# Deep RGB-D Saliency Detection with Depth-Sensitive Attention and Automatic Multi-Modal Fusion — Supplemental Material

**Overview** In the supplementary material, the content is organized as follows. We first depict the searched multi-modal fusion module and provide some interesting observations in Section A. Then, we provide the feature map visualizations of the RGB branch with or without our depth-sensitive attention module (DSAM) in Section B.

## A. Searched Fusion Module Visualization

The searched four types of cells, *i.e.* the MM, MS, GA and SR cell, are depicted in Fig.A, B, C, D, respectively.

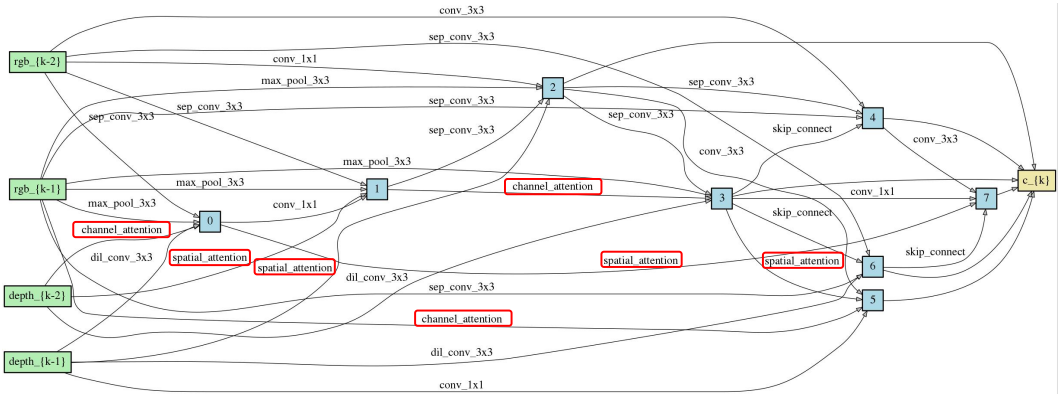



Figure A. The detailed architecture of the searched MM cell for multi-modal fusion. The highlighted red boxes represent the attention operations, which have a large proportion in MM cell, and the connections to RGB and depth features are asymmetric.

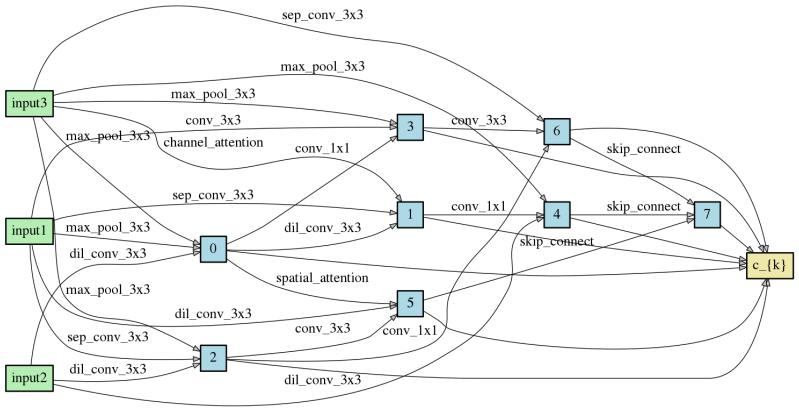



Figure B. The detailed architecture of the searched MS cell for multi-scale fusion.

We also notice some interesting observations as follows: 1) From Fig.A, we observe that the numbers of operations connected to RGB features (12 operations) are more than those connected to depth features (7 operations), which indicates that the connections to RGB and depth features are asymmetric. 2) The MM cell tends to select many attention operations

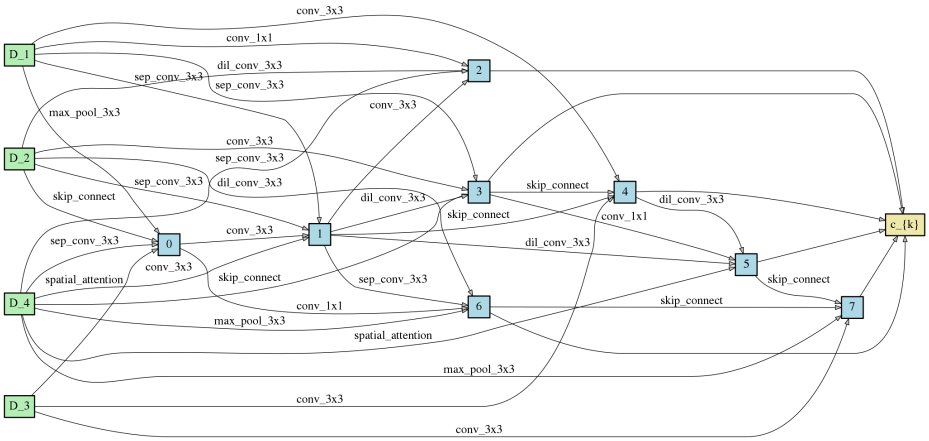



Figure C. The detailed architecture of the searched GA cell for global context aggregation, which has a dense and diverse operation selections.

