JSAI CONFERENCE
江苏省人工智能大会
智领江苏 • 融创未来

The Journey to the SOD Family

——Trip 2020

Runmin Cong (丛润民)
Beijing Jiaotong University
2020-12-30@Jiangsu

数字媒体信息处理研究中心
Center of Digital Media Information Processing

北京交通大学数字媒体信息处理研究中心肇始于1998年，入选科技部“重点领域创新团队”、教育部“创新团队发展计划”。该中心现有教师12人，博、硕士研究生100余人。其中教授8人，副教授3人，包括教育部长江学者特聘教授1人，国家杰出青年基金获得者人1人，教育部新世纪优秀人才2人，北京市杰出青年基金获得者1人，北京市科技新星3人，香江学者1人。该中心的研究领域为数字媒体信息处理，研究方向主要包括图像\视频编码与传输、数字水印与数字取证、媒体内容分析与理解等。

赵 耀 教授

教育部长江学者特聘教授

国家杰出青年基金获得者

万人计划科技创新领军人才

教 授：赵 耀、朱振峰、倪蓉蓉、白慧慧、韦世奎
李晓龙、林春雨、张淳杰

副教授：常冬霞、刘美琴、丛润民

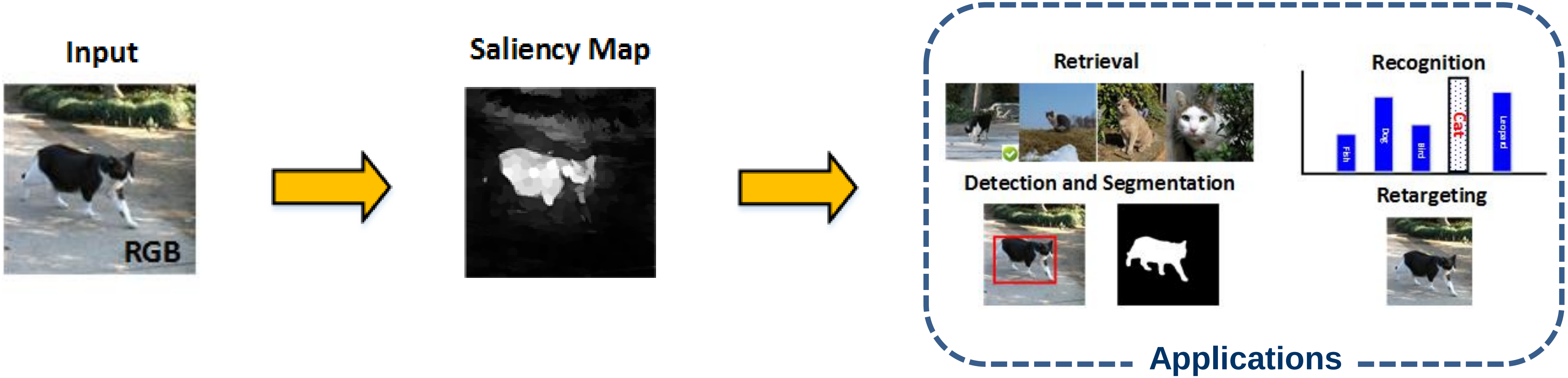
助 理：宋亚男

http://mepro.bjtu.edu.cn/index.html

Outline

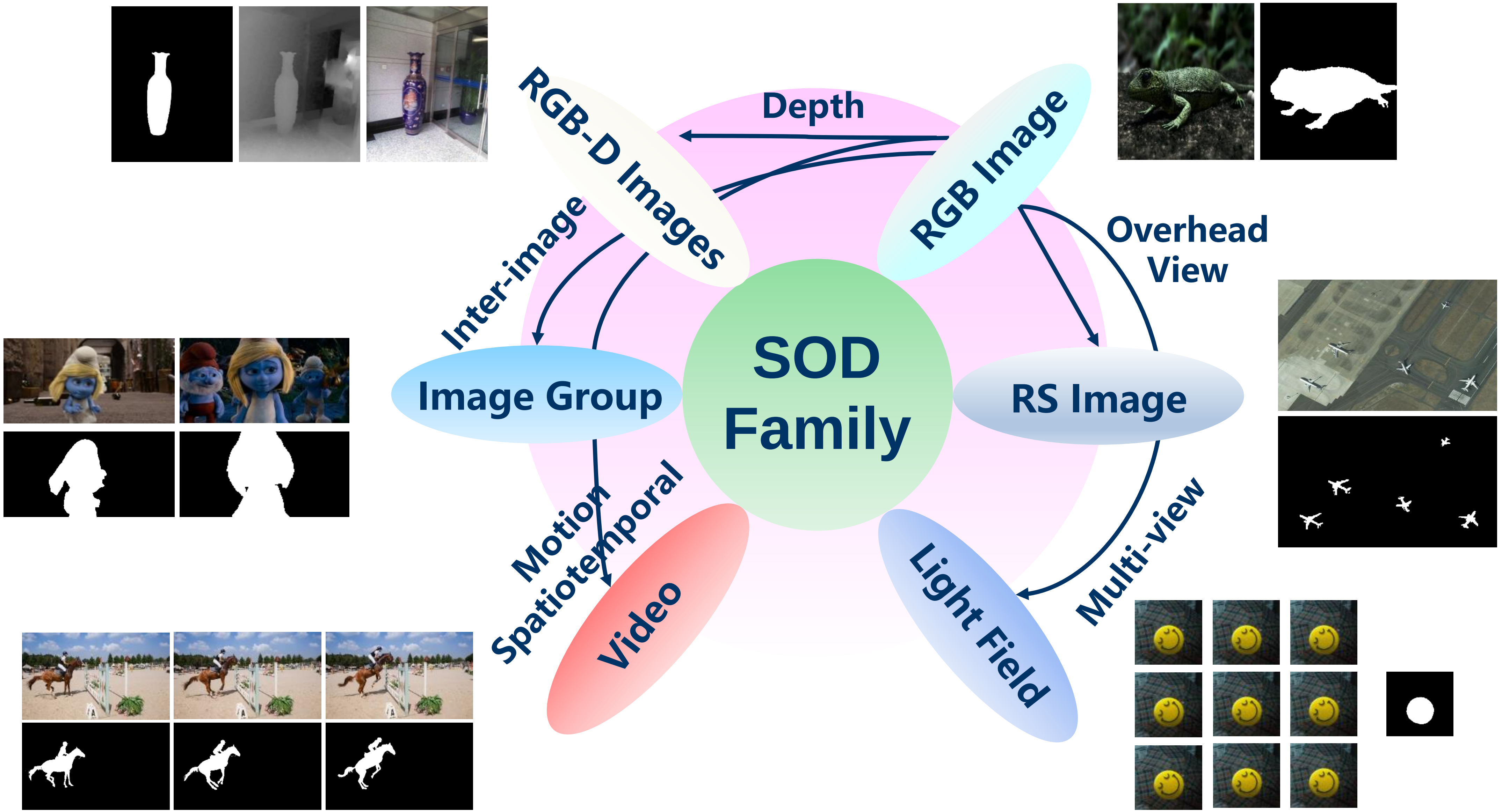
- Introduction
- RGB-D Salient Object Detection
  - ➢ DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection
- Co-salient Object Detection
  - ➢ CoADNet: Collaborative Aggregation-and-Distribution Networks for Co-Salient Object Detection
- Salient Object Detection in Optical RSIs
  - ➢ Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images
- Future Work

