An attention-driven hierarchical multi-scale representation for visual recognition

Zachary Wharton
zachary.wharton@go.edgehill.ac.uk

Ardhendu Behera
https://computing.edgehill.ac.uk/~abehera/

Asish Bera
beraa@edgehill.ac.uk

Department of Computer Science
Edge Hill University
St Helens Road
Ormskirk, Lancashire
L39 4QP, United Kingdom

Abstract

Convolutional Neural Networks (CNNs) have revolutionized the understanding of visual content. This is mainly due to their ability to break down an image into smaller pieces, extract multi-scale localized features and compose them to construct highly expressive representations for decision making. However, the convolution operation is unable to capture long-range dependencies such as arbitrary relations between pixels since it operates on a fixed-size window. Therefore, it may not be suitable for discriminating subtle changes (e.g. fine-grained visual recognition). To this end, our proposed method captures the high-level long-range dependencies by exploring Graph Convolutional Networks (GCNs), which aggregate information by establishing relationships among multi-scale hierarchical regions. These regions consist of smaller (closer look) to larger (far look), and the dependency between regions is modeled by an innovative attention-driven message propagation, guided by the graph structure to emphasize the neighborhoods of a given region. Our approach is simple yet extremely effective in solving fine-grained and generic visual classification problems. It outperforms the state-of-the-art with a significant margin on three and is very competitive on another two datasets.

1   Introduction

The convolution operation is the lifeblood of modern Convolutional Neural Networks (CNNs), used in computer vision to advance popular tasks such as image/video recognition, object detection, visual-question answering, video forecasting, image-to-image translation and many more [30, 31]. It has brought a significant breakthrough in most of these topics and is mainly due to its ability to effectively capture hidden patterns in the Euclidean space since an image can be represented as a regular grid. As a result, it captures meaningful local features by exploiting the shift-invariance and local connectivity. However, a well-known drawback of the convolution operation is its inability to capture long-range dependencies between pixels in the image space as it operates on a fixed-size window. This is inappropriate in discriminating subtle variations in images, especially in fine-grained visual classification (FGVC) involving in recognition of different species of animals, various car/aeroplane models, different kinds of retail products, etc. To address this, researchers have explored informative object parts/regions [5, 22, 25, 40, 53], attention mechanisms to identify these salient regions [4, 36, 37, 45, 56], local and non-local operations [28, 57] to capture discrimina-

tive information. However, these approaches still focus on the Euclidean space while modeling/exploring/identifying salient regions/parts and aggregating interactions/similarities to discriminate subtle variations. In this work, we advance this by taking advantage of Graph Convolutional Networks (GCN), which collectively aggregate key information from graph structure by modeling long-range dependencies in a non-Euclidean domain.

Motivation: To discriminate subtle variations in visual features, many recent works have focused on both single-scale and multi-scale object parts and regions [4, 18, 26, 47, 68]. These are mainly focused on coarse-to-fine exploration by jointly integrating feature representation of regions at different scales. These have notably enhanced recognition accuracy. However, for a better representation of visual-spatial structural relationships among regions, the hierarchical connection between regions should be considered so that the larger regions (see from far) pay more attention to the high-level shape and appearance, and the smaller ones (closer look) concentrate on detailed texture and parts information to capture subtle variations. As a result, it can provide a rich representation by jointly learning meaningful complementary information from multi-scale hierarchical regions that apply to both FGVC (coarse-to-fine) and generic visual classification (fine-to-coarse). We achieve this by a novel multi-scale hierarchical representation learning to boost the recognition accuracy by jointly integrating local (within a region) and non-local (between regions) information to capture long-range dependencies by exploring the graph structure to propagate information between regions within a layer, as well as between layers in the hierarchical structure. In the above SotA methods, the joint feature representation between regions is learned as a part of features. Whereas, we learn this using a graph structure-guided information propagation between regions represented as graph nodes. The propagation is emphasized by learning the "importance" of neighboring regions of a given region. Furthermore, our approach does not require objects or parts bounding-box annotations.

Contributions: Our main contributions are: 1) An innovative multi-scale hierarchical representation learning is proposed for improving visual recognition. 2) A more abstract and coarser representation of multiple graphs denoting the hierarchical structures is considered via spectral clustering-based regions aggregation. 3) A novel gated attention mechanism is proposed to aggregate the cluster-level class-specific confidence. 4) Analysis of the model on five datasets consisting of FGVC and generic visual classification, obtaining competitive results. 5) Visual analysis and ablation studies justify the effectiveness of our model.

2 Related work

Our work involves attention mechanism, multi-scale hierarchical representation, and GCN to improve visual recognition. A precise study on these key aspects is presented.

Generic visual recognition: CNNs have remarkably enhanced the performance of large-scale and generic image classification [11, 24, 49]. Recently, wider and deeper networks with learned data augmentation strategies have been used for performance gain [13]. Researchers have also explored Knowledge Distillation (KD) for visual classification to avoid complex and cumbersome models involving heavy computational overload and memory requirements [75, 79]. A group of small student networks jointly learn by optimizing three loss functions [79]. Similarly, the class-wise self-KD regularizes the dark knowledge and generates effective predictions for generic image classification and FGVC [75]. We address the diverse visual recognition task by exploring hierarchical relations by modeling long-range dependencies among multi-scale regions that self-guide the network to learn jointly.

