SEMICON: A Learning-to-hash Solution for Large-scale Fine-grained Image Retrieval (Supplementary Materials)

Yang Shen[1,2], Xuhao Sun[1], Xiu-Shen Wei[1,2,3⋆], Qing-Yuan Jiang, and Jian Yang[1]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] State Key Laboratory of Integrated Services Networks, Xidian University, China
[3] State Key Laboratory for Novel Software Technology, Nanjing University, China
{shenyang_98,sunxh,weixs,csjyang}@njust.edu.cn, qyjiang24@gmail.com

In the supplementary materials, we provide more detailed ablation studies for the proposed SEMICON and related hyper parameters.

1 Effectiveness of Suppression-Enhancing Mask based Attention

We demonstrate the effectiveness of the proposed SEM module in this section. In concretely, we utilize a mask based attention module which is used for erasing the most discriminative regions as comparisons. As presented in Figure 1, the proposed SEM module performs better among all settings. Note that "Baseline" is degenerated to the ADSH learned with our proposed hash learning framework without performing our SEMICON. Additionally, we sample some attention maps after performing the proposed SEM module and overlay them on the original images for better visualization. As can be seen from Figure 2, those discriminative regions (highlighted in warm colors) in $\boldsymbol{M}_i$ is suppressed but not erased in the following attention map $\boldsymbol{M}_{i+1}$ while some of the other activated regions within the previous attention maps have been further enhanced. For example, the activated regions in $\boldsymbol{M}_1$ on the left of Figure 2 (c) focuses more on the "baklava" in the

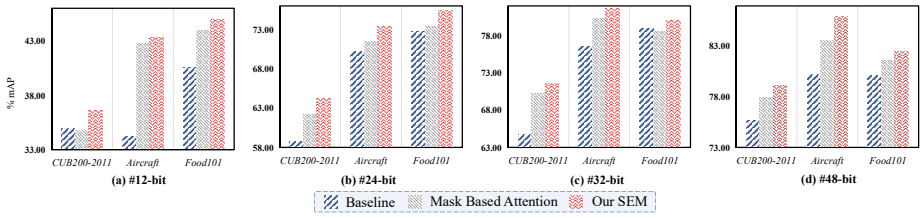

Figure 1. Effectiveness of our suppression-enhancing mask based attention module.

⋆ Corresponding author.

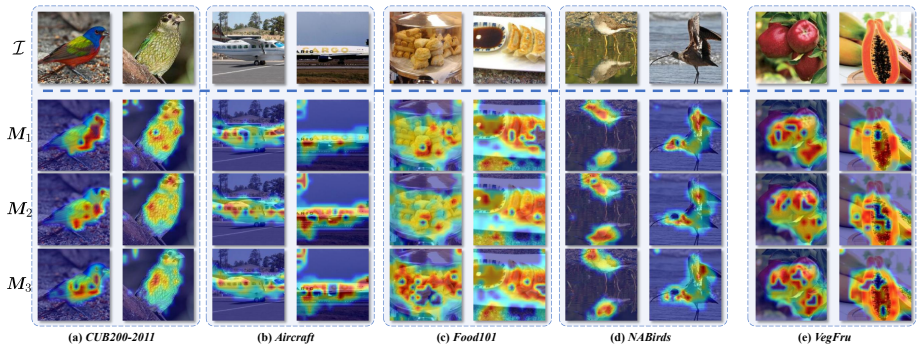

Figure 2. Visualization examples on five fine-grained datasets. The number of m is set as 3. The attention map $\boldsymbol{M}_1$ is obtained after ϕ_{att} in the first stage while $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ is generated according to the proposed SEM (cf. Section 3.2 of the paper) module in the remaining $m-1$ stages respectively.

center of the image while $\boldsymbol{M}_2$ begins to focus on the right part of the image. Additionally, $\boldsymbol{M}_3$ expands the highlighted regions on the surrounding "baklava" in the image based on $\boldsymbol{M}_2$.

2 Effectiveness of Interactive Channel Transformation

As illustrated in Section 3.3 of the paper, the proposed ICON module consists of two steps. In order to demonstrate the effectiveness of the two-step ICON module, we therefore conduct ablation studies on different steps. As presented in Figure 3, the retrieval results are steadily improved after each step on these three fine-grained datasets.

