Supplementary Materials for CLASS: Cross-Level Attention and Supervision for Salient Objects Detection

Lv Tang[1] and Bo Li*[2]

[1] State Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
[2] Youtu Lab, Tencent, Shanghai, China
luckybird1994@gmail.com
libraboli@tencent.com

1 Introduction

This supplemental material contains three parts:

- Section 2 gives more quantitative and qualitative experimental results to compare our CLASS net with the state-of-the-art methods.
- Section 3 gives an investigation of failure cases.
- Section 4 provides more comprehensive analyses of the proposed cross-level attention and cross-level supervision to further demonstrate the novelty of our method.

We hope this supplemental material can help you get a better understanding of our work.

2 More Quantitative and Qualitative Comparison

Due to the limitation of the paper length, we provide more quantitative and qualitative experimental results in this section.

2.1 Qualitative Comparison

As shown in Fig.1, we provide a comprehensive qualitative comparison of our method with other 13 methods on challenging cases. These visual examples can further demonstrate that our method is able to handle various challenging cases and produce accurate salient objects with high quality structure details.

* Correspondence should be addressed to Bo Li.

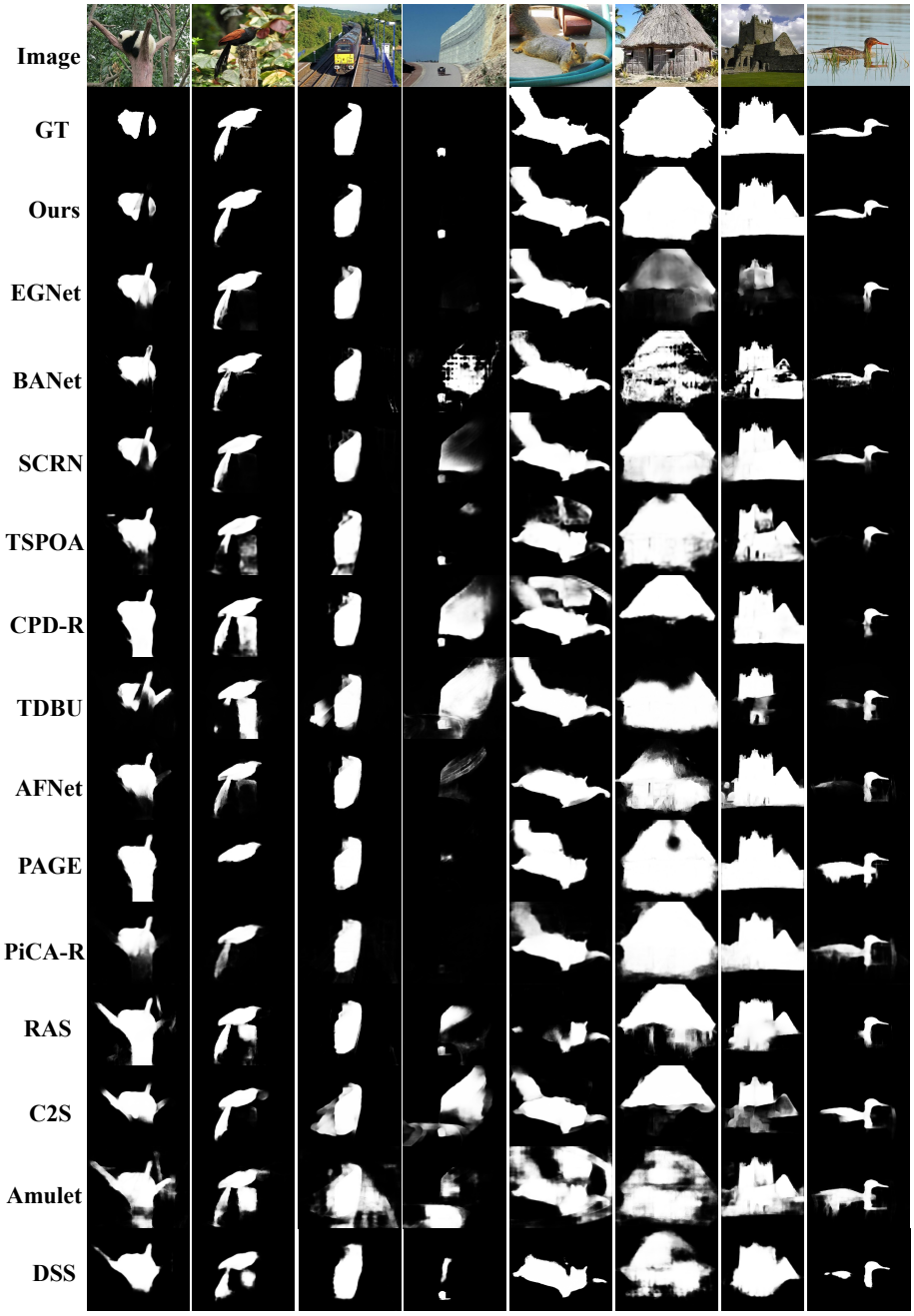

Fig. 1. More examples of 13 state-of-the-art methods and our approach.

