# Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification
# Supplementary Material

Haowei Zhu*, Wenjing Ke*, Dong Li, Ji Liu, Lu Tian, Yi Shan
Advanced Micro Devices, Inc., Beijing, China
{haowei.zhu, wenjing.ke, d.li, lu.tian, yi.shan}@amd.com

## 1. Overview

In this supplementary material, we present more experimental results and analysis.

- We test different inference architectures.
- We provide additional ablation study on effect of ratio of local query selection.
- We show more visualization results of generated attention maps on different benchmarks.
- We conduct experiments on more Transformer baselines.

## 2. Different Inference Architectures

Our default inference architecture is that all the PWCA modules are removed and only SA and GLCA modules are used. For FGVC, we add class probabilities output by classifiers of SA and GLCA for prediction. For Re-ID, we concat two final class tokens of SA and GLCA as the output feature for prediction. We also test two different inference architectures: (1) "SA": using the last SA module for inference. (2) "GLCA": using the GLCA module for inference. Table 1 and 2 present the detailed performance with different baselines on all the FGVC and Re-ID benchmarks, respectively. The results show that only using the SA or GLCA module can obtain similar performance with our default setting. It is also noted that "SA" has the same inference architecture with the baseline by removing all the PWCA and GLCA modules for inference, which does not introduce extra computation cost.

## 3. Ablation Study on Effect of $R$

We test different choices of the ratios of selecting high-response regions as local query. Figure 7 shows that different choices of $R$ can obtain similar performance. We set $R = 10\%$ for all the FGVC benchmarks and set $R = 30\%$ for all the Re-ID benchmarks as default in our method.

## 4. More Visualization Results

We show more visualization results by comparing self-attention and our cross-attention method. Figure 1, 2, 3 present the generated attention maps on different FGVC benchmarks. Figure 4, 5, 6 present the generated attention maps on different Re-ID benchmarks. The results show that our DCAL can reduce misleading attentions and diffuse the attention response to discover more complementary parts for recognition.

## 5. More Transformer Baselines

We conduct two more experiments on CaiT [2] and Swin Transformer [1]. CaiT-XS24 obtains 88.5% while our method obtains 89.7% top-1 accuracy on CUB. Swin-T obtains 84.9% while our method obtains 85.8% top-1 accuracy on CUB. For Re-ID on MSMT, Swin-T achieves 55.7% while we achieve 56.7% mAP. As locality has been incorporated by windows in Swin Transformer, we only apply PWCA into it.

## References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[2] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 1

*Equal contribution.

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 84.4 | 83.6 | 84.6 |
| DeiT-Small | 87.6 | 87.4 | 87.6 |
| DeiT-Base | 88.7 | 88.5 | 88.8 |
| ViT-Base | 91.3 | 91.4 | 91.4 |
| R50-ViT-Base | 91.5 | 91.9 | 92.0 |

(a) CUB-200-2011

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 89.2 | 87.8 | 89.4 |
| DeiT-Small | 92.4 | 91.8 | 92.3 |
| DeiT-Base | 93.9 | 93.5 | 93.8 |
| ViT-Base | 93.5 | 92.9 | 93.4 |
| R50-ViT-Base | 95.3 | 94.8 | 95.3 |

(b) Stanford-Cars

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 86.9 | 86.7 | 87.4 |
| DeiT-Small | 90.1 | 89.8 | 90.0 |
| DeiT-Base | 92.5 | 92.3 | 92.6 |
| ViT-Base | 91.4 | 91.1 | 91.5 |
| R50-ViT-Base | 93.3 | 93.1 | 93.3 |

(c) FGVC-Aircraft

Table 1. Ablation study on different inference architectures for FGVC in terms of accuracy. SA: using SA as the last layer to output class probabilities. GLCA: using GLCA as the last layer to output class probabilities. SA+GLCA: combine the output of SA and GLCA for inference.

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 44.8 / 68.1 | 44.8 / 68.1 | 44.9 / 68.2 |
| DeiT-Small | 54.9 / 77.4 | 55.1 / 77.2 | 55.1 / 77.3 |
| DeiT-Base | 62.2 / 83.1 | 62.3 / 83.1 | 62.3 / 83.1 |
| ViT-Base | 63.9 / 83.2 | 63.9 / 83.1 | 64.0 / 83.1 |

(a) MSMT17

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 71.6 / 85.1 | 71.7 / 84.9 | 71.7 / 84.9 |
| DeiT-Small | 77.4 / 88.0 | 77.4 / 87.8 | 77.4 / 87.9 |
| DeiT-Base | 80.2 / 89.9 | 80.2 / 89.6 | 80.2 / 89.6 |
| ViT-Base | 80.1 / 89.1 | 80.1 / 89.0 | 80.1 / 89.0 |