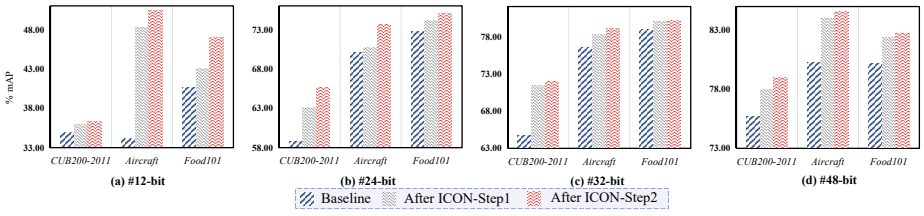

Figure 3. Effectiveness of our two-step interactive channel transformation module.

3 Contributions of m Part-level Hash Codes

As illustrated in implementation details (cf. Section 4.2 of the paper), the final learnt hash codes consist of one global-level hash codes and m part-level hash

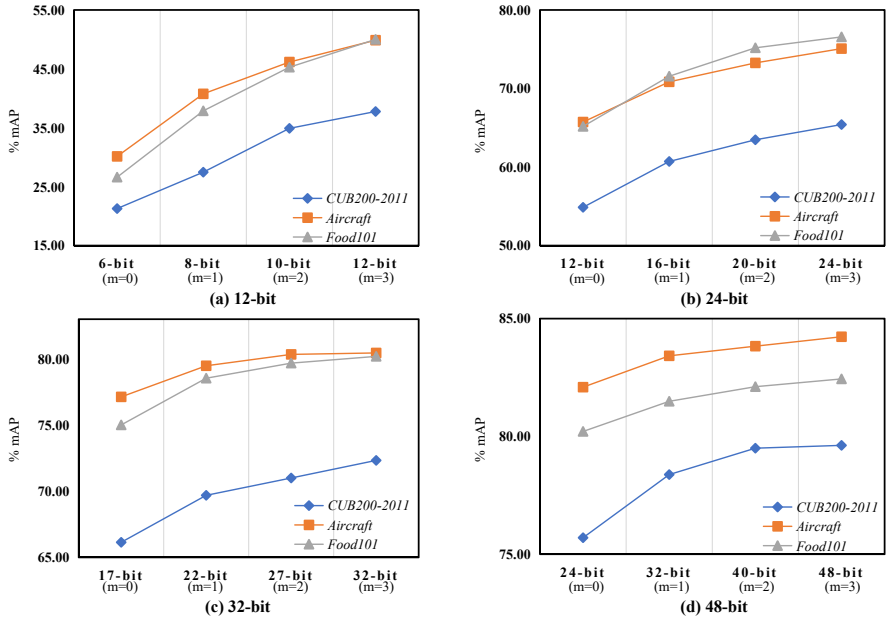

Figure 4. The contributions of m part-level hash codes on three fine-grained datasets.

codes ($m = 3$). Specifically, for those hash codes contain 12, 24, 48 bits, the length of the global-level hash codes is $\frac{k}{2}$ while the length of each part-level hash codes is $\frac{k}{6}$. Particularly, the hash code containing 32 bits is split into a 17-bit global-level hash code and three 5-bit part-level hash codes. As presented in Figure 4 of the supplementary materials, the retrieval results steadily improved with the connection of each part-level hash codes.

4 Sensitivity to Hyper Parameter m

In our SEMICON, we use m to denote the number of attention maps $\boldsymbol{M}$, which is also the number of part-level hash codes. In this section, we present the influence of m by ablation studies only by stacking the suppression-enhancing mask based attentio (SEM) module (cf. Section 3.2 of the paper). As presented in Figure 5 of the supplementary materials, we vary m as 1, 2, 3 and 4. We conduct ablation studies on the 24-bit and 48-bit settings. The length of the global-level hash code is $\frac{k}{2}$ while the length of each part-level hash codes is $\frac{k}{2m}$ (k presents the length of the final hash codes). As analyzed, the lack of attention maps generated from the proposed SEM module may result in that fine-grained images are under-represented for distinguishing subtle visual differences. As can be seen from Figure 5 of the supplementary materials, most of the optimal fine-grained retrieval accuracy is obtained when $m = 3$.