Hierarchical and Multi-scale methods for FGVC: Weakly supervised methods are widely investigated to enhance discriminability by exploring the hierarchical and multi-scale struc-
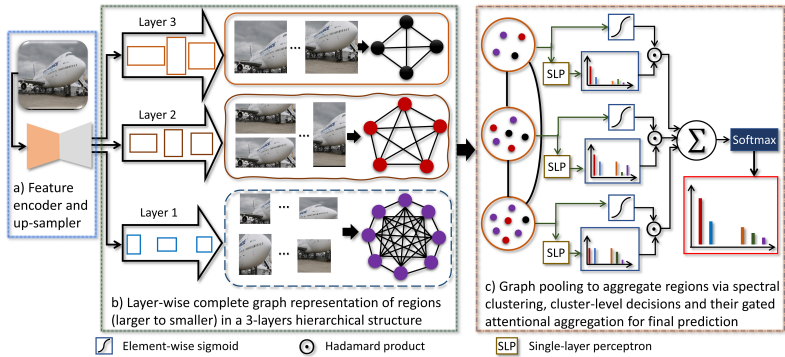
Figure 1: a) CNN backbone as an encoder followed by an up-sampler to increase spatial resolution. b) Layer-wise graph representation using multi-scale regions to capture long-range dependencies using GCN's aggregation scheme in which the feature of a node is computed by recursively aggregating and transforming features from its neighboring nodes. c) Graph coarsening by grouping the similar nodes using spectral clustering, followed by cluster-level predictions and their gated attentional aggregation for making a final decision.

tures. These include but not limited to hierarchical pooling [7, 9, 16, 74], multi-scale representation [19, 59, 87], and context encoding [4]. In [7], higher-order intra- and inter-layer relations are exploited to integrate hierarchical convolutional features from various semantic parts at multiple scales. A cross-layer hierarchical bilinear pooling is explored in [74] to learn complementary information from multiple intermediate convolutional layers. Similarly, a hierarchical semantic embedding is proposed in [9] to model structured correlation at various levels in the hierarchy. In [16], a spatial pyramidal hierarchy-driven top-down and bottom-up attention is described to combine high-level semantics and low-level features. Multi-agent cooperative learning (Navigator-Teacher-Scrutinizer) is aimed to locate informative regions at multi-scale via self-supervision [59]. Inspired by these progresses, we establish an attention mechanism to capture dependencies between multi-scale regions hierarchically at multiple granularities to enhance discriminative feature representation.

Graph convolutional networks (GCN): Motivated by CNN, there is a growing interest to learn high-level representation using GCN in non-Euclidean space [62]. GCN in conjunction with attention mechanism has proven its effectiveness in modeling long-range dependency. Focusing on the node's most relevant hidden representations in a graph, a multi-head-attention [52] is exploited to stabilize the learning task, leveraging the second-order hierarchical pooling [58]. In [84], a self-attention is introduced for hierarchical pooling to learn structural information using the node features and graph topology. Similarly, multi-layer differentiable pooling is proposed to learn hierarchical structures of graphs [72]. Recently, spectral clustering has been adapted in [6] for node clustering to improve the accuracy. A graph-based relation discovery (GRD) approach that applies a collaborative learning strategy is proposed in [85] to learn higher-order feature relationships to enhance the FGVC accuracy. Similarly, GCN is also adapted to advance FGVC by learning the category-specific semantic correspondence using the latent attributes between image-regions [56]. Lately, a criss-cross graph propagation sub-network has been built to find the correlation between discriminative regions [51] in solving FGVC. However, both methods compute fewer regions per image (e.g. 4), which may render a sub-optimal solution. To address this, we propose a multi-scale hierarchical representation learning, guided by attention and graph structure to enhance image recognition tasks through weighted message aggregation.

3 Proposed approach

Our model's architecture is shown in Fig. 1. It pools high-level convolutional features from a CNN by exploring hierarchical (e.g. small → medium → large → full image) regions. These regions are linked using a graph structure, leveraging GCN to propagate information in addressing the shortcomings of prior region-based approaches to enhance the discriminability.

3.1 Problem formulation

A set of N images $I = \{I_n | n = 1 \ldots N\}$ and their respective class labels are given to train an image classifier. The aim is to learn the classifier's mapping function $\mathbf{F}$ that predicts $\hat{y}_n = \mathbf{F}(I_n)$ to match the true label y_n. During training, it learns $\mathbf{F}$ by minimizing a loss $\mathcal{L}(y_n, \hat{y}_n)$ between the true and the predicted label. In our case, the function $\mathbf{F}$ is an end-to-end deep network in which we introduce a simple yet innovative modification to advance the visual recognition. It aggregates information from a given region by attending over its neighbors by exploring a novel attention mechanism. Thus, $\mathbf{F}$ consists of:

$$\mathbf{F} = \text{Softmax}\left(\sum_{l=1}^{L}\sum_{r=1}^{R_l} \phi_l(I_n(A_{r,l}); \theta_l)\right), \tag{1}$$

where $l = 1 \ldots L$ is the layers within the hierarchical structure, and R_l is the number of regions in l^{th} layer. ϕ_l measures the contribution of image-region $A_{r,l}$, representing the rth region located in the lth layer i.e., $I_n(A_{r,l})$ is the hierarchical representation of image I_n. θ_l is the corresponding l^{th} layer's parameter and is learned via end-to-end fashion as follows.