(b) DukeMTMC-ReID

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 79.7 / 91.8 | 79.7 / 91.8 | 79.8 / 91.8 |
| DeiT-Small | 85.2 / 94.1 | 85.2 / 94.0 | 85.3 / 94.0 |
| DeiT-Base | 87.2 / 94.5 | 87.2 / 94.4 | 87.2 / 94.5 |
| ViT-Base | 87.5 / 94.8 | 87.5 / 94.7 | 87.5 / 94.7 |

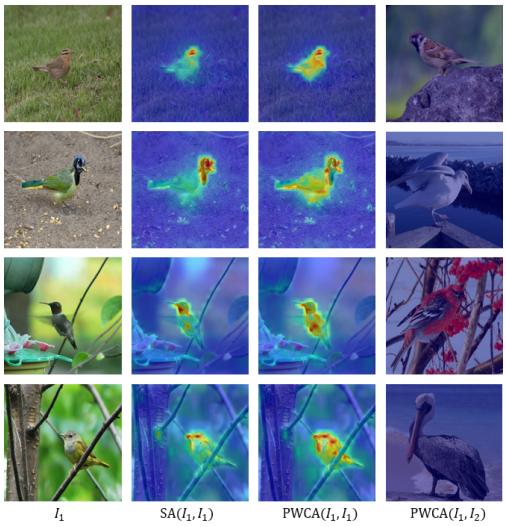
(c) Market1501

| Model | SA (%) | GLCA (%) | SA+GLCA (%) |
|---|---|---|---|
| DeiT-Tiny | 74.1 / 94.6 | 74.0 / 94.6 | 74.1 / 94.7 |
| DeiT-Small | 78.0 / 95.9 | 78.0 / 95.9 | 78.1 / 95.9 |
| DeiT-Base | 79.9 / 96.6 | 80.0 / 96.6 | 80.0 / 96.5 |
| ViT-Base | 80.1 / 96.9 | 80.2 / 96.9 | 80.2 / 96.9 |

(d) VeRi-776

Table 2. Ablation study on different inference architectures for object Re-ID in terms of mAP and rank-1 accuracy. SA: using SA as the last layer to output final feature. GLCA: using GLCA as the last layer to output final feature. SA+GLCA: combine the output of SA and GLCA for inference.

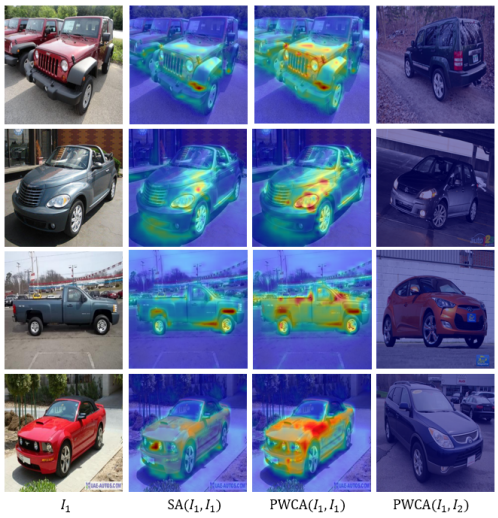

(a) SA vs. GLCA




(b) SA vs. PWCA

Figure 1. Visualization of the generated attention map for self-attention learning and our cross-attention learning on CUB-200-2011.

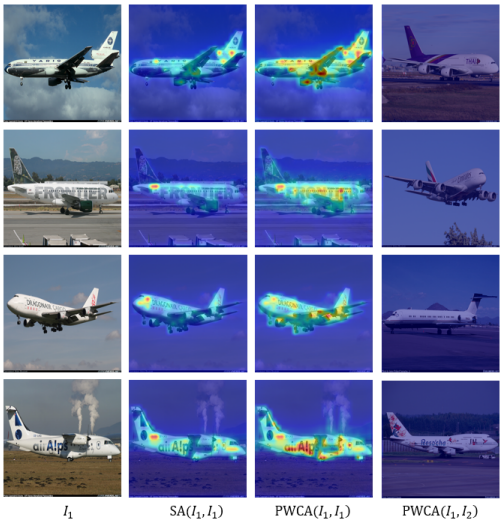

(a) SA vs. GLCA




(b) SA vs. PWCA

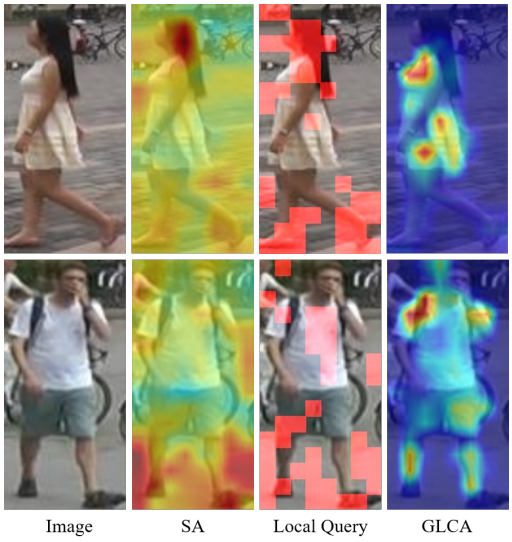
Figure 2. Visualization of the generated attention map for self-attention learning and our cross-attention learning on Stanford-Cars.


