Context-aware Attentional Pooling (CAP) for Fine-grained Visual Classification

In this document, we have included the remaining quantitative and qualitative results, which we could not include in the main document.

Remaining results of Table 2: The performance comparison (accuracy in %) using the remaining two datasets (Stanford Dogs and Oxford Flowers) for Table 2 in the main paper. It is presented in Table 6 below.

Table 6: Performance comparison with the recent top-five SotA approaches on each dataset. Methods marked with * involve transfer/joint learning strategy for objects/patches/regions consisting more than one dataset (primary and secondary)

Stanford Dogs		**Oxford Flowers**	
Method	Accuracy (%)	Method	Accuracy (%)
FCANs (Liu et al. 2016)	89.0	InterAct (Xie et al. 2016)	96.4
SJFT* (Ge and Yu 2017)	90.3	SJFT* (Ge and Yu 2017)	97.0
DAN (Hu et al. 2019)	92.2	OPAM* (Peng, He, and Zhao 2018)	97.1
WARN (Rodríguez et al. 2020)	92.9	DSTL* (Cui et al. 2018)	97.6
CPM* (Ge, Lin, and Yu 2019)	**97.1**	MC_{Loss}* (Chang et al. 2020)	**97.7**
Proposed	96.1	**Proposed**	**97.7**

Remaining results of Table 3: The accuracy of the proposed method is evaluated on the **NABirds** dataset using six different SotA base CNNs for Table 3 in the main paper. It is presented in Table 7 below.

Table 7: Our model's accuracy (%) on the **NABirds** dataset with different SotA base CNN architectures. Previous best accuracy is 86.4% (Luo et al. 2019) for primary only and 87.9% (Cui et al. 2018) for combined primary and secondary datasets.

Base CNN	Accuracy(%)
ResNet-50	88.8
Inception V3	89.1
Xception	**91.0**
Densenet	88.3
NASNet-M	88.7
Mobile-Net V2	89.1

Remaining results of Table 4: In ablation study (Table 4 of the main paper), we have presented the performance of the proposed model (with the addition of our novel context-aware attentional pooling (+C) and classification (+E) module) on the Aircraft, Stanford Cars and Oxford-IIIT Pets datasets. The same evaluation procedure is performed on the Stanford Dogs, Oxford Flowers and Caltech Birds (CUB-200) datasets and the recognition accuracy (%) is presented in Table 8. Like in Table 4, a similar trend is observed in the improvement of accuracy when our context-aware attentional pooling (+C) and classification (+E) modules are added to various SotA base CNN architectures (B).

Table 8: Accuracy (%) of the proposed model with the addition of our novel context-aware attentional pooling (+C) and classification (+E) module to various SotA base (B) CNN architectures. It presents the remaining evaluation of Table 4.

	Stanford Dogs			**Oxford Flowers**			**Caltech Birds: CUB-200**		
Base CNN	B	B+C	B+C+E	B	B+C	B+C+E	B	B+C	B+C+E
Inception-V3	78.7	94.2	95.7	92.3	94.9	97.6	76.0	87.1	91.4
Xception	82.7	94.8	96.1	91.9	94.9	97.7	75.6	87.4	91.8
DenseNet121	79.5	94.5	95.5	94.4	95.1	97.6	79.1	87.2	91.6
NASNetMobile	79.5	94.7	96.0	90.7	95.0	97.7	73.0	86.8	89.7
MobileNetV2	76.5	94.3	95.9	92.3	95.0	97.4	74.5	87.0	89.2
Previous Best	(Ge et al. 2019)		93.9	(Xie et al. 2016)		96.4	(Ge et al. 2019)		90.3

Remaining results of Table 5: The performance is evaluated using a different number of integral regions on the Aircraft and Stanford Cars datasets (Table 5). The same experiment is also carried out on the Stanford Dogs dataset, and the results are given in Table 9 below.

Table 9: Accuracy (%) of our model with numbers of 9, 27, and 36 integral regions on **Stanford Dogs** dataset.