Introduction

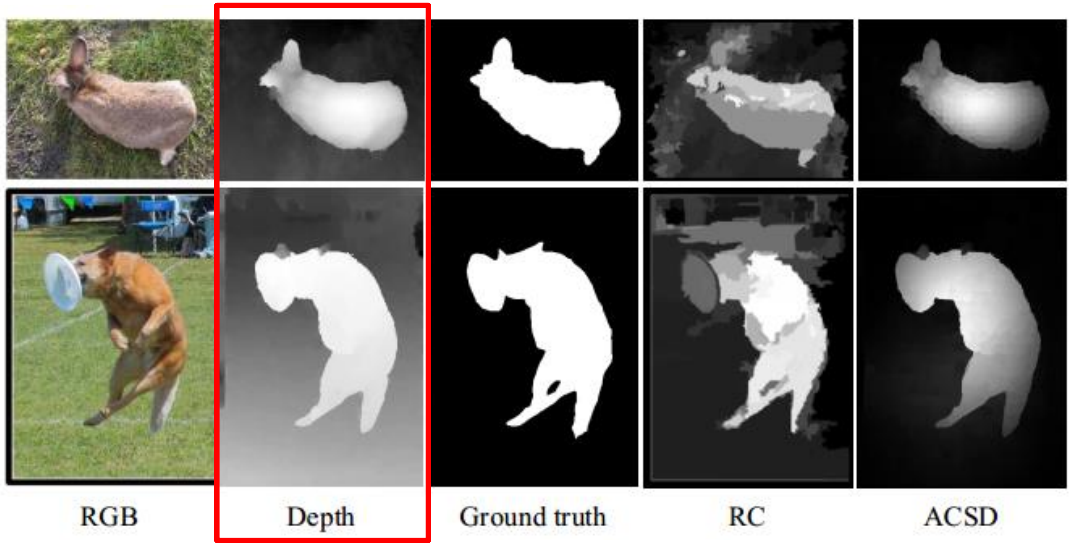


Simulating the human visual attention mechanism, salient object detection aims at detecting the salient regions automatically, which has been applied in image/video segmentation, image/video retrieval, image retargeting, video coding, quality assessment, action recognition, and video summarization.

Introduction



RGB-D Salient Object Detection



- shape
- contour
- internal consistency
- surface normal
- ……

depth quality perception

DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection

Zuyao Chen‡, Runmin Cong‡, Qianqian Xu, and Qingming Huang

IEEE Transaction on Image Processing, 2021

https://rmcong.github.io/proj_DPANet.html

Motivations

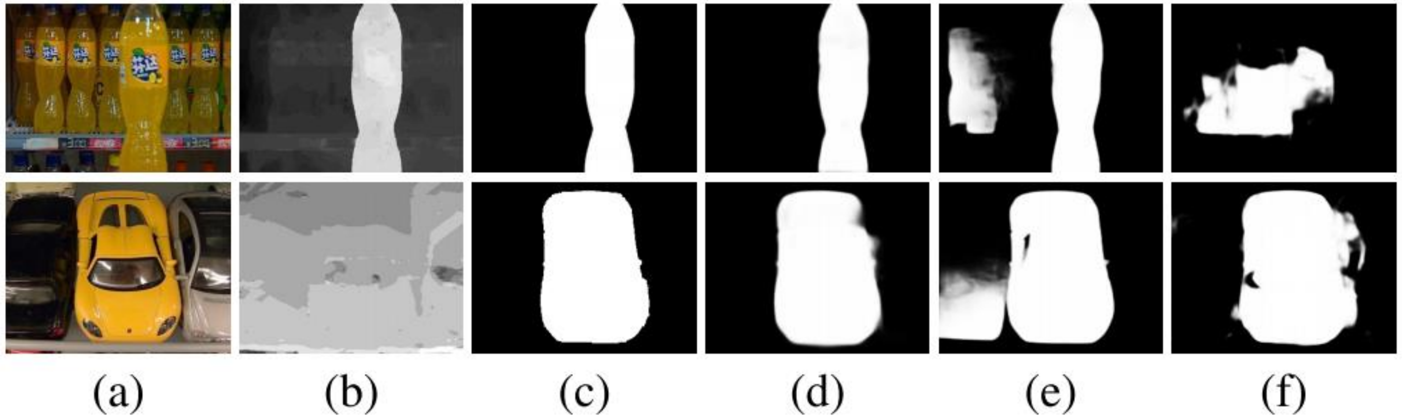


Fig. 1. Sample results of our method compared with others. RGB-D methods are marked in **boldface**. (a) RGB image; (b) Depth map; (c) Ground truth; (d) **Ours**; (e) BASNet [14]; (f) **CPFP** [33].

