev D, Loskovska S, Dzeroski S. Hierarchical annotation of medical images. *Pattern Recognit.*, 2011, 44(10-11): 2436–2449.

[14] Chen T, Lin L, Hui X, Chen R, Wu H. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[15] Li L, Zhou T, Wang W, Li J, Yang Y. Deep Hierarchical Semantic Segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2022, 1246–1257.

[16] Chen H, Wang Y, Hu Q. Multi-Granularity Regularized Re-Balancing for Class Incremental Learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[17] Wang Y, Hu Q, Zhu P, Li L, Lu B, Garibaldi JM, Li X. Deep fuzzy tree for large-scale hierarchical visual classification. *IEEE Transactions on Fuzzy Systems*, 2019, 28(7): 1395–1406.

[18] Wang Y, Wang Z, Hu Q, Zhou Y, Su H. Hierarchical semantic risk minimization for large-scale classification. *IEEE Transactions on Cybernetics*, 2021.

[19] Wang Y, Hu Q, Chen H, Qian Y. Uncertainty instructed multi-granularity decision for large-scale hierarchical classification. *Information Sciences*, 2022, 586: 644–661.

[20] Min W, Jiang S, Liu L, Rui Y, Jain RC. A Survey on Food Computing. *ACM Computing Surveys*, 2019, 52(5): 92:1–92:36.

[21] Ge W, Lin X, Yu Y. Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up. In *Conference on Computer Vision and Pattern Recognition*, 2019, 3034–3043.

[22] Jiang S, Min W, Liu L, Luo Z. Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition. *IEEE Transactions on Image Processing*, 2020, 29: 265–276.

[23] Lin T, RoyChowdhury A, Maji S. Bilinear Convolutional Neural Networks for Fine-Grained Visual Recognition. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1309–1322.

[24] Chen Y, Bai Y, Zhang W, Mei T. Destruction and Construction Learning for Fine-Grained Image Recognition. In *Conference on Computer Vision and Pattern Recognition*, 2019, 5157–5166.

[25] Sun G, Cholakkal H, Khan S, Khan FS, Shao L. Fine-Grained Recognition: Accounting for Subtle Differences between Similar Classes. In *Conference on Artificial Intelligence*, 2020, 12047–12054.

[26] Zhuang P, Wang Y, Qiao Y. Learning Attentive Pairwise Interaction for Fine-Grained Classification. In *Conference on Artificial Intelligence*, 2020, 13130–13137.

[27] Zou DN, Zhang SH, Mu TJ, Zhang M. A new dataset of dog breed images and a benchmark for finegrained classification. *Computational Visual Media*, 2020, 6(4): 477–487.

[28] Chen L, Yang M. Semi-supervised dictionary learning with label propagation for image classification. *Computational Visual Media*, 2017, 3(1): 83–94.

[29] Chen KX, Wu XJ. Component SPD matrices: A low-dimensional discriminative data descriptor for image set classification. *Computational Visual Media*, 2018, 4(3): 245–252.

[30] Ren JY, Wu XJ. Vectorial approximations of infinite-dimensional covariance descriptors for image classification. *Computational Visual Media*, 2017, 3(4): 379–385.

[31] Huang S, Xu Z, Tao D, Zhang Y. Part-Stacked CNN for Fine-Grained Visual Categorization. In *Conference on Computer Vision and Pattern Recognition*, 2016, 1173–1182.

[32] Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *International Conference on Machine Learning*, volume 32, 2014, 647–655.

[33] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is All you Need. In *Conference on Neural Information Processing Systems*, 2017, 5998–6008.

[34] Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, Zhang SH, Martin RR, Cheng MM, Hu SM. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 2022: 1–38.

[35] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the Association for Computational Linguistics*, 2019, 4171–4186.

[36] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al.. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, 2020, 1877–1901.

[37] Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In *Conference on computer vision and pattern recognition*, 2018, 7794–7803.

[38] Cao Y, Xu J, Lin S, Wei F, Hu H. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *International Conference on Computer Vision Workshops*, 2019, 0–0.

[39] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In *Conference on computer vision and pattern recognition*, 2018, 7132–7141.

