Supplementary for Inharmonious Region Localization via Recurrent Self-Reasoning

Penghao Wu
wupenghaocraig@sjtu.edu.cn
Li Niu*
ustcnewly@sjtu.edu.cn
Jing Liang
leungjing@sjtu.edu.cn
Liqing Zhang
zhang-lq@cs.sjtu.edu.cn

MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University
Shanghai, China

In this supplementary, we conduct comparison of computational complexity in Sec. 1, and study the effects of different loss components in Sec. 2. We provide visualization results in Sec. 3. At last, we discuss our limitations in Sec. 4.

1 Computational Complexity Comparison

We report the statistics related to computational complexity in Table 1. We first compare the whole RSRNet with our RSR module. It can be seen that our RSR module is rather light-weighted and only adding small overhead to the backbone. We also compare with baselines DIRL, SegFormer, and MiNet, which shows that the computational complexity of our method is comparable with baselines.

2 Effects of Loss Functions

We investigate the effectiveness of different loss terms ℓ_{bce}, ℓ_{ssim}, and ℓ_{iou} in Table 2. We can see that ℓ_{bce} is essential for our model, and the training process can not converge without it. We also observe that ℓ_{ssim} and ℓ_{iou} are both beneficial for the overall performance. We also set $\lambda = 1$ so that the weights for all the masks are the same. The results show that it is beneficial and reasonable to set exponentially increasing weights from coarse to fine outputs.

3 Visualization Results

First, we visualize the inharmonious region mask $\bar{\boldsymbol{M}}^k$ and the similarity matrix $\boldsymbol{S}^{k-1,l}$ ($l = 0, 3$) in each iteration in our RSR module.

From Figure 1, we can see that the inharmonious mask and the similarity matrix at the early stage miss part of inharmonious region and miss-classify some background region as

*Corresponding author.

	RSR module	RSRNet	DIRL	SegFormer	MINet
Number of parameters	5.52M	54.28M	53.46M	48.28M	68.28M
Inference time	52.8ms	90.6ms	63.80ms	61.36ms	63.95ms
GFlops	4.33	99.27	104.67	30.97	116.01

Table 1: Comparison between model size and speed. All the tests run on a single GeForce GTX TITAN X GPU.

	Loss			
Evaluation	w/o ℓ_{bce}	w/o ℓ_{ssim}	w/o ℓ_{iou}	$\lambda = 1$
AP(%)↑	42.82	81.59	82.10	81.38
F_1 ↑	0.4255	0.7444	0.7524	0.7529
IoU(%)↑	35.34	69.95	70.37	69.93

Table 2: Ablation study on different loss terms. λ is the hyper-parameter in Eqn. (3) in the main paper.

inharmonious. As the iteration goes, the detected inharmonious region is gradually recovered with higher confidence, with the similarity map providing more complete and accurate information.

When comparing similarity map and inharmonious mask, similarity map is able to provide the general location of inharmonious region but rather coarse, so we need a GRU cell to further process it. By comparing the similarity map of different scales ($l = 0, 3$), the one of smaller scale provides more complete yet more noisy inharmonious region, while the the one of larger scale is more conservative but less noisy.

Then, we visualize the masks $\bar{\boldsymbol{M}}_{rsr}$, $\boldsymbol{M}_{dec}$, $\boldsymbol{M}_{fnl}$ and the adaptive combination map $\boldsymbol{G}$ in Figure 2. The goal of our adaptively combined mask is utilizing the advantages of both $\bar{\boldsymbol{M}}_{rsr}$ and $\boldsymbol{M}_{dec}$. We can see that the mask $\bar{\boldsymbol{M}}_{rsr}$ from our RSR module is more confident about the general shape and location of the inharmonious object, so it is more compact without holes or uncertain areas in the inharmonious region. Nevertheless, it is less accurate with the edges and lacking in many details. For the mask $\boldsymbol{M}_{dec}$ from decoder, it has utilized more information from multi-scale encoder features. Hence, it can better segment the inharmonious region with sharp and accurate edges, for example, tree branches in row 1 and the zebra ear in row 2. However, it may contain some holes or uncertain areas in the mask, like the human body in row 4. The adaptively combined mask $\boldsymbol{M}_{fnl}$ generally chooses the edges of the inharmonious region from $\boldsymbol{M}_{dec}$ and the inner part from $\bar{\boldsymbol{M}}_{rsr}$, which can be seen from the combination mask $\boldsymbol{G}$. Therefore, the combined mask $\boldsymbol{M}_{fnl}$ can have a compact mask without holes while keep the detailed and sharp edge information at the same time.

We also provide more qualitative comparisons between our model and other baselines in Figure 3.

4 Limitation

We have found that for a few cases when the inharmonious region is separated into several parts, our method may fail to detect some parts which are very small (see Figure 4) and only detect the part which appears to be the most inharmonious. In such cases, context information may need to be considered to segment the region completely, which is left for future work.