- how to effectively **integrate** the complementary information from RGB image and its corresponding depth map;
- how to **prevent** the contamination from unreliable depth information;

Contributions

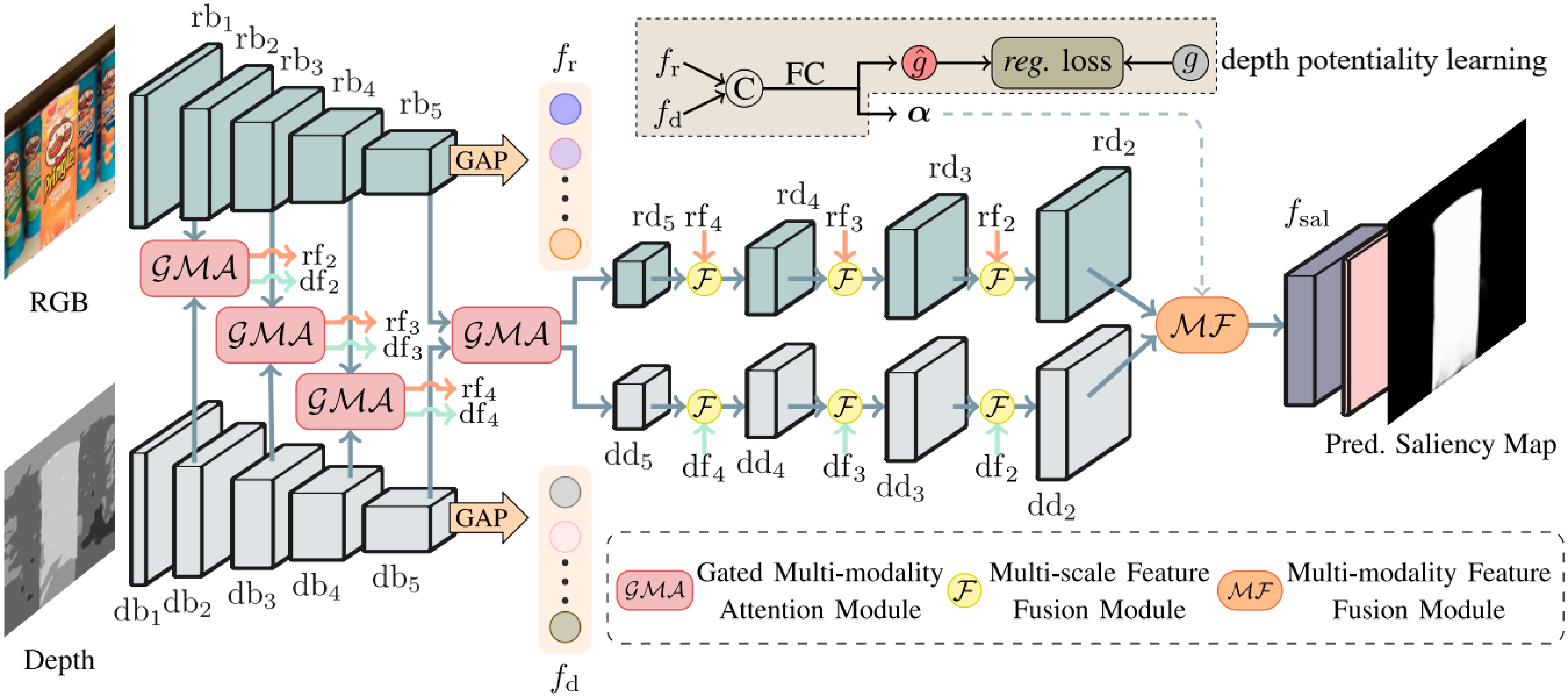
a) **For the first time, we address the unreliable depth map in the RGB-D SOD network in an end-to-end formulation**, and propose the DPANet by incorporating the depth potentiality perception into the cross-modality integration pipeline.

b) **Without increasing the training label** (i.e., depth quality label), we model a **task-orientated depth potentiality perception module** that can adaptively perceive the potentiality of the input depth map, and further weaken the contamination from unreliable depth information.

c) We propose a **gated multi-modality attention (GMA) module** to effectively aggregate the cross-modal complementarity of the RGB and depth images.

d) Without any pre-processing or post-processing techniques, the proposed network **outperforms 16 state-of-the-art methods on 8 RGB-D SOD datasets** in quantitative and qualitative evaluations.

