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

Supplementary Document

BMVC 2021 Submission # 1518

In this supplementary document, the remaining quantitative and qualitative results are presented. A few additional supporting experimental results are also included.

Dataset Description: Details about the datasets with the state-of-the-arts (SotA), and the accuracy of proposed method are given in Table 5 (cf. lines: 252-253, and 437-438 in the paper).

Dataset	#Train / #Test	#Class	SotA	Proposed
Aircraft-100 [36]	6,667 / 3,333	100	CAP [4]: **94.9**	**94.9**
Flowers-102 [38]	2,040 / 6,149	102	CAP [4]: 97.7	**98.7**
Oxford-IIIT Pets-37 [39]	3,680 / 3,669	37	CAP [4]: 97.3	**98.1**
CIFAR-100 [30]	50,000 / 10,000	100	BOT [77]: 83.5	**83.8**
Caltech-256 [19]	15,360 / 14,420	256	CPM [18]: 94.3	**96.2**

Table 5: For evaluation, datasets consisting of fine-grained (Aircraft-100, Flowers-102, and Pets-37) and generic (CIFAR-100 and Caltech-256) visual classification are used. Accuracy (%) of our model in comparison to the best SotA.

Attention Type	Attention Heads	Aircraft	Flowers	Pets
Concatenate	**3**	**94.9**	**98.7**	**98.1**
Average	2	85.5	97.8	97.3
Average	3	90.2	98.5	97.6
Average	4	90.8	98.7	98.0

Table 6: More results of Table 3 in the main paper using average of different attention head's outputs versus their concatenation. The concatenation result is presented in Table 3, and the best accuracy is achieved using 3 attention heads with output dimension of 512. In this table, the accuracy with *averaging* is presented. It is observed that the performance using averaging increases with the number of attention heads. However, the model complexity (number of trainable parameters and GFLOPs) also increases with the number of attention heads as shown in Table 7. Thus, concatenation using an optimal number of attention heads (H=3) is preferred. This has been specified in the main paper (cf. lines: 207-208).

Additional results of Table 3 (concatenation vs averaging) in the main article: The remaining results of Table 3 by comparing concatenation with averaging the outputs from *multi-head attention* in (2). It is found that the concatenation is better than the averaging.

The results are provided in Table 6 (cf. lines: 438-439 in the paper). The performance of average aggregation increases with the number of heads. However, the computational complexity (number of trainable parameters and GFLOPs) also increases with the number of attention heads as shown in Table 7.

Clusters K	Channels	Attention Heads	Trainable Parameters	GFLOPs	Per-frame inference time in milliseconds (ms)
8	256	2	27,473,088	13.206	8.0
8	256	3	29,428,928	13.208	8.5
8	256	4	31,515,840	13.210	8.6
8	512	2	31,515,840	13.210	8.5
8	512	3	36,082,880	13.215	8.5
8	512	4	41,174,208	13.220	8.6
16	512	3	36,095,176	13.219	8.5
20	512	3	36,101,324	13.222	8.6
32	512	3	36,119,768	13.229	8.6
36	512	3	36,125,916	13.231	8.6
40	512	3	36,132,064	13.233	8.6
48	512	3	36,144,360	13.238	8.7

Table 7: Statistics about how the various hyper-parameters (#K,#H, and the dimension of H) affect the complexity of our model. This has been mentioned in line-360-361 of the main article. The number of clusters K in soft clustering-based graph pooling does have a little impact on the model complexity (bottom six rows). The number of attention heads and their output dimensions (256 or 512) influence the complexity i.e., higher number of attention heads combined with larger dimension increase the complexity. However, there is a little impact of these values on GFLOPs and inference time in milliseconds.

Dataset	Top-1 Acc	Top-2 Acc	Top-5 Acc	Top-10 Acc
Aircraft-100	94.9	98.8	99.6	99.8
Flowers-102	98.7	99.6	99.9	100.0
Pets-37	98.1	99.8	100.0	100.0
Caltech-256	96.2	99.0	99.7	99.8
CIFAR-100	83.8	89.3	92.0	93.6

Table 8: Top-N accuracy (in %) of the proposed model using optimal number of attention heads H=3 with output dimensions of 512 and L=3 layers in the hierarchical representation. The top-2 accuracy is around 99% except CIFAR-100. Similarly, the top-5 accuracy is nearly 100% (except CIFAR-100). This shows the effectiveness of the proposed model.

