Revisiting Multi-level Feature Fusion: A Simple yet Effective Network for Salient Object Detection

Yu Qiu[1], Yun Liu[2], Xiaoxu Ma[2], Lei Liu[1], Hongcan Gao[2], Jing Xu[1]

[1]College of Artificial Intelligence, [2]College of Computer Science, Nankai University, Tianjin, China

Introduction

- This paper presents an Automatic Top-Down Fusion (ATDF) model which is able to automatically flow the global information at the top sides of CNNs into bottom sides. Each side adds a novel valve module to receive the specifically useful and instructive global information to guide its learning.
- The top semantic information can guide the learning of bottom layers, and the bottom side outputs can accurately predict both the location and details of salient objects.

Automatic Top-Down Fusion method

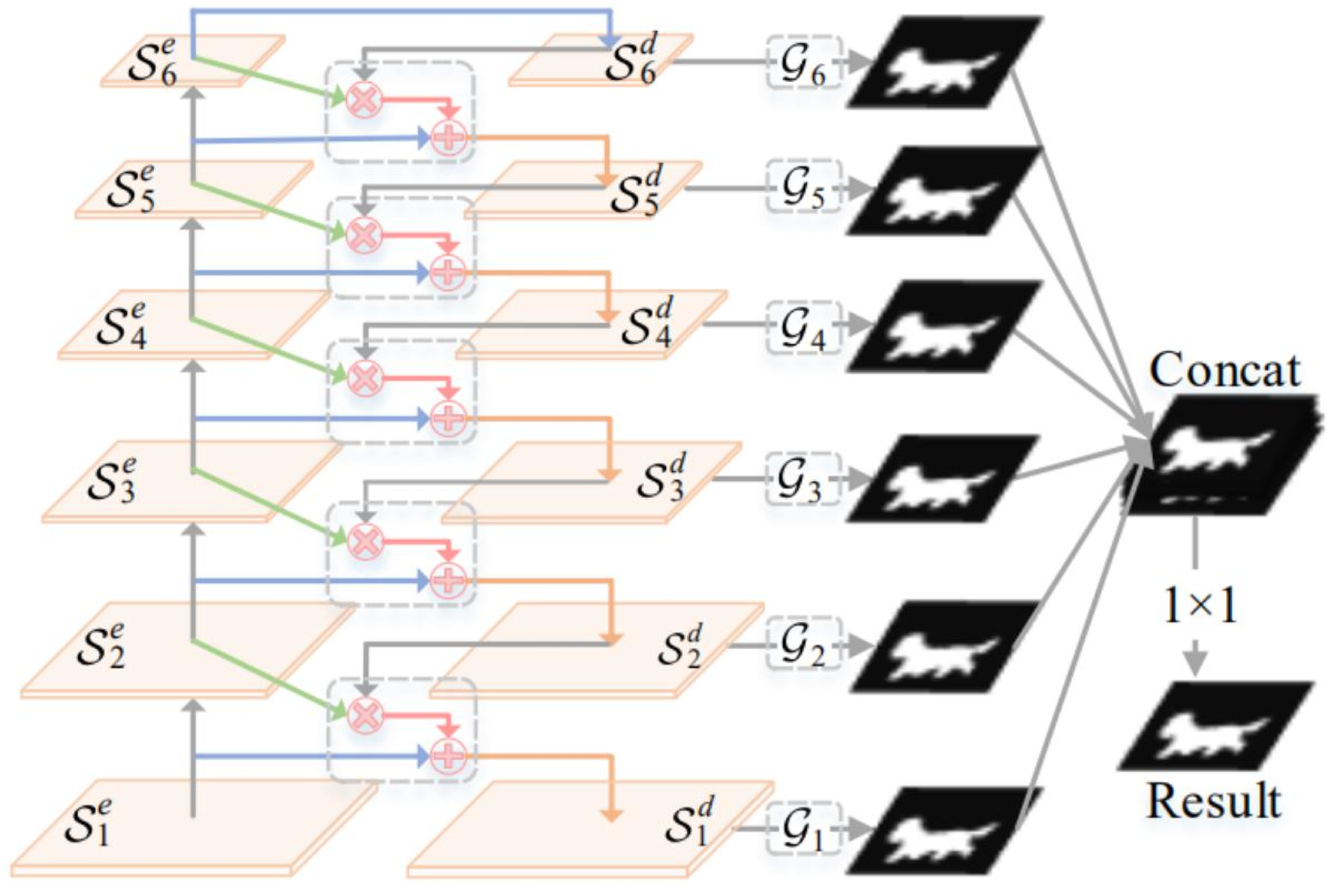

Fig.1. Overall Framework

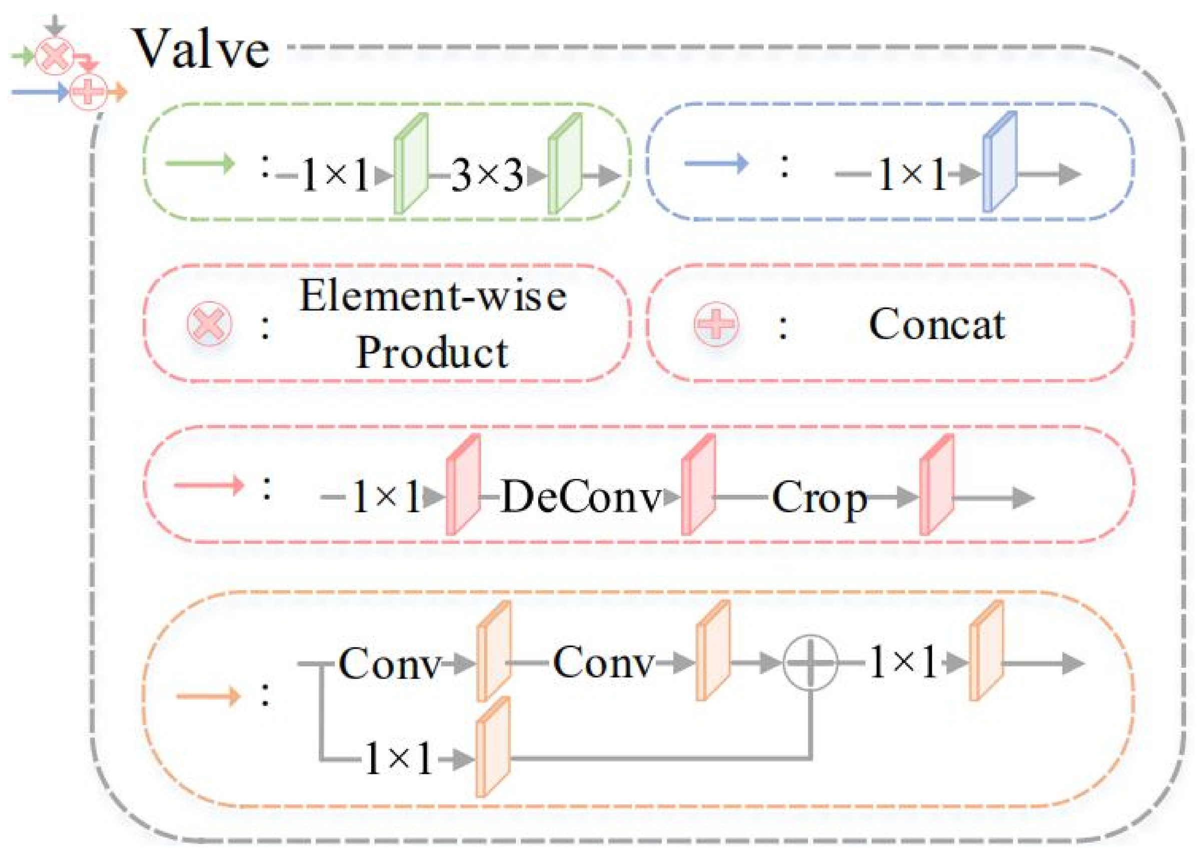

Fig.2. The valve module

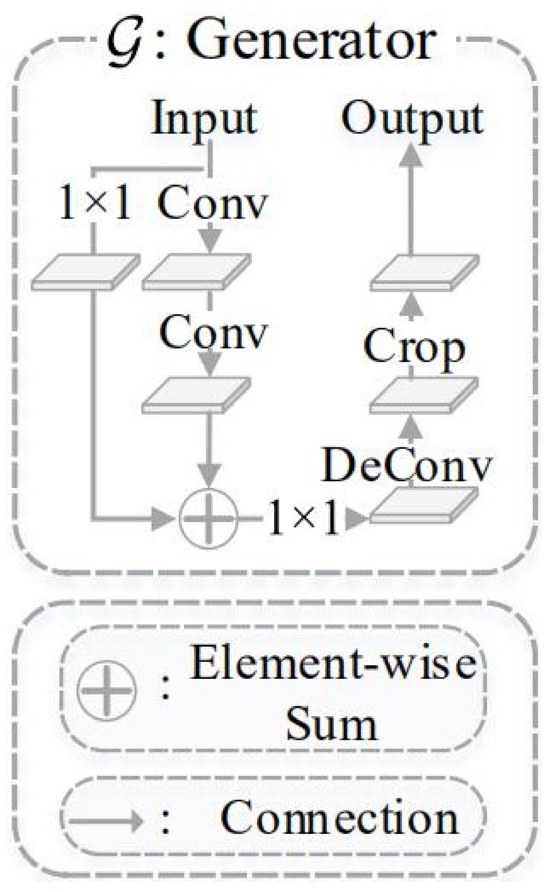

Fig.3. The generator

