Supplementary Material: Summarize and Search: Learning Consensus-aware Dynamic Convolution for Co-Saliency Detection

Ni Zhang[1] Junwei Han[1] Nian Liu[2*] Ling Shao[2]
[1]Northwestern Polytechnical University [2]Inception Institute of Artificial Intelligence
{nnizhang.1995, junweihan2010, liunian228}@gmail.com, ling.shao@ieee.org

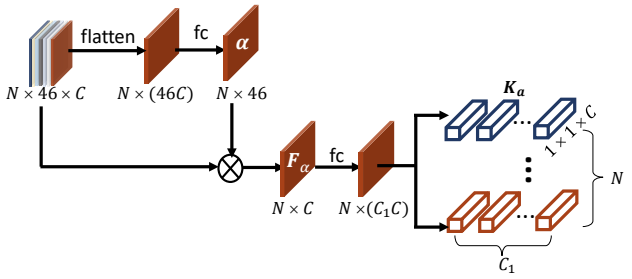

Figure 1. The detailed process for generating $\boldsymbol{K_a}$. $\otimes$ means element-wise multiplication and summation.

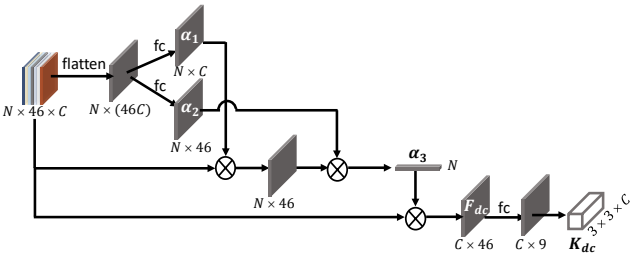

Figure 2. The detailed process for generating $\boldsymbol{K_{dc}}$. $\otimes$ means element-wise multiplication and summation.

1. Consensus-aware Kernel Construction

For better understanding, we show the detailed processes for generating the adaptive kernels $\boldsymbol{K_a}$ with 1×1 size and the depthwise common kernel $\boldsymbol{K_{dc}}$ in Figure 1 and Figure 2, respectively.

2. The Necessity of Designing Two Kinds of Kernels

We argue that it is necessary to design two kinds of kernels, *i.e.*, adaptive kernels and the common kernel, in our model. **1)** Using two different transformations to generate two kinds of kernels mimics the multi-branch architecture widely used in CNNs, hence increasing the transformation complexity and model capability. **2)** Doing so can disentangle the learning of image-wise adaptive information and group-wise common knowledge, thus better conforming to the nature of co-saliency detection. **3)** Doing so makes it possible to explore the relationship between the image-wise adaptive information and group-wise common knowledge encoded in our two kinds of kernels. This is very important for co-saliency detection. However, only using a single kind of kernel can not achieve this goal.

The effectiveness of the aforementioned two kinds of kernels has been demonstrated in Table 1 in our paper. To further verify the necessity of the large common kernel (LCK), we report experimental results on four benchmark datasets in Table 1. We can see that using large adaptive kernels (LAK) alone at multiple levels can not obtain obvious improvement ("+LAK+ML" vs. "+LAK"). However, using both kinds of kernels ("+LAK+LCK+ML") at multiple levels largely outperforms "+LAK+ML", thus further verifying the necessity of LCK.

3. Model Complexity Analysis

We supplement the FLOPs and the number of parameters of the models in Table 1 in our paper, shown in Table 2. By comparing the proposed large kernels with vanilla kernels ("+LAK" vs. "+VAK" or "+LCK" vs. "+VCK"), we can find that large kernels incur larger computational costs due to the introduction of extra depthwise kernels to increase kernel sizes. However, if we use the same method as the vanilla kernels to generate large kernels ("+Vanilla LAK + Vanilla LAK"), it can dramatically increase computational costs compared to our proposed method ("+LAK+LCK"), thus demonstrating the superiority of using the depthwise separable convolution operation for increasing kernel sizes. Finally, from Table 1 in our paper and Table 3, we can see that "+LAK+LCK+ML" can bring significant performance gains with acceptable computational costs growth compared to "+LAK+LCK". Hence, we use "+LAK+LCK+ML" as our final model.

