

# SR-GNN: Spatial Relation-aware Graph Neural Network for Fine-Grained Image Categorization

Asish Bera\*, Member, IEEE, Zachary Wharton\*, Yonghuai Liu, Senior Member, IEEE,  
Nik Bessis, Senior Member, IEEE, and Ardhendu Behera†, Member, IEEE

**Abstract**—Over the past few years, a significant progress has been made in deep convolutional neural networks (CNNs)-based image recognition. This is mainly due to the strong ability of such networks in mining discriminative object pose and parts information from texture and shape. This is often inappropriate for fine-grained visual classification (FGVC) since it exhibits high intra-class and low inter-class variances due to occlusions, deformation, illuminations, etc. Thus, an expressive feature representation describing global structural information is a key to characterize an object/ scene. To this end, we propose a method that effectively captures subtle changes by aggregating context-aware features from most relevant image-regions and their importance in discriminating fine-grained categories avoiding the bounding-box and/or distinguishable part annotations. Our approach is inspired by the recent advancement in self-attention and graph neural networks (GNNs) approaches to include a simple yet effective relation-aware feature transformation and its refinement using a context-aware attention mechanism to boost the discriminability of the transformed feature in an end-to-end learning process. Our model is evaluated on eight benchmark datasets consisting of fine-grained objects and human-object interactions. It outperforms the state-of-the-art approaches by a significant margin in recognition accuracy.

**Index Terms**—Attention mechanism, Convolutional Neural Networks, Graph Neural Networks, Human action, Fine-grained visual recognition, Relation-aware feature transformation.

## I. INTRODUCTION

THE advent of deep convolutional neural networks (CNN) has significantly enhanced image recognition performance in the past decade. It is achieved mainly due to their abilities to provide a high-level description (*e.g.*, global shape and appearance) of image content by capturing discriminative object-pose and -parts information from texture and shape. This high-level description is more apposite for the large-scale visual classification (LSVC) tasks consisting of distinctive categories (*e.g.*, ImageNet and COCO datasets). However, their performance in solving fine-grained visual classification (FGVC) problems is not at the same level as in LSVC. This is mainly due to the subtle changes between hard-to-distinguish object classes in FGVC, but most often visually measurable by humans. Common datasets in FGVC include different types of birds [1], flowers [2], dogs [3], aircraft [4], car models [5], etc. A typical observation in FGVC is that objects from different classes share visually similar structures (large inter-class similarities), and objects in the same class often exhibit significant variations due to different

A. Bera, A. Behera, Z. Wharton, Y. Liu, and N. Bessis are with the Department of Computer Science, Edge Hill University, UK.

\* Equal contribution, † Corresponding author, beheraa@edgehill.ac.uk  
Manuscript received Month 08, 2021; revised Month 02 and 04, 2022.

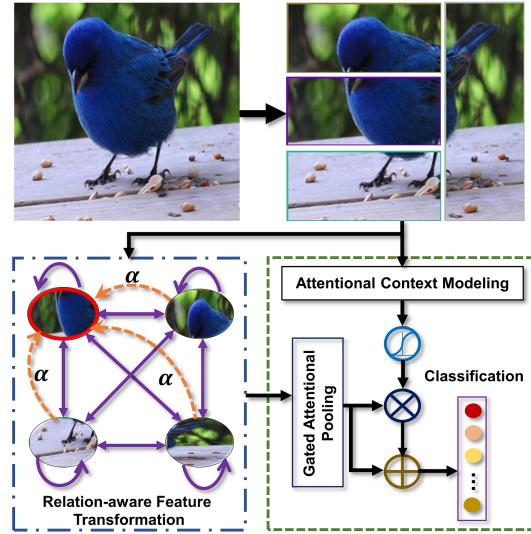


Fig. 1. Our SR-GNN consists of a GNN-based relation-aware feature transformation by propagating information between image regions and an attentional context modeling to refine these transformed features. They jointly tackle the challenge of describing and discriminating subtle variations in FGVC by exploring the visual-spatial relationships among regions and aggregating the context-aware features. For clarity, 4 different regions are shown here.

structures, lighting, clutter and viewpoints (large intra-class variations). As a result, it is a challenging task to learn a unified and discriminative representation for each class. A key step to address this challenge is to extract discriminative features from vital object-parts and combine them for the representation of a consistent distinctive global structure of a given class. The current state-of-the-art (SotA) approaches are mainly craftily designed to extract such discriminative features and structures by exploring 1) part annotations from humans, and 2) automatically finding these discriminative parts from the whole image. We refer the interested readers to [6] for a detailed survey. Most of the earlier works belong to the first category in which the locations of discriminative object-parts are given (*e.g.*, bounding box or mask). Some methods learn part-based detectors, and some leverage semantic segmentation to localize distinct parts. The parts annotation is a cumbersome and expensive human labeling task that is often prone to human errors and requires expert knowledge. Moreover, part-based methods limit both scalability and practicality of real-world FGVC applications. Thus, many recent methods have used image-level labels to guide their models in identifying the key object parts to discriminate the sub-categories by exploring attention mechanisms in the image space or feature space [7]–[10] to automatically mine discriminative features.

**Motivation:** In this work, we propose a simple yet effective connection between the image space and feature space to discriminate subtle variances in FGVC. Our approach is motivated by the recent success of Graph Neural Networks (GNN) [11] and attention mechanisms [12], [13] in deep learning. Many SotA methods for FGVC use a pre-trained object/part detector (or proposal from mask R-CNN) in a weakly supervised manner, resulting in the absence of detailed description, which is indispensable to capture better object-part and part-part relationships to model the subtle changes. These parts can be affected by occlusions, noisy backgrounds, pose variations and ambiguous repetitive patterns. Thus, multiple partial descriptors for a part are potentially useful in disambiguating and discriminating subtle changes since the model learns meaningful complementary information to provide a rich representation of an object. Our method neither considers object-part proposal/bounding-box nor tries to localize them. Instead, it automatically learns a richer representation of an object by exploring the attention-driven visual-spatial relationships among a pool of geometrically-constrained regions. These regions are generated using the region proposal in the Regional Attention Network (RAN) [8], which uses cells and blocks in computing histograms of oriented gradients (HOG).

To describe a richer representation with the discrimination power for subtle variations, we design a spatial relation-aware GNN (SR-GNN) to model visual-spatial relations between regions. These relationships are captured using a novel relation-aware feature transformation and its refinement via attentional context modeling, as conceptually shown in Fig. 1. Firstly, a backbone CNN is used to extract high-level visual features. These are upsampled for feature pooling using geometrically-constrained regions of various sizes and positions (Fig. 2(a)). Secondly, it transforms these pooled features using relation-aware feature transformation leveraging GNN that captures the visual-spatial relationships via propagating information between regions represented as the nodes of a connected graph (Fig. 2(b)) to enhance the discriminative power of features. To address the limitations of over-smoothing and a large number of learnable parameters in GNN [11], it adapts the topic sensitive PageRank [14] using the approximate personalized propagation of neural predictions (APPNP) message-passing algorithm via power iteration, achieving linear computational complexity. Then it applies a novel gated attentional pooling to the learned graph nodes for final feature representation. Finally, it employs an attentional context modeling (Fig. 2(c)) that explores self-attention [12] and weighted attention [13] in an innovative way to learn a weight vector, which is multiplied with the final relation-aware transformed feature extracted in the previous step as a refined feature before classification.

Performance-wise, CAP [7] is currently top of the leaderboard for many FGVC datasets. It uses attention to accumulate features from integral regions in a context-aware fashion. Then it uses an LSTM to learn the spatial arrangements (context encoding) of these integral regions for subtle discrimination to tackle FGVC tasks. Finally, it aggregates information by grouping similar responses of the LSTM’s hidden states to generate locally aggregated descriptors using NetVLAD in the classification step. Our proposed SR-GNN is significantly

different from CAP in the following aspects: (a) introduction of a relation-aware spatial graph with the APPNP message-passing algorithm to extract more expressive features by capturing visual-spatial relationships via propagating information between regions, (b) a graph-based gated attentional pooling to aggregate features from graph nodes, and (c) an attentional context modeling that consists of self-attention and weighted attention to compute a weight vector for the refinement of the final relation-aware features for classification. The only similarity in both approaches is the feature extraction using a CNN backbone. Although the context-aware attention in CAP and self-attention in SR-GNN (Section III-D) are inspired by the same self-attention mechanism in natural language processing [12], they are explored differently to solve the specific problem in hand. In CAP, it accumulates features from various regions and an LSTM is then applied for context encoding by considering sequential information. Whereas in SR-GNN, it is investigated in a novel way to compute a weight vector (Fig. 2(c)) by exploring contextual information via adapting self-attention [12] and weighted attention [13]. Unlike in CAP, the weighted attention does not consider sequential information, but learns the weight vector from multiple regions by joint learning. To the best of our knowledge, we are the first to investigate the efficiency of the PageRank algorithm leveraging APPNP to advance the FGVC accuracy. These key concepts are also novel in comparison to other SotA methods, including more recent vision Transformers (ViT) [15]–[18].

The main contributions of this paper are: 1) A novel relation-aware visual representation and its refinement via attentional spatial context for enriching region-level description to capture the subtle changes and eventually enhance the FGVC performance; 2) An easy-to-use end-to-end FGVC deep network that does not require object/parts bounding boxes annotation or proposal and thus has an advantage of easy implementation; 3) A proposal of a gated attentional pooling for the automatic aggregation of the relation-aware features; and 4) Ablation studies and visual analysis of the performance of SR-GNN.

The rest of this paper is organized as follows: Section II summarizes related works on FGVC. Section III describes the proposed framework. The experimental results are discussed in Section IV, and an in-depth ablation study is presented in Section V, followed by a conclusion in Section VI.

## II. RELATED WORK

Our work is closely related to weakly-supervised object-parts, attentional and GNN methods for FGVC, including human actions. We present a concise survey of these approaches.

### A. Object-Parts Based Methods

Informative object-parts are crucial and are explored [19]–[21] for robust subtle discrimination. Distinct object-parts are selected at multiple scales from object proposals in [19] to distinguish subtle variations. In [20], the objectness map is generated using deep features for part-level and object-level descriptions and their fusion for visual discrimination. Object detection and instance segmentation pipeline are iterated in

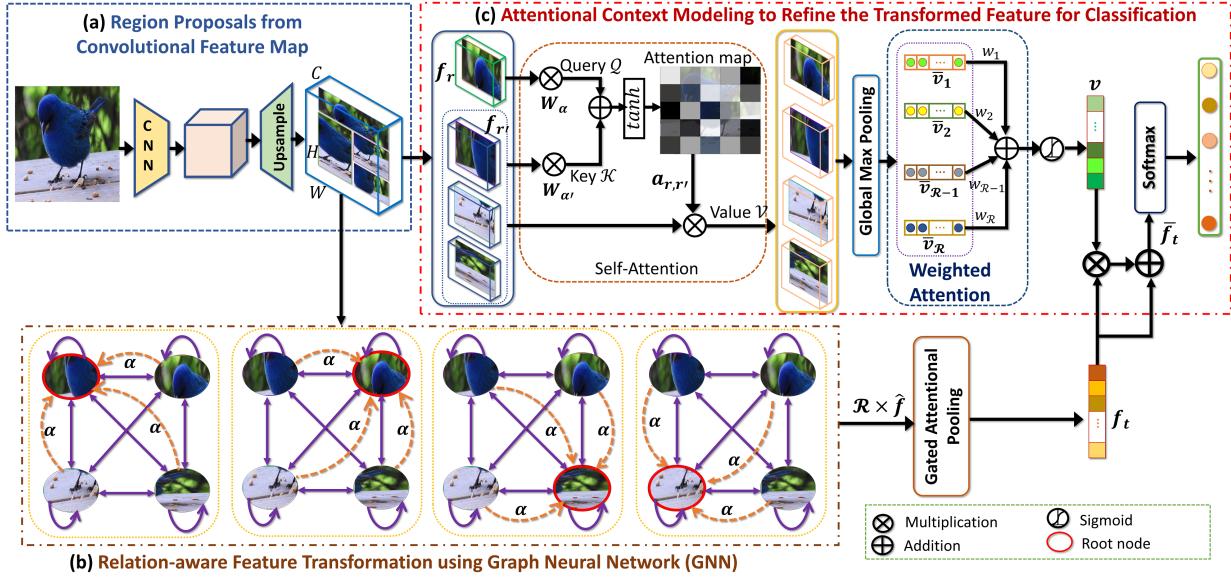


Fig. 2. The architecture of our SR-GNN. (a) Extract features using a set of regions from the upsampled CNN features of a given input image. (b) **Relation-aware Feature Transformation:** it updates each region's visual-spatial relationships by propagating information between them using message propagation in GNN. Then, the transformed features from all regions are used by the gated attentional pooling to produce the final transformed features  $f_t$ . (c) **Attentional Context Refinement:** firstly, it computes an attention-focused context vector  $v_r$  from region-pooled features using self-attention. Next, a context refinement weight-vector  $v$  is computed over the weighted summation of all  $v_r$ . Finally, the context vector  $v$  is used to refine the transformed feature  $f_t$  and then feeds it to the Softmax layer for classification.

[21] for complementary part localization, and then LSTM is used to encode contextual details. Similarly, local details are learned from distinct patches which are generated by shuffling the whole image into smaller patches [22]–[24]. The global image structure is randomly disrupted by a region confusion technique in [22] to learn finer details and semantic correlation within sub-regions. Whereas random erasing in [23] introduces additional noise by object-part occlusion to select informative patches. A region grouping sub-network in [24] learns correlation weight coefficients between regions to select and refine discriminatory patches. Similarly, vision and language modalities are combined in [25]. The vision localizes objects using saliency and co-segmentation, while the language applies cross-modal analysis to correlate natural language descriptions and discriminative object parts. A multi-scale and multi-granularity deep reinforcement learning in [26] finds hierarchical discriminative regions in multiple granularities and automatically determines the number of such regions to boost the accuracy. Likewise, a hierarchical representation of image-regions enhances action classification accuracy in [27]. Most of these approaches focus on locating the informative object-parts and then extract expressive feature descriptors. In sharp contrast, our method learns distinct features by mining visual-spatial correlations using contextual cues from a set of regions, relying on cells and blocks used in HOG [8].