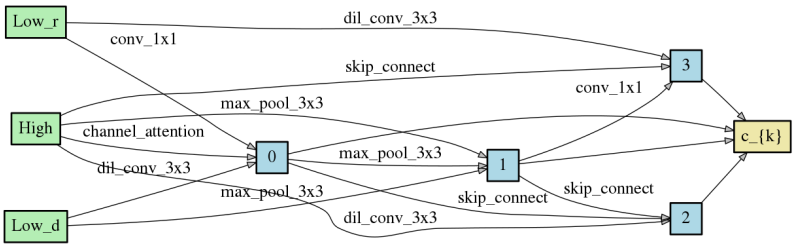



Figure D. The detailed architecture of the searched SR cell for spatial information restoration. Low_r and low_d denote the low-level RGB and depth features, respectively. The "High" represents the high-level features.

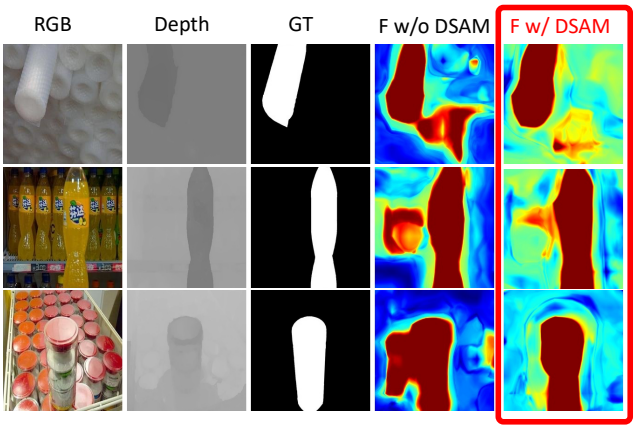
(*i.e*. spatial or channel attention) when performing multi-modal feature fusion. Such a observation is consistent with the third design principle mentioned in Section 3.1 of our manuscript. 3) As shown in Fig.D, in the SR cell, the operation numbers connected to low-level RGB and depth features are both 2, which indicates that low-level features play the role of complementary feature fusion, which is consistent with the second design principle in Section 3.1.

## B. Feature Map Visualization

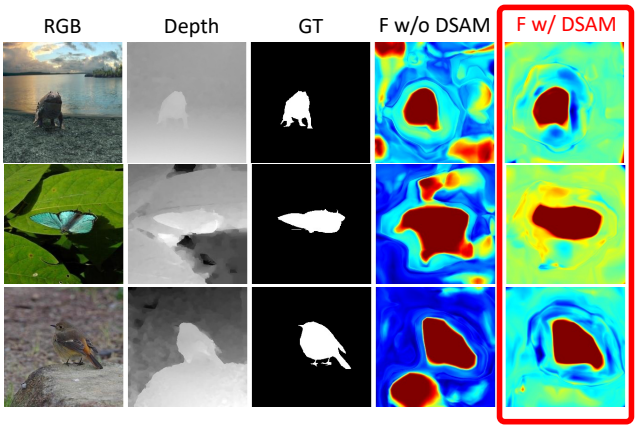
To analyze the effects of our DSAM, Fig.E shows some visualization results of the RGB feature maps with or without the DSAM (*i.e*. the $r5$ features in our manuscript) by CAM [1]. From these results, we can observe that the proposed DSAM can significantly pay more attention to the salient objects and eliminate the background distraction under various challenging scenarios, for example, cluttered objects in Fig.E (a), similar texture in Fig.E (b), other complex scenarios in Fig.E (c).

## References

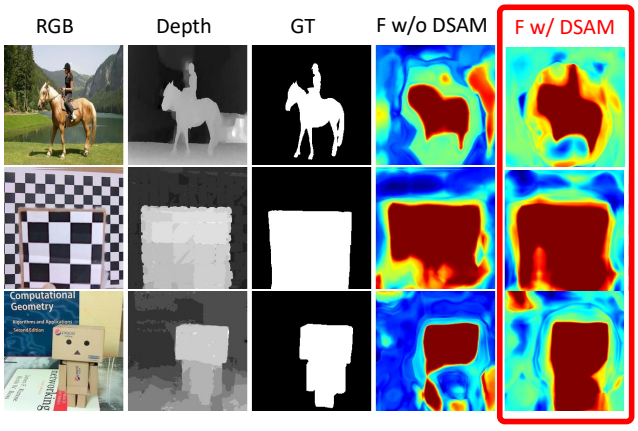
[1] B. Zhou, A. Khosla, Àgata Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2929, 2016. 2, 3




(a) Cluttered objects. We can observe that our DSAM effectively reduces the background distraction.




(b) Similar texture. We discover that our DSAM pays more attention to the salient objects.




(c) Other complex scenarios. We can observe that DSAM ensures the structural integrity of salient objects to some extent.

Figure E. The class activation mapping (CAM) [1] visualization results of the RGB feature maps $r5$. F w/ DSAM denotes the feature map results of ours.