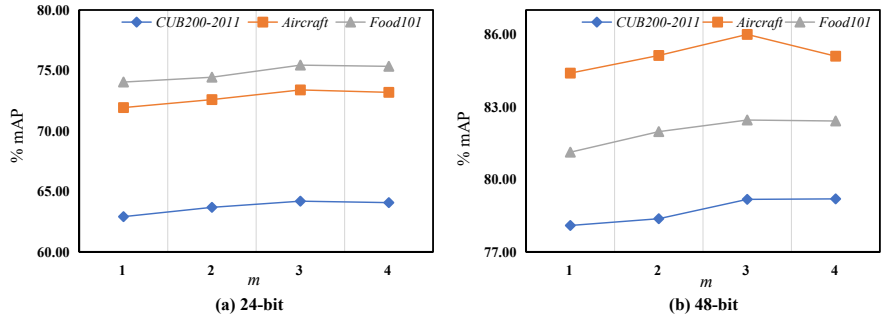

Figure 5. Comparison results of hyper parameter m which denotes the number of part-level hash codes.

5 Sensitivity to Hyper Parameter α

In this section, we conduct ablation studies on the value of the hyper parameter α. We perform these ablation studies by stacking the suppression-enhancing mask based attention (SEM) module (cf. Section 3.2 of the paper) and the interactive channel transformation (ICON) module (cf. Section 3.3 of the paper). The hyper parameter α is used to regularize the degree of suppression ratio of discriminative regions and the enhance ratio of other activated regions in the proposed SEM module (cf. Section 3.3 of the paper). As presented in Figure 6 of the supplementary materials, comparable retrieval results of different α values show that the SEM module is not sensitive to α.

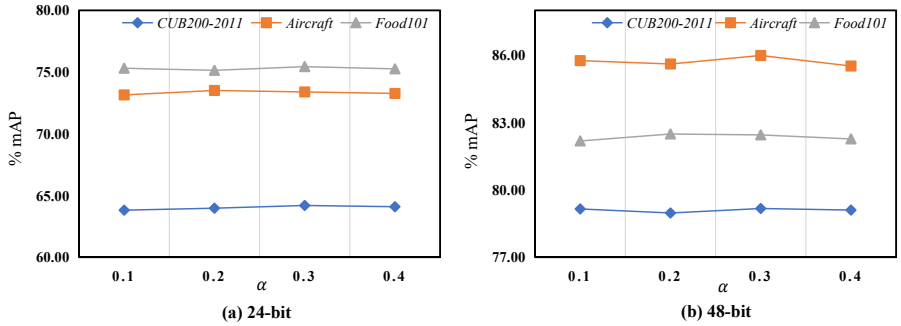

Figure 6. Comparison results of hyper parameter α which is used to regularize the degree of suppression ratio of discriminative regions and the enhance ratio of other activated regions in the proposed SEM module (cf. Section 3.3 of the paper).

6 Visualization of Deep Channels

We randomly sample some channels before and after performing interactive channel transformation (ICON) module and overlay them on the original images for better visualization on the *CUB200-2011* dataset. As can be seen from Figure 7 of the supplementary materials, the activated regions of the sampled feature maps (highlighted in warm colors) after performing interactive channel transformation are actually more semantically meaningful. For example, the activated regions within the right sub-image of the first example pair in Figure 7 (a) tends to correspond more to the leg-part of a bird.

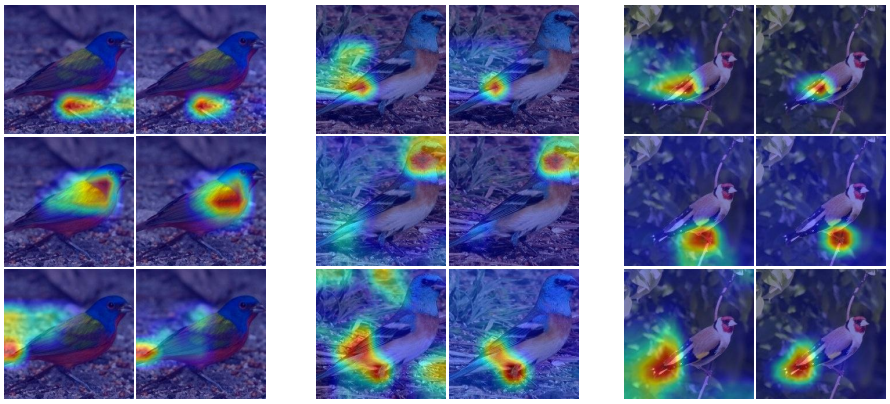

Figure 7. Visualization of deep channels. The left parts within each sub-figure are the visualization of some channels extracted before performing ICON while the right parts are those after performing ICON.