### B. Attention-based Approaches

The attention mechanism is proliferated to identify salient regions and/or subtle discriminatory features to attain superior performance [7]–[9], [28], [29]. A trilinear attention sampling in [29] learns features from hundreds of part proposals and then applies knowledge distillation to integrate them. Top-down and bottom-up attentions are combined in an attentional

pyramid CNN [30] to aggregate high-level semantic and low-level finer features. In [9], a feedback path is connected from a recognition agent to an attention agent to optimize region proposals. Regional attention network (RAN) [8] presents a hybrid attention method that focuses on semantic informativeness from multiple regional contexts for fine-grained gesture/action recognition. Attend and guide network (AG-Net) [10] applies to scale-invariant feature transform (SIFT) keypoints and Gaussian mixture model to propose regions that are guided by the attention mechanism for fine-grained visual categorization of objects and human actions. Modular attention in [28] applies multiple attention modules to focus on region-based predictions refined by attention gates. Attention on feature channels is explored in [31] to focus on discriminatory regions. A sparse attentional framework in [32] follows a selective sampling technique to estimate finer details. A counterfactual attention learning is proposed in [33] to measure the visual attention quality that guides the learning process via counterfactual intervention to learn more useful attention for enhanced FGVC accuracy. Similarly, object extent learning and spatial context learning are integrated in look-into-object [34] to understand the object structure by automatically modeling the context information among regions. In [35], attentive pairwise interaction network discovers contrastive cues from a pair of images, and discriminates them with pairwise attentional interaction in an end-to-end manner. More recently, a sequence of image patches with positional embedding and multi-head self-attention are integrated in vision Transformers [15]–[18] to enhance FGVC. Swin Transformer [16] exploits a hierarchical shifting window-based self-attention with linear computational complexity. A part selection module is adapted to improve over pure ViT in [17] by integrating raw attention weights of the Transformer into an attention map to guide

the ViT. A complementary attention module and multi-feature fusion module are combined in [18] using Swin Transformer. Inspired by these, we propose a simple yet effective attention mechanism to refine the GNN-driven features at multi-scale and their aggregation for further performance improvement.

### C. Graph Neural Networks (GNN)

Following the CNN concept, GNN is proposed to explore problems consisting of non-Euclidean data. It is powerful for smoother messages passing between neighboring nodes to enhance performance [11]. Recently, it has been explored in zero-shot recognition, multi-label image recognition, image captioning, visual question answering, and the others [36]. However, its efficacy in FGVC is yet to be fully explored. In [37], GNN is used to learn latent attributes by modeling semantic correspondence between discriminative regions within the same sub-category. In [38], region correlation is explored to discover informative regions using the criss-cross graph propagation sub-network and correlation feature via a unified framework. However, both methods are limited to a few regions per image (*e.g.*, 4) which may be sub-optimal for building and propagating information within sub-networks for effective context modeling to address FGVC. A graph-based relation discovery (GaRD) method [39] learns the positional and semantic feature relationships and adopts a feature grouping strategy to tackle FGVC. We propose a GNN-based spatial relation-aware feature aggregation by considering multiple partial descriptors to propagate and capture finer complementary information between neighboring regions by approximating topic-sensitive PageRank [14].

## III. PROPOSED APPROACH

The proposed SR-GNN architecture is shown in Fig. 2. It takes as input an input image, extracts a high-level convolutional feature map, applies region-based visual-spatial feature selection and refines the transformed features using an attentional spatial context modeling to advance FGVC.

### A. Problem Formulation

To train an image classifier, a set of  $N$  images  $I = \{I_n | n = 1, 2, \dots, N\}$  and their respective class labels are given. The classifier learns a mapping function  $\mathcal{F}$  that predicts  $\hat{y}_n = \mathcal{F}(I_n)$ , which matches the true label  $y_n$ . During training, it learns  $\mathcal{F}$  by minimizing a loss  $\mathcal{L}(y_n, \hat{y}_n)$  between the true and the predicted labels. In this work,  $\mathcal{F}$  is an end-to-end deep network in which we introduce a simple yet effective network modification to advance the SotA in FGVC. The mechanism focuses on two main components to capture the fine-grained changes in images: 1) relation-aware feature selection and transformation, and 2) an attentional context modeling to refine these transformed features. Therefore, the mapping function  $\mathcal{F}$  consists of:

$$\mathcal{F} = \text{Softmax} \left( \underbrace{\mathcal{F}_1(I_n; \theta_t)}_{\text{Feature Transform}} \otimes \overbrace{\sigma(\mathcal{F}_2(I_n; \theta_c))}^{\text{Attentional Context}} \right), \quad (1)$$

where  $\theta_t$  and  $\theta_c$  are the learnable parameter sets for the feature transformation from the given image  $I_n$  to a high-level

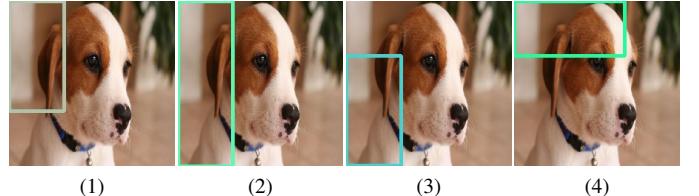


Fig. 3. Various regions (4 shown for clarity) from the region proposal. These regions are used for bilinear pooling from the upsampled CNN features.

descriptor and the attentional context refinement, respectively.  $\sigma(\cdot)$  is an element-wise sigmoid function to regulate how much of the transformed feature should be considered in decision making.

### B. CNN Feature Map and Region Proposals

We use the lightweight Xception [40] backbone for extracting CNN features like CAP [7] and upsample it for region proposals as in [8] by exploring cells and blocks in the HOG computation. The region proposal generates  $\mathcal{R}$  possible regions of different aspect ratios and areas. For clarity, 4 regions are shown in Fig. 7. Each region is then represented with a feature vector  $f$  of dimension  $w(\text{width}) \times h(\text{height}) \times C(\text{channels})$  via bilinear interpolation to implement differentiable image transformations (conceptualized in Fig. 2(a)).

### C. Relation-Aware Feature Transformation

We represent an image  $I_n$  using  $\mathcal{R}$  regions. The essential aim is to update each region's visual representation  $f$  by propagating information between regions to characterize their visual-spatial relationships, which capture the subtle variations between them. Thus, the first step to representing these relationships is to build a graph  $G = (\mathcal{R}, E)$  with nodes  $\mathcal{R}$  and the connections between them via edges  $E$ . As a result, GNN can then be used to learn and reason visual-spatial relationships by propagating messages from one region to its connected neighbors in the graph. The nodes are described by a set  $\mathbf{X} = \{f\}$  consisting of a number  $\mathcal{R}$  of input features  $f$ , and the respective output  $\mathbf{Y} = \{\hat{f}\}$  with transformed feature  $\hat{f}$  per node. The graph  $G$  is described by the adjacency matrix  $\mathbf{A} \in \mathbb{R}^{\mathcal{R} \times \mathcal{R}}$ . The adjacency matrix  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_{\mathcal{R}}$  denotes  $\mathbf{A}$  with added self-loops and  $\mathbf{I}_{\mathcal{R}}$  is the identity matrix (Fig. 2(b)). A well-known message passing algorithm is the GNN [11] in which a simple layer-wise propagation rule is used:  $\mathbf{H}^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)})$ , with  $\mathbf{H}^{(0)} = \mathbf{X}$  and  $\mathbf{H}^{(L)} = \mathbf{Y}$ ,  $l = 0, 1, \dots, L - 1$  being the number of layers,  $\mathbf{W}^{(l)}$  is a weight matrix for the  $l$ -th layer and  $\sigma(\cdot)$  is a non-linear activation function (*e.g.*, ReLU).  $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$  is the symmetrically normalized adjacency matrix, and  $\tilde{\mathbf{D}}$  is the diagonal node degree matrix of  $\tilde{\mathbf{A}}$ . The GNN message passing algorithm is limited to a smaller neighborhood mainly due to 1) aggregation by averaging causes over-smoothing if too many layers are used and thus, loses its focus on the local neighborhood; and 2) a larger neighborhood significantly increases the depth and number of learnable parameters since the common aggregation schemes use learnable weight matrices in each layer. To address these, we adapt the approximate personalized propagation of neural predictions (APPNP) message-passing algorithm [14]. It achieves the linear computational

complexity by approximating topic-sensitive PageRank via power iteration, relating to a random walk with restarts. Each power iteration step is calculated as:

$$\begin{aligned} \mathbf{H} &= \text{MLP}(\text{GAP}(\mathbf{X}); \theta), \\ \mathbf{Y}^{(0)} &= \mathbf{H}, \\ \mathbf{Y}^{(k+1)} &= (1 - \alpha) \hat{\mathbf{A}} \mathbf{Y}^{(k)} + \alpha \mathbf{H}, \\ \mathbf{Y}^{(K)} &= \text{Sigmoid} \left( (1 - \alpha) \hat{\mathbf{A}} \mathbf{Y}^{(K-1)} + \alpha \mathbf{H} \right), \end{aligned} \quad (2)$$

where *global average pooling* (GAP) at each node reduces the dimension of  $f: w \times h \times C \rightarrow 1 \times 1 \times C$ ;  $\alpha \in (0, 1]$  is the teleport (or restart) probability influencing the size of the neighborhood for each node; and  $K$  is the number of power iteration steps ( $k \in [0, K-2]$ ). MLP is a multi-layer perceptron with a parameter  $\theta$  for predicting  $\mathbf{H}$  that allows preserving the node's local neighborhood, and acts as both the starting vector and the teleport set. For example, every column of  $\mathbf{H}$  defines a distribution over regions that acts as a teleport set. Note that MLP operates on each node's feature  $f$  independently, allowing for parallelization.

Each node transforms a region into a feature vector  $\hat{f}$ . Now the aim is to aggregate all nodes' features (*i.e.*,  $\mathcal{R} \times \hat{f}$ ) into a single image-level descriptor  $f_t$ . We achieve this by adapting the gated attentional pooling [41] that is computed as:

$$f_t = \sum_{i=1}^{\mathcal{R}} \sigma(\hat{f}_i \mathbf{W}_1 + \mathbf{b}_1) \odot (\hat{f}_i \mathbf{W}_2 + \mathbf{b}_2), \quad (3)$$

where weight matrices  $\mathbf{W}_1$  and  $\mathbf{W}_2$ , and biases  $\mathbf{b}_1$  and  $\mathbf{b}_2$  are learnable parameters.  $\hat{f}_i \in \mathbf{Y}$  represents the  $i^{th}$  node output feature of the graph  $G$  in (2).  $\sigma(\cdot)$  is an element-wise sigmoid and acts as a soft attention mechanism that decides which regions are more relevant to the current graph-level task, and  $\odot$  is the Hadamard product. The learnable feature transformation parameter in (1) is thus  $\theta_t = \{\theta, \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$ .

#### D. Transformed Feature Refinement

It is inspired by the self-attention mechanism in natural language processing [12]. The self-attention handles a long path-length contextual modeling by a lightweight gating mechanism in which the attention matrix is generated using a simple dot-product. In self-attention, *query*  $\mathcal{Q}$ , *key*  $\mathcal{K}$  and *value*  $\mathcal{V}$  are learned from the same input, and are different from the traditional attention-based sequence-to-sequence models. Often,  $\mathcal{Q}$ ,  $\mathcal{K}$  and  $\mathcal{V}$  are learned by three independent transformation layers. The dot product of  $\mathcal{Q}$  and  $\mathcal{K}$  results in the attention weight matrix, which is multiplied with  $\mathcal{V}$  to produce the desired transformed feature representation. We adapt this principle to compute attention within a given region  $r$  (self-loop) as well as between other regions  $r$  and  $r'$  ( $r, r' \in \mathcal{R}$  and  $r' \neq r$ ). The aim is to generate an attention-focused context vector (*i.e.*, *value*  $\mathcal{V}$ ) that enables our model to selectively focus on more relevant regions to generate holistic context information. Thus, in our self-attention although  $\mathcal{Q}$ ,  $\mathcal{K}$  and  $\mathcal{V}$  vectors are learned from the same input image but focus on different regions *i.e.*,  $\mathcal{Q}$  is learned from  $r$  whereas,  $\mathcal{K}$  and  $\mathcal{V}$  are learned from  $r'$ . Let  $f_r$  and  $f_{r'}$  be the high-level convolutional features representing the regions  $r$  and  $r'$ ,

TABLE I  
STATISTICS OF THE DATASETS USED IN OUR EXPERIMENTS. ACCURACY (%) OF THE BEST SOTA AND OUR SR-GNN.

| Dataset     | #Train / #Test  | #Class | SotA      | SR-GNN      |
|-------------|-----------------|--------|-----------|-------------|
| Aircraft    | 6,667 / 3,333   | 100    | 94.9 [7]  | <b>95.4</b> |
| CUB-200     | 5,994 / 5,794   | 200    | 91.8 [7]  | <b>91.9</b> |
| Cars        | 8,144 / 8,041   | 196    | 95.7 [7]  | <b>96.1</b> |
| Dogs        | 12,000 / 8,580  | 120    | 97.1 [21] | <b>97.3</b> |
| Flowers     | 2,040 / 6,149   | 102    | 97.7 [31] | <b>97.9</b> |
| NABirds     | 23,929 / 24,633 | 555    | 91.0 [7]  | <b>91.2</b> |
| Stanford-40 | 4,000 / 5,532   | 40     | 97.8 [10] | <b>98.8</b> |
| PPMI-24     | 2,110 / 2,099   | 24     | 98.6 [8]  | <b>98.9</b> |

respectively. The attention-focused context vector  $\mathbf{v}_r \in \mathcal{V}$  for region  $r$  is computed as (Fig. 2(c)):

$$\begin{aligned} \mathbf{v}_r &= \sum_{r'=1}^{\mathcal{R}} a_{r,r'} f_{r'}, \quad a_{r,r'} = \text{Softmax}(\mathbf{W}_a \alpha_{r,r'} + \mathbf{b}_a) \\ \alpha_{r,r'} &= \tanh(\overbrace{\mathbf{W}_\alpha f_r}^{\text{query } \mathcal{Q}} + \overbrace{\mathbf{W}_{\alpha'} f_{r'}}^{\text{key } \mathcal{K}} + \mathbf{b}_\alpha), \end{aligned} \quad (4)$$

where  $\mathbf{W}_\alpha$  and  $\mathbf{W}_{\alpha'}$  are weight matrices for computing  $\mathcal{Q}$  and  $\mathcal{K}$  from the respective regions  $r$  and  $r'$ ;  $\mathbf{W}_a$  is their nonlinear combination;  $\mathbf{b}_a$  and  $\mathbf{b}_\alpha$  are the biases. The attention-driven context vector  $\mathbf{v}_r$  infers the *strength of  $f_r$  conditioned on itself and its neighborhood* (Fig. 2(c)). The final context refinement weight vector  $\mathbf{v}$  representing all the regions  $\mathcal{R}$  is computed using an element-wise sigmoid activation function  $\sigma(\cdot)$  over a weighted summation of all  $\mathbf{v}_r \in \mathcal{V}$  using an attention importance weight  $w_r$ .

$$\begin{aligned} \mathbf{v} &= \sigma \left( \sum_{r=1}^{\mathcal{R}} \bar{\mathbf{v}}_r w_r \right), \quad \text{where } \bar{\mathbf{v}}_r = \text{GMP}(\mathbf{v}_r) \text{ and} \\ w_r &= \text{Softmax}(\mathbf{W}_\beta \bar{\mathbf{v}}_r + \mathbf{b}_\beta), \end{aligned} \quad (5)$$

where GMP is the *global max-pooling*, weight matrix  $\mathbf{W}_\beta$  and bias  $\mathbf{b}_\beta$  are learnable parameters. It is similar to the approach in [13] for solving machine translation problems where the model searches for parts of a source sentence relevant to predicting a target word. However, our model does not consider the sequential information but learns to emphasize latent representations of multiple regions by joint learning. To improve the gradient flow, this refinement of weight vector  $\mathbf{v}$  is used to enhance the relation-aware image-level feature  $f_t$  in (3) via a skip connection *i.e.*,  $\tilde{f}_t = f_t + f_t \otimes \mathbf{v}$  before passing it to a Softmax layer for estimating the target class probability  $\hat{y}_n$  for the image  $I_n$ . The learnable attentional context refinement parameter in (1) is thus  $\theta_c = \{\mathbf{W}_\alpha, \mathbf{W}_{\alpha'}, \mathbf{W}_a, \mathbf{b}_a, \mathbf{b}_\alpha, \mathbf{W}_\beta, \mathbf{b}_\beta\}$ .

## IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

We first present the datasets, experimental details followed by comparison to the SotA. Then, we analyze our model's complexity followed by qualitative analysis to get an insight into the decision-making process. Finally, we conduct ablation studies to evaluate its key components and parameters.

### A. Datasets

Our model avoids object/part bounding box labels for evaluation on eight benchmark datasets (fine-grained objects/

TABLE II

ACCURACY (%) COMPARISON WITH THE MOST RECENT TOP-10 SOTA METHODS. \* INVOLVES TRANSFER/JOINT LEARNING STRATEGY FOR OBJECTS/PATCHES/REGIONS INVOLVING MORE THAN ONE DATASET (TARGET AND SECONDARY). § APPLIES VISION TRANSFORMER. † USES ADDITIONAL TEXT DESCRIPTION. THE LAST THREE ROWS SHOW THE ACCURACY OF BASE CNN (XCEPTION [40]), SR-GNN WITHOUT THE ATTENTIONAL REFINEMENT (“W/o REFINE”) MODULE (FIG. 2(C)), AND FULL SR-GNN MODEL. THE FOLLOWING ABBREVIATIONS ARE USED TO DENOTE VARIOUS CNN BACKBONES: RN34/RN50/RN101/RN152 FOR RESNET-34/50/101/152; IN-V3 FOR INCEPTION-V3; BCNN FOR BILINEAR CNN; XCEP FOR XCEPTION, DN161/DN201 FOR DENSENET-161/201; ViT-B FOR VISION TRANSFORMER-B-16; SWIN FOR SWIN TRANSFORMER WITH SWIN-BASE-224; GN FOR GOOGLENET; WRN FOR WIDE RESIDUAL NETWORKS; SE FOR SQUEEZE-AND-EXCITATION NETWORKS. CODING IMPLIES ENCODING/CODEBOOK; PARAM AS PARAMETRIC, AND FUSION FOR MULTIPLE CNNs.

| Aircraft      |       |             | CUB-200       |       |             | Cars                |        |             | Dogs          |            |             |
|---------------|-------|-------------|---------------|-------|-------------|---------------------|--------|-------------|---------------|------------|-------------|
| Method        | CNN   | Acc         | Method        | CNN   | Acc         | Method              | CNN    | Acc         | Method        | CNN        | Acc         |
| DCL [22]      | RN50  | 93.0        | CSC [37]      | RN50  | 89.2        | DCL [22]            | RN50   | 94.5        | Cross-X [45]  | RN50       | 88.9        |
| GCL [38]      | RN50  | 93.2        | DAN [46]      | In-v3 | 89.4        | S3Ns [32]           | RN50   | 94.7        | MRDMN [47]    | RN50       | 89.1        |
| CAMF§ [18]    | Swin  | 93.3        | BARM [9]      | DN161 | 89.5        | TrnFG§ [17]         | ViT-B  | 94.8        | APIN [35]     | RN101      | 90.3        |
| PMG [48]      | RN50  | 93.6        | GaRD [39]     | RN50  | 89.6        | CSC [37]            | RN50   | 94.9        | ViT§ [15]     | ViT-B      | 91.7        |
| SCAP [49]     | RN50  | 93.6        | PMG [48]      | RN50  | 89.9        | GaRD [39]           | RN50   | 95.1        | DAN [46]      | In-v3      | 92.2        |
| CSC [37]      | RN50  | 93.8        | APIN [35]     | DN161 | 90.0        | PMG [48]            | RN50   | 95.1        | TrnFG§ [17]   | ViT-B      | 92.3        |
| APIN [35]     | DN161 | 93.9        | CPM* [21]     | GN    | 90.4        | APIN [35]           | DN161  | 95.3        | CAMF§ [18]    | Swin       | 92.8        |
| APCN [50]     | RN50  | 94.1        | CAMF§ [18]    | Swin  | 91.2        | CAMF§ [18]          | Swin   | 95.3        | WARN [28]     | WRN50      | 92.9        |
| GaRD [39]     | RN50  | 94.3        | TrnFG§ [17]   | ViT-B | 91.7        | APCN [50]           | RN50   | 95.4        | CAP [7]       | Xcep       | 96.1        |
| CAP [7]       | RN50  | 94.9        | CAP [7]       | Xcep  | 91.8        | CAP [7]             | Xcep   | 95.7        | CPM* [21]     | GN         | 97.1        |
| Base CNN      | Xcep  | 79.5        | Base CNN      | Xcep  | 75.6        | Base CNN            | Xcep   | 84.8        | Base CNN      | Xcep       | 82.7        |
| W/o Refine    |       | 93.5        | W/o Refine    |       | 90.2        | W/o Refine          |        | 93.7        | W/o Refine    |            | 96.5        |
| <b>SR-GNN</b> |       | <b>95.4</b> | <b>SR-GNN</b> |       | <b>91.9</b> | <b>SR-GNN</b>       |        | <b>96.1</b> | <b>SR-GNN</b> |            | <b>97.3</b> |
| Flowers       |       |             | NABirds [42]  |       |             | Stanford-40 [43]    |        |             | PPMI-24 [44]  |            |             |
| Method        | CNN   | Acc         | Method        | CNN   | Acc         | Method              | CNN    | Acc         | Method        | CNN        | Acc         |
| MGE [51]      | RN50  | 95.9        | Cross-X [45]  | SE    | 86.4        | CAM [52]            | GN     | 72.6        | LLC [53]      | Coding     | 39.7        |
| PBC [54]      | GN    | 96.1        | SPA [55]      | Param | 87.6        | ProCRC [56]         | VGG19  | 80.9        | ScSPM [57]    | Coding     | 41.5        |
| IntAct [58]   | VGG19 | 96.4        | DSTL* [59]    | In-v3 | 87.9        | Introsop [60]       | VGG16  | 81.7        | CSDL [61]     | Coding     | 48.8        |
| SJFT* [62]    | RN152 | 97.0        | GaRD [39]     | RN50  | 88.0        | PKPCR [63]          | VGG19  | 82.4        | Exemplr [64]  | Dictionary | 49.3        |
| OPAM* [65]    | VGG   | 97.1        | APIN [35]     | DN161 | 88.1        | Concepts [66]       | VGG16  | 83.1        | VLAD [67]     | Coding     | 50.7        |
| Cos.Ls* [68]  | RN50  | 97.2        | CSPE [69]     | In-v3 | 88.5        | Color [70]          | Fusion | 84.2        | Color [70]    | Fusion     | 65.9        |
| PMA† [71]     | VGG16 | 97.4        | MGE [51]      | RN101 | 88.6        | $\alpha$ -pool [72] | VGGM   | 86.0        | DSFNet [73]   | RN34       | 72.3        |
| DSTL* [59]    | In-v3 | 97.6        | ViT§ [15]     | ViT-B | 89.9        | Implicit [74]       | RN50   | 87.7        | Coding [27]   | NASNet     | 82.3        |
| MCL* [31]     | BCNN  | 97.7        | TrnFG§ [17]   | ViT-B | 90.8        | RAN [8]             | RN50   | 97.4        | AG-Net [10]   | RN50       | 98.2        |
| CAP [7]       | Xcep  | 97.7        | CAP [7]       | Xcep  | 91.0        | AG-Net [10]         | RN50   | 97.8        | RAN [8]       | DN201      | 98.6        |
| Base CNN      | Xcep  | 91.9        | Base CNN      | Xcep  | 68.1        | Base CNN            | Xcep   | 80.0        | Base CNN      | Xcep       | 79.3        |
| W/o Refine    |       | 97.1        | W/o Refine    |       | 89.9        | W/o Refine          |        | 97.9        | W/o Refine    |            | 97.9        |
| <b>SR-GNN</b> |       | <b>97.9</b> | <b>SR-GNN</b> |       | <b>91.2</b> | <b>SR-GNN</b>       |        | <b>98.8</b> | <b>SR-GNN</b> |            | <b>98.9</b> |

human-actions, detailed in Table I): Aircraft [4], Caltech-UCSD Birds (CUB-200) [1], Stanford Cars [5], Stanford Dogs [3], Oxford Flowers [2], NABirds [42], Stanford-40 actions [43], and People Playing Musical Instruments (PPMI-24) [44]. The top-1 accuracy (%) is used for the evaluation.

### B. Implementation Details

TensorFlow 2.0 is used for implementation. Like CAP [7], we use Xception [40] as a backbone CNN. The output dimension  $7 \times 7 \times 2048$  is upsampled to  $42 \times 42 \times 2048$  for region pooling (Fig. 2(a)). Region proposal in [8] is used with a HOG cell-size of  $14 \times 14$  to generate 27 optimal region proposals ( $\mathcal{R}$ ), consisting of a minimum region of 2 cells to a maximum of the full image. The region-pooling size of  $w=h=7$  is used. The feature transformation module (Fig. 2(b)) consists of two GNN layers with an optimal output size of 1024. Each layer contains a single-layer MLP with the teleport probability  $\alpha=0.3$ . The number of channels is kept the same ( $C=2048$ ) as Xception output. Source codes of SR-GNN will be available via the GitHub repository at <https://github.com/ArdhenduBehera/SR-GNN>.

### C. Experimental Settings

Pre-trained ImageNet weights are used to initialize base CNN for faster convergence with the size of images being  $256 \times 256$ . We apply the data augmentation of random rotation ( $\pm 15$  degrees), random scaling ( $1 \pm 0.15$ ), and then random cropping to select the image size of  $224 \times 224$ . The Stochastic Gradient Descent (SGD) is used to optimize the categorical cross-entropy loss function with an initial learning rate of  $10^{-3}$  and multiplied by 0.1 after every 50 epochs. The model is trained for 150 epochs with a mini-batch size of 8 using NVIDIA Titan V GPU (12GB).

### D. Performance Comparisons with State-of-the-Art Methods

The accuracy (%) of SR-GNN over eight datasets and its comparisons to the previous best results (according to the best of our knowledge) are given in Table I. The accuracy of SR-GNN over each dataset is compared with those of the top-10 SotA methods in the literature in Table II. These SotA methods are based on attention mechanism [7]–[9], [28], [31], [32], [51], discriminative object-part localization [21], [22], [24], mutual reinforcement learning [9], GNN [37]–[39], vision transformers [15]–[18], etc. SR-GNN clearly outperforms all previous methods over eight datasets and their accuracy

TABLE III

ACCURACY (%) OF OUR SR-GNN USING RESNET-50 BASE CNN WITH DIFFERENT SIZES OF IMAGES FOR FGVC.

|      | 224×224, CAP [7] |             |             | 448×448 size |             |             |
|------|------------------|-------------|-------------|--------------|-------------|-------------|
|      | Airc.            | CUB         | NAB         | Cars         | Flowers     | Dogs        |
| SotA | <b>94.9</b>      | 90.9        | <b>88.8</b> | 95.4 [50]    | 96.8 [31]   | 88.9 [45]   |
| Ours | 94.8             | <b>91.0</b> | <b>88.8</b> | <b>95.8</b>  | <b>98.0</b> | <b>97.1</b> |

gain over each dataset is given in parenthesis: Aircraft (0.5%), CUB-200 (0.1%), Cars (0.4%), Dogs (0.2%), Flowers (0.2%), NABirds (0.2%), Stanford-40 (1.0%), and PPMI-24 (0.3%). These margins of improvements are very significant since FGVC is a challenging task to discriminate various subcategories. This is evident from the top-10 SotA accuracies over each dataset (Table II) that achieve the successive marginal gain between 0.1-0.3% over the past 2-3 years. For example, a cumulative gain of 1.2% is achieved by the top-10 SotA methods over the Cars dataset within the past 3 years with an average of 0.13% (9 successive differences). Our gain of 0.4% over CAP is thus significantly higher. Similarly, a cumulative gain of 1.9% (DCL to CAP) is achieved over the Aircraft dataset (average: 0.21%). Our SR-GNN gains 0.5% in comparison to CAP and is thus also significant. Moreover, some methods attain similar accuracy, *e.g.*, GaRD [39] and PMG [48] on Cars: 95.1%. SR-GNN outperforms over eight datasets with a gain of between 0.1% to 1.0%. Moreover, our accuracy gain is 0.1% - 1.2% over six FGVC datasets over CAP currently at the top of the leaderboard. Many SotA methods are weakly-supervised such as localization of objects/parts using pre-trained object/part detector and/or proposals using semantic segmentation (*e.g.*, mask R-CNN or Grad-CAM). The process often includes at least two steps: firstly, detect the weakly-supervised regions and then apply the fine-grained recognition. Moreover, additional secondary datasets (*e.g.*, COCO in [21] for Dogs: 97.1%, and ImageNet in [31] for Flowers: 97.7%) are used for further training to achieve SotA accuracy [21]. In sharp contrast, our SR-GNN is a single-step process that is trained end-to-end using only the target datasets and is thus computationally efficient and easy to implement.

We have explicitly compared the performance of our method with the SotA ones implemented with ResNet-50 backbone using image sizes of 224×224 and 448×448. The results are given in Table III. It is evident that CAP performs the best among the existing methods on Aircraft (94.9%), CUB (90.9%), and NABirds (88.8%) with an image size of 224×224 and in this case, our method achieves a very competitive results ( $\pm 0.1\%$ ). Alternatively, AP-CNN (Cars: 95.4%) [50], MCL (Flowers: 96.8%) [31], and Cross-X (Dogs: 88.9%) [45] use an image size of 448×448 with ResNet-50 instead. With such image size, our SR-GNN achieves 95.8% on Cars, and 98.0% on Flowers; and gains a margin of 8.2% over Cross-X on Dogs (SR-GNN: 97.1%). Though, we attain an accuracy of 97.1% over Dogs as CPM [21], the latter applies a complex training process using GoogleNet backbone. Clearly, our SR-GNN outperforms many SotA methods with an image size of 224×224 over all the datasets using Xception or ResNet-50 backbone.