3.2 Hierarchical multi-scale regions

Our hierarchical multi-scale regions approach is motivated by recent region-based approaches [3, 4, 22, 25, 70, 83] in solving FGVC. However, it is different since we use smaller (look closer) to larger (look from far) regions in a hierarchical fashion. Whereas the regions in [22, 25, 70, 83] are object proposals from a detector, and the approaches in [3, 4] use multi-scale regions by exploring HOG cells and blocks. Our regions within a layer in the hierarchy have the same area but with different aspect ratios (Fig. 1b). For example, given width w and height h, there are at least 3 different regions with the same area: 1) width = w and height = h, 2) width = h and height = w, and 3) width = height = $\sqrt{w \times h}$. The aim is to represent an image with a rich discriminative descriptor by considering multiple hierarchical overlapped regions that are not only applicable for advancing FGVC involving large inter-class similarities and great intra-class variations but also pertinent to solve the generic visual classification with distinctive categories (e.g. Caltech-256 [23], CIFAR-100 [53]).

3.3 Graph-based region's feature representation

For a given image I_n, there are L layers within the hierarchical structure. In each layer l ($l = 1 \ldots L$), there are R_l regions representing a complete graph, resulting in L numbers of graphs ($G = \{G_1, G_2, \ldots, G_L\}$) as shown in Fig. 1b. A node in graph G_l denotes a region $r = 1 \ldots R_l$ with feature $f_{r,l}$, resulting in R_l nodes. The aim is to update $f_{r,l}$ of the region/node r by considering a non-local operation, capturing long-range dependencies in the image space. This way, it can capture interactions between regions regardless of their locations in the image space. We achieve this objective by exploring the graph convolution that transforms feature $f_{r,l}$ to $f'_{r,l}$, leveraging different weights to different neighborhood regions to address the drawbacks of the existing approaches in encoding long-range dependencies. We revisit the message passing algorithm in GCN [32] to get a feature representation of a given node i by considering its neighbors using a simple propagation rule i.e. $f_i = \sigma\left(\sum_j \frac{1}{c_{i,j}} f_j W\right)$,

where j represents neighboring nodes of i and $c_{i,j}$ is a normalization constant for the edge (i, j), which originates from the symmetrically normalized adjacency matrix $D^{-\frac{1}{2}} \mathcal{M} D^{-\frac{1}{2}}$ of the GCN. D is the diagonal node degree matrix of the graph adjacency matrix $\mathcal{M}$, $\sigma(.)$ is a nonlinear activation (e.g. Sigmoid or ReLU), and W is a learnable weight matrix. GCN aggregates information from the neighboring nodes that neither "remember" important nodes nor "emphasize" them by giving higher weights throughout the learning. We address this shortcoming by specifying different weights to the neighborhood nodes via the attention mechanism [1, 52]. Especially, our method is inspired by the shared attention mechanism in [53] to perform self-attention in graph nodes. For a given graph G_l, it is computed as:

$$
\alpha_{r,r'} = \frac{\exp(e_{r,r'})}{\sum_{k \in R_l} \exp(e_{r,k})}; \; e_{r,r'} = \text{LeakyReLU}\left(\mathbf{W_a}^T([Wf_r || Wf_{r'}])\right);
$$

$$
f_r' = \Big\|_{h=1}^{H} \sigma\left(\sum_{r' \in R_l} \alpha_{r,r'}^h W_h f_{r'} + b_h\right); \; f_r = \psi(I_n(A_r)),
\tag{2}
$$

where ψ is the feature representation of region A_r in an image I_n, and σ is a nonlinear activation function. Here, ψ refers to local operation (within a region) of global average pooling, and σ denotes Exponential Linear Unit (ELU) [12]. The normalized attention coefficients $\alpha_{r,r'}$ indicate the "importance" of region r' to region r ($r, r' \in \{R_l\}$). W is the shared weight matrix and is used for the linear transformation of each region. $\mathbf{W_a}$ is the weight matrix of the attention mechanism [53] consisting of a single-layer feedforward neural network with the LeakyReLU as a nonlinear activation function. $\|$ represents concatenation to extend single attention to *multi-head attention* (similar to [52]) to stabilize the learning process [53]. We follow this suggestion to use H independent attention heads, where $\alpha_{r,r'}^h$ are the normalized attention coefficients from the hth attention head with the respective linear transformation's weight matrix W_h and bias b_h. We compare the concatenation performance with averaging and experimentally find that the former is better than the latter aggregation.

The above process is applied to each graph G_l within the hierarchical representation of L graphs corresponding to L layers within the structure. The learnable parameters in (2) are graph-specific i.e., not shared among graphs representing different layers, but shared between nodes within a graph. The next objective is how to combine these graphs to make the region-level decision. We achieve this by exploring graph pooling, which allows a graph to learn more abstract representations of the input graphs by summarizing local components and is similar to the pooling operations in CNNs. The recent advancement in model-based pooling methods [6, 20, 34, 72] inspires us learnable pooling. In this work, we adapt the spectral clustering-based graph pooling [6], which computes the soft clustering of the input graphs by aggregating regions belonging to the same cluster using a multi-layered perceptron (MLP) with softmax activation function. The MLP maps each region feature $f_{r,l}' \in \{f'\}$ in graph G_l into the ith row of a soft cluster assignment matrix $\mathbf{S} \in R^{R \times K}$ i.e. $\mathbf{S} = \text{MLP}(f'; \theta)$, where $R = \sum_{l=1}^{L} R_l$ is the total number of regions within the hierarchical structure, and K is the target number of clusters ($K = 3$ in Fig. 1c). The softmax activation guarantees the value of $s_{i,j} \in \mathbf{S}$ within $[0, 1]$, resulting in soft cluster assignment. As a result, we are able to combine L number of graphs into a single complete graph (Fig. 1c) with K nodes in which each node represents a cluster of closely related regions for making the final prediction.