Base CNN	#9	#27	#36
ResNet-50	90.5	95.8	92.1
Xception	95.3	96.1	95.2
NASNet-M	91.7	96.0	93.3

Top-N Accuracy (%): We have also evaluated the proposed approach using top-N accuracy metric on Oxford-IIIT Pets, Stanford Cars and Aircraft datasets. The performance of our modules on top of various base architectures is presented in Table 10 below. On all three datasets, the top-2 accuracy is around 99% and is independent of the type of base CNN architecture used. Moreover, the top-5 accuracy is nearly 100%. This justifies the significance of our novel attentional pooling and encoding modules in enhancing performance and their wider applicability.

Table 10: Top-N accuracy (in %) of the proposed model using different base architectures on Oxford-IIIT Pets, Stanford Cars and Aircraft datasets. The top-2 accuracy is around 99% and is independent of the type of base CNN architecture used. The top-5 accuracy is nearly 100%. This shows the significance of the proposed attentional pooling and encoding modules.

Dataset	Base CNN architecture	Top 1	Top 2	Top 3	Top 5
Oxford-IIIT Pets	Inception-V3	96.2	99.0	99.5	99.9
	Xception	97.0	99.7	99.9	99.9
	DenseNet121	96.9	99.2	99.6	99.7
	NASNetMobile	97.3	99.4	99.8	99.9
	MobileNetV2	96.4	98.9	99.5	99.6
Stanford Cars	Inception-V3	94.8	99.4	99.7	99.8
	Xception	95.7	99.3	99.7	99.8
	DenseNet121	93.6	98.7	99.5	99.9
	NASNetMobile	93.7	99.1	99.7	99.8
	MobileNetV2	94.0	99.3	99.8	99.9
Aircraft	Inception-V3	94.8	99.1	99.7	99.8
	Xception	94.1	98.9	99.2	99.5
	DenseNet121	94.6	98.8	99.3	99.4
	NASNetMobile	93.8	99.4	99.8	99.8
	MobileNetV2	94.4	99.1	99.7	99.8

Additional Qualitative Analysis:
We have provided the additional qualitative analysis of our model's performance by selecting a few example images, which are wrongly classified against the label they are mistaken for (selected from the mistaken subcategories). This is presented in Figure 5. It is evident that the mistaken labels come from classes with extremely similar features, often being from the same manufacturer (Boeing 747, Audi, etc.). We have also noticed that subcategories can have very specific defining features that are not clearly visible in every image due to poor angles or lighting conditions (e.g. The chin of a Ragdoll and legs of a Birman cat shown in Fig. 5g).

We have also included an additional qualitative analysis of discriminating ability (Figure 6 to Figure 10) of our model using t-SNE to visualize class separability and compactness on the different datasets as well as various backbone CNNs.