Model complexity: We could not include more details about the model complexity of our method in the main paper (Section 4.2). It is presented here in Table 7 (cf. lines: 360-361 & 439-440).

Top-N Accuracy (%): We have also evaluated the proposed approach using top-N accuracy metric on Aircraft-100 [36], Oxford-Flowers-102 [38], Oxford-IIIT Pets [39], CIFAR-100 [30], and Caltech-256 [19] datasets. Our model's performance is presented in Table 8 (cf. line: 440 in the paper). All datasets except CIFAR-100, the top-2 accuracy is around 99%. Moreover, their top-5 accuracy is nearly 100%. It clearly reflects the efficiency of our proposed method to enhance the performance of both FGVC and generic object recognition.

Visualization and Analysis

We have provided additional qualitative results of our method which are mentioned in the Section 4.2, cf. lines: 389 - 390 and Section 4.3 lines: 440 - 443.

1) Example of the regions linking various layers to visualize the hierarchical structure is shown in Fig. 4-5 (cf. lines: 440 - 441 in the paper).

2) Cluster-specific contributions of the graph-based regions are shown in Fig. 6-8 (cf. lines: 441 - 442 in the paper)

3) t-SNE [50] analysis of layer-wise attention heads are shown in Fig. 9-12 (cf. lines: 442 - 443 in the paper)

(a) An example image from the Aircraft dataset

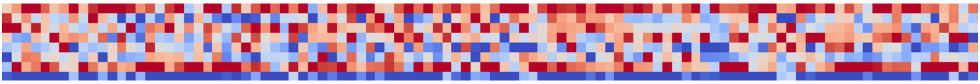
(b) Layer 1 regions in our hierarchical structure

(c) Layer 2 regions in our hierarchical structure

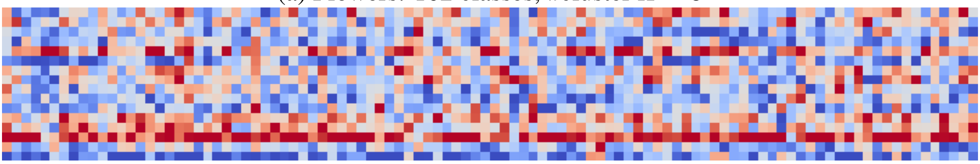
Figure 4: Layer-wise regions of fixed area but with different aspect ratios corresponding to a given hierarchical layer are generated using the region proposal algorithm in [3]. In this example, we consider 3-layer hierarchical structure consisting of 52 regions. The original image is shown in (a).

(d) Layer 3 regions in our hierarchical structure

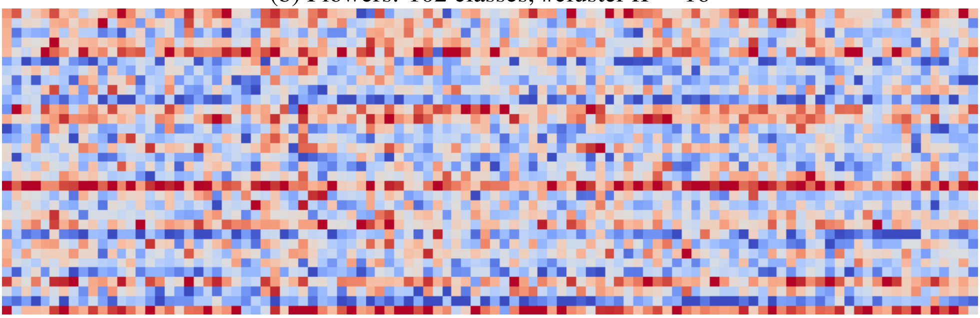
Figure 5: Layer-wise regions of fixed area but with different aspect ratios corresponding to a given hierarchical layer are generated using the region proposal algorithm in [3]. In this example, we consider 3-layer hierarchical structure consisting of 52 regions. The original image is shown in Fig. 4. (a).



