Supplementary for Inharmonious Region Localization with Auxiliary Style Feature

Penghao Wu
wupenghaocraig@sjtu.edu.cn
Li Niu*
ustcnewly@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 compare our model with other baselines on model size and speed in Sec. 1. We conduct more experiments to explore and discuss the effectiveness of our design in Sec. 2. Last, we test our model on synthetic images with multiple inharmonious regions in Sec. 3.

1 Comparison of Model Size and Speed

In this section, we compare the number of parameters, GFlops, and inference time between our methods and 4 competitive baselines in Table 1. All the evaluations are conducted on a single GTX TITAN X GPU. We can see that the number of parameters and the inference time of our AustNet are comparable with others, though the GFlops is larger. It can also be seen that the choice of semantic segmentation model of our AustNet-S largely determines the number of parameters and inference time. Therefore, we can choose between AustNet and AustNet-S according to our need when dealing with performance-speed/parameter trade-off.

2 Additional Experiments

2.1 More Qualitative Comparison

We provide more qualitative comparison between our method with other baselines in Fig 1. Our method can accurately detect the inharmonious regions in challenging cases.

2.2 Study on Style Feature

To investigate the discriminativeness of extracted style feature, we calculate the average inter-region feature similarity s_{inter} and average intra-region feature similarity s_{intra} over test images for models with and without our color-mapping module and ℓ_{sty}. The results are shown in Table 2. We observe that when our color mapping module or ℓ_{sty} is missing,

*Corresponding author.