2.2 Quantitative Comparison

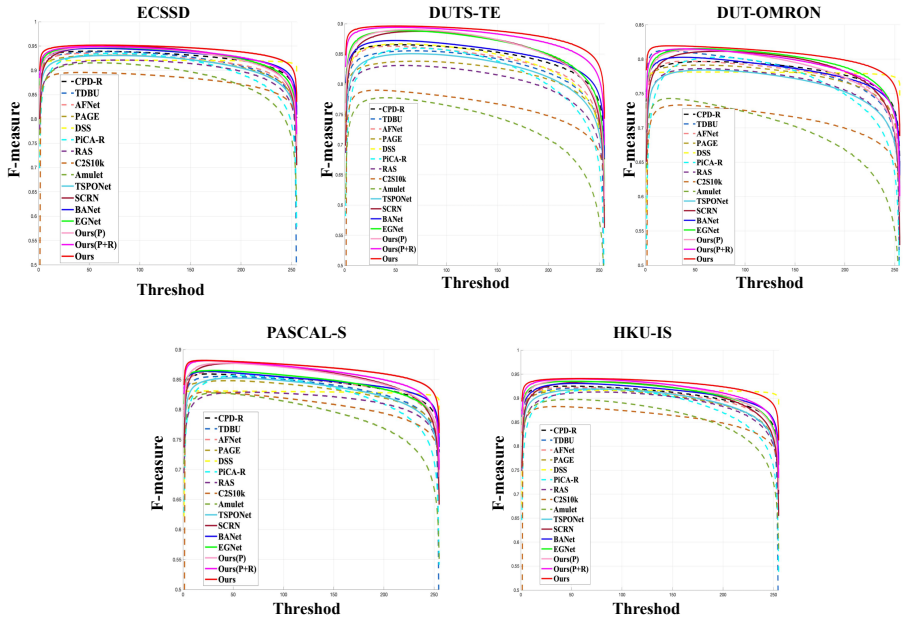

Fig. 2. Comparison of the F-measure curves across five benchmark datasets.

Table 1. Performace comparison between our approach (ResNet-50) and new state-of-the-art models.

Models	ECSSD			DUTS-TE			DUT-OMRON			PASCAL-S			HKU-IS		
	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
SCRN(ICCV2019)	0.918	0.927	0.037	0.808	0.885	0.04	0.746	0.837	0.056	0.827	0.842	0.062	0.896	0.916	0.034
BANet(ICCV2019)	0.923	0.924	0.035	0.808	0.885	0.04	0.746	0.832	0.059	0.823	0.845	0.069	0.9	0.913	0.032
EGNet(ICCV2019)	0.92	0.925	0.037	0.815	0.879	0.04	0.755	0.841	0.053	0.817	0.846	0.073	0.901	0.918	0.031
F3N(AAAI2020)	0.925	0.924	0.033	0.84	0.888	0.035	0.766	0.838	0.053	0.84	0.855	0.062	0.91	0.917	0.028
MINet(CVPR2020)	0.924	0.925	0.033	0.828	0.884	0.037	0.755	0.833	0.055	0.829	0.85	0.063	0.909	0.919	0.029
Ours	0.933	0.928	0.033	0.856	0.894	0.034	0.774	0.838	0.052	0.849	0.863	0.059	0.921	0.923	0.028

F-measure curves of different methods are displayed in Fig. 2, for overall comparisons. One can observe that our approach noticeably outperforms all the other state-of-the-art methods. These observations demonstrate the efficiency and robustness of our CLASS net across various challenging datasets.

To further demonstrate the efficiency and robustness of our CLASS net, we compare our method with two new state-of-the-art methods, including F3N [1] and MINet [2]. The results are reported in Table.1. It can be seen that our

Table 2. Performance on SOC of different attributes. The last row shows the whole performance on the SOC dataset.The best two results are in red and green fonts.

