Supplementary Material: Visual Saliency Transformer

Nian Liu[1*] Ni Zhang[2*] Kaiyuan Wan[2] Ling Shao[1] Junwei Han[2†]
[1]Inception Institute of Artificial Intelligence [2]Northwestern Polytechnical University
{liunian228, nnizhang.1995, kaiyuan.wan0106, junweihan2010}@gmail.com, ling.shao@ieee.org

Table 1: Ablation studies of our proposed model on RGB SOD datasets. "RC" means RGB convertor. "Bili" denotes bilinear upsampling and "F" means multi-level token fusion. "TMD" denotes our proposed token-based multi-task decoder, while "C2D" means using the conventional two-stream decoder to perform saliency and boundary detection without using task-related tokens. The best results are labeled in **blue**.

Settings	DUTS [25]				HKU-IS [10]				PASCAL-S [12]				SOD [18]			
	S_m	maxF	$E_\xi^{\max}$	MAE	S_m	maxF	$E_\xi^{\max}$	MAE	S_m	maxF	$E_\xi^{\max}$	MAE	S_m	maxF	$E_\xi^{\max}$	MAE
Baseline	0.824	0.780	0.909	0.071	0.858	0.854	0.938	0.075	0.826	0.795	0.878	0.096	0.802	0.803	0.880	0.100
+RC	0.827	0.785	0.913	0.070	0.860	0.856	0.939	0.074	0.830	0.797	0.879	0.095	0.804	0.805	0.880	0.100
+RC+Bili	0.867	0.835	0.929	0.048	0.901	0.901	0.956	0.044	0.856	0.827	0.891	0.074	0.833	0.836	0.891	0.077
+RC+RT2T	0.881	0.856	0.934	0.043	0.914	0.918	0.961	0.037	0.864	0.838	0.896	0.070	0.844	0.850	0.894	0.069
+RC+RT2T+F	0.895	0.874	**0.939**	0.039	0.925	0.932	0.966	0.032	0.871	0.845	0.897	0.068	0.851	0.861	0.899	0.068
+RC+RT2T+F+TMD	**0.896**	**0.877**	**0.939**	**0.037**	**0.928**	**0.937**	**0.968**	**0.030**	**0.873**	**0.850**	**0.900**	**0.067**	**0.854**	**0.866**	**0.902**	**0.065**
+RC+RT2T+F+C2D	0.891	0.870	0.937	0.040	0.924	0.931	0.966	0.033	0.869	0.844	0.896	0.069	0.852	0.860	0.898	0.067

1. Ablation Study on RGB SOD Datasets

We further report the results of ablation studies on four RGB SOD datasets, *i.e.*, DUTS, HKU-IS, PASCAL-S, and SOD, in Table 1 to demonstrate the effectiveness of our VST model components.

The baseline model is using transformer encoder to extract patch tokens $\boldsymbol{T}_r^{\mathcal{E}}$ and then directly using $\boldsymbol{T}_r^{\mathcal{E}}$ to predict the saliency map with 1/16 scale by using MLP on each patch token. Based on the baseline, we insert RGB convertor right after the transformer encoder, shown as "+RC" in Table 1. Compared to the baseline, RC brings performance gains especially on the DUTS and PASCAL-S datasets, which demonstrates its effectiveness. For other components, *i.e.*, RT2T, multi-level token fusion, and multi-task transformer decoder, we get consistent conclusions with the ablation studies on RGB-D SOD datasets as follows.

First, using bilinear upsampling ("+RC+Bili") can significantly improve the model performance while using our proposed RT2T ("+RC+RT2T") can further bring performance gains, hence demonstrating the effectiveness of our proposed RT2T. Second, based on "+RC+RT2T", multi-level token fusion ("+RC+RT2T+F") can lead to better performance on all four datasets, which verifies its effectiveness. Third, using the multi-task transformer decoder ("+RC+RT2T+F+TMD") can improve the model performance on all four datasets and it is also superior to the conventional two-stream decoder ("+RC+RT2T+F+C2D").

To this end, the results of ablation studies on both RGB and RGB-D SOD datasets strongly demonstrate the effectiveness of our proposed VST components.

