Supplementary Materials for Selectivity or Invariance: Boundary-aware Salient Object Detection

Jinming Su[1,3], Jia Li[1,3*], Yu Zhang[1], Changqun Xia[3] and Yonghong Tian[2,3*]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]National Engineering Laboratory for Video Technology, School of EE&CS, Peking University
[3]Peng Cheng Laboratory, Shenzhen, China

In this document, we provide additional implementation details of important components in the proposed **BANet**. In addition, we also investigate failure cases and compare run time with other methods to analyze our method more comprehensively. More representative results will also be provided.

1. Additional Implementation Details

1.1. Settings of TCS

Transitional compensation stream (TCS) adaptively provides the feature compensation for boundary and interior streams, and mediates the selectivity and invariance of these two branches. In other words, the features of TCS have medium selectivity and invariance, which ensures that detailed structures of salient objects to be correctly detected. In the implementation of TCS, we empirically adopt the fused features of $\theta_2(\pi_2)$ and $\theta_5(\pi_5)$ as input of TCS, which aggregates localization-aware fine-level features and semantic-aware coarse-level features to ensure the representation capability of features in TCS. The experimental validation is listed in Tab. 1. From Tab. 1, we can observe that combining $\theta_5(\pi_5)$ with the previous features $\theta_i(\pi_i)(i = 1, 2, ..., 4)$ is able to improve the performance, especially $\theta_2(\pi_2)$, which indicates that the combination of high-level ($\theta_5(\pi_5)$) and low-level ($\theta_2(\pi_2)$) features are useful for TCS to compensate features and promote performance of SOD.

Table 1. Performance of BANet with different fusion of features in TCS on ECSSD dataset. "θ_i and θ_j" $(i, j = 1, 2, ..., 5)$ means the features of $\theta_i(\pi_i)$ and $\theta_j(\pi_j)$ are fused as the the input of TCS. "θ_2 and θ_5" is used in BANet.

	MAE	F_β^w	F_β
θ_5	0.038	0.897	0.918
θ_1 and θ_5	0.036	0.906	0.925
θ_2 and θ_5	**0.035**	**0.908**	**0.929**
θ_3 and θ_5	0.037	0.901	0.921
θ_4 and θ_5	0.038	0.899	0.921

Table 2. Performance of BANet with different ISD-N in IPS and TCS on ECSSD dataset. "ISD-i and ISD-j" $(i, j = 1, 2, ...)$ means that IPS uses ISD-i and TCS uses ISD-j. "ISD-5 and ISD-3" is used in BANet.

	MAE	F_β^w	F_β
ISD-1 and ISD-1	0.044	0.886	0.918
ISD-3 and ISD-3	0.037	0.901	0.924
ISD-5 and ISD-5	**0.035**	0.907	0.927
ISD-7 and ISD-7	0.037	0.905	**0.929**
ISD-5 and ISD-1	0.036	0.904	0.924
ISD-5 and ISD-3	**0.035**	**0.908**	**0.929**

1.2. Settings of ISD in IPS and TCS

The features of different regions have different properties. The features of interior perception stream (IPS) need more invariance, while TCS needs invariance and selectivity both. In order to enhance the capability of extracting invariant features, we design integrated successive dilation module (ISD-N) that has the capability of perceiving various local contexts with the smallest dilation rate and the largest dilation rate of $2^N - 1$ for TCS and IPS. In our experiments, IPS utilizes ISD-5, which is enough to perceive the receptive field of the whole input image whose resolution is less than 500×500. In addition, we adopt ISD-3 in TCS because of the medium capability of extracting invariant features. The experimental validation is shown in the Tab. 2.

2. Failure Cases

We investigate the failure cases in our method as displayed in Fig. 2. Although sometimes our method is successful to deal with refections or shadows of salient objects, we find that our method may sometimes fail especially when the boundary is difficult to be detected. This problem exists widely in the existing methods as shown in Fig. 2. Although our method has a better boundary that can alleviate this problem to some extent, this problem still exists. This problem may need to consider more surrounding environ-

