Cross-X Learning for Fine-Grained Visual Categorization Supplementary Materials

Wei Luo[1,2] Xitong Yang[2] Xianjie Mo[1] Yuheng Lu[2,5] Larry S. Davis[2]
Jun Li[3] Jian Yang[4] Ser-Nam Lim[5]
[1]South China Agricultural University [2]University of Maryland, College Park
[3]MIT [4]Nanjing University of Science and Technology [5]Facebook AI
{cswluo,yangxitongbob,cedricmo.cs,junl.mldl,sernam}@gmail.com
{ylu,lsd}@umiacs.umd.edu csjyang@njust.edu.cn

1. Hyper-parameters

params	NABirds	CUB-Birds	Cars	Dogs	Aircraft
$\#P$	2	2	2	3	2
γ_1	1	1	1	1	0.5
γ_2	0.25	0.25	0.25	0.5	0.1
γ_3	1	1	1	1	0.1
λ_1	1	1	1	1	1
λ_2	1	1	1	1	1

Table 1. Hyper-parameters of Cross-X with SENet-50 backbone.

params	NABirds	CUB-Birds	Cars	Dogs	Aircraft
$\#P$	2	2	2	2	2
γ_1	0.5	0.5	1	0.01	0.5
γ_2	0.25	0.25	0.25	0.01	0.1
γ_3	0.5	0.5	1	1	0.5
λ_1	1	1	1	1	1
λ_2	1	1	1	1	1

Table 2. Hyper-parameters of Cross-X with ResNet-50 backbone.

Cross-X learning involves 6 hyper-parameters — P, $\gamma_1, \gamma_2, \gamma_3, \lambda_1, \lambda_2$. Among them, P is the number of excitations employed in OSME; γ_1, γ_2 and γ_3 are used to balance the effects of C^3S for different layers (see Eq. (10)); λ_1 and λ_2 are adopted to adjust the effects of CL (see Eq. (11)). These hyper-parameters are determined by evaluating models on hold-out validation datasets. The hyper-parameters for various datasets are presented in Tab. 1 and 2.

2. Training details

All experiments in ablation studies are implemented on the SENet backbone (Sec. 4.3). On all datasets, images are resized to 448×448 for training and testing. OSMEs with 2 excitations are used in all experiments on all datasets except that on Stanford Dogs where 3 excitations are employed.

To present the state-of-the-art performance (Sec. 4.4), images on CUB-Birds, NABirds, and VGG-Aircraft are first resized to 600×600, and then image patches of size 448×448 from random cropping and center cropping are used for training and testing, respectively. We did not observe any advantage of this trick on Stanford Cars and Stanford Dogs, thus default

operations as that implemented in the ablation study are employed on these two datasets. The re-implementation of SENet-50 and ResNet-50 in Sec. 4.4 also obeys these operation rules.

3. Visualization

We display additional activation maps in this section for images from birds (Fig. 1), cars (Fig. 2), aircraft (Fig. 3) and dogs (Fig. 4). The images shown here are consistent with the analysis presented in Section 4.5 of the paper.