(a) 747-200 vs 747-100


(b) 747-300 vs 747-400


(c) C-47 vs DC-3


(d) Audi TTS Coupe 2012 vs Audi TT RS Coupe 2012


(e) Bentley Continental GT Coupe 2012 vs Bentley Continental GT Coupe 2007


(f) Chevrolet Express Cargo Van 2007 vs Chevrolet Express Van 2007


(g) Birman vs Ragdoll


(h) American Pitbull Terrier vs American Bulldog


(i) Staffordshire Bull Terrier vs American Pitbull Terrier


(j) Spotted Catbird vs Gray Catbird


(k) Red Winged Blackbird vs Brewer Blackbird

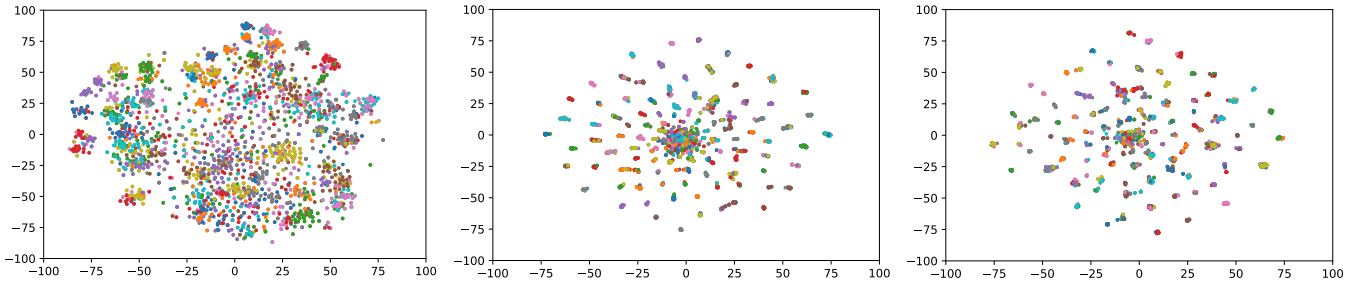

(l) Laysan Albatross vs Sooty Albatross


(m) English Marigold vs Dandelion

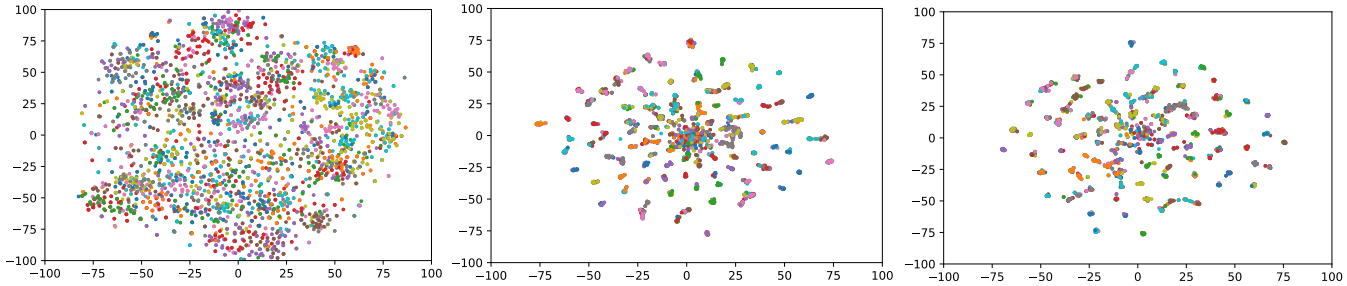

(n) Sweet Pea vs Lenten Rose


(o) Clematis vs Hibiscus

Figure 5: Some of the example images, which are incorrectly classified by our model (left) against the label they are mistaken for (right - selected from the mistaken subcategories): Aircraft (a-c), Stanford Cars (d-f), Oxford-IIIT Pets (g-i), Caltech-UCSD Birds - CUB-200 (j-l), and Oxford Flowers (m-o). It can be seen that the mistaken labelling comes from classes with extremely similar appearance features and/or perspective changes, often being from the same manufacturer (Boeing 747, Audi, etc.). We have also noticed that subcategories can have very specific defining features that are not clearly visible in every image due to poor angles or lighting conditions (e.g. The chin of a Ragdoll and legs of a Birman cat).

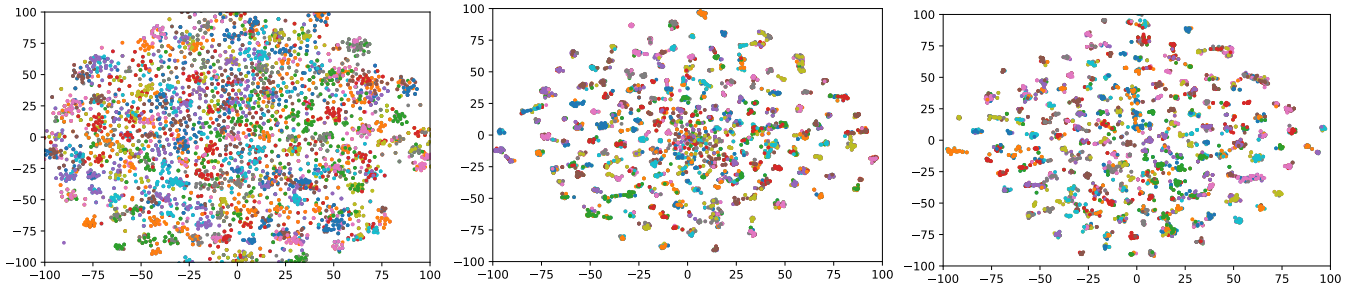


(a) Base CNN (Inception-V3) (b) Attentional Pooling (Inception-V3) (c) CAP + Encoding (Inception-V3)