3.4 Gated attention for prediction

Given a graph representation, the image classification can be linked to either node-level [32] or graph-level [71]. At node-level, a single label to a node is assigned and is based on its

high-level feature. Whereas a compact representation of the graph by combining pooling and readout operations is often used in graph-level [6, 34, 72, 80] for classifying the entire graph. A significant drawback of the node-level classification is its inability to propagate hierarchical information to facilitate decision-making. Similarly, graph-level approaches aggregate node representations before applying classification. As a result, the node-level decision-making capability is overlooked. It is worth mentioning that node-level decisions have some relationships with the graph-level predictions [27]. Thus, we use a novel gated attention mechanism to aggregate the node-level class-specific confidence instead of node-level feature representation for the graph-level classification. First, we compute the node-level class-specific confidence using a shared single-layered perceptron (SLP) applied to each node k i.e., $\beta_k = f_k W_1 + b_1$, parameterized by a weight matrix W_1 and bias b_1, resulting in output $\beta_k \in R^{1 \times C}$ where C is the number of classes. The dimension of f_k is the same as the $f'_{r,l}$ as f_k represents a cluster of regions using soft clustering, as mentioned before. Second, for graph-level predictions, we define a soft attention mechanism that decides which node's class-specific confidences are relevant to the current graph-level task and is computed as:

$$\hat{y}_n = \text{Softmax}\left(\sum_{k=1}^{K} \beta_k \odot \sigma(W_2 f_k + b_2) \right), \tag{3}$$

where $\sigma(.)$ is the sigmoid activation and $\odot$ is a Hadamard product. The shared weight matrix W_2 and bias b_2 are the parameters of the corresponding linear transformation.

4   Experiments

We consider both fine-grained (Aircraft-100 [58], Oxford Flowers-102 [40], and Oxford-IIIT Pets-37 [41]) and generic (CIFAR-100 [33] and Caltech-256 [23]) visual classification tasks to demonstrate our method under general situations of data diversity using five benchmark datasets. It is worth mentioning that CIFAR-100 dataset consists of tiny (32×32 pixels) RGB images in compared to the rest. We compare our approach with a wide variety of strong baselines and past methods. We follow the standard train/test split described in the respective datasets, and use the conventional top-1 accuracy (%) for the evaluation. We implement our model in TensorFlow 2.0. We use Stochastic Gradient Descent (SGD) optimizer with a learning rate of 10^{-5}. The models are trained for 200 epochs with a batch size of 8 on a NVIDIA Quadro RTX 8000 GPU (48 GB memory).

Implementation details: For a fair comparison [4, 73], we use the lightweight Xception [11] as a backbone CNN, and the output is $6\times$ upsampled using a bilinear upsampler. Pre-trained ImageNet weights are used to initialize Xception for quicker convergence [24]. Regions of the fixed area but with different aspect ratios corresponding to a given hierarchical layer are generated using the region proposal algorithm in [3]. A single-layer GCN (Fig. 1b) is used for the layer-wise graph representation of a set of regions. The GCN employs $H = 3$ attention heads in (2) with a per head output feature size of 512, resulting in a final feature dimension of $512 \times 3 = 1536$. A dropout rate of 0.2 is applied to the normalized attention coefficients $\alpha_{r,r'}$ in (2). This signifies that each region in a given layer is exposed to a stochastically sampled neighborhood in every training iteration. A single-layer MLP is used in spectral clustering-based soft clustering. The number of cluster K is empirically evaluated and is found to be dataset-dependent. For a 3-layer hierarchical structure (Fig.1b), there are 52 possible regions (layer 1: 25, layer 2: 22, and layer 3: 5) generated using the region proposal in [3]. Input images are resized to 256×256, and data augmentation of random rotation and random zoom are used. Finally, random cropping is applied to select 224×224.

Aircraft [58]		Flowers [40]		Pets [41]		CIFAR-100 [33]		Caltech-256 [24]	
Method	Acc	Method	Acc	Method	Acc	Method	Acc	Method	Acc±std
S3N[15]	92.8	InAt[65]	96.4	One[64]	90.0	OCN[64]	80.1	IFK[43]	47.9±0.4
MC*[8]	92.9	SJF*[21]	97.0	FOA[81]	91.4	DML†[82]	80.3	CLM[55]	53.6±0.2
DCL[10]	93.0	OPA*[42]	97.1	NAC[48]	91.6	SD†[66]	81.5	FV[46]	57.3±0.2
GCL[51]	93.2	CL*[9]	97.2	TL*[53]	92.5	WRN40[76]	81.7	ZFN[7]	74.2±0.3
PMG[17]	93.4	PMA§[50]	97.4	InAt[65]	93.5	WRN28[88]	82.3	VGG‡[49]	86.2±0.3
CSC[56]	93.8	DST*[14]	97.6	OPA*[42]	93.8	BOT†[79]	83.5	L²SP[57]	87.9±0.2
GRD[85]	94.3	MC*[8]	97.7	GPipe*[29]	95.9	Augment[13]	89.3	VSVC[78]	91.4±0.4
CAP[4]	**94.9**	CAP[4]	97.7	CAP[4]	97.3	GPipe*[29]	**91.3**	CPM*[24]	94.3±0.2
Baseline	79.5	*Baseline*	91.9	*Baseline*	91.0	*Baseline*	80.9	*Baseline*	72.2±0.4
Ours	**94.9**	**Ours**	**98.7**	**Ours**	**98.1**	**Ours**	83.8	**Ours**	**96.2±0.1**

Table 1: Accuracy (%) comparison with the most recent top eight SotA methods. * involves transfer/joint learning strategy for objects/patches/regions involving more than one dataset (target and secondary). § uses additional textual description. † applies self-knowledge distillation. ‡ combines VGG-16 and VGG-19. The *baseline* accuracy is without our attention-driven hierarchical multi-scale representation.