TABLE IV

COMPARISON OF OUR SR-GNN WITH VISION TRANSFORMERS. MODEL COMPLEXITY IS GIVEN IN PARAMETERS (MILLION) AND GFLOPS (BILLION), FOR INPUT-SIZE 384×384, AS PROVIDED IN [16]. TOP: INPUT-SIZE: 224×224, MID: 448×448, AND BOTTOM: 224×224.

| Method        | Transformer | CUB         | Car         | Dog         | NAB         | Air         | Parm (GFlop)      |
|---------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------------|
| Swin [16]     | Swin-224    | 89.7        | 94.2        | 91.8        | -           | 91.0        | 88 (15.4)         |
| CAMF [18]     | Swin-224    | 90.9        | 94.8        | 92.6        | -           | 92.9        | -                 |
| ViT [15]      | ViT-B-16    | 90.3        | 93.7        | 91.7        | 89.9        | -           | 86 (55.4)         |
| TrnFG [17]    | ViT-B-16    | 91.7        | 94.8        | 92.3        | 90.8        | -           | -                 |
| Swin [16]     | Swin-224    | 90.7        | 94.8        | 92.5        | -           | 93.0        | 88 (47.0)         |
| CAMF [18]     | Swin-224    | 91.2        | 95.3        | 92.8        | -           | 93.3        | -                 |
| <b>SR-GNN</b> | -           | <b>91.9</b> | <b>96.1</b> | <b>97.3</b> | <b>91.2</b> | <b>95.4</b> | <b>30.9 (9.8)</b> |

TABLE V

ACCURACY (%) OF SR-GNN WITH OTHER SOTAs BASE CNNS INSTEAD OF XCEPTION USING THE SAME TEST SETUP (224×224 IMG-SIZE,  $\alpha = 0.3$  &  $\mathcal{R} = 27$ ). CAP IS USED AS BACKBONE BY REPLACING THE CNN FEATURE MAP AND REGION PROPOSALS (FIG. 2(A)), AND RELATION-AWARE FEATURE TRANSFORMATION BY GNN (FIG. 2(B)).

| Dataset  | ResNet-50 | Inception-V3 | NASNetMobile | CAP  |
|----------|-----------|--------------|--------------|------|
| Aircraft | 94.8      | 94.4         | 94.4         | 95.1 |
| CUB      | 91.0      | 90.7         | 90.8         | 91.9 |
| Cars     | 93.1      | 94.1         | 95.6         | 95.8 |
| Dogs     | 93.2      | 95.3         | 95.9         | 96.6 |
| Flowers  | 97.4      | 97.4         | 97.3         | 97.8 |
| NABirds  | 88.8      | 89.2         | 88.6         | 90.7 |

More recently, vision Transformer such as ViT [15] uses fixed-size patches, and Swin Transformer [16] applies a shifted window scheme to construct a hierarchical representation of patches. In contrast, we use multi-scale regions leveraging GNN for subtle discrimination. ViT often requires large-scale training datasets (*e.g.*, JFT-300M, ImageNet-22K, etc.) and then fine-tuning on a target dataset to perform well for FGVC. Unlike CNNs, a Transformer is built with a relatively complex, larger, and heavier architecture. For example, ViT base model consists of 86M parameters and 55.4B GFLOPs (Table IV). Recently, CAMF [18] has demonstrated that a Swin Transformer can achieve better performance than pure ViT with an image size of 448×448. Whereas our method (224×224) outperforms vision Transformers with both sizes of 448×448 and 224×224 with a clear margin on five FGVC datasets (Table IV). Our gain (in parenthesis) on each dataset is: Dogs (4.5%), Aircraft (2.1%), Cars (0.8%), CUB (0.7%), and NABirds (0.4%). Moreover, our method incurs significantly less computational overload than the Transformers. For example, SR-GNN (224×224) consists of 30.9M parameters and 9.8B GFLOPs. This is 57.1M parameters and 37.2B GFLOPs lesser than the Swin Transformer. Furthermore, SR-GNN expeditiously outperforms these SotA models with a notable margin with an end-to-end training and simple evaluation protocol avoiding additional secondary data and resource constraints, justifying its wider adaptability.

**Performance using other SotA base CNNs:** SR-GNN uses the lightweight Xception [40] as a backbone to extract CNN features for further processing. It can easily be integrated into other CNN backbones with a little computational overhead. In order to verify this, we have evaluated our SR-GNN using three different SotA CNN backbones: ResNet-50 [75], Inception-V3 [76], and NASNetMobile [77], with an image resolution of 224×224 over the six FGVC datasets. The results are provided

TABLE VI  
SR-GNN’S CAPACITY AND COMPUTATIONAL OVERHEAD FOR DIFFERENT REGIONS USING AN NVIDIA TITAN V GPU (12GB).

| #No. of Regions | #Trainable params in millions (~M) | GFLOPs in billions (~B) | Per-image inference time in millisecond (~ms) | Training time (batch size 8) in ~hours |      |      |         |
|-----------------|------------------------------------|-------------------------|---|--|------|------|---------|
|                 |                                    |                         |   | Aircraft                               | Cars | Dogs | Flowers |
| 11              | 30.9                               | 9.4                     | 3.9   | 3.5                                    | 8.4  | 13.4 | 1.9     |
| 19              | 30.9                               | 9.6                     | 5.0   | 4.1                                    | 10.2 | 15.2 | 2.6     |
| 27              | 30.9                               | 9.8                     | 5.0   | 4.5                                    | 11.2 | 17.0 | 2.8     |
| 36              | 30.9                               | 10.1                    | 6.0   | 5.0                                    | 12.6 | 18.3 | 3.1     |

in Table V. Our method using these backbones is very similar to the one using Xception and consistently outperforms the SotA approaches in Table II with the same backbones. However, SR-GNN’s accuracy using Xception is slightly higher than the similar backbones such as Inception-V3 and ResNet-50, and is thus our optimal choice. The main reason could be the architectural design of Xception in which depth-wise separable convolutions are used within the Inception module. It is built with a linear stack of depth-wise separable convolutional layers with residual connections. The design leads to a better representation of high-level CNN features in comparison to the ResNet-50 and Inception-V3 architectures. NASNetMobile [77] is a lightweight model that is designed for mobile and embedded vision systems. It involves significantly less computational cost. From the performance (Table V), the accuracy using this mobile architecture is as competitive as the standard CNNs. Generally, many approaches consider ResNet-50 and our method significantly outperforms those using the same ResNet-50 backbone, as evident from Table II-III. A similar trend can be observed for Inception-V3.

We have also evaluated our model by replacing the region proposals (Fig. 2(a)) and feature transformation using GNN (Fig. 2(b)) with the CAP [7] model. The results are given in Table V. The performance is aligned with the original CAP *i.e.*, the accuracy (Aircraft: 95.1%, CUB: 91.9%, Cars: 95.8%, Dogs: 96.6%, Flowers: 97.8% and NABirds: 90.7%) is superior to SotA methods including CAP, except NABirds on which CAP’s accuracy is 91.0% [7]. Nevertheless, the accuracy on NABirds (90.7%) is still superior to the other approaches in Table II. Moreover, SR-GNN surpasses these results over Aircraft (0.3%), Cars (0.3%), Dogs (0.7%), Flowers (0.1%), and NABirds (0.5%) with a clear margin and achieves the same accuracy of 91.9% over CUB. This justifies the benefits of our novel GNN-driven relation-aware feature transformation (Fig. 2(b)) and attentional context refinement (Fig. 2(c)) modules, and their significance in enhancing FGVC accuracy. Also, SR-GNN is lighter than CAP requiring 3.3M and 0.4B fewer (Table VII) parameters and GFLOPs, respectively, implying its computational efficiency.

**Performance on Human-Object Interactions:** To demonstrate our method under general data diversity, we tested our SR-GNN on the Stanford-40 actions [43] and People Playing Musical Instruments (PPMI-24) [44] datasets, representing fine-grained human-object interactions. Its accuracy is 98.8% on Stanford-40 and 98.9% on PPMI-24. It outperforms the best results attained by AG-Net (Stanford-40: 97.8%) [10] and RAN (PPMI-24: 98.6%) [8]. Our model also *learns the importance* (weight) of a region via a novel attentional context to refine the transformed features. Whereas, CAP uses LSTM

to learn spatial arrangement between regions, and an LSTM-driven feature encoding to aggregate the information from its hidden states. AG-Net and RAN learn features from each region independently without feature interaction and use a Squeeze-and-Excitation block to extract features followed by an attention module.

A generalized conventional average and bilinear pooling, namely  $\alpha$ -pooling [72], achieves an accuracy of 86.0% over the Stanford-40 actions. The  $\alpha$ -pooling enhances the performance in implicit pose normalization (87.7%) [74] over this dataset, and achieves SotA accuracy compared to other prior works. However, our method attains an impressive margin (11.1%) over this work. Even without feature refinement (W/o Refine), our model achieves the best result over this dataset.

Some prior methods have extracted traditional/hand-crafted feature descriptors (*e.g.*, SIFT) and applied bag-of-feature encoding techniques [61], [67] over which deep features attain better performance. A reinforcement learning method, DSFNet [73] captures the global discriminative information and fine-grained representations on PPMI-24. Hierarchical learning based on the spatial pyramid is presented in [27]. Their pre-trained networks achieve better performance than the other existing approaches on this dataset. However, our method gains a high margin of 16.6% over their approach.

**Comparison using mAP evaluation metric:** Many works consider the mAP (mean average precision) as an evaluation metric on the above-mentioned two datasets. For a fair comparison, we have evaluated the performance of SR-GNN using mAP. Our approach achieves 96.6% mAP on Stanford-40 which is 0.4% improved over AG-Net (96.2%) [10]. Similarly, we have attained higher mAP in comparison to the human mask loss (94.1%) [78], part-action network (91.2%) [79], and many recent works on Stanford-40. We have achieved improved mAP (95.3%) on PPMI-24 over existing works such as VLAD spatial pyramid (81.3%) [80], 10-model color fusion (65.9%) [70], and the others. While, the mAP of SR-GNN (95.3%) on PPMI-24 is 1.4% lower than RAN (96.7%), it attains 0.3% gain in accuracy.

The accuracy of our model is compared without the attentional context refinement module (Fig. 2(c)) and is given in the 2nd-last row of Table II. A notable observation is that even without context refinement (“W/o Refine”), SR-GNN outperforms many methods tested on Dogs (96.5%, 2nd-best), NABirds (89.9%, 3rd-best, same as ViT), Stanford-40 (97.9%, best), and PPMI-24 (97.9%, 3rd-best). Also, the accuracies on Aircraft (93.5%), CUB (90.2%), and Flowers (97.1%) are competitively retained within the accuracies of the top-10 SotA methods. Our attentional context refinement module enhances the overall accuracy on diverse datasets, while avoid-

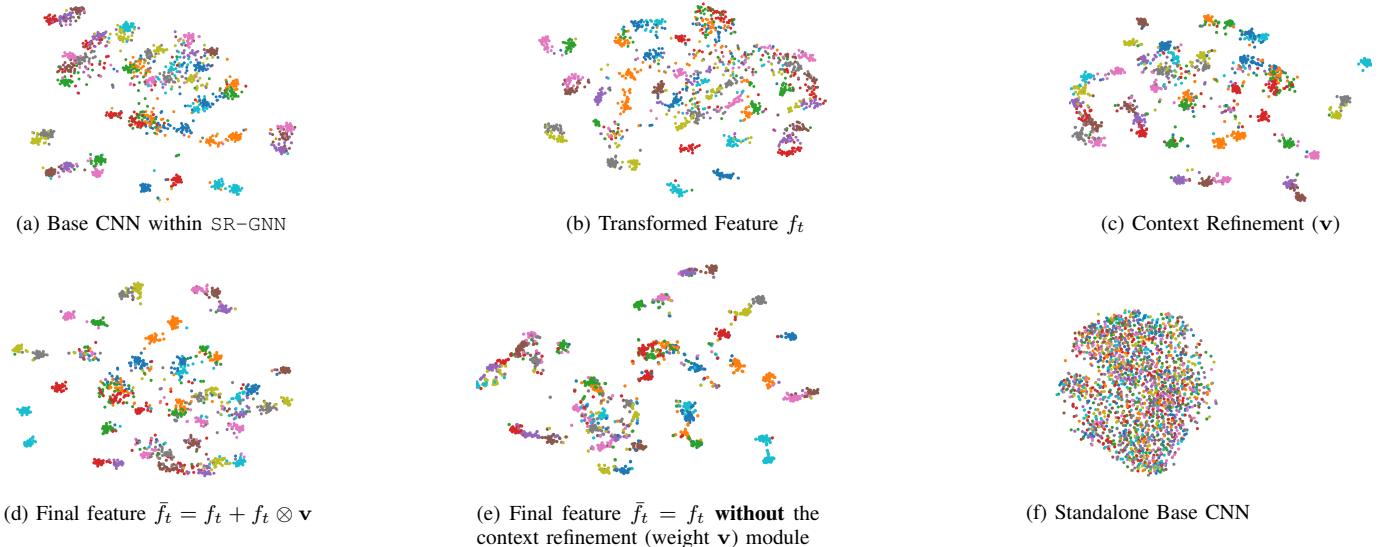


Fig. 4. SR-GNN’s discriminability of the Aircraft test-set using t-SNE [81] to visualize class separability and compactness using features from a) base CNN (Xception, Fig. 2(a)) within our model, b) relation-aware transformed feature using GNN (Fig. 2(b)), c) attentional context refinement weight-vector  $v$  (Fig. 2(c)), and d) the final image-level feature map  $\bar{f}_t$  for classification (Fig. 2(c)). Each color represents a particular class. There are 50 classes chosen randomly from the Aircraft’s test set. e) SR-GNN without the context refinement module, and f) Standalone Xception base CNN without our modules (re-trained on the Aircraft dataset).



Fig. 5. Visualization of the relation-aware transformation using **cosine similarity** to measure pairwise relationships (cool to warm  $\Rightarrow$  weak to strong) between nodes in the graph. Top (Aircraft): 707-320, 737-400 and 737-200 (left to right). Bottom (Dogs): Japanese Spaniel, Shih Tzu and Toy Terrier.