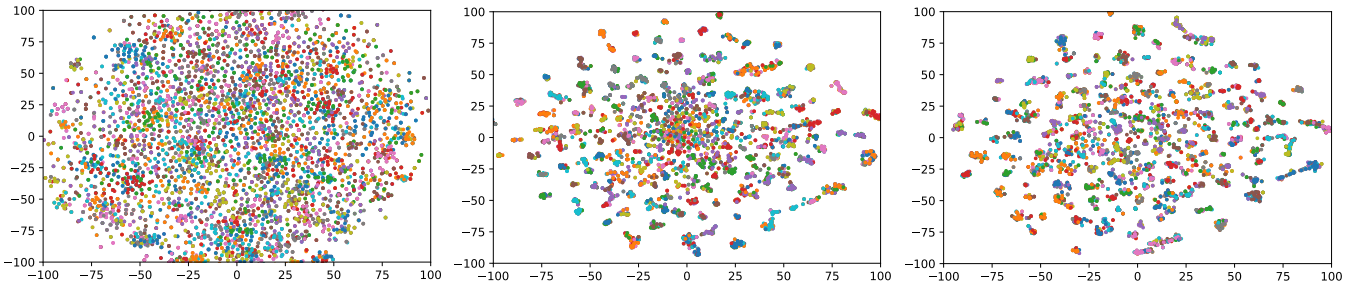


(d) Base CNN (NASNetMobile) (e) Attentional Pooling (NASNetMobile) (f) CAP + Encoding (NASNetMobile)

Figure 6: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Aircraft** test images using Inception-V3 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.

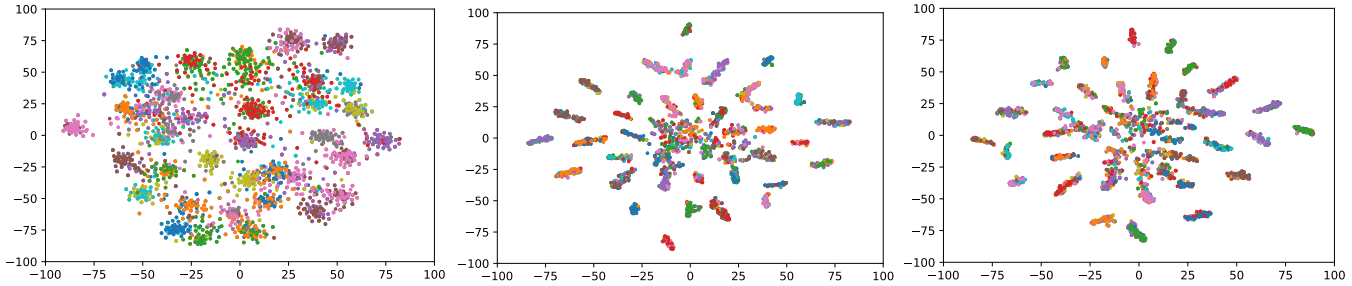


(a) Base CNN (Inception-V3) (b) Attentional Pooling (Inception-V3) (c) CAP + Encoding (Inception-V3)