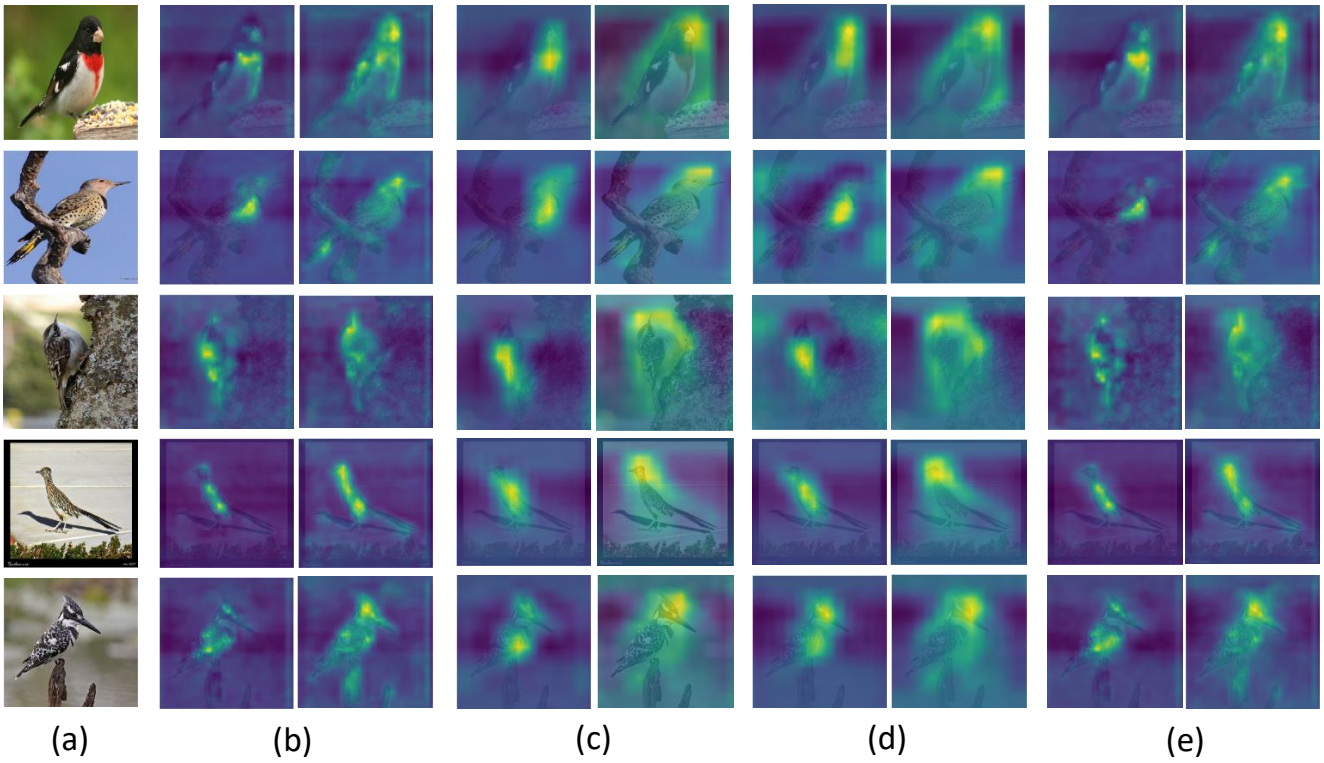

Figure 1. Superimposed display of activation maps (b) $\mathbf{U}_p^{L-1}$, (c) $\mathbf{U}_p^L$ and (d) $\mathbf{U}_p^G$ for images from CUB-Birds. The first column (a) shows original images and the last two columns (e) are combined activation maps from corresponding columns of $\mathbf{U}_p^{L-1}$, $\mathbf{U}_p^L$ and $\mathbf{U}_p^G$. Each of (b)∼(e) shows the activations of two excitation modules in corresponding layers. Best viewed in color.