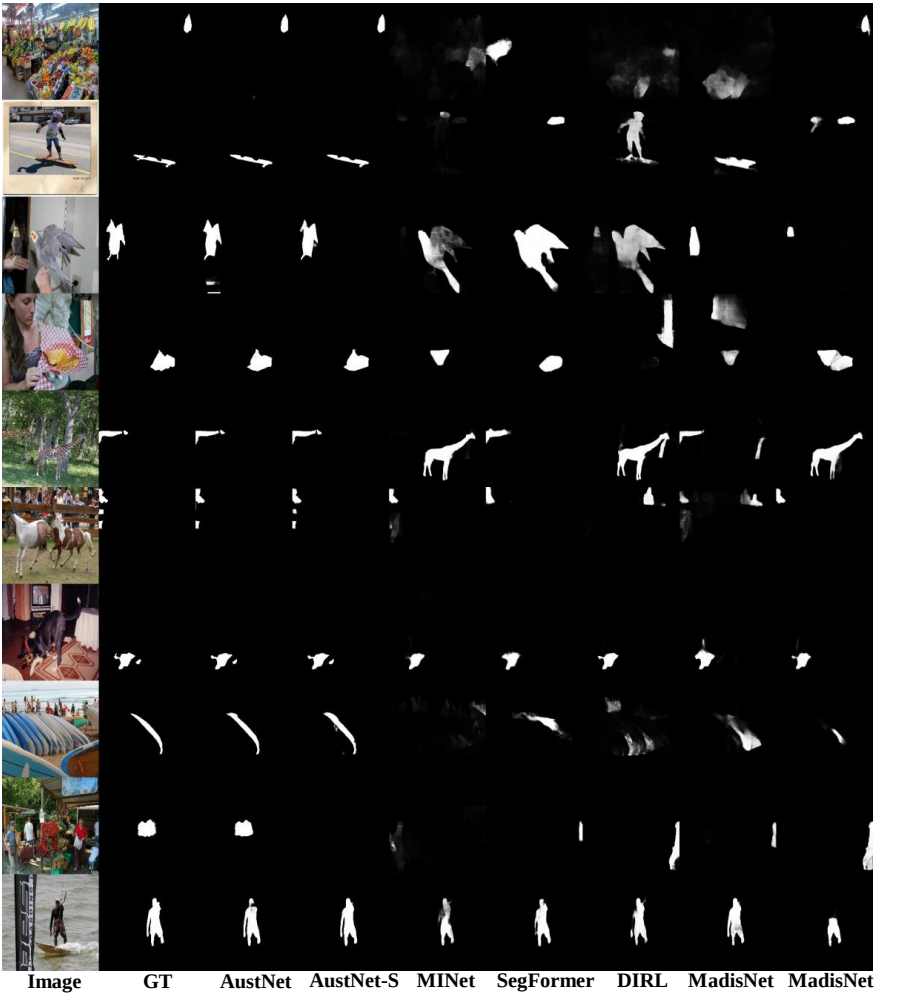

Figure 1: Qualitative comparison with baseline methods. GT is ground-truth mask.

	Number of Parameters	GFlops	Inference Time
SegFormer	48.28M	30.97	61.36ms
MINet	68.28M	116.01	63.95ms
DIRL	53.46M	104.67	63.80ms
MadisNet(UNet)	59.94M	82.90	45.60ms
MadisNet(DIRL)	56.69M	105.12	71.80ms
AustNet	50.05M	167.01	70.90ms
AustNet-S	120.61M	198.70	148.30ms

Table 1: Comparison of number of parameters, GFlops, and inference time between our model and other baselines.

	RGB+YUV	RGB+YUV+Color-mapping	RGB+YUV+Color-mapping + ℓ_{sty}
s_{inter}	0.3503	0.3435	0.1588
s_{intra}	0.4792	0.6212	0.7402

Table 2: Study on the discriminativeness of extracted style feature in terms of s_{inter} and s_{intra}.

s_{inter} and s_{intra} are close, indicating that the style features become less discriminative. The discrepancy between s_{inter} and s_{intra} is enlarged after adding the color mapping, and further enlarged after adding ℓ_{sty}, which means that our extracted style features are discriminative and informative enough to separate the inharmonious region from the background.

2.3 Study of the Voting Process in Different Stages

We visualize the voting score map and estimated inharmonious region mask in each decoder stage of our AustNet in Fig. 2. We can see that initial inharmonious region mask contains some mis-detected regions, and the voting score map is also not very accurate with some harmonious pixels receiving relatively low scores. However, as the decoding process goes, we introduce the estimated inharmonious region mask into the voting process, making both voting score maps and inharmonious region masks more and more accurate.

2.4 Benefit of Semantic Information in the Voting Process

When some harmonious objects have very different color style from the main background in the image, AustNet may mis-detect them as inharmonious region since the style features may not be perfectly learned. In this case, our AustNet-S with semantic information is designed to remedy this problem. To verify the benefit of semantic information in the voting process, we show several examples of the comparison between AustNet and AustNet-S in Fig. 3. For the first case, both our AustNet and AustNet-S successfully localize the inharmonious giraffe, but AustNet mis-detects a part of cobblestone as inharmonious. We choose a mis-detected point p (marked with yellow box), and visualize the style similarity matrix and semantic similarity matrix at point p. As we can see from the style similarity map $V^{sty}_{p,:}$, since the main background of grassland and woods is green, only a small part of cobblestone region has similar style features to this point p. Therefore, in the style voting module of AustNet, this point receives low scores from many points in the green background, which causes it to be viewed as inharmonious. However, with semantic information in the style voting module, the weights of score it receives from other cobblestone region will be large (as shown in $W^{sem}_{p,:}$), so it would receive a relatively high score and not be classified as inharmonious. Similarly, in