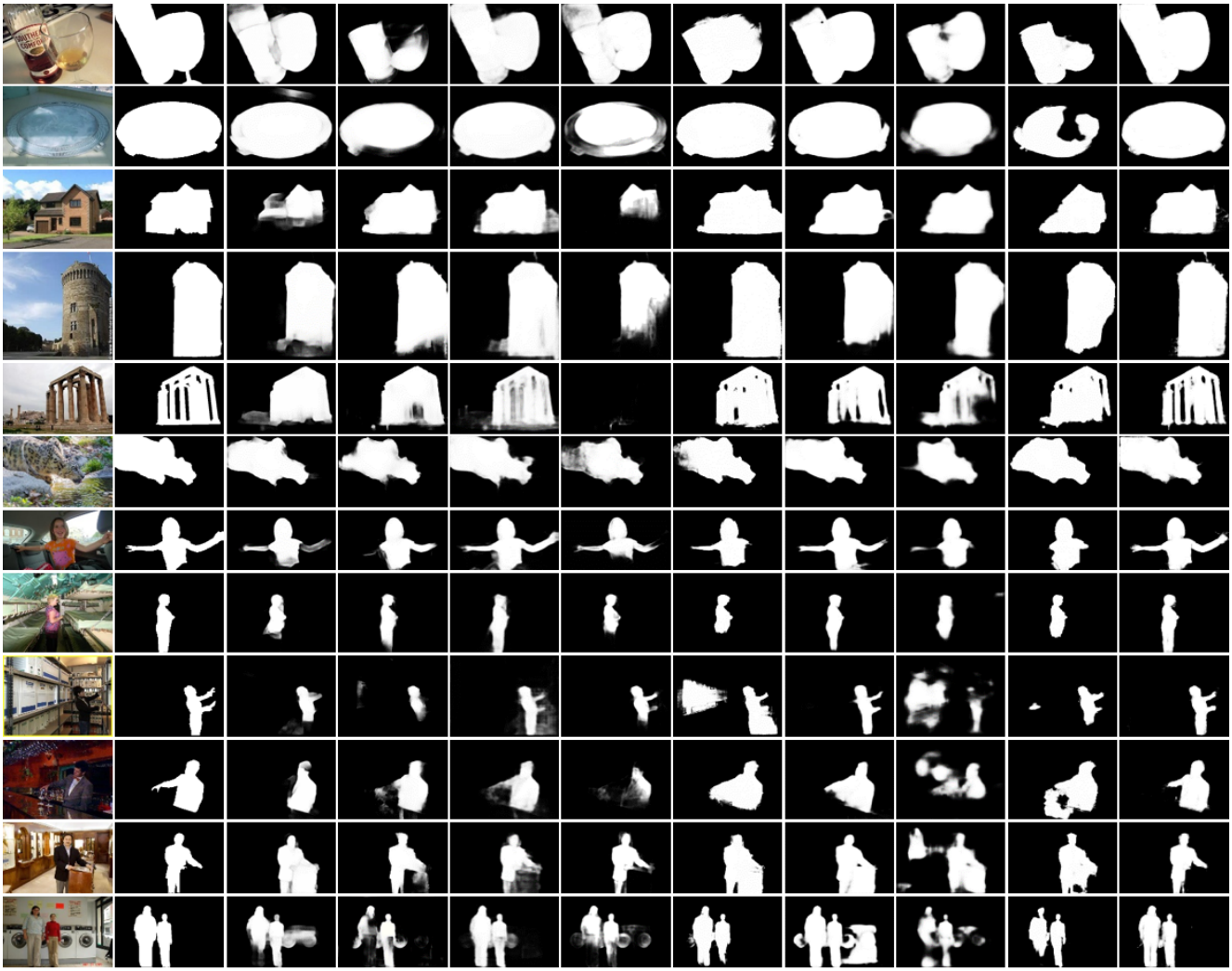

Figure 1. More examples of eight state-of-the-art methods and our approach.

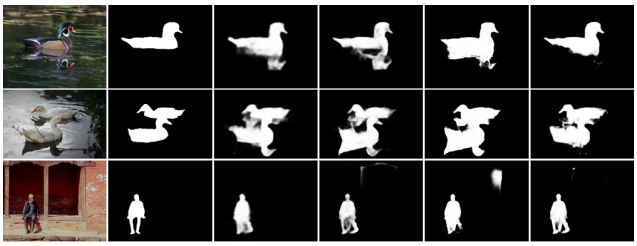

Figure 2. Representative failure cases.

ment and even semantic information to be solved well.

3. Run Time

To analyze our method more comprehensively, we show the run time of BANet and other approaches in Tab. 3. The table shows the average time for different methods to predict an image on ECSSD dataset. The evaluation is conducted with an unloaded NVIDIA RTX 2080ti GPU. BANet takes 0.048 second to produce a saliency map. We can see that the proposed method is faster than most state-of-the-art methods. Although C2SNet and RA are faster than our method in prediction speed, our performance far exceeds these two methods as shown in Tab. 1 of our paper.

Table 3. Average run time comparison with other methods on ECSSD. DSS and R3Net employ dense CRF.

	C2SNet	RA	PicaNet	R3Net	DGRL	RFCN	DSS	BANet
time/s	0.026	0.033	0.136	0.290	0.110	0.271	0.235	0.048

4. More Representative Results

We present additional representative results in Fig. 1 that we could not include in the paper due to space limitation.