(a) Flowers: 102 classes, #cluster $K = 8$

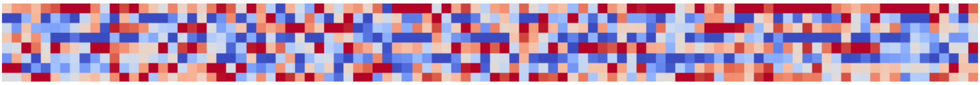


(b) Flowers: 102 classes, #cluster $K = 16$

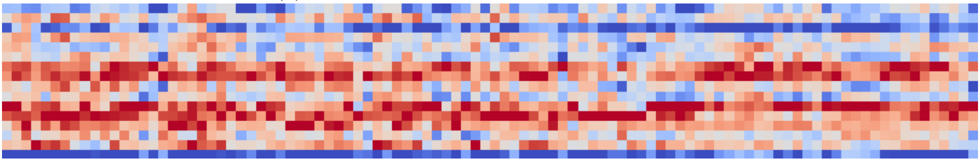


(c) Flowers: 102 classes, #cluster $K = 32$

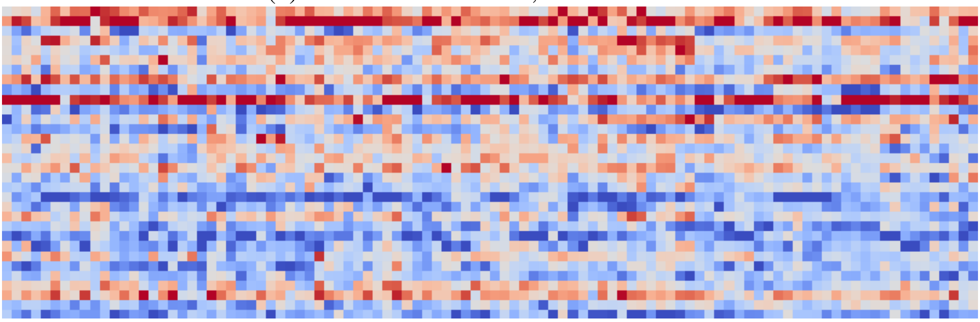
Figure 6: Visualization of the cluster-specific contributions (i.e. weights, cool to warm $\Rightarrow$ less to more) from the graph representation of regions towards a given category during the spectral clustering-based graph pooling. The y-axis (rows) represents K (coarser representation) and the x-axis (cols) shows the number of classes. Each column is different, representing the feature discriminability during the decision making process. All test images from the **Oxford-Flowers-102** dataset are used to compute weights.



