Saliency Hierarchy Modeling via Generative Kernels for Salient Object Detection

Wenhu Zhang[1], Liangli Zheng[2], Huanyu Wang[3], Xintian Wu[3], and Xi Li[3,4,5]

[1] Polytechnic Institute, Zhejiang University
[2] School of Software Technology, Zhejiang University
[3] College of Computer Science and Technology, Zhejiang University
[4] Shanghai Institute for Advanced Study, Zhejiang University
[5] Shanghai Al Laboratory
{wenhuzhang, lianglizheng, huanyuhello, hsintien, xilizju}@zju.edu.cn

1 Quantitative comparisons with other variants.

Table 1: Quantitative comparisons among variants given the same prior (i.e., Grad-Cam).

#	Grad-Cam Prior	DUTS-TE ($\mathcal{F}_\beta \uparrow$)	ECSSD ($\mathcal{F}_\beta \uparrow$)
1	Baseline(w/o Prior)	.830	.909
2	Baseline(Sup.)	.835	.918
3	Baseline+HKG(Sup.)	.849	.925
4	Baseline+SHM(Ours)	.854	.927
5	Baseline+SHM+HKG(Ours)	.867	.933

Given the same prior (i.e., Grad-Cam), we study the variants (w/ and w/o SHM or HSG) to show the improvements of each contribution on DUTS-TE and ECSSD based on ResNet-50. The quantitative results are shown in Table 1. 'w/o Prior' indicates not using the prior guidance. 'Sup' refers to using the prior guidance as auxiliary supervision. 'Ours' is the proposed region-level saliency hierarchy modeling in SHM.

Table 2: Quantitative comparisons among variants using the ground-truth supervision for the sub-saliency masks.

#	DUTS-TE ($\mathcal{F}_\beta \uparrow$)	Ground-truth	Grad-Cam(Ours)
1	w/o Prior	.830	.830
2	Baseline+SHM	.848	.854
3	Baseline+SHM+HKG	.855	.867

In our design, the generated sub-saliency masks are supervised by divided ground truth. We divide the Grad-Cam map into several regions and take the intersection of the ground-truth with each regions as the supervision. Here, we conduct the experiment that supervising all sub-saliency masks with the entire ground as shown in Table 2.

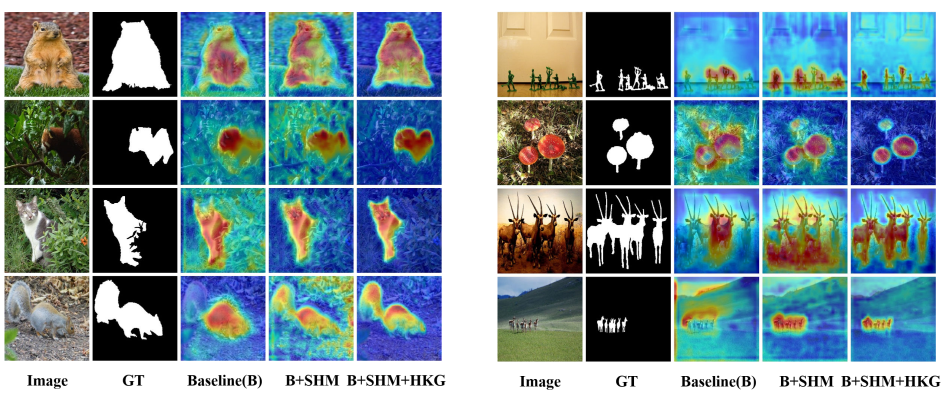
2 Sensitivity analyses on the hyper-parameters ρ and K

Table 3: Ablation Studies on hyper-parameters ρ and K.