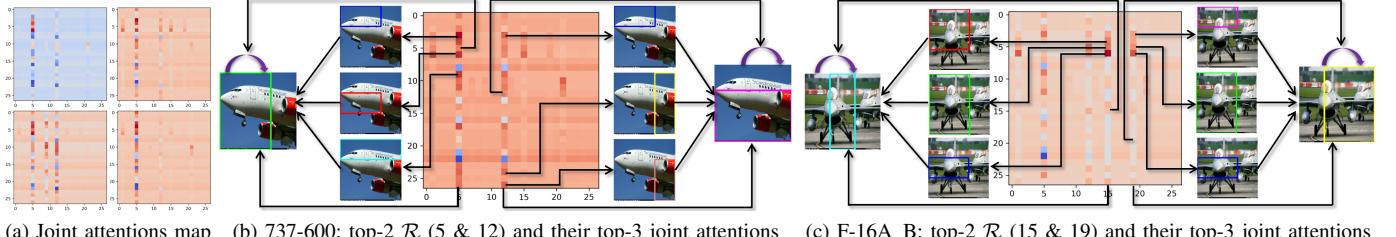


Fig. 6. Visualization of attentional context refinement (Fig. 2(c)) in our SR-GNN over the sub-types in the Aircraft dataset: a) joint attentional maps for ‘A340-200’, ‘ATR-72’, ‘DC-10’ and ‘ERJ 135’ aircraft sub-types. b) Top-2 regions (cols 5 & 12) contributing towards sub-type ‘Boeing 737-600’ conditioned on the respective other top-3 regions (rows) in joint decision-making. The self-attention (self-loop) is also shown in the top-2 regions. c) Similarly, top-2 regions (cols 15 & 19) contributing towards sub-type ‘F-16A\_B’ conditioned on the respective other top-3 regions (rows). Regions are shown in the respective original images.

ing additional parts-level annotations, vision Transformers, secondary datasets and/or pre-trained subnetworks to enhance the accuracy.

TABLE VII  
COMPUTATIONAL COMPLEXITY COMPARISON OF THE PROPOSED SR-GNN WITH STATE-OF-THE-ARTS.

| Method        | Param (M)   | GFLOPs (B) | Inf. Time/img (ms) |
|---------------|-------------|------------|--------------------|
| AG-Net [10]   | 54.8        | 10.4       | 5.2                |
| MRDMN-L [47]  | 51.2        | 14.0       | 4.9                |
| TASN [29]     | 37.3        | 21.9       | 7.5                |
| CAP [7]       | 34.2        | 10.2       | 4.2                |
| Base CNN      | 20.9        | 9.2        | 2.7                |
| W/o Refine    | 24.4        | 9.3        | 4.9                |
| <b>SR-GNN</b> | <b>30.9</b> | <b>9.8</b> | 5.0                |

### E. Model Complexity

SR-GNN’s capacity and computational complexity is assessed using GFLOPs (Giga floating point operations), model size as a number of trainable parameters in millions (M) and per-image inference time in milliseconds (ms) [83]. These values over four datasets is given in Table VI. Its comparison to the SotA methods is also provided in Table VII. For  $\mathcal{R}=27$ , its complexity in terms of trainable parameters (base CNN: 20.9M, W/o Refine: 24.4M and full model: 30.9M), per-image inference time (base CNN: 2.7ms, W/o Refine: 4.9ms and SR-GNN: 5.0ms) and GFLOPs (base CNN: 9.2B, W/o Refine: 9.3B and full model: 9.8B) are given in Table VI.

TABLE VIII

DAVIES-BOULDIN [82] INDEX (LOWER IS BETTER) TO QUANTIFY CLUSTER SIMILARITIES USING THE t-SNE [81] OUTPUTS OVER ALL TEST IMAGES FROM AIRCRAFT DATASET. FOR THE TRAINING AND VALIDATION OF THE BACKBONE CNN (XCEPTION), WE USE THE STANDARD TRANSFER LEARNING BY FINE-TUNING IT ON THE TARGET DATASET USING THE SAME DATA AUGMENTATION AND HYPER-PARAMETERS (SEC. IV). THE CLUSTERS GENERATED BY SR-GNN ARE MORE COMPACT AND SEPARATED THAN THE BASELINE XCEPTION (LAST ROW). THE FINAL FEATURE DESCRIPTION OF SR-GNN IS BETTER THAN THE INDIVIDUAL RELATION-AWARE FEATURE TRANSFORM AND CONTEXT REFINEMENT MODULES.

| Feature Extraction Point  | Aircraft                | Cars                    | Flowers                 |
|---|-------------------------|-------------------------|-------------------------|
| SR-GNN's different extraction points                                |                         |                         |                         |
| Base CNN (Fig. 2(a))  | 4.00 (Fig. 4(a))        | 9.89 (Fig. 8(a))        | 1.52 (Fig. 9(a))        |
| Transformed feature $f_t$ (Fig. 2(b))                               | 3.34 (Fig. 4(b))        | 2.49 (Fig. 8(b))        | 1.35 (Fig. 9(b))        |
| Attentional Context Refinement Weight $v$ (Fig. 2(c))               | 2.65 (Fig. 4(c))        | 2.25 (Fig. 8(c))        | 1.12 (Fig. 9(c))        |
| Final feature $\bar{f}_t = f_t + f_t \otimes v$ (Fig. 2(c))         | <b>2.07</b> (Fig. 4(d)) | <b>2.25</b> (Fig. 8(d)) | <b>1.02</b> (Fig. 9(d)) |
| SR-GNN's <b>without</b> the context refinement (weight $v$ ) module |                         |                         |                         |
| Final feature w/o refinement $\bar{f}_t = f_t$                      | 3.42 (Fig. 4(e))        | 3.49 (Fig. 8(e))        | 1.19 (Fig. 9(e))        |
| Base CNN (Xception) trained on target dataset (Transfer Learning)   |                         |                         |                         |
| Xception (baseline)   | 77.22 (Fig. 4(f))       | 95.85 (Fig. 8(f))       | 38.67 (Fig. 9(f))       |

The key modules only add a little overhead to the base CNN in terms of trainable parameters and GFLOPs: 1) relation-aware feature transformation (Fig. 2(b)): 3.4M and 0.18B; and 2) attentional context modeling (Fig. 2(c)): 6.4M and 0.51B. GFLOPs and model's trainable parameters are widely used by the community to compare the computational efficiency of various deep models [83]. By considering this, our SR-GNN (param: 30.9M, GFLOPs: 9.8B) is computationally a lighter model than CAP (param: 34.2M, GFLOPs: 10.2B) using Xception that is more lightweight than the other SotA models in Table VII. Likewise, using ResNet-50 as a base CNN, SR-GNN (param: 33.4M, GFLOPs: 8.4B) is lighter than RAN (param: 49.0M, GFLOPs: 8.5B) and AG-Net (param: 54.8M, GFLOPs: 10.4B, and per-image inference time: 5.2 ms).

The inference time is dependent on types of GPU, hardware and software environments used. For example, both our SR-GNN and CAP [7] use the same Titan V GPU (12 GB) to run the model, but CAP is implemented using TensorFlow 1.x whereas, SR-GNN runs using TensorFlow 2.x. It is well-known that the TensorFlow 2.x is significantly slower<sup>1</sup> than TensorFlow 1.x, resulting in increase of the inference time for SR-GNN (5.0ms) in comparison to CAP (4.2ms) even though the former is more lightweight (param: 30.9M, GFLOPs: 9.8B) than the latter (param: 34.2M, GFLOPs: 10.2B). Moreover, the per-image inference time of our SR-GNN is 0.8ms and 0.1ms higher than CAP and MRDMN-L [47], respectively. SR-GNN without refinement (4.9ms) shares the same inference time with MRDMN-L, but gains 7.5% higher accuracy on Dogs. A precise comparison with the existing top-10 SotA methods focusing on the inference time is given in the supplementary document, irrespective of the GPU and hardware configuration, deep learning tools (*e.g.*, TensorFlow, PyTorch, MXNet, etc.) and related experimental constraints used in those works. Moreover, the Transformers are computationally more complex than SR-GNN as shown in Table IV.

Some works [21], [31], [62] improve the accuracy by exploring secondary data. Also, such methods involve multiple steps and are resource-intensive. For example, there are three steps in [21]: 1) object detection and segmentation using Mask R-CNN and a conditional random field (CRF); 2) complementary

part mining using 512 regions; and 3) classification using context gating. Their model is trained using 4 GPUs (12GB each), per-image inference time is 27ms for Step 3 and extra 227ms in Step 2. Our model is trained on a single GPU (12GB) with per-image inference time of 5ms only. So, SR-GNN is faster and lighter than most of the existing methods.

#### F. Qualitative Analysis and Visualization

To get insight into our model's decision-making process, we visualize the feature maps at key steps. Each step provides the discriminability of our model by visualizing the class separability and compactness. To achieve this, we use t-SNE [81], which is shown in Fig. 4. Randomly selected 50 classes are chosen from the Aircraft test-set. The test images are processed to extract features from base CNN (Fig. 2(a)), relation-aware transformed feature (Fig. 2(b)), context refinement weight vector (Fig. 2(c)) and the final refined feature descriptor. The respective visualization (unique color per class) is presented in Fig. 4. It is evident that clusters representing both relation-aware features (Fig. 4(b)) and context weight vector (Fig. 4(c)) are further apart and more compact compared to the base CNN features (Fig. 4(a)). Moreover, the clusters representing the final refined feature map (Fig. 4(d)) is further enhanced, resulting in a clearer distinction between various clusters representing different classes. In addition, the importance of the context refinement task is visualized by avoiding it from the final feature vector (Fig. 4(e)). Lastly, standalone Xception is fine-tuned by discarding our proposed modules, and its impact is shown in Fig. 4(f). Overall, these qualitative visualizations evince the essence of key components and superior performance of SR-GNN.

We have further computed the Davies-Bouldin index [82] to quantitatively evaluate the cluster similarities using the t-SNE outputs given in Table VIII. This index signifies the similarity between clusters, where the similarity is the ratio of within-cluster distances to between-cluster distances. As a result, a lower value implies better clustering. For the Aircraft test-set, these values are 4.00 (Fig. 4(a)), 3.34 (Fig. 4(b)), 2.65 (Fig. 4(c)), and 2.07 (Fig. 4(d)). Whereas, the value increases without feature refinement 3.42 (Fig. 4(e)), and it is very high

<sup>1</sup><https://github.com/tensorflow/tensorflow/issues/33487>

TABLE IX

ACCURACY (%) OF OUR SR-GNN WITH VARIED KEY MODULES. 1) UNIFORM PATCH-SIZE AS AN ALTERNATIVE TO GENERATE REGIONS 2) PERFORMANCE OF VARIOUS GRAPH POOLING TECHNIQUES IN OUR FEATURE TRANSFORM MODULE (FIG. 2(B)), AND 3) EFFECTIVENESS OF MAJOR COMPONENTS OF SR-GNN.

| Key Modules                | Aircraft    | Cars        | Dogs        | Flowers     |
|----------------------------|-------------|-------------|-------------|-------------|
| Uniform $4 \times 4$ grid  | 94.3        | 92.9        | 96.3        | 97.0        |
| Global average pooling     | 95.0        | 95.1        | 96.2        | 97.5        |
| Global max pooling         | 94.9        | 95.5        | 96.4        | 97.8        |
| Global sum pooling [84]    | 94.9        | 95.1        | 96.3        | 97.7        |
| Sort pooling [85]          | 95.2        | 95.3        | 96.5        | 97.8        |
| W/o GNN                    | 93.5        | 94.5        | 96.0        | 94.9        |
| W/o Refine                 | 93.5        | 93.7        | 96.5        | 97.1        |
| W/o self-attention         | 93.7        | 95.1        | 96.1        | 96.4        |
| W/o weighted-attention     | 94.4        | 95.5        | 96.0        | 95.1        |
| <b>SR-GNN (full-model)</b> | <b>95.4</b> | <b>96.1</b> | <b>97.3</b> | <b>97.9</b> |

TABLE X

ACCURACY (%) OF SR-GNN WITH DIFFERENT NUMBERS OF REGION PROPOSALS (FIG. 2(A)).

| Dataset  | #11  | #19  | #27         | #36  |
|----------|------|------|-------------|------|
| Aircraft | 90.6 | 90.3 | <b>95.4</b> | 89.4 |
| CUB      | 86.9 | 85.8 | <b>91.9</b> | 82.8 |
| Cars     | 93.4 | 90.9 | <b>96.1</b> | 92.5 |
| Dogs     | 94.5 | 94.3 | <b>97.3</b> | 95.5 |
| Flowers  | 95.2 | 97.5 | <b>97.9</b> | 95.8 |

(77.22) using Xception backbone only (Fig. 4(f)). The results on Cars and Flowers are given in the supplementary document.

In order to understand how the relation-aware transformation is exploited (Fig. 2(b)) in our SR-GNN, we also visualize the cosine similarity to measure the pairwise relationships between each pair of nodes (regions) in the graph as is shown in Fig. 5. It is not easy to visualize the graph structure associating object parts with categories. Thus, pairwise cosine similarities of nodes representing regions are explored to reflect how each object is overall represented and distinguished from each other. It is evident that the graph indeed captures the relational structure to discriminate subordinate categories. The main reason is that  $\alpha$  in (2) preserves locality to avoid over-smoothing by staying close to the root node and leveraging the information from a large neighborhood.

We have also looked inside our attentional context refinement module (Fig. 2(c)) to visualize the class-specific visual relationships jointly learned during the training. These relationships are learned as an  $\mathcal{R} \times \mathcal{R}$  joint attention map and are shown in Fig. 6(a) for the ‘A340-200’, ‘ATR-72’, ‘DC-10’, ‘ERJ 135’ aircraft sub-types where  $\mathcal{R}=27$ . Each column represents a region conditioned on itself and other regions linking rows. **Blue to red** signifies the class-specific *less to more* attention towards that region. We further link the top-2 regions (cols) to their respective top-3 joint visual attentions (rows) by exploring the attention map. These are shown in Fig. 6(b) and 6(c) for the respective ‘Boeing 737-600’ and ‘F-16A\_B’ Aircraft sub-type. These regions are drawn in the original image to show their joint relationships. From both figures, it is evident that our model learns to focus on the key context information for discriminating subtle variations. More results for visualization are given in the supplementary.

## V. ABLATION STUDY

The ablative study is conducted from several important aspects: suitability of using uniform-grid as an alternative to our adopted region proposals, exploring SotA graph-based pooling techniques [84], [85], efficacy of each key components of SR-GNN, impact of the number of regions ( $\mathcal{R}$ ) on recognition accuracy, influence of GNN layer’s output dimensionality in performance, and the number of power-iterations in GNN layers.