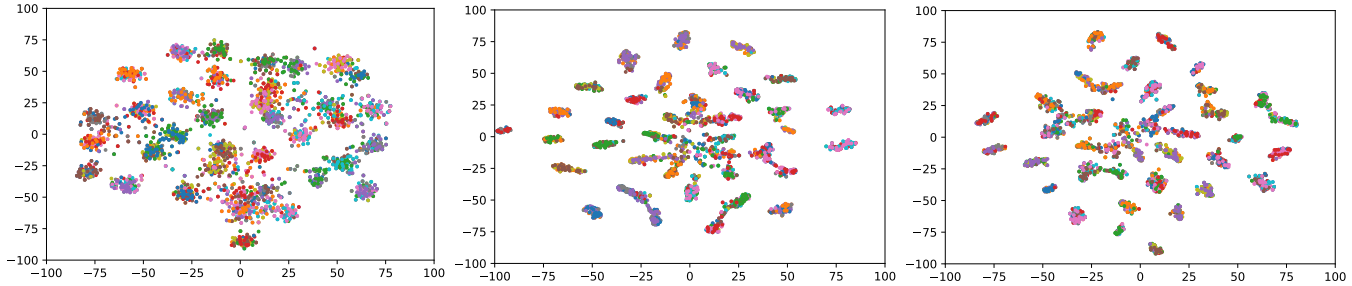


(d) Base CNN (NASNetMobile) (e) Attentional Pooling (NASNetMobile) (f) CAP + Encoding (NASNetMobile)

Figure 7: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Stanford Cars** test images using Inception-V3 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.

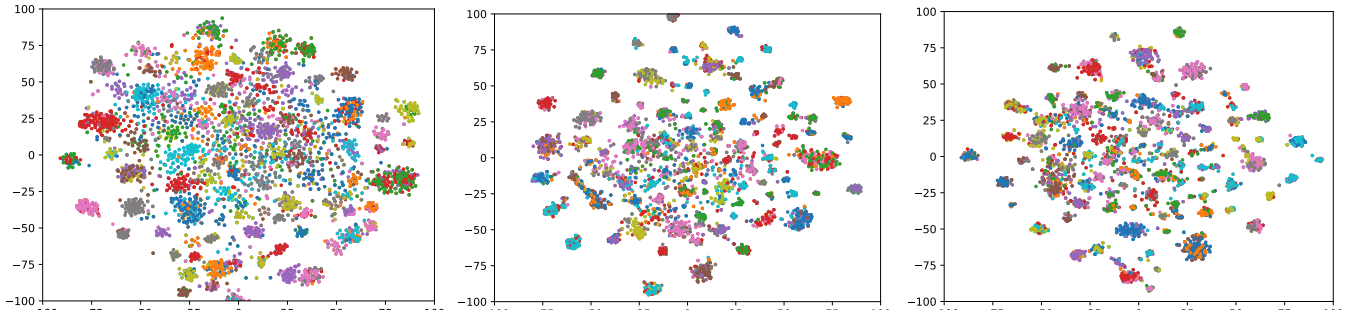


(a) Base CNN (Inception-V3) (b) Attentional Pooling (Inception-V3) (c) CAP + Encoding (Inception-V3)

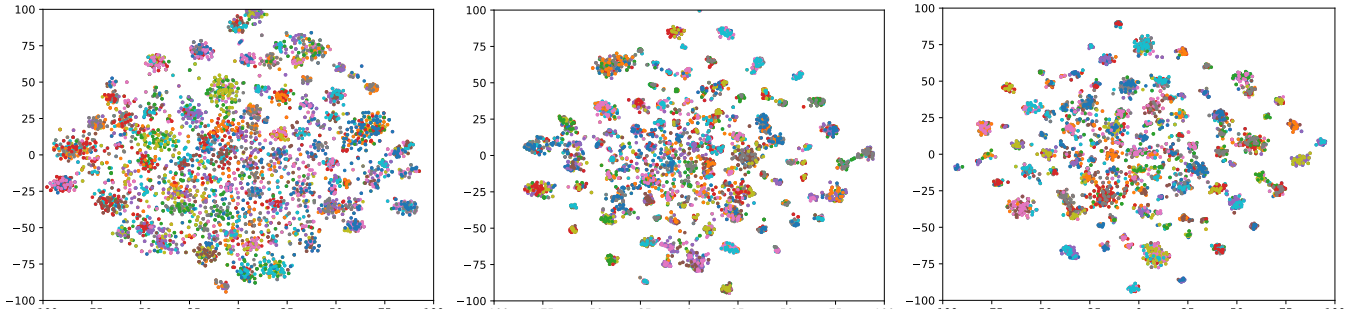


(d) Base CNN (NASNetMobile) (e) Attentional Pooling (NASNetMobile) (f) CAP + Encoding (NASNetMobile)

Figure 8: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Oxford-IIIT Pets** test images using Inception-V3 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.