(a) Aircraft - 100 classes, #cluster $K = 8$

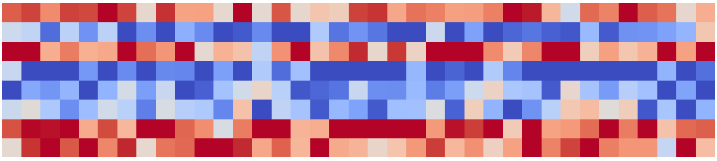


(b) Aircraft - 100 classes, #cluster $K = 16$

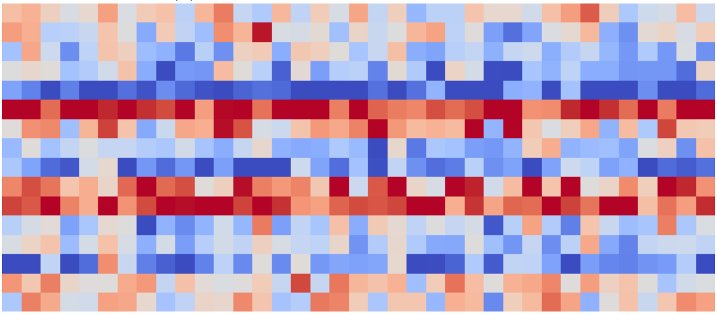


(c) Aircraft - 100 classes, #cluster $K = 32$

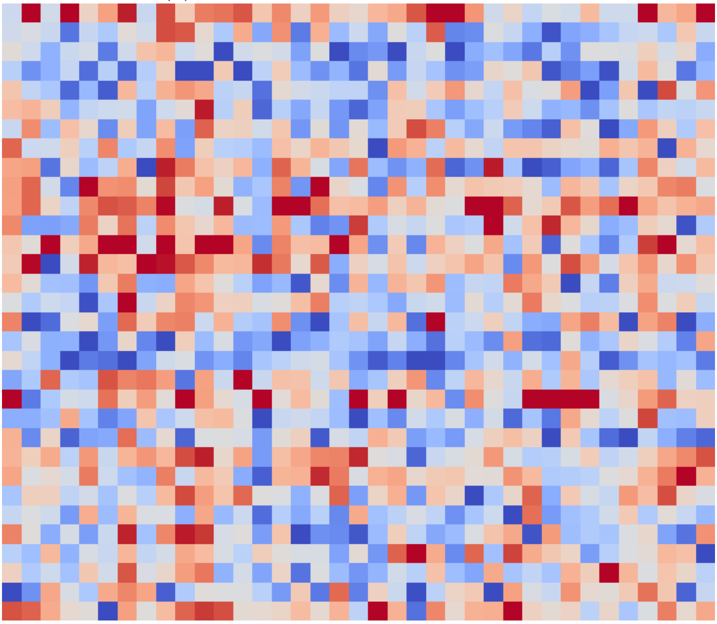
Figure 7: Visualization of the cluster-specific contributions (i.e. weights, cool to warm $\Rightarrow$ less to more) from the graph representation of regions towards a given category during the spectral clustering-based graph pooling. The y-axis (rows) represents K (coarser representation) and the x-axis (cols) shows the number of classes. Each column is different, representing the feature discriminability during the decision making process. All test images from the **Aircraft-100** dataset are used to compute weights. Figures (a)-(b) are shown in Fig.3 in the main paper.



(a) Pets - 37 classes, #cluster $K = 8$

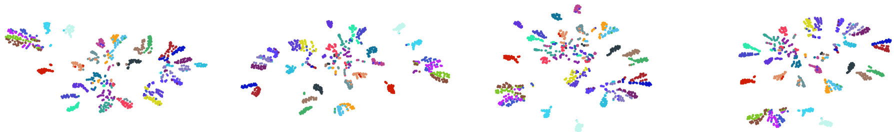


(b) Pets - 37 classes, #cluster $K = 16$

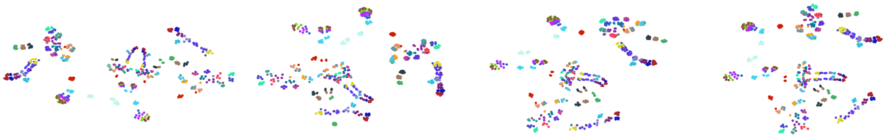


(c) Pets - 37 classes, #cluster $K = 32$

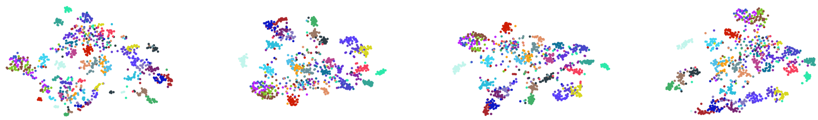
Figure 8: Visualization of the cluster-specific contributions (i.e. weights, cool to warm ⇒ less to more) from the graph representation of regions towards a given category during the spectral clustering-based graph pooling. The y-axis (rows) represents K (coarser representation) and the x-axis (cols) shows the number of classes. Each column is different, representing the feature discriminability during the decision making process. All test images from the **Oxford-IIIT Pets-37** dataset are used to compute weights.



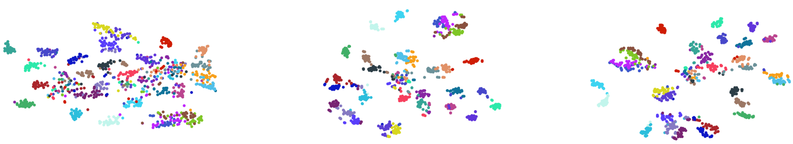
(a) Layer 1 (left to right): head_1, head_2, head_3, and their concatenation