### 1) Formation of Region Proposals and Key Modules:

The regions with variable areas and aspect ratios that are akin to computing the HOG cells and blocks are preferred here. ViT uses uniform regions such as  $16 \times 16$  or  $14 \times 14$  for an image resolution of  $224 \times 224$ . Inspired by this, our method is tested with a uniform grid-structure as an alternative for generating region proposals. The results with  $4 \times 4$  grid (region-size is  $16 \times 16$  for up-sampled feature resolution of  $64 \times 64$ ) on four datasets are given in Table IX (row 1), which is the best among other grid-sizes of  $2 \times 2$  (region-size:  $32 \times 32$ ),  $3 \times 3$  (region-size:  $21 \times 21$ ), and  $5 \times 5$  (region-size:  $13 \times 13$ ). Even though the accuracy with a regular grid of  $4 \times 4$  is better than many existing approaches (reported in Table II), it is not a suitable choice for generating regions to learn finer details. This is evident from the accuracy gain of SR-GNN using regions of different aspect ratios and areas in comparison to the  $4 \times 4$  uniform grid. These gains are: 3.2% over Cars, 1.1% over Aircraft, 1.0% over Dogs, and 0.9% over Flowers. These results show that regular regions are not pertinent enough to capture subtle variations for spatial relation modeling among the regions. Moreover, SR-GNN outperforms vision Transformers (Table II) that use regular regions but fail to capture the overall object structure, and hence, do not achieve superior results as ours. Also, Mask-RCNN is used for region proposals in CPM [21] which has attained 1.5% and 0.2% lower accuracy than ours over the respective CUB and Dogs datasets (Table II). Thus, all these results justify the benefits of our method in exploring multi-scale regions.

We have evaluated the efficacy of the existing SotA graph-based pooling methods to compare the performance with the chosen gated attentional pooling to pool features from the nodes of the relation-aware GNN. These are the global average pooling, global max pooling, global sum pooling [84], and sort pooling [85]. The results are shown in Table IX. It is evident that the gated attentional pooling performs better than these alternative methods. This is mainly because the gated attentional pooling uses element-wise sigmoid and acts as a soft attention mechanism that decides which nodes (regions) are more relevant to the current graph-level classification by selecting the most discriminative features from the regions and is thus more suitable to capture their subtle variances than the other pooling approaches. Next, the impact of varied key modules (shown in Fig. 2) are evaluated over 4 datasets (Table IX). It includes only self-attention to refine local features *i.e.*, without (W/o) GNN, without (W/o) self-attention and self-attention without (W/o) weighted-attention. These results justify the importance of each key component in our SR-GNN, without which accuracy degrades significantly.

TABLE XI  
ACCURACY (%) OF SR-GNN WITH 512 AND 1024 OUTPUT DIMENSIONS AT DIFFERENT TELEPORT (OR RESTART) PROBABILITY  $\alpha \in [0.1, 0.8]$  IN (2).

| Dataset  | GNN output dimension = 512 |      |      |      |      |      |      |      | GNN output dimension = 1024 |      |             |      |      |      |      |      |
|----------|----------------------------|------|------|------|------|------|------|------|-----------------------------|------|-------------|------|------|------|------|------|
|          | 0.1                        | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  | 0.1                         | 0.2  | 0.3         | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  |
| Aircraft | 91.6                       | 94.6 | 94.8 | 94.7 | 94.7 | 92.1 | 91.5 | 92.3 | 90.7                        | 92.4 | <b>95.4</b> | 91.1 | 91.7 | 92.1 | 92.3 | 90.9 |
| Cars     | 95.5                       | 95.5 | 95.6 | 94.9 | 93.9 | 93.9 | 93.5 | 93.3 | 95.4                        | 95.7 | <b>96.1</b> | 95.8 | 96.0 | 95.8 | 95.8 | 95.8 |
| Dogs     | 96.7                       | 96.7 | 96.9 | 96.5 | 96.4 | 96.5 | 96.5 | 96.1 | 96.8                        | 97.0 | <b>97.3</b> | 97.1 | 96.7 | 96.7 | 96.7 | 96.6 |
| Flowers  | 97.6                       | 97.8 | 97.9 | 97.8 | 97.6 | 97.6 | 97.7 | 97.7 | 97.5                        | 97.5 | <b>97.9</b> | 97.8 | 97.7 | 97.7 | 97.7 | 96.6 |

TABLE XII  
ACCURACY (%) WITH VARIOUS PROPAGATION STEPS IN GNN LAYERS.

| Iteration | #1          | #2   | #3   | #4   | #5   | #8   | #10  |
|-----------|-------------|------|------|------|------|------|------|
| Aircraft  | <b>95.4</b> | 94.7 | 94.9 | 94.7 | 94.8 | 94.8 | 94.4 |
| Dogs      | <b>97.3</b> | 97.1 | 97.1 | 97.0 | 96.9 | 97.1 | 96.9 |
| Flowers   | <b>97.9</b> | 97.7 | 97.3 | 97.8 | 97.6 | 97.8 | 97.6 |

2) *Number of region proposals:* The impact of different numbers ( $\mathcal{R}$ ) of regions on the accuracy of our SR-GNN is given in Table X. The regions are generated by varying the cell size (Section III-B) as suggested in [8]. Four regions are shown in Fig. 7, and the rest are given in the supplementary document. Different regions are generated by controlling the HOG’s cell size. The best accuracy is achieved for cell size of  $14 \times 14$  *i.e.*,  $\mathcal{R}=27$ . Moreover, our model complexity and per-image inference time with increasing numbers of regions are presented in Table VI. The number of trainable parameters does not depend on the number of regions, whereas the GFLOPs and the per-image inference time increase with the number of regions, as expected.

3) *Impact of the neighborhood size of a given node on accuracy:* The neighborhood size of a given node in our SR-GNN (Fig. 2(b)) is controlled by  $\alpha$  in (2). In order to measure its impact on accuracy, we evaluate various values of  $\alpha \in [0.1, 0.8]$  on the Aircraft, Cars, Dogs and Flowers datasets. The results for GNN output dimensions of 512 and 1024 are shown in Table XI. It is clear that the accuracy increases with the increasing value of  $\alpha$  and reaches a maximum of around 0.3, suggesting the optimal size of the local neighborhood of a given node. We observe a similar trend for the GNN layers with output dimensions of 512 and 1024, respectively. However, for the Dogs and Cars datasets, the accuracy is slightly higher for the latter. This is because different graphs characterize different neighborhood structures.

4) *Number of power iteration steps in GNN:* We have assessed the performance with various power iteration steps  $K$  in (2) in the GNN layers. The iteration steps are varied from  $K=1$  to  $K=10$ , and the results are given in Table XII. For the Aircraft, the accuracy slightly decreases as  $K$  increases. This could be because our SR-GNN advances closer to the global PageRank solution after the first iteration. However, the accuracy variations are marginal for the Dogs and Flowers datasets, and we achieve the best performance with a single propagation step in GNN for all the datasets. This is desired in real-world applications for computational efficiency without loss of accuracy.

## VI. CONCLUSION

We have proposed a novel end-to-end deep network called SR-GNN to enhance the recognition accuracy of fine-

grained objects and human-actions, avoiding any object-parts bounding-box annotation. The model introduces an innovative relation-aware visual feature transformation and its refinement via attentional spatial context modeling to enrich region-level description to capture subtle variations observed and required in FGVC. The model has also proposed a gated attentional pooling to automatically aggregate the relation-aware transformed features. Ultimately, our model’s SotA quantitative and qualitative results on eight benchmark datasets and ablation study show the efficacy of SR-GNN.

In the near future, we will advance our SR-GNN focusing on following key aspects: 1) adapting it to a Graph Transformer Network (GTN) for generating new graph structures to learn a soft selection of connected regions and composite relations for generating useful multi-hop connections to further enhance the recognition accuracy, 2) evaluating SR-GNN on LSVC datasets consisting of distinctive categories (*e.g.*, ImageNet and COCO), and 3) optimizing and extending it to recognize fine-grained actions and activities in videos.

## ACKNOWLEDGEMENT

This research is supported by the UKIERI-DST grant CHARM (UKIERI-2018-19-10), and Research Investment Fund at Edge Hill University. The GPU used in this research was donated by the NVIDIA.

## REFERENCES

- [1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” 2011.
- [2] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in *Indian Conf. on Computer Vision, Graphics & Image Processing*, 2008, pp. 722–729.
- [3] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *Proc. CVPR Workshop on Fine-Grained Visual Categorization*, vol. 2, no. 1, 2011.
- [4] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, “Fine-grained visual classification of aircraft,” *arXiv preprint:1306.5151*, 2013.
- [5] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in *Proc. of the IEEE Int’l Conf. on Computer Vision Workshops*, 2013, pp. 554–561.
- [6] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, “Fine-grained image analysis with deep learning: A survey,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2021.
- [7] A. Behera, Z. Wharton, P. Hewage, and A. Bera, “Context-aware attentional pooling (cap) for fine-grained visual classification,” in *Proc. 35th AAAI Conference on Artificial Intelligence*, pp. 929–937, 2021.
- [8] A. Behera, Z. Wharton, Y. Liu, M. Ghahremani, S. Kumar, and N. Bessis, “Regional attention network (ran) for head pose and fine-grained gesture recognition,” *IEEE Trans. on Affective Computing*, 2020.
- [9] C. Liu, H. Xie, Z.-J. Zha, L. Yu, Z. Chen, and Y. Zhang, “Bidirectional attention-recognition model for fine-grained object classification,” *IEEE Transactions on Multimedia*, 2019.
- [10] A. Bera, Z. Wharton, Y. Liu, N. Bessis, and A. Behera, “Attend and guide (ag-net): A keypoints-driven attention-based deep network for image recognition,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3691–3704, 2021.

- [11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Int. Conf. Learning Representations*, 2017.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
- [14] J. Klicpera, A. Bojchevski, and S. Günnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," in *International Conf. on Learning Representations (ICLR)*, 2019.
- [15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.
- [16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
- [17] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, and A. Yuille, "Transfg: A transformer architecture for fine-grained recognition," *arXiv preprint arXiv:2103.07976*, 2021.
- [18] Z. Miao, X. Zhao, J. Wang, Y. Li, and H. Li, "Complemental attention multi-feature fusion network for fine-grained classification," *IEEE Signal Processing Letters*, vol. 28, pp. 1983–1987, 2021.
- [19] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Processing*, vol. 25, no. 4, pp. 1713–1725, 2016.
- [20] H. Yao, S. Zhang, C. Yan, Y. Zhang, J. Li, and Q. Tian, "Autobd: Automated bi-level description for scalable fine-grained visual categorization," *IEEE Trans. Image Processing*, vol. 27, no. 1, pp. 10–23, 2017.
- [21] W. Ge, X. Lin, and Y. Yu, "Weakly supervised complementary parts models for fine-grained image classification from the bottom up," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3034–3043.
- [22] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 5157–5166.
- [23] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13001–13008.
- [24] Z. Wang, S. Wang, P. Zhang, H. Li, W. Zhong, and J. Li, "Weakly supervised fine-grained image classification via correlation-guided discriminative learning," in *Proc. of the 27th ACM International Conf. on Multimedia*, 2019, pp. 1851–1860.
- [25] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5994–6002.
- [26] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *Int. Journal of Computer Vision*, vol. 127, no. 9, pp. 1235–1255, 2019.
- [27] R. Li, Z. Liu, and J. Tan, "Reassessing hierarchical representation for action recognition in still images," *IEEE Access*, vol. 6, pp. 61386–61400, 2018.
- [28] P. Rodríguez, D. Velazquez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. González, "Pay attention to the activations: a modular attention mechanism for fine-grained image recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 502–514, 2019.
- [29] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 5012–5021.
- [30] Y. Ding, S. Wen, J. Xie, D. Chang, Z. Ma, Z. Si, and H. Ling, "Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *arXiv preprint arXiv:2002.03353*, 2020.
- [31] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, and Y.-Z. Song, "The devil is in the channels: Mutual-channel loss for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4683–4695, 2020.
- [32] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proc. of the IEEE International Conf. on Computer Vision*, 2019, pp. 6599–6608.
- [33] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF International Conf. Computer Vision*, 2021, pp. 1025–1034.
- [34] M. Zhou, Y. Bai, W. Zhang, T. Zhao, and T. Mei, "Look-into-object: Self-supervised structure modeling for object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recognit.*, 2020, pp. 11774–11783.
- [35] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13130–13137.
- [36] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [37] S. Wang, Z. Wang, H. Li, and W. Ouyang, "Category-specific semantic coherency learning for fine-grained image recognition," in *Proc. of the 28th ACM Int'l Conf. on Multimedia*, 2020, pp. 174–183.
- [38] Z. Wang, S. Wang, H. Li, Z. Dou, and J. Li, "Graph-propagation based correlation learning for weakly supervised fine-grained image classification," in *AAAI*, 2020, pp. 12289–12296.
- [39] Y. Zhao, K. Yan, F. Huang, and J. Li, "Graph-based high-order relation discovery for fine-grained recognition," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 15079–15088.
- [40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conf. Comput. Vis. Patt. Recognit.*, 2017, pp. 1251–1258.
- [41] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in *Int. Conf. Learn. Represent. (ICLR)*, 2016.
- [42] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *IEEE Conf. Comput. Vis. Patt. Recog.*, 2015, pp. 595–604.
- [43] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Int. Conf. Computer Vision*. IEEE, 2011, pp. 1331–1338.
- [44] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 9–16.
- [45] W. Luo, X. Yang, X. Mo, Y. Lu, L. S. Davis, J. Li, J. Yang, and S.-N. Lim, "Cross-x learning for fine-grained visual categorization," in *Proc. of the IEEE Int'l Conf. on Computer Vision*, 2019, pp. 8242–8251.
- [46] T. Hu, H. Qi, Q. Huang, and Y. Lu, "See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification," *arXiv preprint arXiv:1901.09891*, 2019.
- [47] K. Xu, R. Lai, L. Gu, and Y. Li, "Multiresolution discriminative mixup network for fine-grained visual categorization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [48] D. Chang, K. Pang, Y. Zheng, Z. Ma, Y.-Z. Song, and J. Guo, "Your 'flamingo' is my 'bird': Fine-grained, or not," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 11476–11485.
- [49] H. Liu, J. Li, D. Li, J. See, and W. Lin, "Learning scale-consistent attention part network for fine-grained image recognition," *IEEE Transactions on Multimedia*, 2021.
- [50] Y. Ding, Z. Ma, S. Wen, J. Xie, D. Chang, Z. Si, M. Wu, and H. Ling, "Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2826–2836, 2021.
- [51] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *Proc. IEEE International Conf. on Computer Vision*, 2019, pp. 8331–8340.
- [52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.
- [53] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Comp. Society Conf. on Comput. Vis. and Patt. Recog.*, 2010, pp. 3360–3367.
- [54] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "Pbc: Polygon-based classifier for fine-grained categorization," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 673–684, 2016.
- [55] M. Ali, J. Gao, and M. Antolovich, "Parametric classification of bingham distributions based on grassmann manifolds," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5771–5784, 2019.
- [56] S. Cai, L. Zhang, W. Zuo, and X. Feng, "A probabilistic collaborative representation based approach for pattern classification," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 2950–2959.
- [57] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Conf. on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1794–1801.
- [58] L. Xie, L. Zheng, J. Wang, A. L. Yuille, and Q. Tian, "Interactive: Inter-layer activeness propagation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 270–279.
- [59] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *IEEE Conf. on Computer Vision and Pattern Recog.*, 2018, pp. 4109–4118.