4.1 Comparison with the state-of-the-art (SotA)

The performance of our model is compared with the recent eight SotA methods in Table 1. Based on the dataset characteristics, we divide our comparison into two groups:

Fine-grained image classification: We compare the recent methods that avoid object/parts bounding-box annotations as followed in our work. Our method outperforms these existing methods on Flowers (1.0%), and Pets (0.8%) datasets. It attains 94.9% accuracy on Aircraft akin to CAP [4]. For a fair comparison with CAP, our method achieves the best performance on all three fine-grained datasets using Xception as a backbone. CAP's accuracy using Xception are: 94.1% on Aircraft, 97.7% on Flowers, and 97.0% on Pets. Using the same backbone, our method gains a clear margin on Aircraft (0.8%), Flowers (1.0%), and Pets (1.1%). We also achieved the SotA results compared to the others based on the attention mechanism [4, 8, 42, 53], GCN [56, 51, 85], and low-rank discriminative bases [59]. Moreover, our approach gives superior performance over these existing methods (marked * in Table 1) without leveraging secondary datasets. For example, ImageNet is used in [8] for joint learning (Flowers: 97.7%), and pre-trained sub-networks or even higher image resolution (e.g. 448×448 in PMG [17], 480×480 in GPipe [29], MC [8], etc.) is considered in improving FGVC accuracy. Moreover, CSC (Aircraft: 93.8%) [56] uses three-modules, and its graph representation interacts with a fewer number of patches (2-4). For a higher number of regions, it fails to learn spatial contextual information resulting in lower accuracy. Likewise, DFG [59] also suffers from the same scalability problem in their graph-structure. Our model is scalable to any number of regions without increasing the model parameters since the graph node's parameters are shared. This clearly reflects our model's ability to capture subtle changes, avoiding additional resource/constraint to enhance accuracy.

Generic image classification: AutoAugment [13] used PyramidNet (26M parameters) with ShakeDrop regularization to achieve 89.3% accuracy on CIFAR-100 following a resource-intensive (4 GPUs) training. It improved to 91.3% using a giant AmoebaNet-B (557M) in GPipe [29]. Whereas, AutoAugment's accuracy degraded to 82.9% using Wide-ResNet WRN-28-10 (36.5M). The method in [76] attains 81.7% with WRN-40-10, and drops to 81.2% with WRN-28-10. Thus, the backbone CNN plays a vital role in performance gain and is susceptible to higher network depth, architecture, and capacity. It is also evaluated in BOT
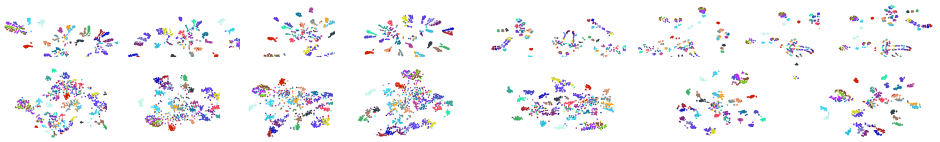
Figure 2: t-SNE visualization [51] of class-specific discriminative feature using $H = 3$ attention heads in (2), and $L = 3$ in (1). (a) top-left: Layer 1 (left to right): $head_1$, $head_2$, $head_3$, and their concatenation; (b) top-right: Layer 2: same order as in (a); (c) bottom-left: Layer 3: same order as in (a); (d) bottom-right: final representation (left to right) using 2, 3 and 4 attention heads. Attention head-specific plots are shown in (Layer 1) → (Layer 2) → (Layer 3), representing layers from smaller regions to larger ones. It is evident that the discriminability of the features representing medium-size regions (Layer 2) > small-size (Layer 1) > large-size (Layer 3), and could be linked to the problem of FGVC. (d) shows the combined layers' representation using $H = 2$, 3, and 4 attention heads.

[79], showing the lighter versions of ResNet/WRN achieve lower performance (ResNet152 (60M): 82.3%, WRN44-8: 82.6%) than their best accuracy (83.5%) using PyramidNet101-240. Contrarily, we include our novel module on the top of Xception (22.9M), attaining better performance (83.8%) with lesser computational overhead and complexity. In addition, we achieve 2.2% improvement over GPipe on both Aircraft (92.7%) and Pets (95.9%). Likewise, the gains over AutoAugment are: Aircraft 2.2%, Flowers 3.3%, and Pets 9.1%. Our method suppresses many SotA methods and attains better accuracy with a novel attention-driven hierarchical multi-scale representation using GCN, with 36.1M parameters that is 0.4M lesser than WRN-28-10. Hence, our end-to-end lightweight model performs comparatively better than the methods used heavier and deeper backbones with resource-intensive training.