(b) Layer 2 (left to right): head_1, head_2, head_3, and their concatenation

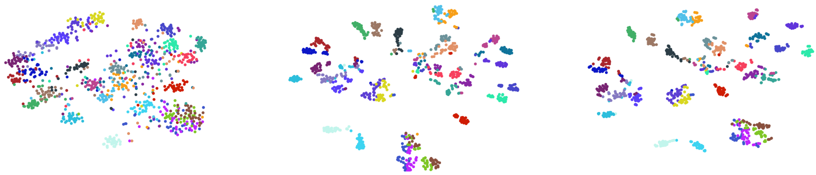


(c) Layer 3 (left to right): head_1, head_2, head_3, and their concatenation

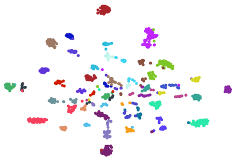


(d) Final representation (left to right): using 2, 3 and 4 attention heads

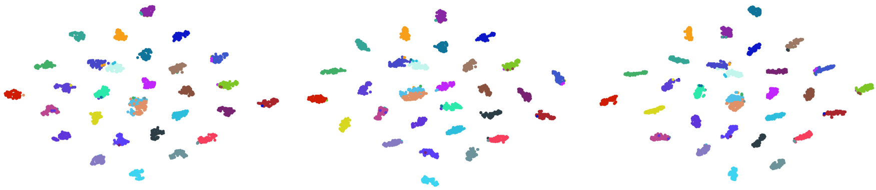
Figure 9: **For clarity, repetition of Fig. 2 in the main article with larger size**. t-SNE [50] visualization of class-specific discriminative feature representation of multi-scale hierarchical regions using $H = 3$ attention heads in (2), and $L = 3$ layers hierarchical structure in (1). All test images from 30 randomly chosen classes within Aircraft dataset are used. Attention head-specific plots are shown in $(a) \rightarrow (c)$, representing layers from smaller regions (a) to larger ones (c). It is evident that the discriminability of the features representing medium-size regions (b) > small-size (a) > large-size (c). (d) shows the combined layers' representation using 2, 3 and 4 attention heads. More than 2 attention heads has shown better discriminability.



(a) Aircraft (left to right): $K = 8, 16, \text{and } 32$

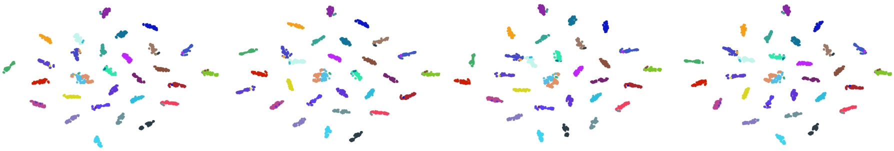


(b) Flowers (left to right): $K = 8, 16, \text{and } 32$

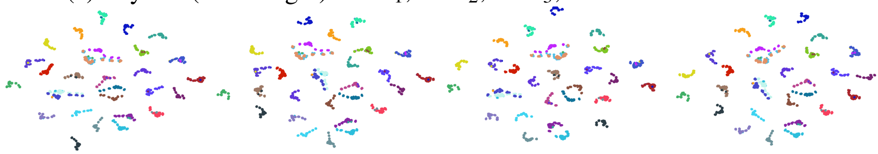


(c) Pets (left to right): $K = 8, 16, \text{and } 32$

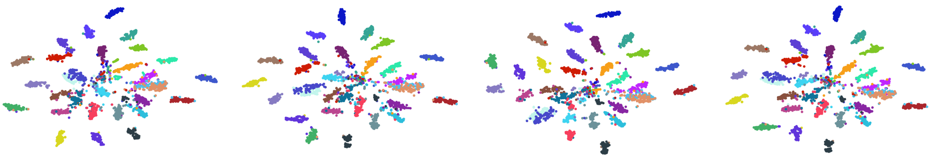
Figure 10: t-SNE [50] visualization of class-specific discriminative feature representing different clusters K (coarser representation) to aggregate graph structure-driven regions via spectral clustering-based graph pooling (Fig. 1c). All test images from 30 randomly chosen classes within a dataset are used for the visualization.