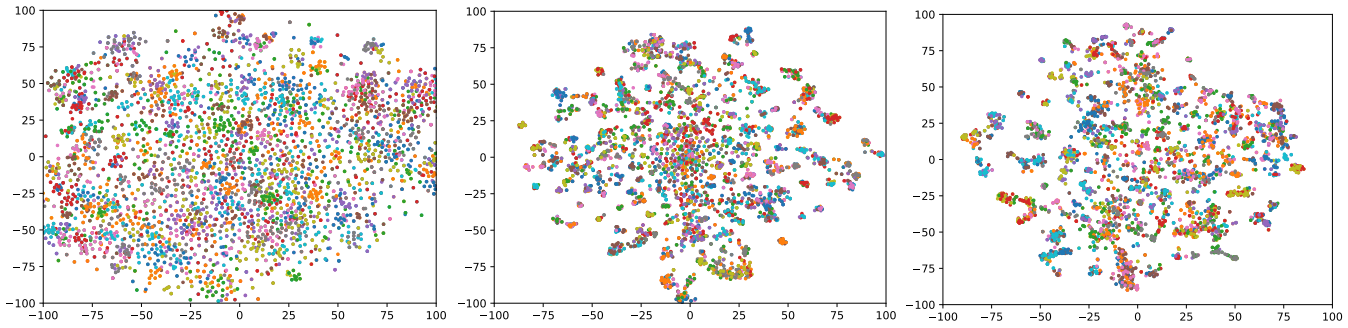


(a) Base CNN (MobileNetV2) (b) Attentional Pooling (MobileNetV2) (c) CAP + Encoding (MobileNetV2)

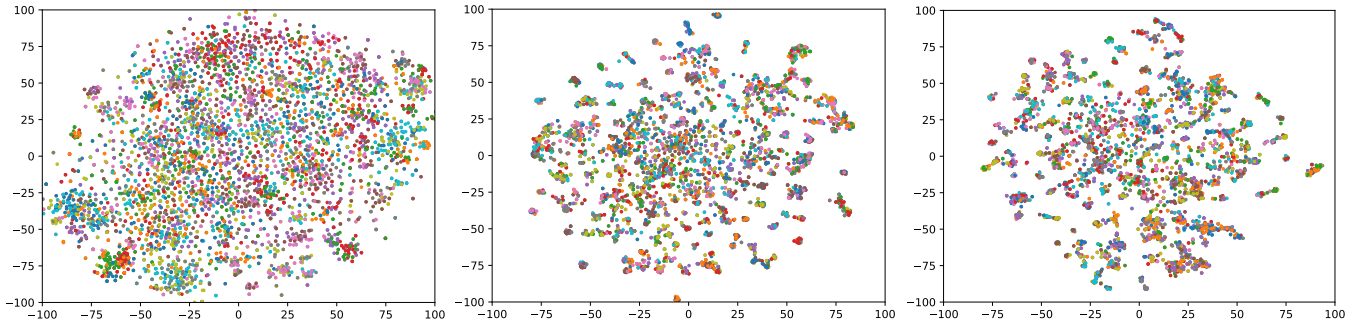


(d) Base CNN (NASNetMobile) (e) Attentional Pooling (NASNetMobile) (f) CAP + Encoding (NASNetMobile)

Figure 9: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Oxford Flowers** test images using MobileNetV2 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.



(a) Base CNN (MobileNetV2) (b) Attentional Pooling (MobileNetV2) (c) CAP + Encoding (MobileNetV2)



(d) Base CNN (NASNetMobile) (e) Attentional Pooling (NASNetMobile) (f) CAP + Encoding (NASNetMobile)

Figure 10: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of the **Caltech-UCSD Birds (CUB-200)** test images using MobileNetV2 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.