- As shown in Fig.1, the main architecture of ATDF is beneficial from the encoder-decoder networks.
- As shown in Fig.2, the valve module can adaptively determine the flow of the useful high-level information from top sides to bottom sides.
- As shown in Fig.3, the generator can further improve the capability of aggregated hierarchical information for saliency prediction.

Datasets and Evaluation Criteria

- **Datasets:** We utilize the **DUTS** training dataset to fine-tune our model. We evaluate our method on the **DUTS test set** and other five popular datasets including **ECSSD**, **SOD**, **HKU-I**, **THUR15K** and **DUT-OMRON**.
- **Evaluation Criteria:** The max F-measure score and mean absolute error (MAE).

Experiments

Methods	SOD		HKU-IS		ECSSD		DUT-OMRON		THUR15K		DUTS-test	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
VGG16 backbone												
LEGS [5]	0.733	0.194	0.766	0.119	0.830	0.118	0.668	0.134	0.663	0.126	0.652	0.137
ELD [6]	0.758	0.154	0.837	0.074	0.866	0.081	0.700	0.092	0.726	0.095	0.727	0.092
RFCN [7]	0.802	0.161	0.892	0.080	0.896	0.097	0.738	0.095	0.754	0.100	0.782	0.089
DCL [8]	0.831	0.131	0.892	0.063	0.895	0.080	0.733	0.095	0.747	0.096	0.785	0.082
Amulet [9]	0.795	0.144	0.897	0.051	0.913	0.061	0.743	0.098	0.755	0.094	0.778	0.085
UCF [10]	0.805	0.148	0.888	0.062	0.901	0.071	0.730	0.120	0.758	0.112	0.772	0.112
NLDF [11]	0.837	0.123	0.902	0.048	0.902	0.066	0.753	0.080	0.762	0.080	0.806	0.065
DSS [13]	0.842	0.122	0.913	0.041	0.915	0.056	0.774	0.066	0.770	0.074	0.827	0.056
PiCA [14]	0.836	0.102	0.916	0.042	0.923	0.049	0.766	0.068	0.783	0.083	0.837	0.054
C2S [15]	0.819	0.122	0.898	0.046	0.907	0.057	0.759	0.072	0.775	0.083	0.811	0.062
RAS [16]	0.847	0.123	0.913	0.045	0.916	0.058	0.785	0.063	0.772	0.075	0.831	0.059
ATDF (ours)	0.859	0.114	0.927	0.032	0.931	0.044	0.795	0.055	0.796	0.066	0.863	0.042
ResNet backbone												
SRM [12]	0.840	0.126	0.906	0.046	0.914	0.056	0.769	0.069	0.778	0.077	0.826	0.059
PiCA [14]	0.852	0.103	0.917	0.043	0.929	0.049	0.789	0.065	0.788	0.081	0.853	0.050
ATDF (ours)	0.862	0.110	0.933	0.031	0.939	0.040	0.814	0.051	0.801	0.064	0.877	0.037