*Corresponding author.

Table 1. Ablation studies for verifying the necessity of the large common kernel (LCK). "LAK" means large adaptive kernels and "ML" means adopting our proposed consensus-aware dynamic convolution (CADC) at multiple decoder levels.

Settings	CoCA				CoSOD3k				CoSal2015				MSRC			
	$S_m \uparrow$	maxF $\uparrow$	$E_\xi \uparrow$	MAE $\downarrow$	$S_m \uparrow$	maxF $\uparrow$	$E_\xi \uparrow$	MAE $\downarrow$	$S_m \uparrow$	maxF $\uparrow$	$E_\xi \uparrow$	MAE $\downarrow$	$S_m \uparrow$	maxF $\uparrow$	$E_\xi \uparrow$	MAE $\downarrow$
+LAK	0.661	0.508	0.735	0.146	0.789	0.741	0.827	0.105	0.852	0.843	0.894	0.073	0.797	0.844	0.867	0.132
+LAK+ML	0.656	0.509	0.724	0.156	0.792	0.745	0.827	0.105	0.857	0.849	0.895	0.072	0.802	0.852	0.871	0.126
+LAK+LCK+ML	**0.681**	**0.548**	**0.744**	**0.132**	**0.801**	**0.759**	**0.840**	**0.096**	**0.866**	**0.862**	**0.906**	**0.064**	**0.821**	**0.873**	**0.895**	**0.115**

Table 2. FLOPs and the number of parameters of different model variants. "VAK" and "VCK" mean vanilla adaptive kernels and the vanilla common kernel, respectively. "LAK" and "LCK" represent large adaptive kernels and the large common kernel. "Vanilla LAK" and "Vanilla LCK" mean vanilla large adaptive kernels and the vanilla common kernel. "ML" means adopting CADC at multiple decoder levels.

	FLOPs (G)	Params (M)
baseline	90.710	24.637
+VAK	90.844	66.406
+VCK	90.796	42.238
+LAK	90.846	66.409
+LCK	90.896	91.053
+LAK+LCK	90.980	132.297
+Vanilla LAK + Vanilla LCK	93.000	352.439
+LAK+LCK+ML	91.800	392.850

4. More Visual Comparison with State-of-the-Art Methods

We show more visual comparisons with state-of-the-art models in Figure 4. It shows that our model can accurately search and segment the co-occurring salient objects in many challenging scenes, *e.g.*, objects with small or big sizes, complex backgrounds, and multiple objects. However, other methods are easily disturbed by the extraneous salient objects or miss the target objects.

5. More Synthesized Examples

We show more synthesized examples generated by our proposed data synthesis strategy in Figure 3.

References

[1] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *TIP*, 22(10):3766–3778, 2013.

[2] Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *ECCV*, pages 485–501, 2018.

[3] Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, and Wei Guo. Icnet: Intra-saliency correlation network for co-saliency detection. *NIPS*, 2020.

[4] Dingwen Zhang, Junwei Han, Jungong Han, and Ling Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *TNNLS*, 27(6):1163–1176, 2015.

[5] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *CVPR*, pages 2994–3002, 2015.

[6] Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, and Junwei Han. A self-paced multiple-instance learning framework for co-saliency detection. In *ICCV*, pages 594–602, 2015.

[7] Kaihua Zhang, Tengpeng Li, Bo Liu, and Qingshan Liu. Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In *CVPR*, pages 3095–3104, 2019.

[8] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, pages 455–472, 2020.

(a) (b) (c) (d) (e) (f)

Figure 3. More synthesized examples of our proposed data synthesis strategies. (a) original images. (b) original ground truth. (c) normally synthesized images. (d) ground truth for the normally synthesized images. (e) reversely synthesized images. (f) ground truth for the reversely synthesized images.