[40] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Conference on Computer Vision and Pattern Recognition*, 2020, 13–19.

[41] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021, 1–9.

[42] Xu Y, Wei H, Lin M, Deng Y, Sheng K, Zhang M, Tang F, Dong W, Huang F, Xu C. Transformers in computational visual media: A survey. *Computational Visual Media*, 2022, 8(1): 33–62.

[43] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021, 10347–10357.

[44] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021, 10012–10022.

[45] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision*, volume 12346, 2020, 213–229.

[46] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*, 2020, 1–9.

[47] Ye L, Rochan M, Liu Z, Wang Y. Cross-Modal Self-Attention Network for Referring Image Segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2019, 10502–10511.

[48] Yang F, Yang H, Fu J, Lu H, Guo B. Learning texture transformer network for image super-resolution. In *Conference on Computer Vision and Pattern Recognition*, 2020, 5791–5800.

[49] He J, Chen JN, Liu S, Kortylewski A, Yang C, Bai Y, Wang C. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022, 852–860.

[50] Zhang Y, Cao J, Zhang L, Liu X, Wang Z, Ling F, Chen W. A free lunch from ViT: adaptive attention multi-scale fusion Transformer for fine-grained visual recognition. In *Conference on Acoustics, Speech and Signal Processing*, 2022, 3234–3238.

[51] Hu Y, Jin X, Zhang Y, Hong H, Zhang J, He Y, Xue H. RAMS-Trans: Recurrent Attention Multi-scale Transformer for Fine-grained Image Recognition. In *ACM International Conference on Multimedia*, 2021, 4239–4248.

[52] Wang J, Yu X, Gao Y. Feature Fusion Vision Transformer for Fine-Grained Visual Categorization. *British Machine Vision Conference*, 2021.

[53] Liu X, Wang L, Han X. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. *Neurocomputing*, 2022, 492: 137–149.

[54] Chou PY, Lin CH, Kao WC. A Novel Plug-in Module for Fine-Grained Visual Classification. *CoRR*, 2022.

[55] Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh AB, Morency LP. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Conference of the Association for Computational Linguistics*, 2018, 2247–2256.

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

[57] Maji S, Kannala J, Rahtu E, Blaschko M, Vedaldi A. Fine-Grained Visual Classification of Aircraft. Technical report, Oxford Active Vision Lab, 2013.

[58] Krause J, Stark M, Deng J, Fei-Fei L. 3D Object Representations for Fine-Grained Categorization. In *IEEE Workshop on 3D Representation and Recognition*, 2013, 554–561.

[59] Min W, Liu L, Luo Z, Jiang S. Ingredient-guided cascaded multi-attention network for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, 1331–1339.

[60] Min W, Liu L, Wang Z, Luo Z, Wei X, Wei X, Jiang S. Isia food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, 393–401.

[61] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, 770–778.

[62] Sheng K, Dong W, Huang H, Chai M, Zhang Y, Ma C, Hu BG. Learning to assess visual aesthetics of food images. *Computational Visual Media*, 2021, 7(1): 139–152.

[63] Zhao T, Zhang B, He M, Zhang W, Zhou N, Yu J, Fan J. Embedding visual hierarchy with deep networks for large-scale visual recognition. *IEEE Transactions on Image Processing*, 2018, 27(10): 4740–4755.

[64] Wang Y, Liu R, Lin D, Chen D, Li P, Hu Q, Chen CP. Coarse-to-fine: progressive knowledge transfer-based multitask convolutional neural network for intelligent large-scale fault diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[65] Fan J, Zhao T, Kuang Z, Zheng Y, Zhang J, Yu J, Peng J. HD-MTL: Hierarchical deep multi-task learning for large-scale visual recognition. *IEEE transactions on image processing*, 2017, 26(4): 1923–1938.

Author Biographies

Xinda Liu is currently working towards a Ph.D. degree at Beihang University, and is a researcher in the State Key Laboratory of Virtual Reality Technology and Systems. His research interests include machine learning and image processing.

Lili Wang is a professor in the School of Computer Science and Engineering, Beihang University, and a researcher in the State Key Laboratory of Virtual Reality Technology and Systems. Her research interests include virtual reality, augmented reality, and rendering.