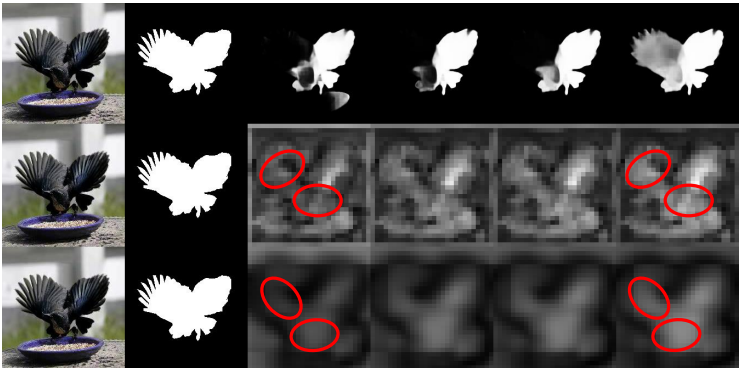

Figure 1: The update process of our RSR module at different iterations. At k-th iteration, we visualize the inharmonious region mask $\bar{\boldsymbol{M}}^k$ in the first row and similarity map $\boldsymbol{S}^{k-1,0}$ (*resp.*, $\boldsymbol{S}^{k-1,3}$) with $l=0$ (*resp.*, $l=3$) in the second (*resp.*, third) row (areas with large changes are highlighted by red circles). GT is the ground-truth inharmonious mask.

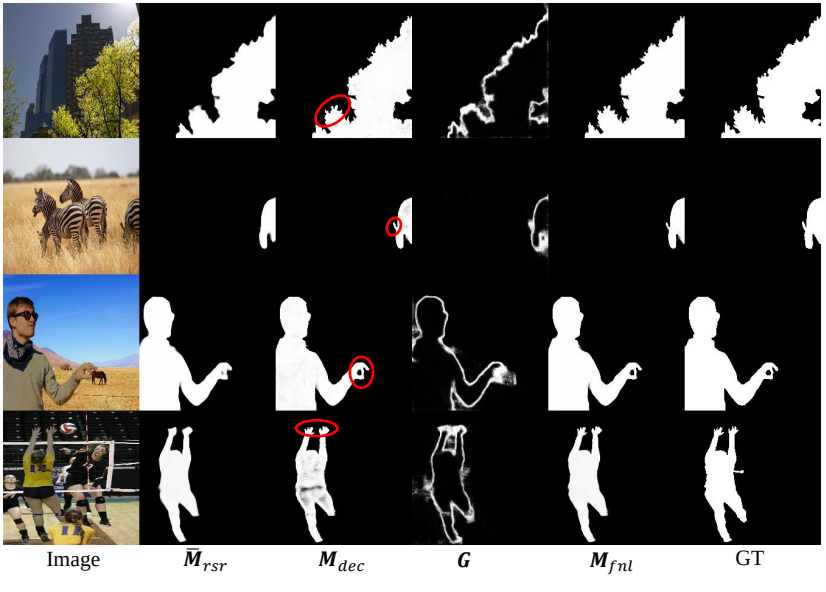

Figure 2: Visualization of three masks $\bar{\boldsymbol{M}}_{rsr}$, $\boldsymbol{M}_{dec}$, $\boldsymbol{M}_{fnl}$ and the combination mask $\boldsymbol{G}$. GT is the ground-truth inharmonious mask. Zoom in to see the edges and details which are highlighted by red circles.

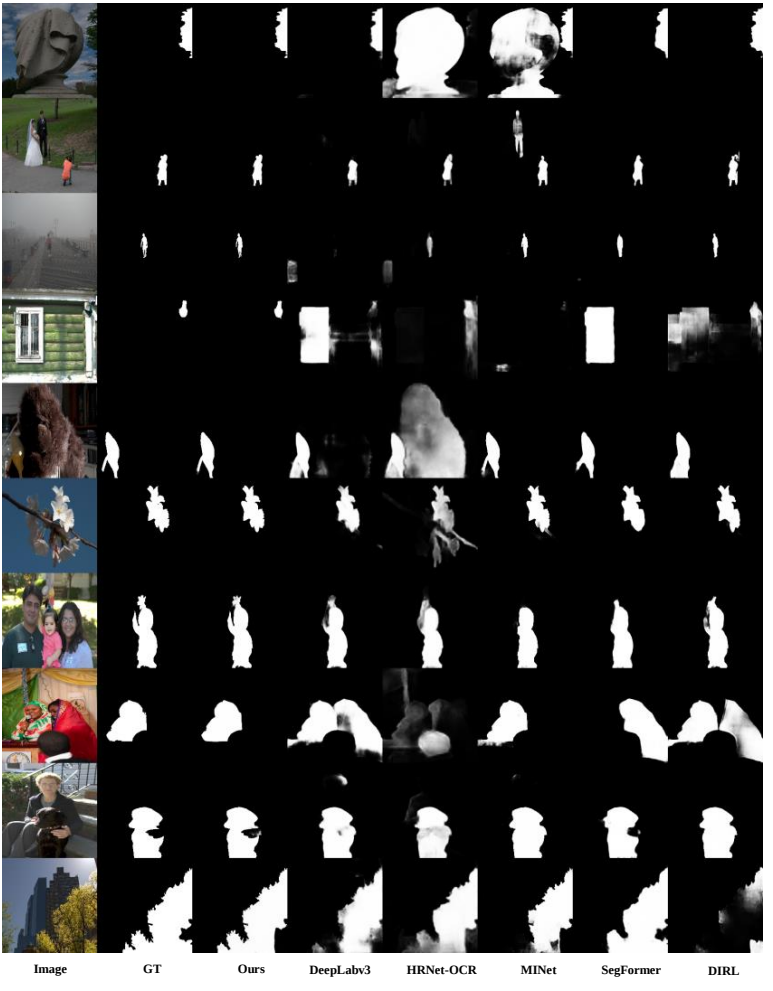

Figure 3: More qualitative comparison of our model with other state-of-the-art methods from related fields. GT is the ground-truth inharmonious mask.

Figure 4: Our method only detects part of the inharmonious region in a few cases when the inharmonious region is separated by background.