Democracy Does Matter: Comprehensive Feature Mining for Co-Salient Object Detection
Supplementary Material

Siyue Yu[1,2], Jimin Xiao[1]*, Bingfeng Zhang[1,2], Eng Gee Lim[1]
[1]XJTLU, [2]University of Liverpool
{siyue.yu,jimin.xiao,bingfeng.zhang,enggee.lim}@xjtlu.edu.cn

1. Overview

In this supplementary material, we will analyze the self-contrastive learning module (SCL) with some visualizations. Besides, we will provide more qualitative comparisons between our model and other state-of-the-art approaches. Additionally, we will compare the complexity of our method with others. We will also analyze the influence of α in Eq.(19) in our paper. Moreover, we will discuss about some failure cases.

2. Self-Contrastive Learning Module Analysis

We display some response maps in different cases on the CoCA dataset [7] in Fig. 1. Note that this dataset is used for evaluation. M^{final} denotes the normal response maps generated by original inputs, M_c^{final} denotes the co-salient response maps generated by inputs where the background regions are erased, and M_b^{final} denotes the background response maps generated by inputs where the co-salient objects are erased. Then, $proto$, $proto_c$ and $proto_b$ can be derived based on the corresponding response maps. As shown in Fig. 1, it can be found that the M^{final} can focus on most regions of the target co-salient objects. Moreover, comparing M_c^{final} and M_b^{final}, the M_c^{final} can highlight all the related co-salient objects. In contrast, the M_b^{final} are sensitive to the surroundings of the co-salient objects. In this case, our assumption of SCL, where $proto$ and $proto_c$ are pulled together while $proto$ and $proto_b$ are pushed away, can be verified. With our SCL, the model can learn to differentiate co-salient features and background features. Thus, the noise information can be suppressed.

3. Qualitative Comparison

We list more qualitative comparisons with previous sate-of-the-art methods in Fig. 2. We use the CoCA dataset [7] for demonstration, as it is a challenging real-world dataset, containing more challenging cases. The compared methods include CSMG [4], GCAGC [5], CoEGNet [1], GICD [7], GCoNet [2], and DeepACG [3]. It is evident that our predictions are closer to the ground truth. Specifically, when the background contains misleading objects, such as the humans in the group 'Binoculars', our model can suppress the noisy information and focus on the targets, compared with GCoNet [2] and GICD [7]. Additionally, when there are complex background clutters, like images in the groups 'Pillow' and 'Tablet', compared with all other methods, ours are robust to this challenging setting.

Table 1. Complexity comparisons. 'param.' denotes the number of parameters. We set 5 inputs to compute FLOPs.

method	FLOPs (G)	param. (M)	runtime (fps)	$F_\beta^{max}\uparrow$
CADC [6]$_{\text{ICCV21}}$	457.9	392.8	18.0	0.548
GICD [7]$_{\text{ECCV20}}$	467.6	278.0	40.8	0.513
GCoNet [2]$_{\text{CVPR21}}$	311.5	142.0	116.2	0.544
DCFM†(ours)	313.0	140.5	101.9	0.592
DCFM (ours)	316.6	142.3	84.4	**0.598**

4. Complexity Analysis with State-of-the-art Methods

The computational complexity of Eq.(2) and Eq.(18) in our paper is $O((NHW)^2)$ and $O((HW)^2)$ respectively. The increment of FLOPs is small since the input size is small. We list the complexity comparisons in Tab. 1, '†' means without DFE. Ours can achieve an impressive performance with fewer FLOPs and parameters compared with CADC [6] and GICD [7]. Besides, ours can obtain a better performance with limited increment of FLOPs and parameters compared with GCoNet [2], especially for DCFM†. Overall, our method has an impressive performance with comparable runtime.

*corresponding author
[1]The work was supported by National Natural Science Foundation of China under 61972323.