Attr	SCRN	EGNet	F3N	MINet	Ours
AC	0.759	0.756	0.784	0.79	0.784
BO	0.747	0.702	0.791	0.813	0.814
CL	0.766	0.726	0.757	0.77	0.773
HO	0.78	0.756	0.79	0.792	0.79
MB	0.719	0.687	0.761	0.708	0.75
OC	0.732	0.702	0.724	0.729	0.725
OV	0.781	0.764	0.793	0.785	0.785
SC	0.709	0.683	0.747	0.726	0.745
SO	0.645	0.614	0.668	0.652	0.68
Avg	0.738	0.71	0.757	0.752	0.761

Table 3. Average speed (FPS) comparisons between our approach (ResNet-50) and the previous state-of-the-art methods.

	Ours	BANet	SCRN	AFNet	PAGE	CPD
Size	352×352	400×300	352×352	224×224	224×224	352×352
FPS	40	13	38	26	25	62
	EGNet	PiCA	RAS	C2SNet	Amulet	DSS
Size	400×300	224×224	400×300	400×300	256×256	224×224
FPS	12	7	45	30	16	12

method consistently outperforms other methods across five benchmark datasets. SOC [3] is a new challenging dataset with nine attributes. In Table.2, we evaluate the mean F-measure score of our method in this dataset. We can see the proposed model achieves the competitive results among most of attributes and the overall score is best.

Average speed (FPS) comparisons among different methods (tested in the same environment) are also reported in Table.3. As can be seen, our approach is one of the fastest methods which can run in real time. Although there is a small gap between our method and two fastest methods CPD [4] and RAS [5] in fps, our method performs much better on other evaluation metrics. This observation can further demonstrate the efficiency of our CLASS net.

3 Failure Cases

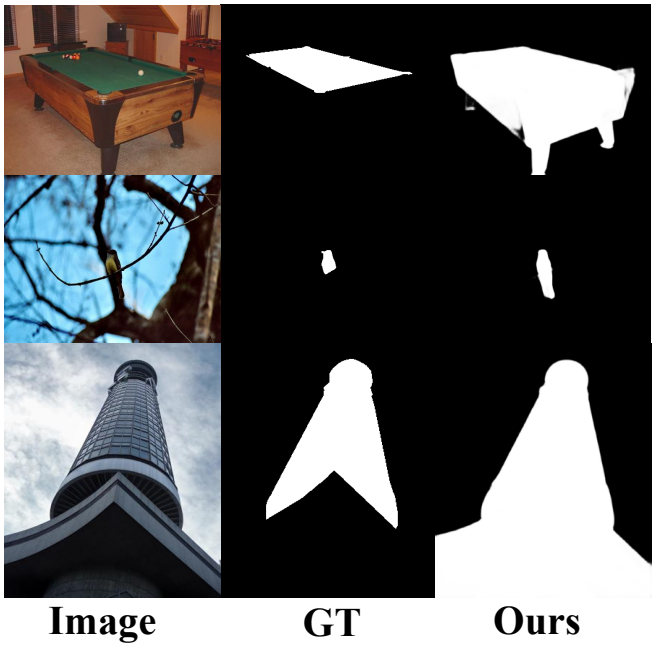

Fig. 3. Examples which correct the ground truth.

As demonstrated, our method has achieved impressive performance in accurate salient object detection. However, there are still some cases where our detection results are inconsistent with the ground truth.

It is noteworthy that being inconsistent with the ground truth does not mean all these cases are necessarily inferior results. As shown in Fig. 3, some of our results can even correct the errors in the ground truth by maintaining the wholeness of salient objects.

Besides, we show several typical failure cases of our method in Fig. 4. From the row of 1 and 2 in Fig. 4, we can observe that in some controversial scenes, our method tend to only segment the top salient object in the image. In the row of 3 and 4, our method labels all relevant regions of the salient objects while the ground truth only labels parts of the salient objects. This situation can be caused by the proposed cross-level attention mechanism, which is designed to keep the wholeness of the salient objects. In the fifth row, our method fails to detect the subjective salient object. In the last row of Fig. 4, our method cannot