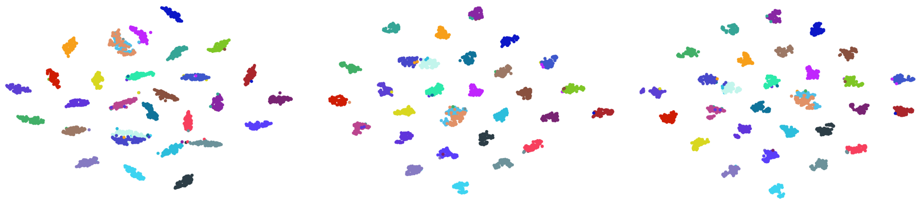
(a) Layer 1 (left to right): head_1, head_2, head_3, and their concatenation



(b) Layer 2 (left to right): head_1, head_2, head_3, and their concatenation

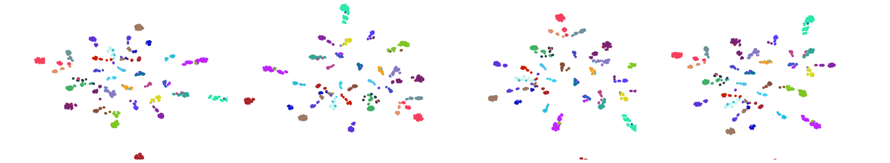


(c) Layer 3 (left to right): head_1, head_2, head_3, and their concatenation

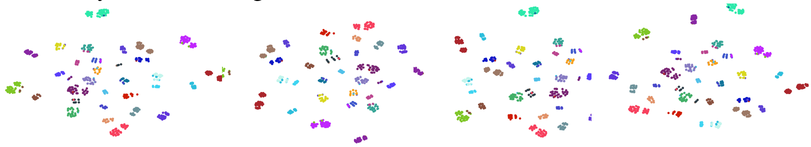


(d) Final representation (left to right): using 2, 3 and 4 attention heads

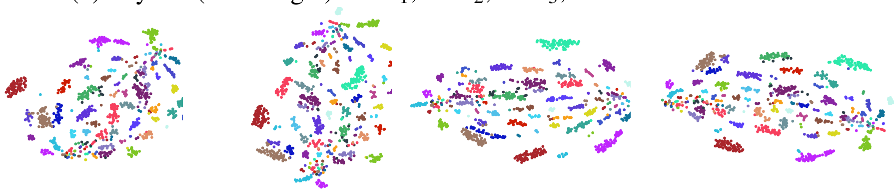
Figure 11: t-SNE [50] visualization of class-specific discriminative feature representation of multi-scale hierarchical regions using H=3 attention heads in (2), and L=3 layers hierarchical structure in (1). All test images from 30 randomly chosen classes within **Oxford-IIIT Pets-37** dataset are used. Attention head-specific plots are shown in $(a) \rightarrow (c)$, representing layers from smaller regions (a) to larger ones (c). It is evident that the discriminability of the features representing medium-size regions (b) > small-size (a) > large-size (c). (d) shows the combined layers' representation using 2, 3 and 4 attention heads. More than 2 attention heads has shown better discriminability.



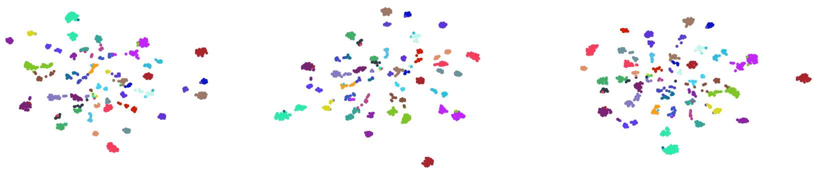
(a) Layer 1 (left to right): head$_1$, head$_2$, head$_3$, and their concatenation



(b) Layer 2 (left to right): head$_1$, head$_2$, head$_3$, and their concatenation



(c) Layer 3 (left to right): head$_1$, head$_2$, head$_3$, and their concatenation



(d) Final representation (left to right): using 2, 3 and 4 attention heads

Figure 12: t-SNE [50] visualization of class-specific discriminative feature representation of multi-scale hierarchical regions using H=3 attention heads in (2), and L=3 layers hierarchical structure in (1). All test images from 30 randomly chosen classes within **Oxford-Flowers-102** dataset are used. Attention head-specific plots are shown in $(a) \rightarrow (c)$, representing layers from smaller regions (a) to larger ones (c). It is evident that the discriminability of the features representing medium-size regions (b) > small-size (a) > large-size (c). (d) shows the combined layers' representation using 2, 3 and 4 attention heads. More than 2 attention heads has shown better discriminability.