We achieve 96.2% on Caltech-256 and is better (1.9%) than the recent best CPM [22] that uses additional training (secondary) data from ImageNet for selective-joint fine-tuning. Also, CPM follows a complex and multi-step training mechanism. However, its accuracy is 93.5% with only the target dataset. Moreover, our model gains 4.8% improvement over the recent VSVC [78] that combines multi-view information. Thus, it clearly reflects the efficacy of our method for generic visual recognition on top of the FGVC.

4.2   Model complexity and qualitative analysis

The complexity is given as Gigaflops (GFLOPS) and trainable parameters in millions (M). For a 3-layer architecture with Xception backbone, 3 attention-heads, and output dimension of 512, our model has 36.1M parameters and 13.2 GFLOPS. The same with an output dimension of 256, these are 29.4M and 13.2 GFLOPS. The number of regions does not impact on the parameters, but a trivial ($\sim 10^{-2}$) increase in GFLOPS. Our model is also comparable to CAP [4] (34.2M) and lighter than RAN [3] (49M). Our per-image inference time is 8.5 milliseconds (ms). It is 27ms for step 3 and an extra 227ms in step 2 of [22]. The inference time for FCANs [37] is reported as 150 ms. Additional information on complexity is included in the supplementary.

We provide qualitative analysis via visualization (t-SNE [51]) to get insight into our model. Firstly, we look into the class-specific discriminability of features representing regions using our layer-wise graph structure, and is shown in Fig. 2. All test images from 30 randomly chosen classes in Aircraft are considered. Each layer is represented as a complete graph with nodes linking regions. Three ($H = 3$) attention heads are attached to a region and output from each head is visualized as well as their concatenation. From Fig. 2, it is evident that multi-scale hierarchical representation does influence the discriminability. Secondly, we
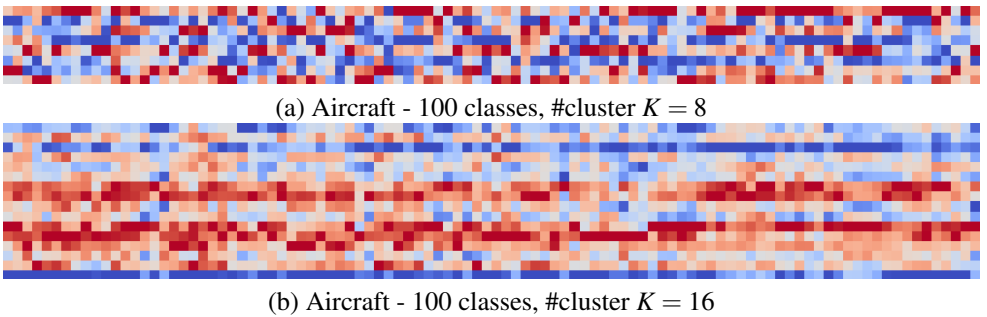
(a) Aircraft - 100 classes, #cluster $K = 8$

(b) Aircraft - 100 classes, #cluster $K = 16$

Figure 3: Visualization of the cluster-specific contributions (i.e. weights, cool to warm $\Rightarrow$ less to more) from the graph representation of regions towards a given category during the spectral clustering-based graph pooling. The y-axis represents K and the x-axis shows the number of classes. Each column is different, representing the feature discriminability during the decision-making process. All test images from the Aircraft are used to compute weights.

look into the cluster-specific "importance" (i.e., higher weights) of regions towards a class during the spectral clustering-based graph pooling, and is shown in Fig. 3. It clearly evinces the effect of a different number of clusters (K) to aggregate contributions from regions during the decision-making. For Aircraft, it is clear that the cluster compactness and separability improve with increasing K, but not much with Flowers and Pets dataset (also presented in Table 2). The explanation is given in the following ablation study.

4.3   Ablation study

We study the impact of the following key aspects using Aircraft, Flowers, and Pets datasets. **1)** Effect of K in (3) while forming a coarser (i.e., more abstract) graph representation to combine multiple graph structures via spectral clustering (Fig.1c). The impact of K on accuracy (Table 2) is dataset-dependent since the value for best accuracy on Aircraft ($K = 40$) is different from Flowers and Pets ($K = 8$). This can be linked to the dataset size and types. For example, the training size of Aircraft (6,667) is larger than the Flowers (2,040) and Pets (3,680). **2)** Performance linking number of regions ($R = \sum_{l=1}^{L} R_l$ in (1)) in our multi-scale representation and is presented in Table 3. The number of regions is controlled by the region proposal algorithm's cell-size parameter in [🔲] (Section 3.2). **3)** The impact of the number of attention heads H in (2) and per-head output dimension on accuracy is shown in Table 3. The model performs better for 3 heads with an output dimension of 512. Finally, we evaluate the influence of the number of layers L in our hierarchical representation and is provided in Table 4. A notable observation in the above ablative studies is the Aircraft's sensitivity to the values of H, L, and regions $R = \{R_l\}$. For example, accuracy drops from 94.9% to 88.1% when $|R|$ increases from 52 to 62, and it climbs from 83.7% to 94.9% for $H = 2$ to $H = 3$ (Table 3). Similarly, the accuracy increases from 88.7% to 94.9% for $L = 2$ to $L = 3$ (Table 4). Moreover, the baseline model (without our attention-driven hierarchical description) accuracy (79.5%) is significantly low in comparison to Flowers (91.9%) and Pets (91%) FGVC datasets (Table 1). This suggests that the Aircraft dataset is a difficult one and the sensitiveness could be due to the aeroplane shapes (higher aspect ratio) and the presence of significant perspective distortions in images (e.g. plane on ground, in mid-air, and landing/take off). As a result, the model struggles to find the optimal H, L, and R parameters.