- [60] A. Rosenfeld and S. Ullman, "Visual concept recognition and localization via iterative introspection," in *Proc. Asian Conf. Comp. Vis.*, 2016, pp. 264–279.
- [61] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 623–634, 2014.
- [62] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 1086–1095.
- [63] R. Lan, Y. Zhou, Z. Liu, and X. Luo, "Prior knowledge-based probabilistic collaborative representation for visual recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1498–1508, 2020.
- [64] J.-F. Hu, W.-S. Zheng, J. Lai, and T. Gong, S. and Xiang, "Recognising human-object interaction via exemplar based modelling," in *Proc. IEEE Conf. Comp. Vis. Patt. Recog. (CVPR)*, 2013, pp. 3144–3151.
- [65] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018.
- [66] A. Rosenfeld and S. Ullman, "Action classification via concepts and attributes," in *24th Int. Conf. Patt. Recognit.*, 2018, pp. 1499–1505.
- [67] L. Zhang, C. Li, P. Peng, X. Xiang, and J. Song, "Towards optimal vlad for human action recognition from still images," *Image and Vision Computing*, vol. 55, pp. 53–63, 2016.
- [68] B. Barz and J. Denzler, "Deep learning on small datasets without pre-training using cosine loss," in *IEEE Winter Conf. on Applications of Computer Vision*, 2020, pp. 1371–1380.
- [69] D. Korsch, P. Bodesheim, and J. Denzler, "Classification-specific parts for improving fine-grained visual categorization," in *German Conf. on Pattern Recognition*. Springer, 2019, pp. 62–75.
- [70] Y. Lavinia, H. Vo, and A. Verma, "New colour fusion deep learning model for large-scale action recognition," *International Journal of Computational Vision and Robotics*, vol. 10, no. 1, pp. 41–60, 2020.
- [71] K. Song, X.-S. Wei, X. Shu, R.-J. Song, and J. Lu, "Bi-modal progressive mask attention for fine-grained recognition," *IEEE Trans. Image Processing*, 2020.
- [72] M. Simon, Y. Gao, T. Darrell, J. Denzler, and E. Rodner, "Generalized orderless pooling performs implicit salient matching," in *Proc. IEEE International Conf. on Computer Vision*, 2017, pp. 4960–4969.
- [73] Z. Li, Y. Ge, J. Feng, X. Qin, J. Yu, and H. Yu, "Deep selective feature learning for action recognition," in *IEEE International Conf. on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.
- [74] M. Simon, E. Rodner, T. Darrell, and J. Denzler, "The whole is more than its parts? from explicit to implicit pose normalization," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 42, no. 3, pp. 749–763, 2020.
- [75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [76] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
- [77] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.
- [78] L. Liu, R. T. Tan, and S. You, "Loss guided activation for action recognition in still images," in *Proc. Asian Conf. Comp. Vis.* Springer, 2018, pp. 152–167.
- [79] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," in *Proc. of the IEEE International Conf. on Computer Vision*, 2017, pp. 3391–3399.
- [80] S. Yan, J. S. Smith, and B. Zhang, "Action recognition from still images based on deep vlad spatial pyramids," *Signal Processing: Image Communication*, vol. 54, pp. 118–129, 2017.
- [81] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *The Jnl. Mach. Learn. Research*, vol. 15, no. 1, pp. 3221–3245, 2014.
- [82] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intelligence*, no. 2, pp. 224–227, 1979.
- [83] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.
- [84] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations (ICLR)*, 2019.
- [85] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *32nd AAAI Conference on Artificial Intelligence*, 2018.
- [86] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8662–8672.
- [87] L. Zhang, S. Huang, and W. Liu, "Enhancing mixture-of-experts by leveraging attention for fine-grained recognition," *IEEE Transactions on Multimedia*, 2021.
- [88] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *European Conf. on Computer Vision (ECCV)*, 2018, pp. 420–435.
- [89] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 4438–4446.



**Asish Bera** received the PhD degree from the Jadavpur University, Kolkata, India, and the M.Tech degree from the IIEST Shibpur, India. He is currently a Post-doctoral research associate at the Computer Science Department, Edge Hill University, UK. His current research interests include computer vision, deep learning, human activity recognition, and biometrics. He is a member of IEEE.



**Zachary Wharton** is currently an MRes student in the Department of Computer Science, Edge Hill University, UK. He obtained his Bachelor's degree in Computing from Edge Hill University in 2019. His interests include computer vision, deep learning, human-robot interaction (HRI) and pattern recognition.



**Yonghuai Liu** is a professor and director of the Visual Computing Lab at Edge Hill University since 2018. He obtained his first PhD in 1997 from Northwestern Polytechnical University, P.R. China and second PhD in 2001 from The University of Hull, UK. He is an area/associate editor or editorial board member for a number of journals and conferences. He has published more than 180 papers in the top-ranked conferences and journals. His research interests lie in 3D computer vision, image processing, pattern recognition, machine learning, AI, and intelligent systems. He is a senior member of IEEE, Fellow of BCS, and Fellow of HEA of the UK.



**Nik Bessis** received his BA from the T.E.I. Athens and his MA and PhD degrees from De Montfort University, UK. He is a full Professor (2010) and since 2015, the Head (Chair) of the Department of Computer Science at Edge Hill University, UK. He is a FHEA, FBCS and a senior member of IEEE. His research is on social graphs for network and big data analytics as well as developing data push and resource provisioning services in IoT, FI and inter-clouds. He is involved in a number of funded research and commercial projects in these areas. Prof Bessis has published over 300 works and won 4 best papers awards.



**Ardhendu Behera** received the PhD degree in Computer Science from the University of Fribourg, Switzerland and MEng degree in System Science and Automation from the Indian Institute of Science (IISc) Bangalore, India. He is currently a Reader in Computer Vision & AI in the Department of Computer Science, Edge Hill University, UK. He has worked as a Research Fellow and Senior Research Fellow in Computer Vision Group at the University of Leeds. He is a Fellow of HEA and member of IEEE, British Machine Vision Association, Applied Vision Association, British Computing Society, affiliated member of IAPR and ECAI. His main interests include computer vision, deep learning, human-robot social interaction, activity analysis and recognition.

## Supplementary Document

The accuracy of our SR-GNN is higher than the state-of-the-art on diverse datasets. To justify the benefits of our model, a precise comparison with the existing top-10 methods focused on the inference time is given in Table XIII, irrespective of the GPU/hardware configuration, deep learning tools (*e.g.*, TensorFlow, PyTorch, MXNet, etc.) and related experimental constraints used in those works. Table XIII is incorporated in Table VII with the model parameters and GFLOPs by comparing with the top-5 SotA approaches. Some approaches in Table XIII do not provide these two metrics. We have also provided the accuracy comparison with those works to reflect a trade-off between the accuracy (%) and inference time in milliseconds (ms). For this purpose, we have specified the best performance of those referred works on a FGVC dataset and our performance and accuracy gain (in parenthesis) on the same dataset.

It shows that our SR-GNN (full-model) stands in the third position and outperforms other eight SotA approaches based on the inference time. From this study, it is evident that SR-GNN requires very competitive inference time with 0.8 ms more than CAP [7]. It is noted that our SR-GNN without Refine module (4.9 ms) shares the second position with MRDMN-L [47] and achieves 7.5% accuracy gain on the Dogs dataset over this approach. On the contrary, SR-GNN computationally lighter and requires lesser parameters and GFLOPs than these two methods, mentioned in Table VII of revised manuscript. Also, the accuracy gain of SR-GNN is the highest on various FGVC datasets compared to these works. In this context, it can be noted that SR-GNN offers an excellent balance to maintain the trade-off between the accuracy, model complexity, and inference time over a diverse category of recent approaches. Therefore, SR-GNN performs the best considering all the aspects of experimental analysis over the existing SotA methods. Particularly, it stands as the second (W/o Refine) and third (full-model) regarding the inference time in comparison with the top-10 SotA methods.

We have included additional visualizations related to our manuscript. (A) Fig. 7 shows all the region proposals ( $\mathcal{R} = 27$ ) and it is related to Fig. 7 in the main paper.

(B) t-SNE plots related to Table VIII (in the main paper) for Cars (Fig. 8) and Flowers (Fig. 10) datasets.

(C) Joint attention maps are shown on Cars (Fig. 10) and Flowers (Fig. 11) datasets, related to Fig. 6 in the main manuscript.

TABLE XIII  
PERFORMANCE COMPARISON BASED ON INFERENCE TIME WITH THE SOTA METHODS

| Sl. No | Method                     | Param (M) | GFLOPs (B) | Infer. time per img. (ms) | Accuracy (%)    | Our accuracy and (+gain) in %               |
|--------|----------------------------|-----------|------------|---------------------------|-----------------|---|
| 1      | CAP [7]                    | 34.2      | 10.2       | 4.2                       | 94.9 (Aircraft) | 95.4 (+ 0.5)                                |
| 2a     | MRDMN-L [47]               | 51.2      | 14.0       | 4.9                       | 89.0 (Dogs)     | 96.5 (+7.5), W/o refine 97.3 (+8.3), SR-GNN |
| 2b     | <b>SR-GNN (W/o Refine)</b> | 24.4      | 9.3        | 4.9                       | Paper Table II  | Section IV-D                                |
| 3      | <b>SR-GNN (Full Model)</b> | 30.9      | 9.8        | 5.0                       |                 |   |
| 4      | AG-Net [10]                | 54.8      | 10.4       | 5.2                       | 97.8 (Stanf.40) | 98.8 (+1.0)                                 |
| 5      | TASN [29]                  | 37.3      | 21.9       | 7.5                       | 87.9 (CUB-200)  | 91.9 (+4.0)                                 |
| 6      | WARN [28]                  | -         | -          | 11.3                      | 85.6 (CUB-200)  | 91.9 (+6.3)                                 |
| 7      | RG [86]                    | -         | -          | 23.8                      | 87.3 (CUB-200)  | 91.9 (+4.6)                                 |
| 8      | SCAPNet [49]               | -         | -          | 24.4                      | 93.6 (Aircraft) | 95.4 (+1.8)                                 |
| 9      | ME-ASN [87]                | -         | -          | 33.9                      | 89.5 (CUB-200)  | 91.9 (+2.4)                                 |
| 10     | NTS-Net [88]               | -         | -          | 35.0                      | 93.9 (Cars)     | 96.1 (+2.2)                                 |
| 11     | RA-CNN [89]                | -         | -          | 36.9                      | 87.3 (Dogs)     | 97.3 (+10.0)                                |

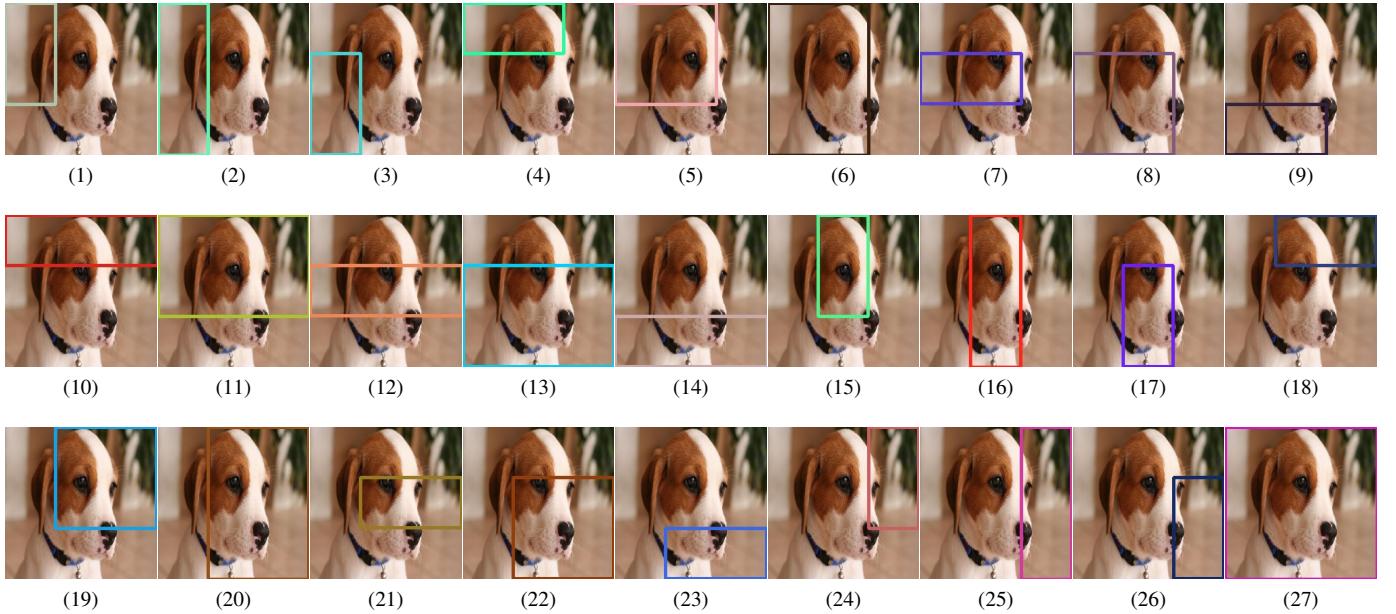


Fig. 7. Bounding box displaying the optimal number ( $\mathcal{R} = 27$ ) of patches/regions in a given input image. These regions are used for bilinear pooling from the upsampled CNN features in Fig. 2(a) (Section III.B). The last region (#27) is the whole image.