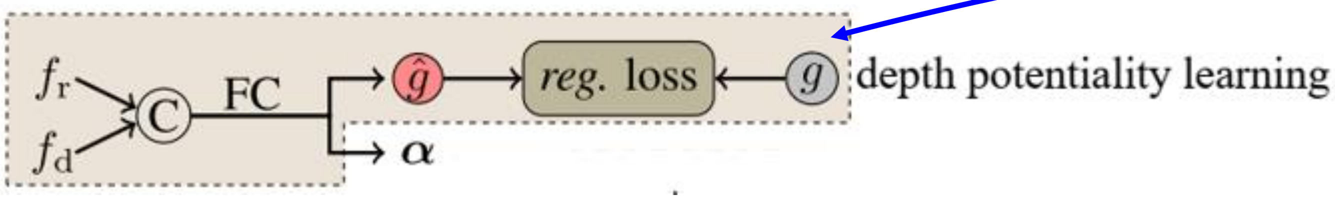
Our Method



RGB
Depth
rb1 rb2 rb3 rb4 rb5
db1 db2 db3 db4 db5
GAP
GAP
f_r
f_d
GMA
rf2 df2
rf3 df3
rf4 df4
f_r f_d C FC $\hat{g}$ reg. loss g depth potentiality learning
α
rd5 rf4 rd4 rf3 rd3 rf2 rd2
dd5 df4 dd4 df3 dd3 df2 dd2
$\mathcal{F}$
$\mathcal{MF}$
f_{sal}
Pred. Saliency Map
$\mathcal{GMA}$ Gated Multi-modality Attention Module
$\mathcal{F}$ Multi-scale Feature Fusion Module
$\mathcal{MF}$ Multi-modality Feature Fusion Module

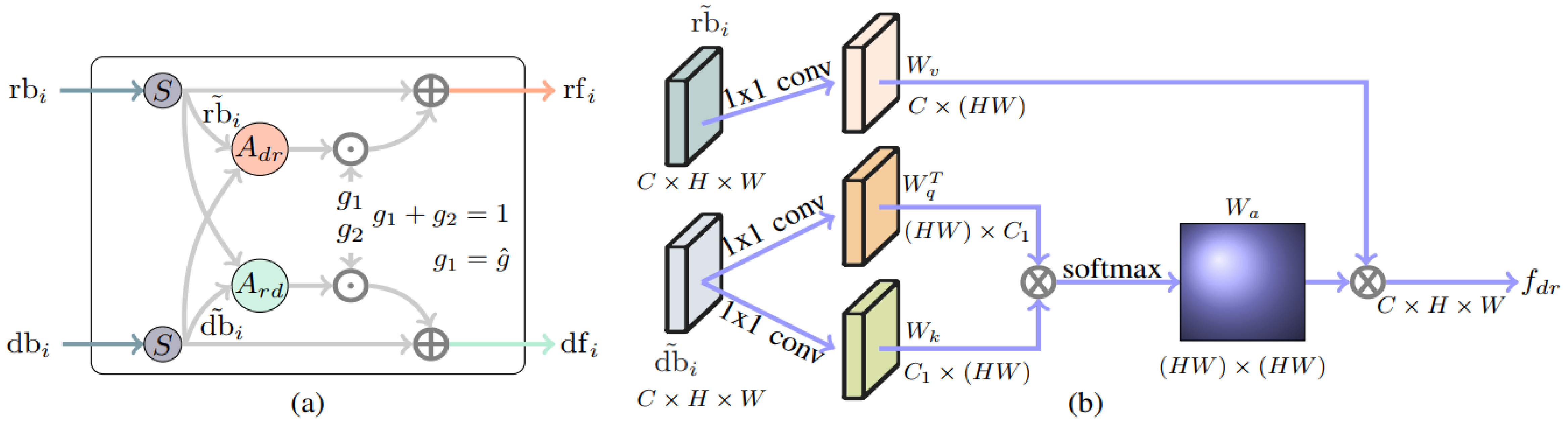
Depth Potentiality Perception

- Most previous works generally integrate the multi-modal features from RGB and corresponding depth information indiscriminately. However, **there exist some contaminations when depth maps are unreliable**.
- Since we do not hold any labels for depth map quality assessment, **we model the depth potentiality perception as a saliency-oriented prediction task**, that is, we train a model to automatically learn the relationship between the binary depth map and the corresponding saliency mask. The above modeling approach is based on the observation that if the binary depth map segmented by a threshold is close to the ground truth, the depth map is highly reliable, so a higher confidence response should be assigned to this depth input.