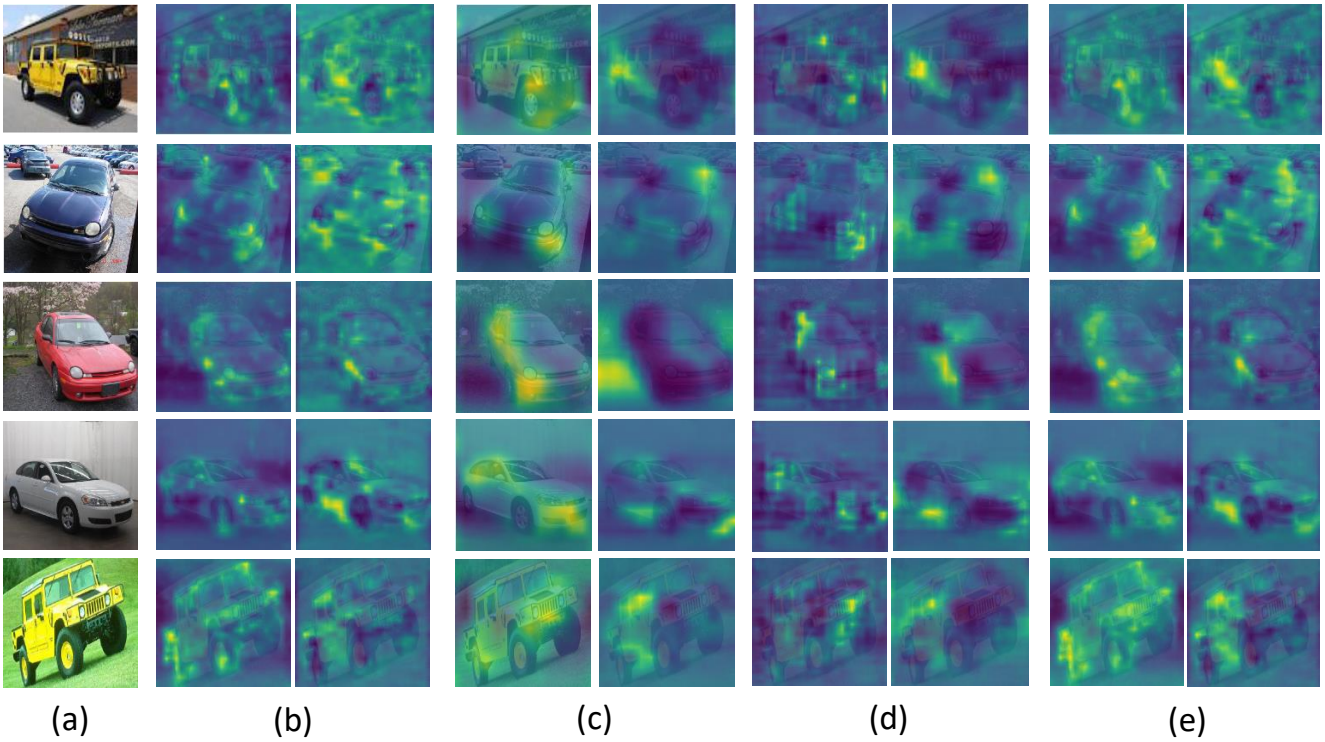

Figure 2. Superimposed display of activation maps (b) $\mathbf{U}_p^{L-1}$, (c) $\mathbf{U}_p^L$ and (d) $\mathbf{U}_p^G$ for images from Stanford Cars. The first column (a) shows original images and the last two columns (e) are combined activation maps from corresponding columns of $\mathbf{U}_p^{L-1}$, $\mathbf{U}_p^L$ and $\mathbf{U}_p^G$. Each of (b)∼(e) shows the activations of two excitation modules in corresponding layers. Best viewed in color.