Supplementary: It contains: 1) dataset description in Table 5; and 2) further results of Table 3 comparing *concatenation with averaging* the outputs from multi-head attention is in Table

#Cluster (K)	2	4	8	16	20	32	36	40	48
Aircraft	-	-	91.6	93.6	94.1	94.6	94.6	**94.9**	94.2
Flowers	95.8	98.3	**98.7**	**98.7**	98.5	98.6	-	-	-
Pets	97.2	97.5	**98.1**	97.7	97.5	97.5	-	-	-

Table 2: Accuracy (%) with varying number of clusters (K).

Dataset	Regions					Head (ch 256)			Head (ch 512)		
	32	43	52	62	74	2	3	4	2	3	4
Aircraft	93.7	94.3	**94.9**	88.1	85.5	94.3	94.3	94.4	83.7	**94.9**	94.2
Flowers	97.2	98.0	**98.7**	96.8	94.9	98.4	98.5	98.2	98.5	**98.7**	98.4
Pets	97.7	97.8	**98.1**	97.6	97.4	97.3	97.3	97.6	97.3	**98.1**	97.8

Table 3: Accuracy (%) with varying number of regions in our multi-scale hierarchical structure and varying number attention heads and output channels per attention head.

#Layers L	Aircraft	Flowers	Pets	CIFAR-100	Trainable Parameters	GFLOPs
1	77.4	89.4	93.5	81.7	27,493,316	11.849
2	88.7	98.0	97.5	82.9	31,691,204	13.213
3	**94.9**	**98.7**	**98.1**	**83.8**	**36,082,880**	**13.215**
4	91.9	94.2	93.8	83.5	40,086,980	13.671
5	91.4	93.8	92.0	82.1	44,284,868	13.673

Table 4: Influence of the number of layers L in (1) in our hierarchical representation. We use the optimal number of attention heads (H=3), and its output dimension is 512. It is evident that the accuracy increases with the number of layers L, but after L=3, it starts decreasing. This trend is in all 4 datasets. The trainable parameters and GFLOPs increase with L as well.

6. It also includes a detailed study on complexity analysis in Table 7, top-N accuracy (%) in Table 8, example of the regions linking various layers to visualize hierarchy in Fig.4-5, cluster-specific contributions of the graph-based regions in Fig.6-8, and t-SNE analysis of layer-wise attention heads are shown in Fig.9-12 with better clarity.

Further Link: A notable characteristic of our model from some recent works [4, 26, 47, 58] is that it does not explicitly localize or search for salient regions. This, in fact, is a forte of our model; making it capable of attending to all possible regions and model their short- and long-range dependencies, since this is possibly the best way to overcome a high intra-class and low inter-class variances in FGVC due to occlusions, deformation, and illuminations. We would like to emphasize that graph-based attention-driven relation-aware expressive representation gained improvement in person re-identification [84], visual question answering [85, 89], social relationship understanding [60], and human-object interaction [44].

5   Conclusion

To address the problem of long-range dependencies to capture the subtle changes, we have proposed an end-to-end deep model by exploring Graph Convolutional Networks to represent regions as a set of complete graph structures. These regions are multi-scale and arranged hierarchically, consisting of smaller (closer look) to larger (far look) size. To emphasize regions, the dependencies among regions are modeled using attention-driven message propagation that explores the finer to coarser graph structures. Our model's performance demonstrates its effectiveness in advancing both FGVC and generic visual classification.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1371–1380, 2020.

[3] Ardhendu Behera, Zachary Wharton, Yonghuai Liu, Morteza Ghahremani, Swagat Kumar, and Nik Bessis. Regional attention network (ran) for head pose and fine-grained gesture recognition. *IEEE Transactions on Affective Computing*, 2020.

[4] Ardhendu Behera, Zachary Wharton, Pradeep Hewage, and Asish Bera. Context-aware attentional pooling (cap) for fine-grained visual classification. In *35th AAAI Conference on Artificial Intelligence*, pages 929–937, 2021.

[5] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. Attend and guide (ag-net): A keypoints-driven attention-based deep network for image recognition. *IEEE Transactions on Image Processing*, 30:3691–3704, 2021.

[6] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International Conference on Machine Learning*, pages 874–883. PMLR, 2020.

[7] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 511–520, 2017.

[8] Dongliang Chang, Yifeng Ding, Jiyang Xie, Ayan Kumar Bhunia, Xiaoxu Li, Zhanyu Ma, Ming Wu, Jun Guo, and Yi-Zhe Song. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29: 4683–4695, 2020.

[9] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 2023–2031, 2018.

[10] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[12] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[13] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[14] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4109–4118, 2018.

[15] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6599–6608, 2019.

[16] Yifeng Ding, Zhanyu Ma, Shaoguo Wen, Jiyang Xie, Dongliang Chang, Zhongwei Si, Ming Wu, and Haibin Ling. Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Transactions on Image Processing*, 30:2826–2836, 2021.

[17] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Yi-Zhe Song, Zhanyu Ma, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. *arXiv preprint arXiv:2003.03836*, 2020.

[18] Amir Erfan Eshratifar, David Eigen, Michael Gormish, and Massoud Pedram. Coarse2fine: A two-stage training method for fine-grained visual classification. *arXiv preprint arXiv:1909.02680*, 2019.