2. Layer Number Study

We conduct experiments to study the optimal numbers of different transformer layers, *i.e.*, $L^{\mathcal{C}}$ in the transformer convertor and $L^{\mathcal{D}}$ in the multi-task transformer decoder, jointly considering computational costs and model performance. Note that there are three decoder modules at three scales in the multi-task transformer decoder, thus we set different transformer layer numbers for them, *i.e.*, $L_3^{\mathcal{D}}$ for 1/16 scale, $L_2^{\mathcal{D}}$ for 1/8 scale, and $L_1^{\mathcal{D}}$ for 1/4 scale. The experimental results on four RGB-D SOD datasets, *i.e.*, NJUD, DUTLF-Depth, STERE, and LFSD, are given in Table 2.

In our initial model setting, we set $L^{\mathcal{C}} = L_3^{\mathcal{D}} = 8$. Since $L_2^{\mathcal{D}}$ and $L_1^{\mathcal{D}}$ are used at relatively large scales, we initially set both of them to 4, as shown in row I in Table 2. Then, we start to change the numbers of different layers.

We first reduce $L_2^{\mathcal{D}}$ and $L_1^{\mathcal{D}}$ from 4 to 2 to save computational costs. The experimental results on row II show that it can get comparable performance with less computational costs compared with row I. Hence, we set $L_2^{\mathcal{D}} = L_1^{\mathcal{D}} = 2$ and start to change $L_3^{\mathcal{D}}$ from 8 to 6, 4, 2, respectively, which are shown in row III, IV, V in Table 2. We find that as $L_3^{\mathcal{D}}$

*Equal contribution.
†Corresponding author.

Table 2: Comparison of using different numbers of transformer layers in our VST model. The final model setting is labeled in **blue**.

ID	Layer Num				MACs	Params	NJUD [7]				DUTLF-Depth [22]				STERE [19]				LFSD [11]			
	$L^{\mathcal{C}}$	$L_3^{\mathcal{D}}$	$L_2^{\mathcal{D}}$	$L_1^{\mathcal{D}}$	(G)	(M)	S_m	maxF	$E_\xi^{\max}$	MAE	S_m	maxF	$E_\xi^{\max}$	MAE	S_m	maxF	$E_\xi^{\max}$	MAE	S_m	maxF	$E_\xi^{\max}$	MAE
I	8	8	4	4	48.35	119.30	0.925	0.925	0.955	0.033	0.940	0.947	0.966	0.026	0.910	0.902	0.948	0.039	0.878	0.884	0.914	0.066
II	8	8	2	2	36.78	113.39	0.923	0.922	0.955	0.035	0.943	0.947	0.968	0.025	0.911	0.904	0.948	0.039	0.874	0.878	0.908	0.069
III	8	6	2	2	36.20	110.43	0.921	0.920	0.952	0.036	0.940	0.945	0.966	0.026	0.910	0.904	0.948	0.040	0.875	0.883	0.911	0.067
IV	8	4	2	2	35.61	107.47	0.921	0.920	0.951	0.036	0.942	0.947	0.968	0.026	0.911	0.904	0.949	0.040	0.876	0.880	0.912	0.068
V	8	2	2	2	35.03	104.52	0.922	0.921	0.952	0.036	0.940	0.944	0.965	0.026	0.912	0.906	0.949	0.039	0.873	0.875	0.908	0.068
VI	6	4	2	2	33.30	95.65	0.923	0.921	0.952	0.036	0.943	0.948	0.968	0.024	0.913	0.906	0.949	0.039	0.875	0.878	0.912	0.067
VII	**4**	**4**	**2**	**2**	30.99	83.83	0.922	0.920	0.951	0.035	0.943	0.948	0.969	0.024	0.913	0.907	0.951	0.038	0.882	0.889	0.921	0.061
VIII	2	4	2	2	28.68	72.00	0.923	0.921	0.953	0.036	0.938	0.943	0.963	0.028	0.912	0.906	0.950	0.039	0.881	0.887	0.917	0.062