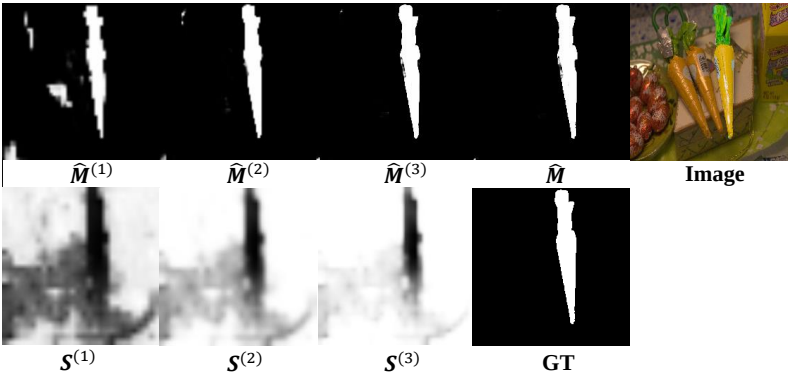

Figure 2: Visualization of the voting score map $S^{(k)}$ and auxiliary inharmonious region mask $\hat{M}^{(k)}$ in each decoder stage from our AustNet. Brighter region in $S^{(k)}$ means higher score.

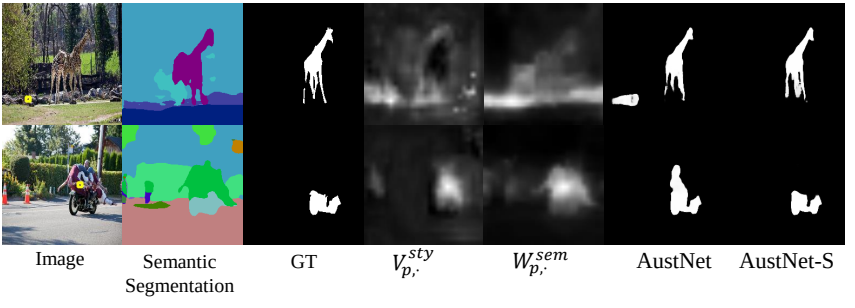

Figure 3: From left to right: input image, predicted segmentation mask, ground-truth inharmonious region mask, style similarity map $V^{sty}_{p,\cdot}$ and semantic similarity map $W^{sem}_{p,\cdot}$ of the specified point (yellow box in the input image), predicted inharmonious region mask from our AustNet and AustNet-S.

the second example, a person is also mis-detected as inharmonious by AustNet. After adding semantic information, the mis-detected point would receive high score from other people in the image, making the final estimation of AustNet-S only contain the motorcycle part.

3 Experiments on Multiple Inharmonious Regions

As introduced in the main paper, we build a set of test images with multiple disjoint inharmonious regions based on the HCOCO subset of iHarmony4. This test set contains 19482 images in total, with the number of inharmonious regions ranging from 2 to 9. We compare our AustNet and AustNet-S with the baseline MadisNet(DIRL) [1]. The evaluation results (AP, F_1, IoU) are (77.39, 0.6761, 54.03) for MadisNet(DIRL) and (87.86, 0.7828, 66.63) for our AustNet and (89.22, 0.7972, 68.71) for AustNet-S. Some visualization results are shown in Fig. 4.

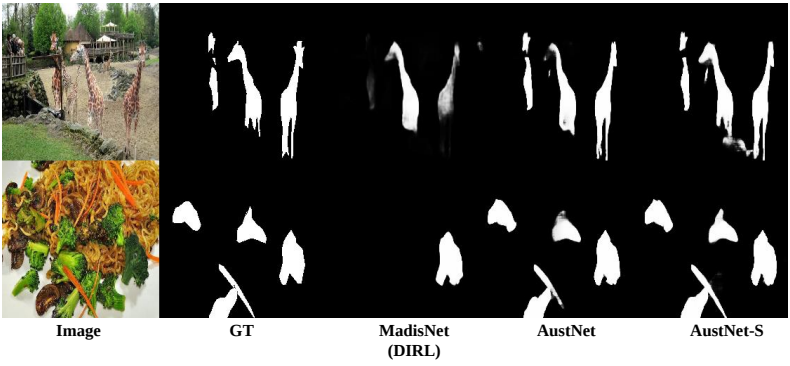

Figure 4: Visualization of the results on test images with multiple disjoint inharmonious regions.

References

[1] Jing Liang, Li Niu, Penghao Wu, Fengjun Guo, and Teng Long. Inharmonious region localization by magnifying domain discrepancy. In *AAAI*, 2022.