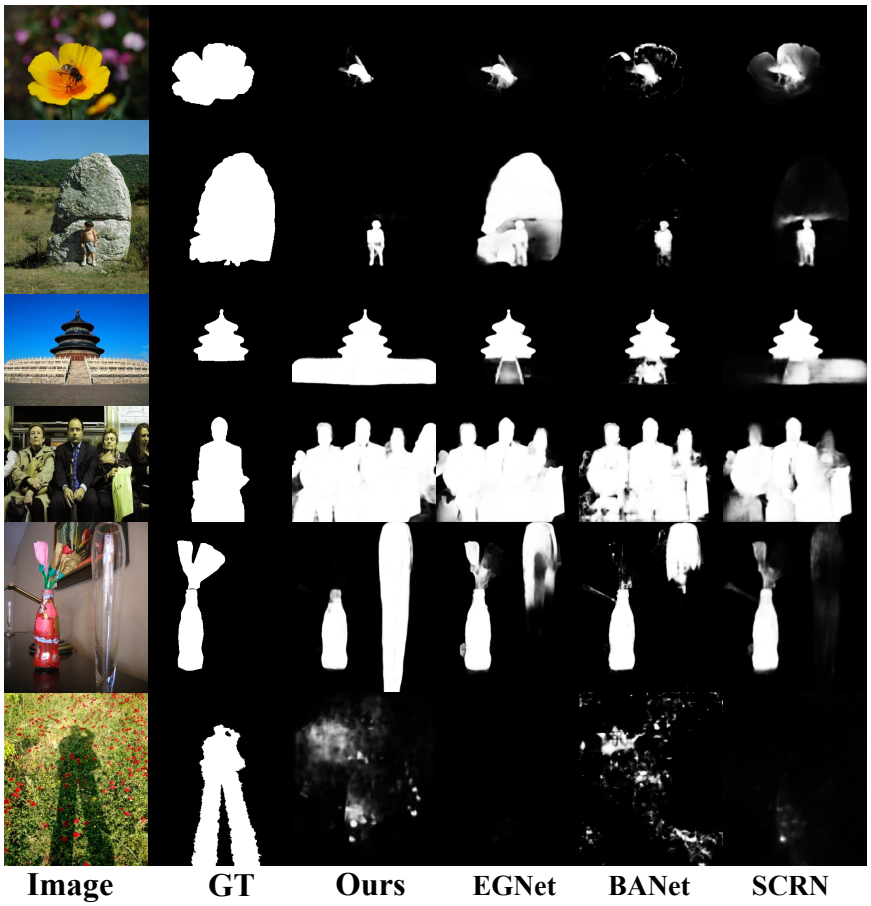

Fig. 4. Failure cases.

detect the true salient object, that can be caused by the bias of the training data. In most training images, shadows are not labeled as salient object. It is worth noting that these failure cases are also hard to most of the other state-of-the-art methods. Therefore, there is still a room for the improvement of our CLASS net.

4 More Analyses of the proposed CLA and CLS

4.1 Analysis of Cross-level Attention

To further demonstrate the novelty of the proposed cross-level attention, we compare our attention module with the common non-local attention [6], which relies on a single layer feature. The quantitative results are shown in Table 4. We first remove all attention module in the proposed model as a baseline. Then we replace the cross-level attention module with the common non-local attention. As can be seen, using the common non-local attention can improve the performance of baseline. However, the common non-local models [6] rely on a single layer feature, they cannot leverage the advantages of features in different levels to capture sufficient long range dependencies. The proposed cross-level attention outperforms the common non-local attention and achieves the best results on all datasets.

Table 4. Performance comparison of different attention settings. The **Baseline** here refers to without any attention module. The **Common Non-Local** means we use common non-local module to replace the proposed cross-level attention module.

Configurations	ECSSD			DUTS-TE			DUT-OMRON			PASCAL-S			HKU-IS		
	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE	F_β	S_m	MAE
Baseline	0.923	0.922	0.037	0.845	0.888	0.036	0.767	0.839	0.055	0.842	0.856	0.062	0.910	0.915	0.031
Common Non-Local	0.928	0.926	0.035	0.849	0.889	0.036	0.769	0.839	0.055	0.844	0.856	0.062	0.915	0.918	0.029
Cross-Level Attention(Ours)	0.933	0.928	0.033	0.856	0.894	0.034	0.774	0.842	0.052	0.849	0.863	0.059	0.921	0.923	0.028