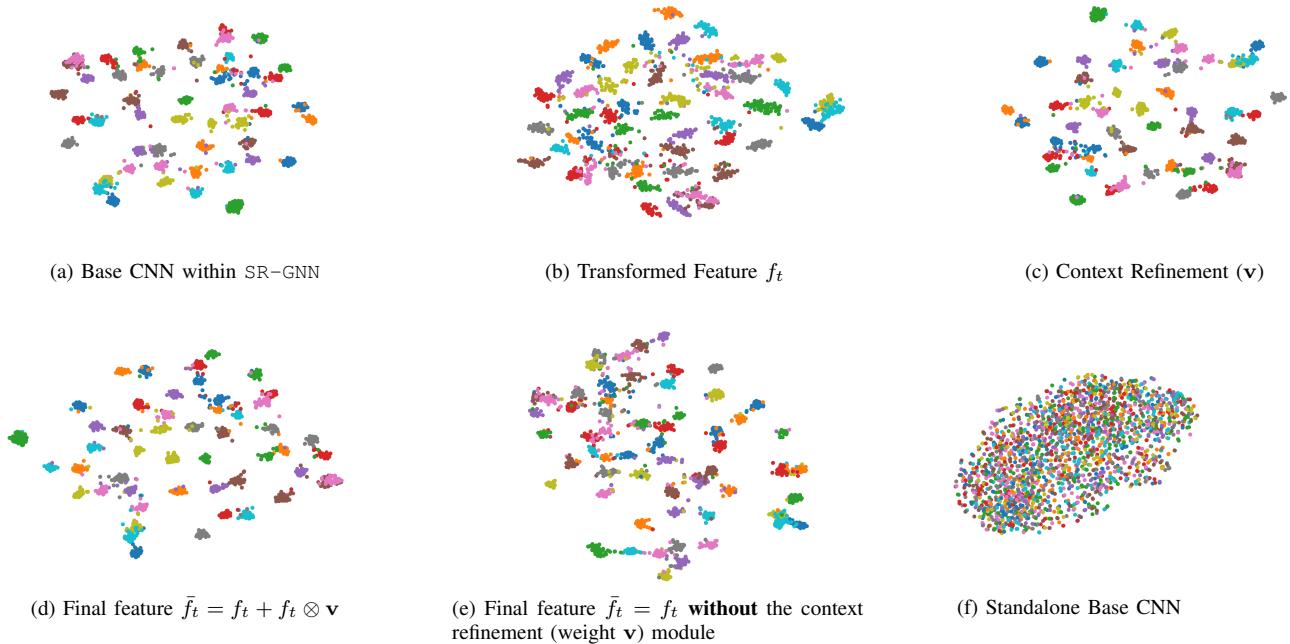


Fig. 8. SR-GNN’s discriminability using t-SNE to visualize class separability and compactness using features from a) base CNN (Xception, Fig. 2(a)) within our model, b) relation-aware transformed feature using GCN (Fig. 2(b)), c) attentional context refinement weight-vector  $\mathbf{v}$  (Fig. 2(c)), and d) the final image-level feature map  $\bar{f}_t$  for classification (Fig. 2(c)). Each color represents a particular class. There are 50 classes chosen randomly from the **Car’s** test set. e) SR-GNN **without** the context refinement module, and f) Standalone Xception base CNN without our modules (re-trained on the Cars dataset).

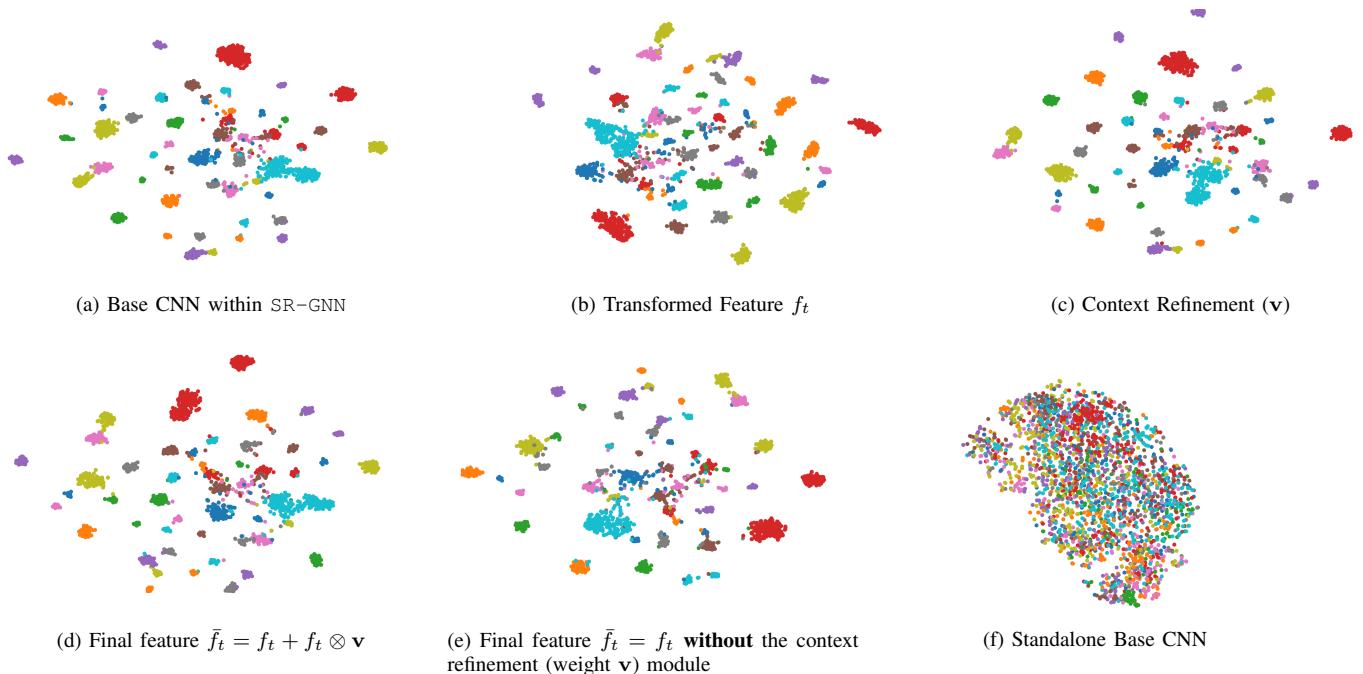
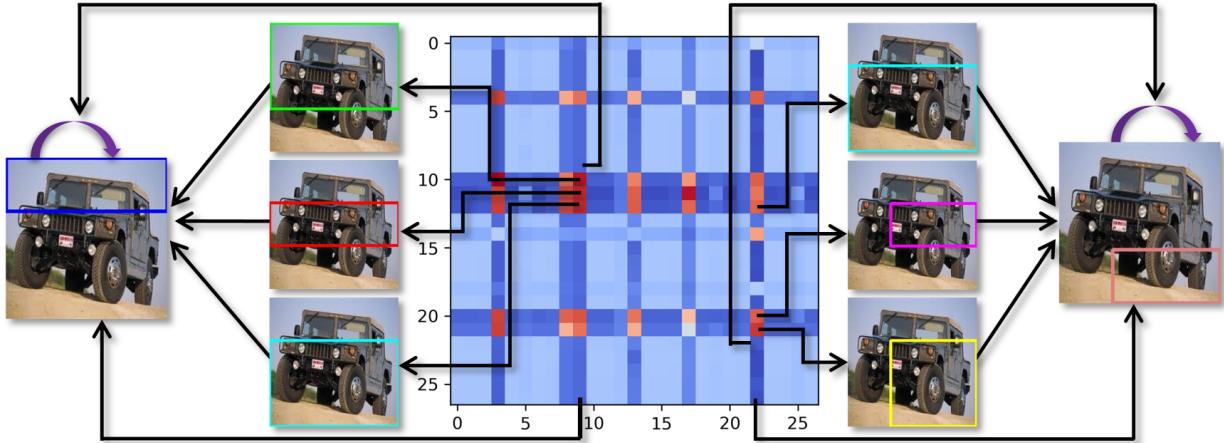
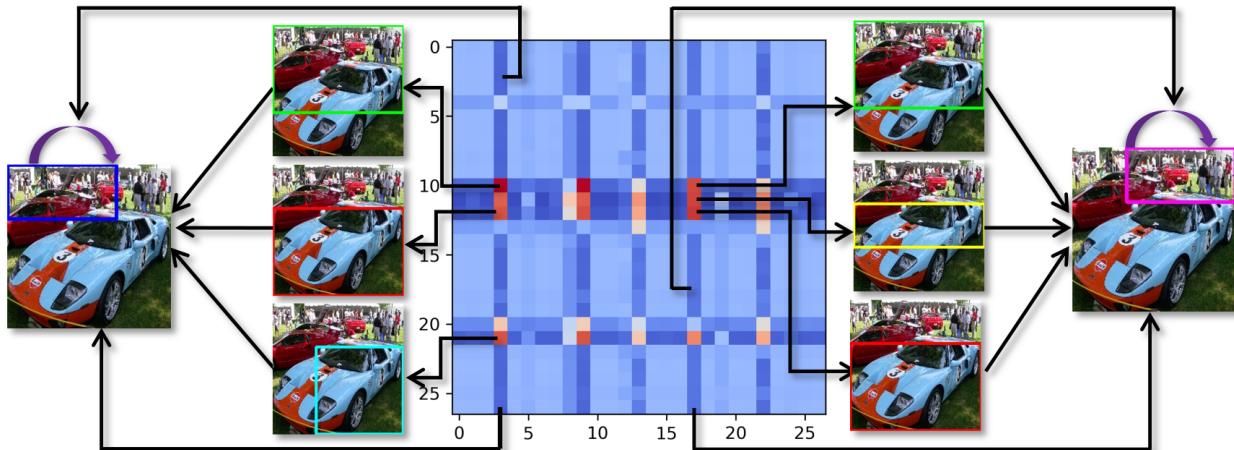


Fig. 9. SR-GNN’s discriminability using t-SNE to visualize class separability and compactness using features from a) base CNN (Xception, Fig. 2(a)) within our model, b) relation-aware transformed feature using GNN (Fig. 2(b)), c) attentional context refinement weight-vector  $\mathbf{v}$  (Fig. 2(c)), and d) the final image-level feature map  $\bar{f}_t$  for classification (Fig. 2(c)). Each color represents a particular class. There are 50 classes chosen randomly from the **Flower’s** test set. e) SR-GNN **without** the context refinement module, and f) Standalone Xception base CNN without our modules (re-trained on the Flowers dataset).



(a) AM General Hummer SUV 2000: top-2 regions (9 &amp; 22) and their top-3 joint attentions (Fig. 2c)



(b) Ford GT Coupe 2006: top-2 regions (3 &amp; 17) and their top-3 joint attentions (Fig. 2c)

Fig. 10. Visualization within attentional context refinement (Fig. 2(c)): a) Top-2 regions (cols 9 & 22) contributing towards sub-type ‘AM General Hummer SUV 2000’ conditioned on the respective other top-3 regions (rows) in joint decision-making. The self-attention (self-loop) is also shown in the top-2 regions. b) Similarly, top-2 regions (cols 3 & 17) contributing towards sub-type ‘Ford GT Coupe 2006’ conditioned on the respective other top-3 regions (rows). Region proposals are shown in the respective original images.

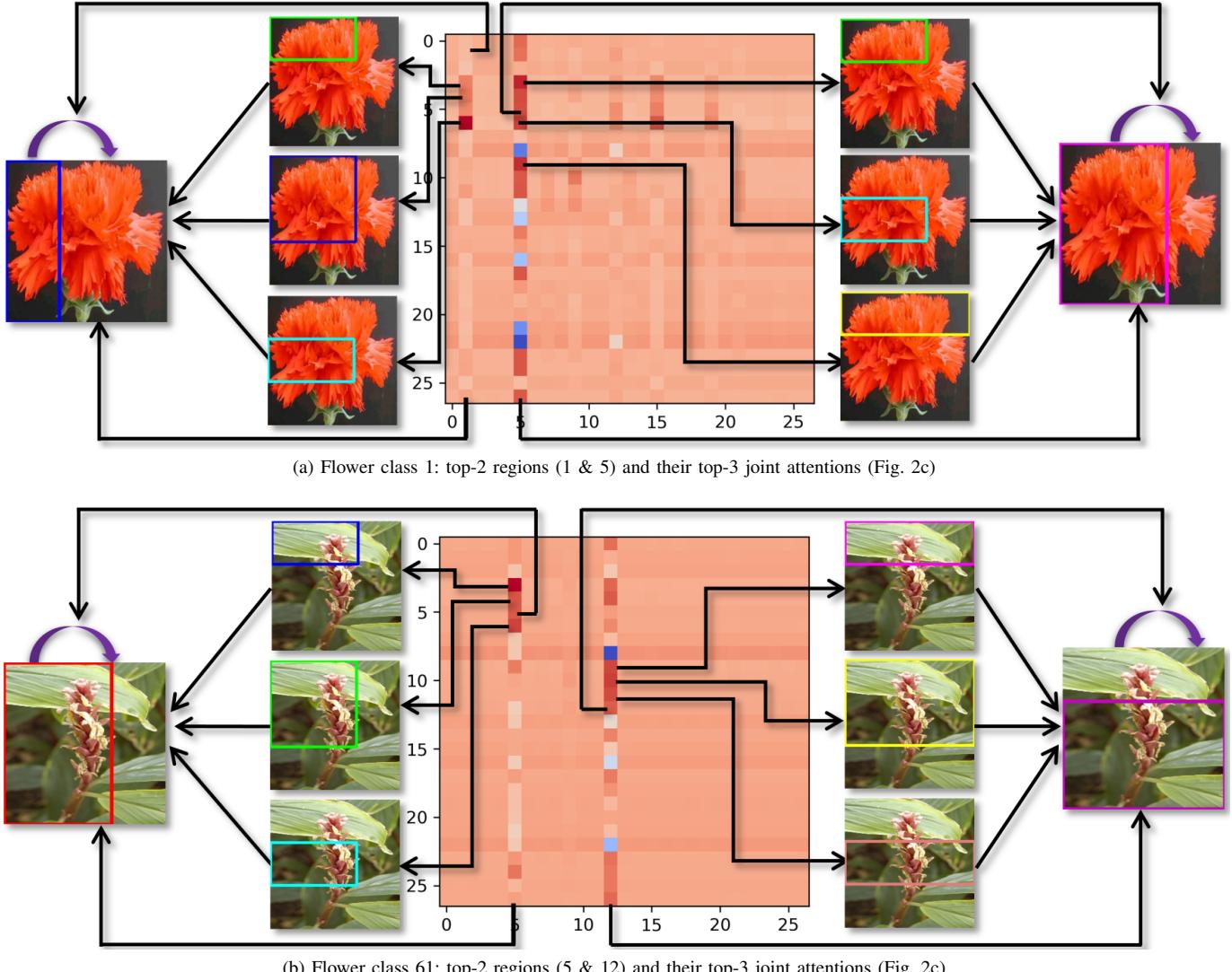


Fig. 11. Visualization within attentional context refinement (Fig. 2(c)): a) Top-2 regions (cols 1 & 5) contributing towards sub-type ‘Flower class 1’ conditioned on the respective other top-3 regions (rows) in joint decision-making. The self-attention (self-loop) is also shown in the top-2 regions. b) Similarly, top-2 regions (cols 5 & 12) contributing towards sub-type ‘Flower class 61’ conditioned on the respective other top-3 regions (rows). Region proposals are shown in the respective original images.