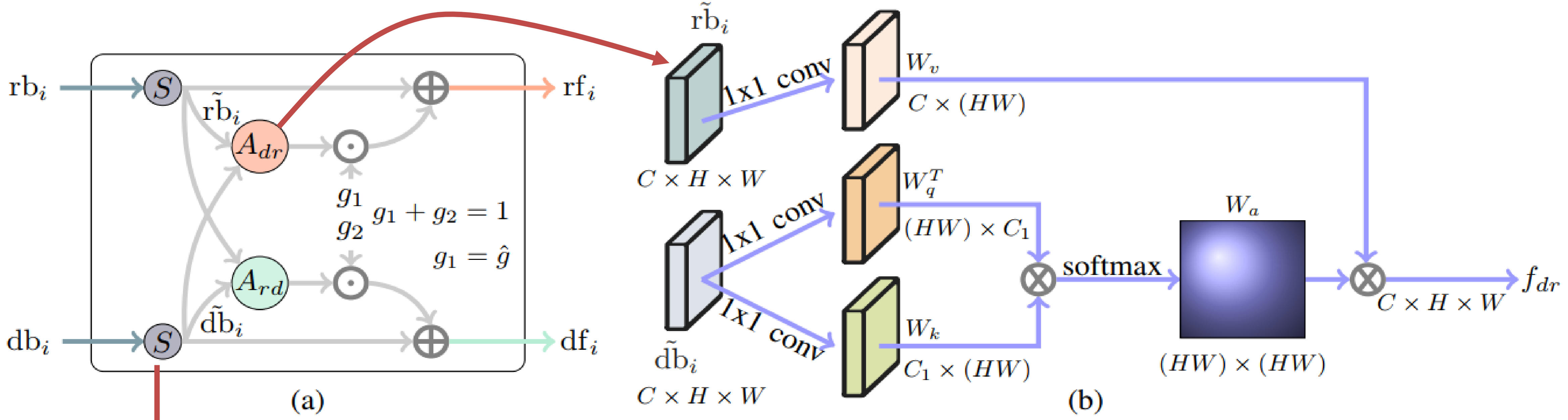
$$D(\tilde{I}, G) = \frac{(1+\gamma)\cdot D_{iou}\cdot D_{cov}}{D_{iou}+\gamma\cdot D_{cov}}$$

Gated Multi-modality Attention Module



- Directly integrating the cross-modal information may induce negative results, such as **contaminations from unreliable depth maps**. Besides, the features of the single modality usually are affluent in spatial or channel aspect with **information redundancy**.
- We design a GMA module that exploits the attention mechanism to **automatically select and strengthen important features** for saliency detection, and **incorporate the gate controller** into the GMA module to prevent the contamination from the unreliable depth map.

Gated Multi-modality Attention Module



single-modal perspective:

spatial attention

reduce the redundancy features and highlight the feature response on the salient regions

cross-modal perspective:

two symmetrical attention sub-modules

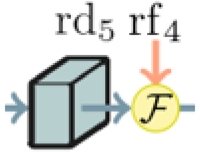
capture long-range dependencies

$$rf_i = \widetilde{rb}_i + g_1 \cdot f_{dr} \qquad g_1 = \hat{g}$$
$$df_i = \widetilde{db}_i + g_2 \cdot f_{rd} \qquad g_1 + g_2 = 1$$

Multi-level Feature Fusion

- Multi-scale Feature Fusion

Low-level features can provide more detail information, such as boundary, texture, and spatial structure, but may be sensitive to the background noises. Contrarily, high-level features contain more semantic information, which is helpful to locate the salient object and suppress the noises. Thus, we adopt a more aggressive yet effective operation, i.e., multiplication.