ρ (loss factor)	0.001	0.01	0.1	1	10
DUTS-TE ($\mathcal{F}_\beta \uparrow$)	.859	.861	.867	.865	.866
K (number of decoder layers)	2	3	4	5	6
DUTS-TE ($\mathcal{F}_\beta \uparrow$)	.848	.859	.864	.867	.867

We study the hyper-parameters on DUTS-TE based on ResNet-50 and show the grid search results in Table 3.

3 Grad-Cam Visualization



(a) Single object. (b) Multiple objects.

Fig. 1: Grad-Cam visualization results of the feature maps H_3. 'Baseline' denotes the vanilla U-Net with ResNet-50 backbone. 'B + SHM' denotes the SHM based decoder with static kernels in the branches. 'B + SHM + HKG' represents our whole SHNet.

We have shown the clear performance gain of the proposed modules in our manuscript (page 14, Tab. 4). To further analyze the effect of our proposals, we visualize the Grad-Cam [1] results on the feature map (i.e., H_3 feature in our manuscript) by adding each component step by step into our SHNet, as shown in Fig. 1.

In the single object scenarios Fig. 1.(a), our proposed SHM modules can attend to more complete salient areas and the HKG module can further eliminate the background distraction. In the multiple object scenarios Fig. 1.(b), our proposed modules significantly pay more attention to the salient objects and sharpen the boundaries under various challenging scenarios.

4 More Qualitative Comparison

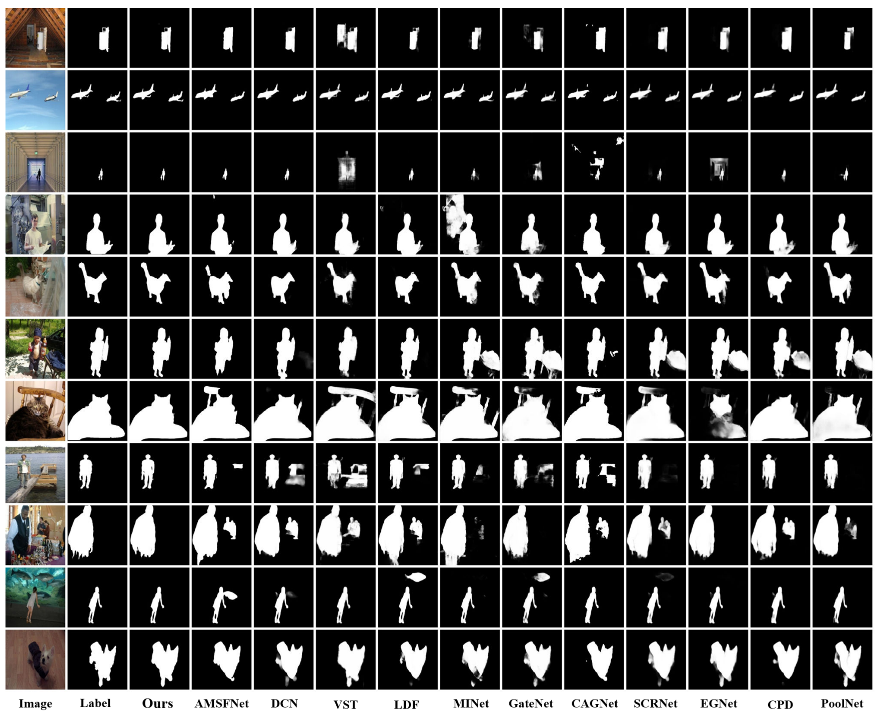


Fig. 2: More qualitative comparisons between the state-of-the-art SOD methods and our SHNet.

To further illustrate the effectiveness of our method, we provide more qualitative comparisons, as shown in Fig. 2. Obviously, the proposed SHNet is able to produce accurate saliency predictions under various challenging scenes, including images with fine structures (1st and 2nd rows), tiny objects (2nd and 3rd rows), low contrast foreground and background (4th and 5th rows), and cluttered distractions (6th $\sim$ 10th rows).

References

1. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Int. Conf. Comput. Vis. pp. 618–626 (2017) 3