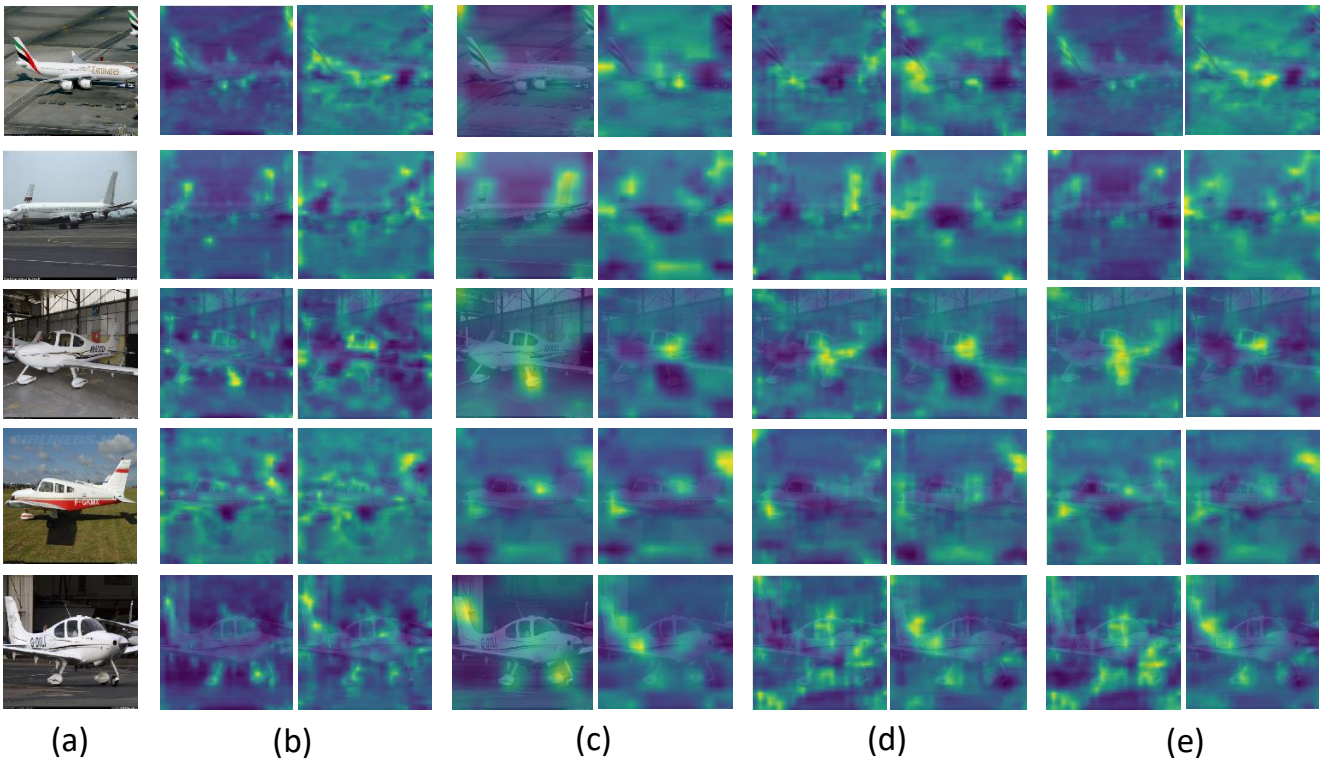

Figure 3. Superimposed display of activation maps (b) $\mathbf{U}_p^{L-1}$, (c) $\mathbf{U}_p^L$ and (d) $\mathbf{U}_p^G$ for images from FGVC-Aircraft. The first column (a) shows original images and the last two columns (e) are combined activation maps from corresponding columns of $\mathbf{U}_p^{L-1}$, $\mathbf{U}_p^L$ and $\mathbf{U}_p^G$. Each of (b)∼(e) shows the activations of two excitation modules in corresponding layers. Best viewed in color.

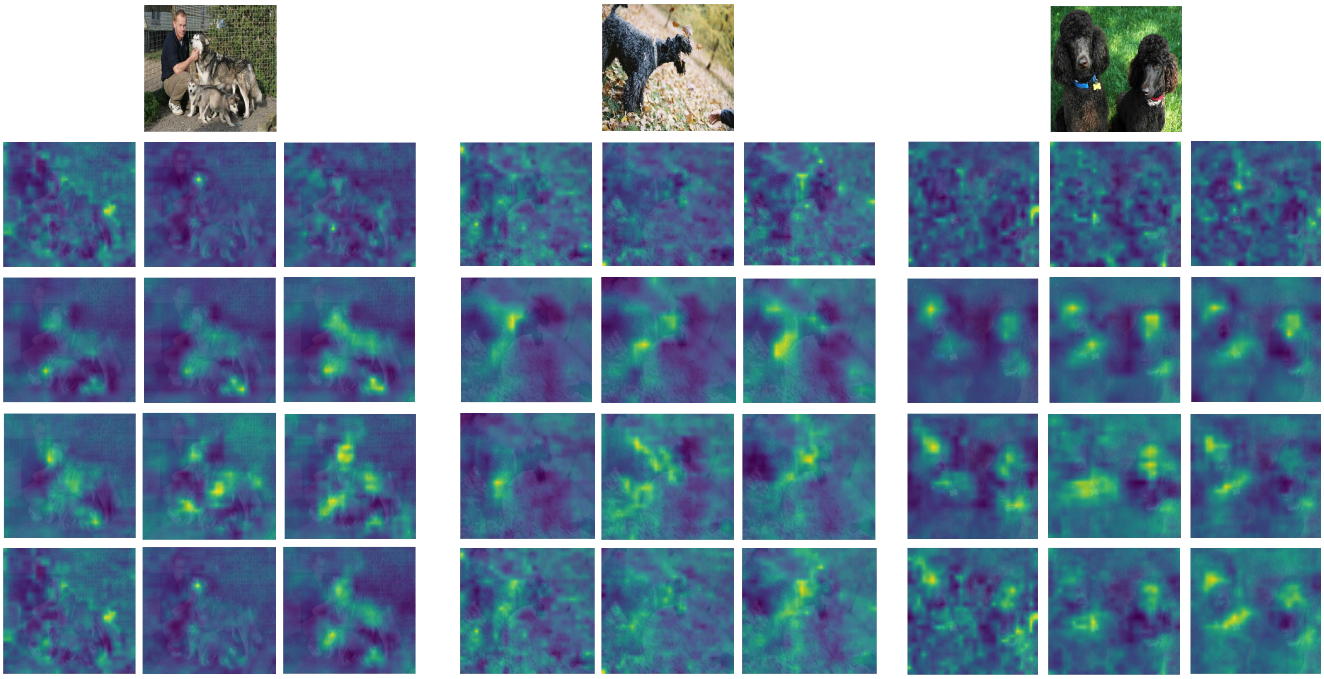

Figure 4. Superimposed display of activation maps $\mathbf{U}_p^{L-1}$ (2nd row), $\mathbf{U}_p^L$ (3rd row) and $\mathbf{U}_p^G$ (4th row) for images from Stanford Dogs. The first row shows original images and the last row are combined activation maps from corresponding rows of $\mathbf{U}_p^{L-1}$, $\mathbf{U}_p^L$ and $\mathbf{U}_p^G$. Each row shows the activations of three excitation modules in corresponding layers. Best viewed in color.