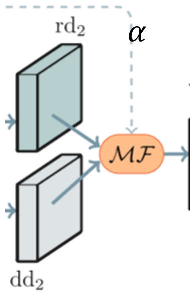
$$f_1 = \delta\big(up\big(conv_3(\mathrm{rd}_5)\big) \odot \mathrm{rf}_4\big)$$

$$f_2 = \delta\big(conv_4(\mathrm{rf}_4) \odot up(\mathrm{rd}_5)\big)$$

$$f_F = \delta\big(conv_5([f_1, f_2])\big)$$

- Multi-modality Feature Fusion

During the multi-modality feature fusion, we consider two issues: (1) How to select the most useful and complementary information from the RGB and depth features. (2) How to prevent the contamination caused by the unreliable depth map during fusing.



$$f_3 = \boldsymbol{\alpha} \odot \mathrm{rd}_2 + \hat{g} \cdot (1 - \boldsymbol{\alpha}) \odot \mathrm{dd}_2$$

$$f_4 = \mathrm{rd}_2 \odot \mathrm{dd}_2$$

$$f_{sal} = \delta\big(conv([f_3, f_4])\big)$$

$\boldsymbol{\alpha}$ is the weight vector learned from RGB and depth information, $\hat{g}$ is the learned weight of the gate as mentioned before.

Loss Function

The final loss is the linear combination of the classification loss and regression loss:

$$\mathcal{L}_{final} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{reg}$$

classification loss:

$$\mathcal{L}_{cls} = \mathcal{L}_{cls} + \sum_{i=1}^{8} \lambda_i \cdot \mathcal{L}_{aux}^{i}$$

regression loss :

$$\mathcal{L}_{reg} = \begin{cases} 0.5(g-\hat{g})^2, & if\ |g-\hat{g}| < 1 \\ |g-\hat{g}| - 0.5, & otherwise \end{cases}$$

Experiments

- Benchmark Datasets: NJUD (1985 RGB-D images), NLPR (1000 RGB-D images), STEREO (797 RGB-D images), LFSD (100 RGB-D images), SSD (80 RGB-D images), and DUT (1200 RGB-D images), RGBD135 (135 RGB-D images), SIP (929 RGB-D images).
- Evaluation Metrics: Precision-Recall (P-R) curve, F-measure, MAE score, and S-measure.
- Following [1], we take 1400 images from NJUD and 650 images from NLPR as the training, and 100 images from NJUD dataset and 50 images from NLPR dataset as the validation set. To reduce the overfitting, we use multi-scale resizing and random horizontal flipping augmentation. During the inference stage, images are simply resized to 256 $\times$ 256, and then fed into the network to obtain prediction without any other post-processing or pre-processing techniques.

[1] H. Chen, et. al: Progressively complementarity-aware fusion network for RGB-D salient object detection. In: CVPR, 2018

