Supplementary Materials
Scene-Aware Label Graph Learning for Multi-Label Image Classification

Xuelin Zhu[1] Jian Liu[3] Weijia Liu[1] Jiawei Ge[1] Bo Liu[2,4] Jiuxin Cao[1,4*]
[1]School of Cyber Science and Engineering, Southeast University
[2]School of Computer Science and Engineering, Southeast University
[3]Ant Group, Hangzhou, China [4]Purple Mountain Laboratories, Nanjing, China
{zhuxuelin,weijia-liu,jiawei_ge,bliu,jx.cao}@seu.edu.cn, rex.lj@antgroup.com

1. Comparison on model efficiency

We compare the efficiency, including the model size and computational complexity as well as FPS (Frame Per Second) value, of our SALGL with existing methods. Results are shown in Table 1 and all FPS values are obtained on Nvidia 3080Ti GPU. Overall, our SALGL achieves a good balance between efficiency and performance, and the considerable FPS value (151 on ResNet101) suggests that our SALGL is able to work well in practical applications.

Method	Backbone	#param.	FLOPs	FPS	mAP
CSRA [2]	ResNet101	43M	31.6G	236	83.5
Q2L [1]	ResNet101	143M	43.3G	137	84.9
SALGL	ResNet101	99M	47.2G	151	85.8
TSFormer [3]	ViT-B16	171M	84.0G	68	88.9
SALGL	ViT-B16	96M	87.9G	65	89.4

Table 1. Comparisons on efficiency and performance (mAP in %).

2. More comparisons about backbone.

We first compare our SALGL with the TSFormer [3] with ImageNet 22k pre-trained ViT-B16 being backbone network. Experimental results are shown in Table 2. Notably, our SALGL achieves better performance than TSFormer on all three datasets.

Method	VOC 2007	NUS-WIDE	MS-COCO
TSFormer	97.0	69.3	88.9
SALGL	97.3	70.1	89.4

Table 2. Comparisons on the ViT-B16 backbone (mAP in %).

Then, we also choose ResNet-cut [2] as the backbone of our SALGL framework to provide a systematic comparison with CSRA [2] and experimental results are presented in Table 3. Notably, our SALGL performs better on both the VOC 2007 and MS-COCO datasets.

*Corresponding author.

Method	Backbone	VOC 2007	MS-COCO
CSRA	ResNet-cut	96.8	85.6
SALGL	ResNet-cut	97.2	86.5

Table 3. Comparisons on the ResNet-cut backbone (mAP in %).

References

[1] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[2] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021.

[3] Xuelin Zhu, Jiuxin Cao, Jiawei Ge, Weijia Liu, and Bo Liu. Two-stream transformer for multi-label image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3598–3607, 2022.