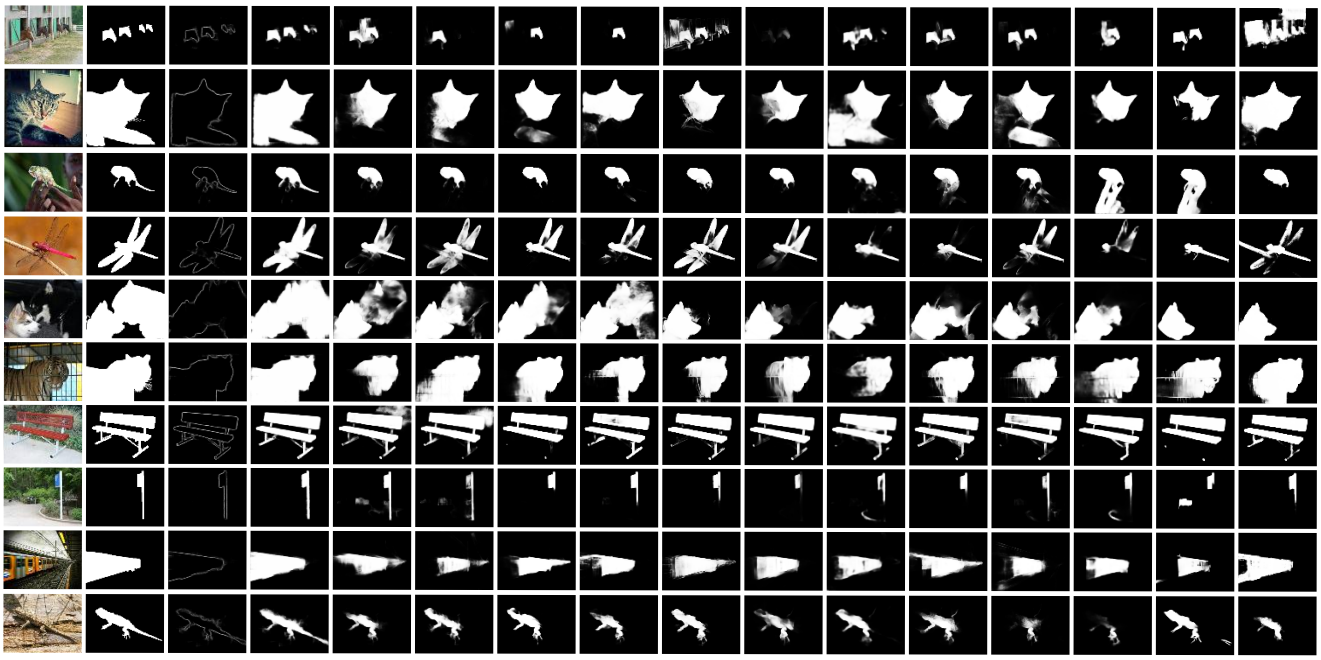

Figure 1: Qualitative comparison against state-of-the-art RGB SOD methods. (GT: ground truth; VST-B: Boundary maps predicted by our VST.)

decreases, the computation costs decrease gradually while the results are generally comparable. However, the model performance on row IV is better than that on row V on DUTLF-Depth and LFSD datasets. Thus, we set $L_3^{\mathcal{D}} = 4$ and start to change $L^{\mathcal{C}}$ from 8 to 6, 4, 2, respectively, which are shown in rows VI, VII, VIII. It can be seen that the performance on row VII is the best and the model has acceptable computational costs. Hence, we set $L^{\mathcal{C}} = L_3^{\mathcal{D}} = 4$ and $L_2^{\mathcal{D}} = L_1^{\mathcal{D}} = 2$ as our final model setting.

3. More Visual Comparison with State-of-the-art Methods

We give more visual comparison results with the state-of-the-art RGB and RGB-D SOD methods in Figure 1 and Figure 2, respectively. It shows that our VST model can handle well in many challenging scenarios, *i.e.*, big salient objects, cluttered backgrounds, foregrounds and backgrounds with very similar appearance, etc, while existing methods are heavily disturbed in these scenarios. Besides, we also show the boundary maps predicted by our RGB VST and RGB-D VST models in Figure 1 and Figure 2, respectively. It can be seen that our models can predict clear boundaries for salient objects.

References

[1] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for rgb-d salient object detection. In *ECCV*, pages 520–538, 2020.

[2] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, pages 275–292, 2020.