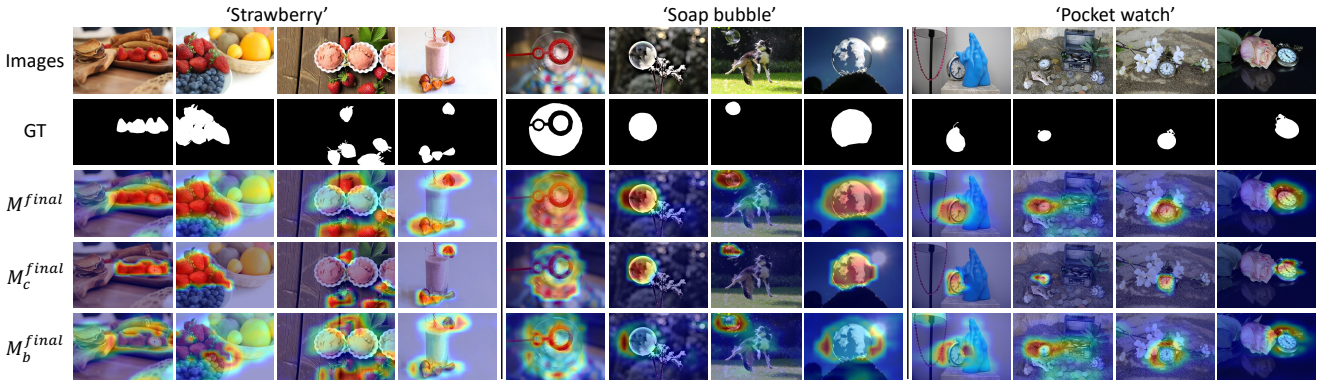

Figure 1. Visualization of the response maps in different cases. The visualizations can verify our assumption of the self-contrastive learning module as M^{final} is consistent with M_c^{final} but different from M_b^{final}.

Figure 2. More visualizations of our predictions and comparisons with previous state-of-the-art approaches. It can be found that our model can better differentiate the co-salient objects and background in complex scenes.

Table 2. Influence of alpha in Eq.(19) in our paper.

α	0.1	1	2	3	4
$F_\beta^{max}\uparrow$	0.578	0.592	0.593	**0.598**	0.587

5. Influence of Alpha in Eq.(19) in Our Paper

We add the ablation study of alpha in Tab. 2. The performance smoothly increases with larger alpha. However, performance decreases when alpha is too big (α=4). When α>4, the model even fails to be trained. This is because in this case, the weight of small positive attention values will be much bigger. Thus, the attention mechanism will be confused and tend to focus on those small values but neglect original high values.

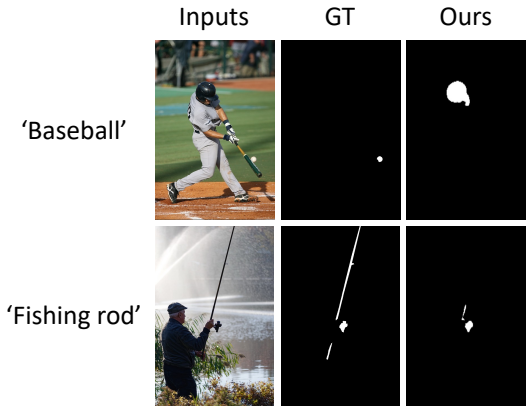

Figure 3. Visualizations of some failed cases.

6. Limitation Discussion

We also report some failure cases in Fig. 3. As shown in the figure, it is difficult for our model to predict small objects precisely. This may be caused by the fact that the inputs are resized into the size of 224×224. Then, with the feature extractor, the size of the output features is 14×14. In this case, it may cause information lost for small objects. Thus, it is difficult for our model to capture the corresponding features. Therefore, how to enhance model robustness for small objects is a direction for our future work.

References

[1] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *arXiv preprint*, 2020. 1

[2] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1

[3] Kaihua Zhang, Mingliang Dong, Bo Liu, Xiao-Tong Yuan, and Qingshan Liu. Deepacg: co-saliency detection via semantic-aware contrast gromov-wasserstein distance. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1

[4] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[5] Kaihua Zhang, Tengpeng Li, Shiwen Shen, Bo Liu, Jin Chen, and Qingshan Liu. Adaptive graph convolutional network with attention graph clustering for co-saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[6] Ni Zhang, Junwei Han, Nian Liu, and Ling Shao. Summarize and search: learning consensus-aware dynamic convolution for co-saliency detection. In *Int. Conf. Comput. Vis.*, 2021. 1

[7] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Eur. Conf. Comput. Vis.*, 2020. 1