(a) SA vs. GLCA

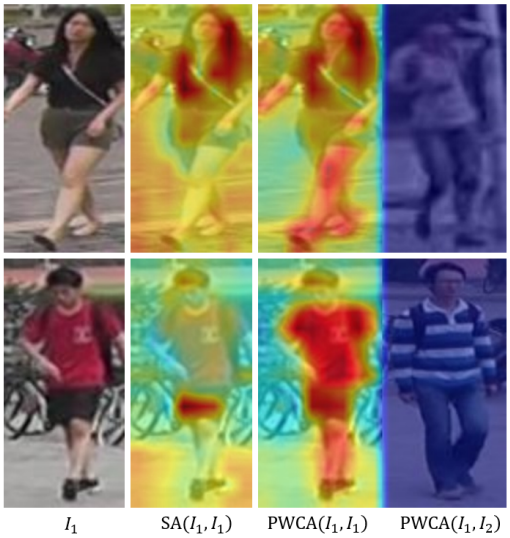



(b) SA vs. PWCA

Figure 3. Visualization of the generated attention map for self-attention learning and our cross-attention learning on FGVC-Aircraft.



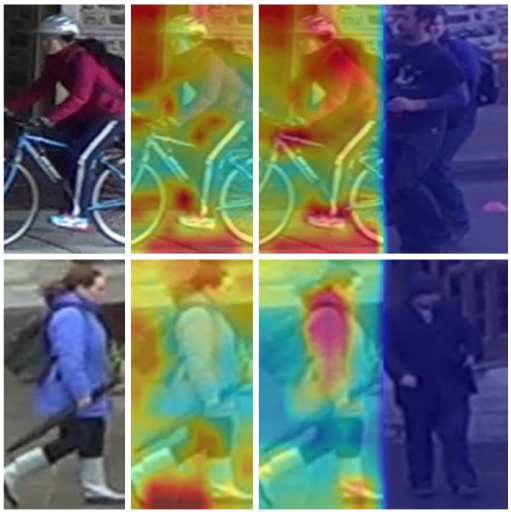

(a) SA vs. GLCA




(b) SA vs. PWCA

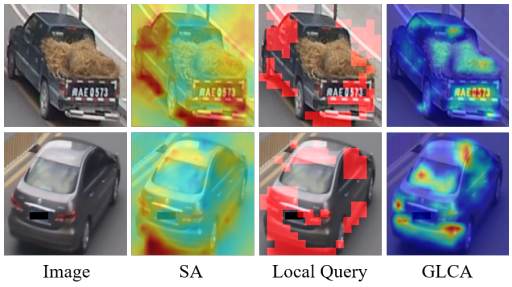
Figure 4. Visualization of the generated attention map for self-attention learning and our cross-attention learning on Market-1501.


(a) SA vs. GLCA

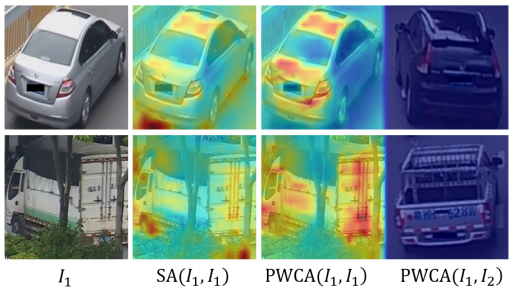



(b) SA vs. PWCA

Figure 5. Visualization of the generated attention map for self-attention learning and our cross-attention learning on DukeMTMC-ReID.




(a) SA vs. GLCA




(b) SA vs. PWCA

Figure 6. Visualization of the generated attention map for self-attention learning and our cross-attention learning on VeRi-776.

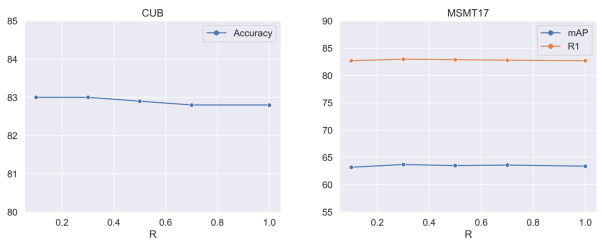



Figure 7. Effect on the ratio of local query selection. DeiT-Tiny is used for CUB and ViT-base is used for MSMT17. We set $R = 10\%$ for all the FGVC benchmarks and set $R = 30\%$ for all the Re-ID benchmarks as default in our method.