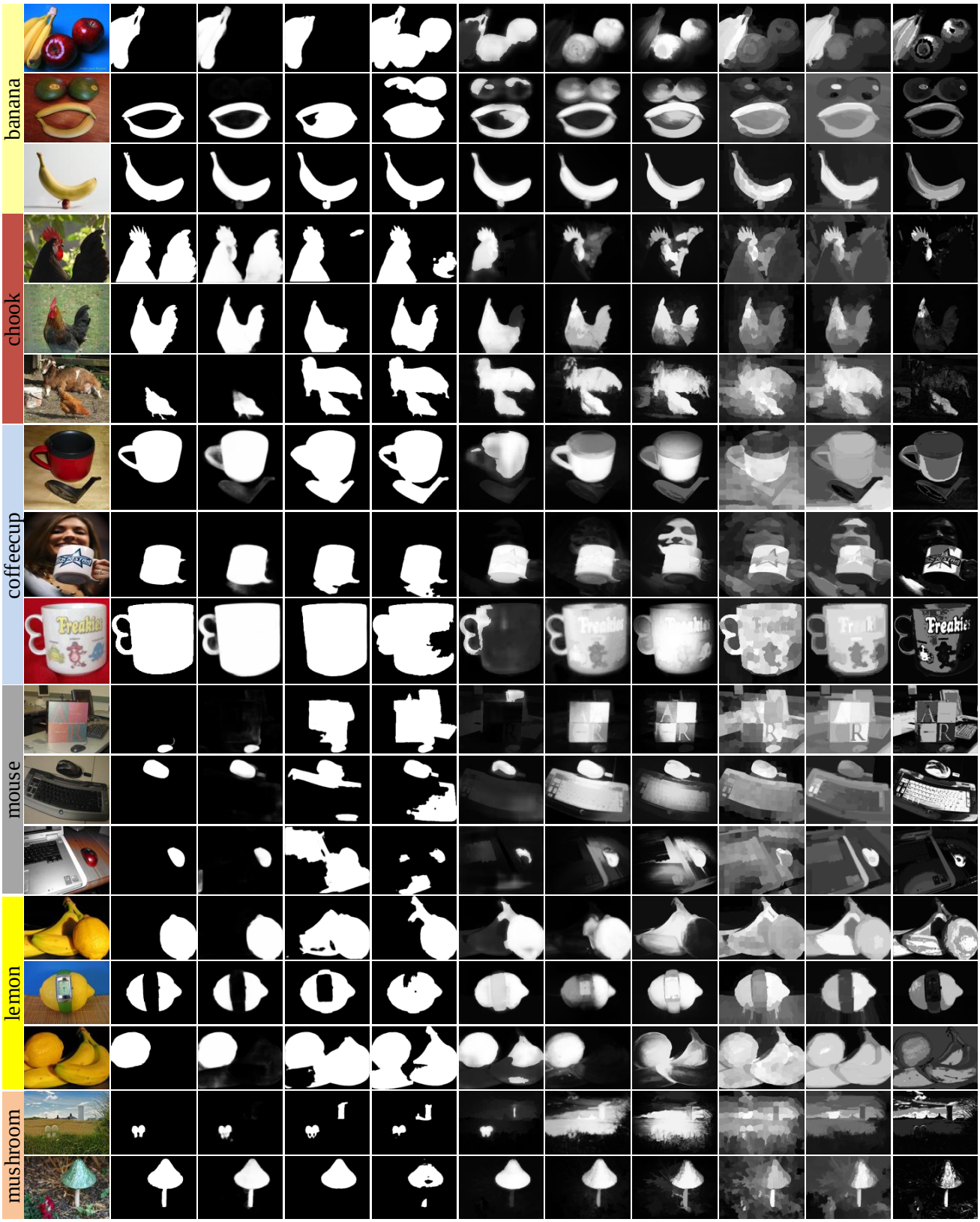

Figure 4. More qualitative comparison of our proposed model with other state-of-the-art methods.