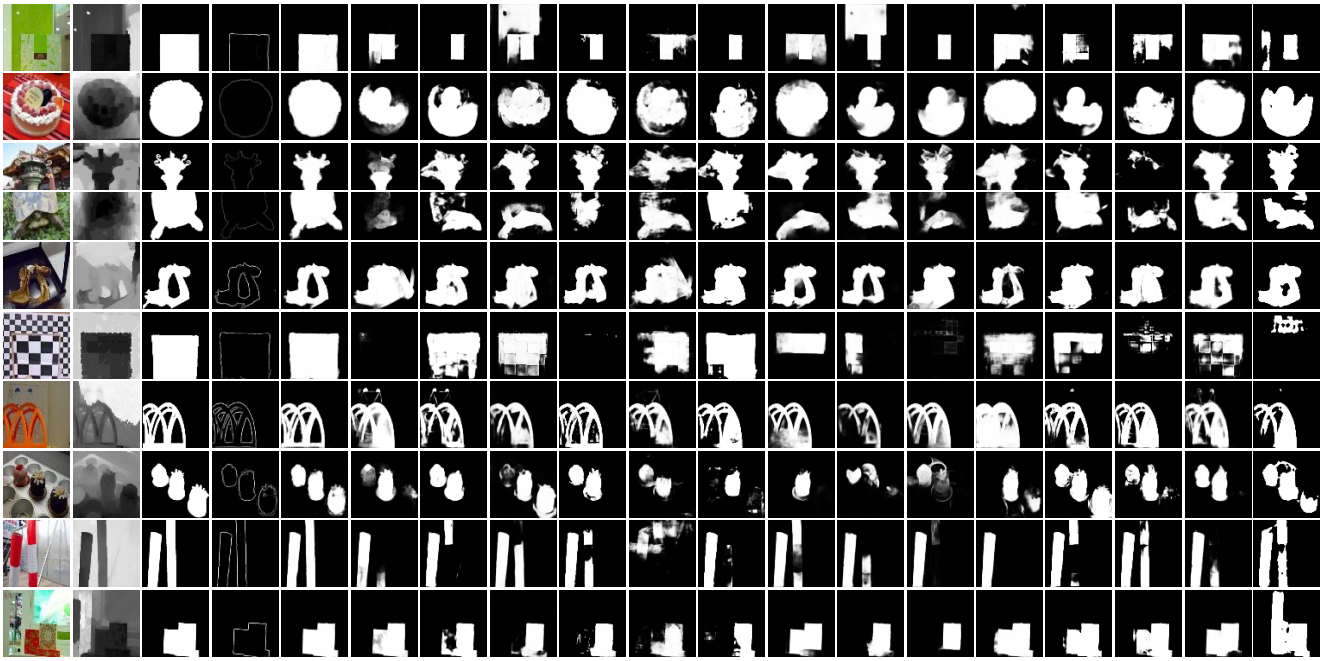

Figure 2: Qualitative comparison against state-of-the-art RGB-D methods. (GT: ground truth; VST-B: Boundary maps predicted by our VST.)

[3] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, pages 1623–1632, 2019.

[4] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, pages 3052–3062, 2020.

[5] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In *ECCV*, pages 702–721, 2020.

[6] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, pages 52–69, 2020.

[7] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119, 2014.

[8] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. Rgb-d salient object detection with cross-modality modulation and selection. In *ECCV*, pages 225–241, 2020.

[9] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for rgb-d salient object detection. In *ECCV*, pages 665–681, 2020.

[10] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[11] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014.

[12] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.

[13] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019.

[14] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018.

[15] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *CVPR*, pages 13756–13765, 2020.

[16] Yi Liu, Qiang Zhang, Dingwen Zhang, and Jungong Han. Employing deep part-object relationships for salient object detection. In *ICCV*, pages 1232–1241, 2019.

[17] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for rgb-d salient object detection. In *ECCV*, pages 346–364, 2020.

[18] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR Workshops*, pages 49–56, 2010.

[19] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012.

[20] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *ECCV*, pages 235–252, 2020.

[21] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.

[22] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019.

[23] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*, pages 9060–9069, 2020.

[24] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019.

[25] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.

[26] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13025–13034, 2020.

[27] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019.

[28] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, pages 8582–8591, 2020.

[29] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In *ECCV*, pages 374–390, 2020.

[30] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for rgb-d saliency detection. In *CVPR*, pages 3472–3481, 2020.

[31] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019.

[32] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020.

[33] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pages 646–662, 2020.

[34] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, pages 9141–9150, 2020.