[19] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4438–4446, 2017.

[20] Hongyang Gao and Shuiwang Ji. Graph u-nets. In *International Conference on Machine Learning*, pages 2083–2092. PMLR, 2019.

[21] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1086–1095, 2017.

[22] Weifeng Ge, Xiangru Lin, and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3034–3043, 2019.

[23] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[24] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4918–4927, 2019.

[25] Xiangteng He, Yuxin Peng, and Junjie Zhao. Fine-grained discriminative localization via saliency-guided faster r-cnn. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 627–635, 2017.

[26] Xiangteng He, Yuxin Peng, and Junjie Zhao. Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization. *International Journal of Computer Vision*, 127(9):1235–1255, 2019.

[27] Chester Holtz, Onur Atan, Ryan Carey, and Tushit Jain. Multi-task learning on graphs with node and graph level labels. *SAGE*, 4(76.84):62–4, 2019.

[28] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.

[29] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems*, pages 103–112, 2019.

[30] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, 2020.

[31] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*.

[32] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[33] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[34] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International Conference on Machine Learning*, pages 3734–3743. PMLR, 2019.

[35] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10313–10322, 2019.

[36] Chuanbin Liu, Hongtao Xie, Zheng-Jun Zha, Lingyun Yu, Zhineng Chen, and Yongdong Zhang. Bidirectional attention-recognition model for fine-grained object classification. *IEEE Transactions on Multimedia*, 2019.

[37] Xiao Liu, Tian Xia, Jiang Wang, Yi Yang, Feng Zhou, and Yuanqing Lin. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*, 2016.

[38] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[39] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G Schwing. Out of the box: reasoning with graph convolution nets for factual visual question answering. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2659–2670, 2018.

[40] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[41] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[42] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, 2018.

[43] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156. Springer, 2010.

[44] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018.

[45] Pau Rodríguez, Diego Velazquez, Guillem Cucurull, Josep M Gonfaus, F Xavier Roca, and Jordi Gonzàlez. Pay attention to the activations: a modular attention mechanism for fine-grained image recognition. *IEEE Transactions on Multimedia*, 22(2):502–514, 2020.

[46] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[47] Prateek Shroff, Tianlong Chen, Yunchao Wei, and Zhangyang Wang. Focus longer to see better: Recursively refined attention for fine-grained image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 868–869, 2020.

[48] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1143–1151, 2015.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[50] Kaitao Song, Xiu-Shen Wei, Xiangbo Shu, Ren-Jie Song, and Jianfeng Lu. Bi-modal progressive mask attention for fine-grained recognition. *IEEE Transactions on Image Processing*, 2020.

[51] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[53] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representation (ICLR)*, 2018.

[54] Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11505–11515, 2020.

[55] Qilong Wang, Peihua Li, Lei Zhang, and Wangmeng Zuo. Towards effective codebookless model for image classification. *Pattern Recognition*, 59:63–71, 2016.

[56] Shijie Wang, Zhihui Wang, Haojie Li, and Wanli Ouyang. Category-specific semantic coherency learning for fine-grained image recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 174–183, 2020.

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[58] Zhengyang Wang and Shuiwang Ji. Second-order pooling for graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[59] Zhihui Wang, Shijie Wang, Shuhui Yang, Haojie Li, Jianjun Li, and Zezhou Li. Weakly supervised fine-grained image classification via guassian mixture model oriented discriminative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9749–9758, 2020.

[60] Zhouxia Wang, Tianshui Chen, Jimmy SJ Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *IJCAI*, 2018.

[61] Zhuhui Wang, Shijie Wang, Haojie Li, Zhi Dou, and Jianjun Li. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *AAAI*, pages 12289–12296, 2020.

[62] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[63] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.

[64] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 3–10, 2015.

[65] Lingxi Xie, Liang Zheng, Jingdong Wang, Alan L Yuille, and Qi Tian. Interactive: Inter-layer activeness propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2016.

[66] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5565–5572, 2019.

[67] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *International Conference on Machine Learning*, pages 2825–2834, 2018.

[68] Shaokang Yang, Shuai Liu, Cheng Yang, and Changhu Wang. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. *arXiv preprint arXiv:2102.09875*, 2021.

[69] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.

[70] Hantao Yao, Shiliang Zhang, Chenggang Yan, Yongdong Zhang, Jintao Li, and Qi Tian. Autobd: Automated bi-level description for scalable fine-grained visual categorization. *IEEE Transactions on Image Processing*, 27(1):10–23, 2017.

[71] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983, 2018.

[72] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *arXiv preprint arXiv:1806.08804*, 2018.

[73] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, 2018.

[74] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *European Conference on Computer Vision*, pages 574–589, 2018.

[75] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13876–13885, 2020.

[76] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[77] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[78] Chunjie Zhang, Jian Cheng, and Qi Tian. Multi-view image classification with visual, semantic and view consistency. *IEEE Transactions on Image Processing*, 29:617–627, 2019.

[79] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3713–3722, 2019.

[80] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[81] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, and Qi Tian. Fused one-vs-all features with semantic alignments for fine-grained visual categorization. *IEEE Transactions on Image Processing*, 25(2):878–892, 2015.

[82] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

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

[84] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020.

[85] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15079–15088, 2021.

[86] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019.

[87] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo, and Tao Mei. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. *IEEE Transactions on Image Processing*, 29:476–488, 2020.

[88] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020.