Table 1. Comparison between our ATDF and 12 state-of-the-art methods in terms of F_β (the larger the better) and MAE (the smaller the better) on six datasets. We highlight the top three results of each column in **red**, **green** and **blue**, respectively.

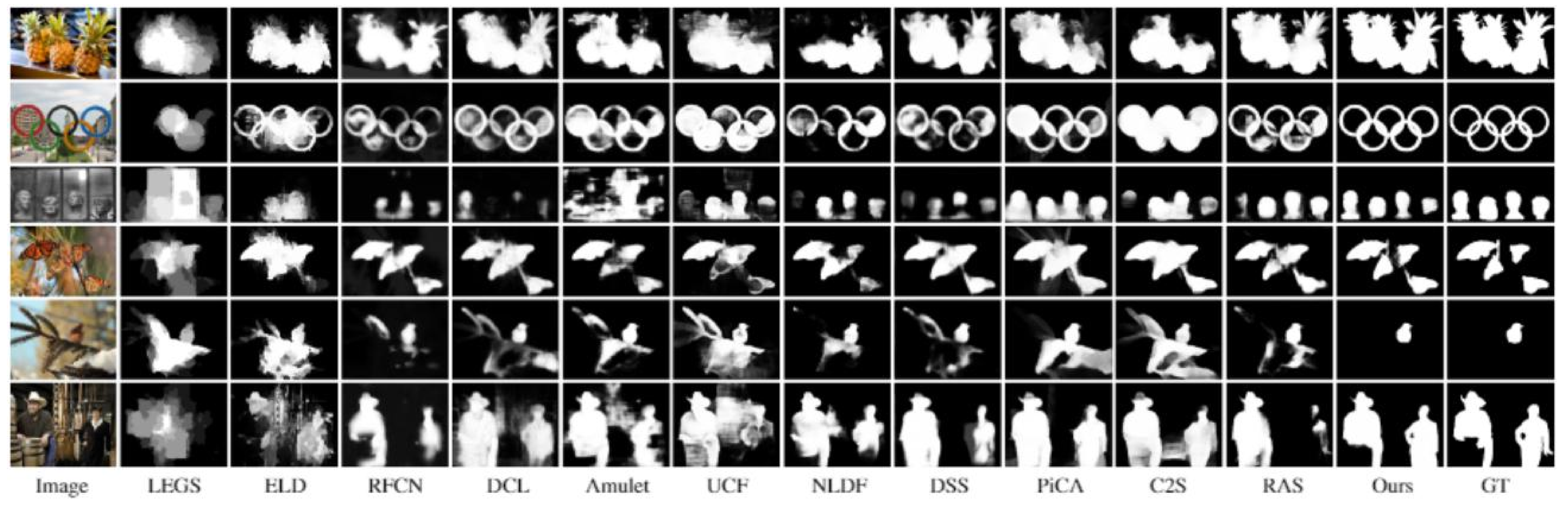

Fig.4. Qualitative comparison of ATDF and 11 methods..

Conclusions

Most of recent salieny detection methods aim at designing effective fusion strategies for side-output features. However, the network architectures become more and more complex. Hence automatically flowing the global information at the top sides into bottom sides to guide the learning of bottom layers is more and more important.