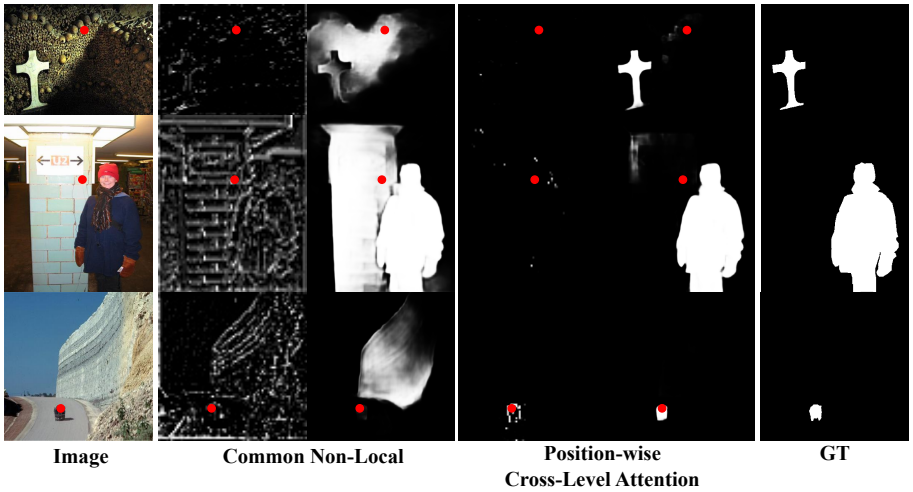

Fig. 5. Visual comparison of position-wise non-local attention.

We also provide some visualization results of attention module for qualitative comparison. For position attention, since the overall attention map is calculated on all positions, there is an corresponding sub-attention map for each specific point in the image. In Fig. 5, for each input image, we select a point (marked by red dot) and show its corresponding sub-attention map as well as the saliency result of the image respectively. We observe that for some "salient-like" positions, common non-local sub-attention map provides a strong connection with real salient regions, which can lead a wrong prediction in these positions. While in the proposed cross-level sub-attention maps, these positions almost have no dependencies on the real salient regions. For the position in the real salient region(third row), the cross-level sub-attention map only highlights the real salient object while the common non-local sub-attention map highlights the interfering region. These visual comparisons show our position-wise cross-level attention can better locate the salient objects and suppress the non-salient regions.

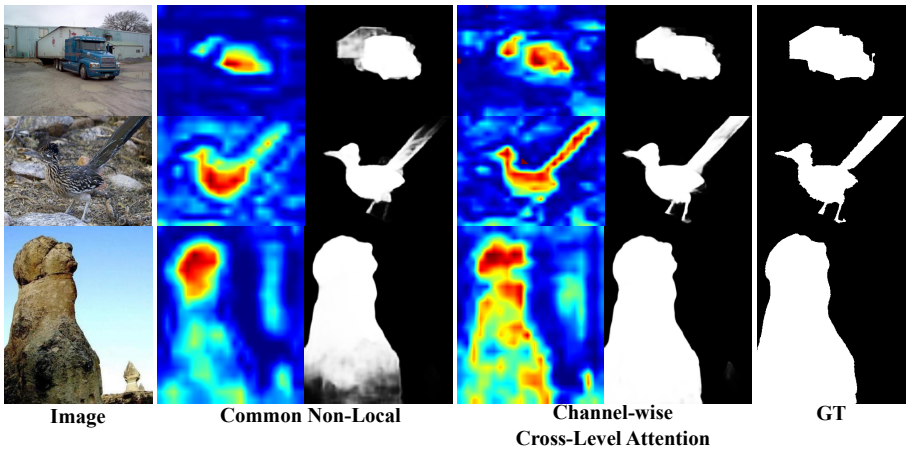

Fig. 6. Visual comparison of channel-wise non-local attention.

For channel attention, it is hard to give comprehensible visualization about the attention map directly. Instead, we fuse the most attended channels provided by common non-local channel attention module and the proposed cross-level channel attention module to see whether they highlight clear semantic areas. In Fig. 6, we can find that the response of salient semantic becomes more noticeable after two kinds channel attention module enhances. However, our cross-level channel-wise attention can better keep the wholeness of salient by highlighting the regions which have different visual appearances (different color, texture and luminance) with the main salient object.

In short, these visualizations further demonstrate the necessity of capturing cross-level long-range dependencies for improving feature representation in SOD.

4.2 Analysis of Cross-level Supervision

In Fig. 7, we provide a visual comparison with different supervision settings. As can be seen, by adding the region-level supervision, our model can better maintain the structural details and boundaries of the salient objects. When add the object-level supervision, our model can highlight the salient object more uniformly. The F-measure curves of different supervision settings are also provided in Fig. 2. these visualizations further shows the effectiveness of our proposed cross-level supervision.

Fig. 7. Visual comparison of different supervision settings.

References

1. Wei, J., Wang, S., Huang, Q.: F3net: Fusion, feedback and focus for salient object detection. CoRR **abs/1911.11445** (2019)
2. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: CVPR, IEEE (2020) 9410–9419
3. Fan, D., Cheng, M., Liu, J., Gao, S., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV (15). Volume 11219 of Lecture Notes in Computer Science., Springer (2018) 196–212
4. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR. (2019) 3907–3916(2019)
5. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: ECCV(2018). Volume 11213., Springer (2018) 236–252
6. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR. (2018) 7794–7803(2018)