Experiments

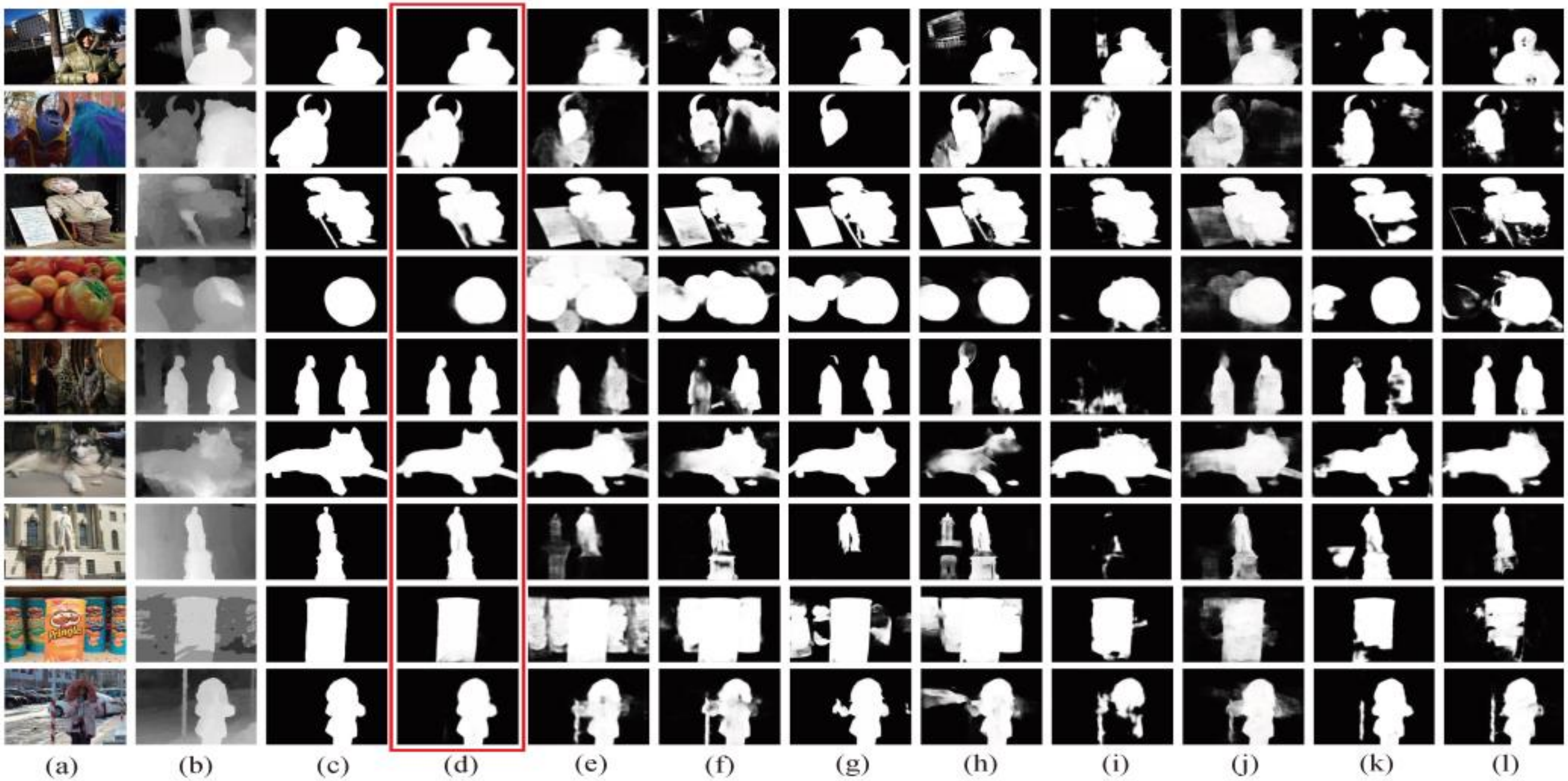


Fig. 4. Qualitative comparison of the proposed approach with some state-of-the-art RGB and RGB-D SOD methods, in which our results are highlighted by a red box. (a) RGB image. (b) Depth map. (c) GT. (d) DPANet. (e) PiCAR. (f) PoolNet. (g) BASNet. (h) EGNet. (i) CPFP. (j) PDNet. (k) DMRA. (l) AF-Net.

Experiments

Method	RGBD135 Dataset			SSD Dataset			LFSD Dataset			NJUD-test Dataset		
	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$
DPANet (ours)	**0.933**	**0.922**	**0.023**	**0.895**	**0.877**	**0.046**	**0.880**	**0.862**	**0.074**	**0.931**	**0.922**	**0.035**
AF-Net (Arxiv19)	0.904	0.892	0.033	0.828	0.815	0.077	0.857	0.818	0.091	0.900	0.883	0.053
DMRA (ICCV19)	0.921	0.911	0.026	0.874	0.857	0.055	0.865	0.831	0.084	0.900	0.880	0.052
CPFP (CVPR19)	0.882	0.872	0.038	0.801	0.807	0.082	0.850	0.828	0.088	0.799	0.798	0.079
PCFN (CVPR18)	0.842	0.843	0.050	0.845	0.843	0.063	0.829	0.800	0.112	0.887	0.877	0.059
PDNet (ICME19)	0.906	0.896	0.041	0.844	0.841	0.089	0.865	0.846	0.107	0.912	0.897	0.060
TAN (TIP19)	0.853	0.858	0.046	0.835	0.839	0.063	0.827	0.801	0.111	0.888	0.878	0.060
MMCI (PR19)	0.839	0.848	0.065	0.823	0.813	0.082	0.813	0.787	0.132	0.868	0.859	0.079
CTMF (TC18)	0.865	0.863	0.055	0.755	0.776	0.100	0.815	0.796	0.120	0.857	0.849	0.085
RS (ICCV17)	0.841	0.824	0.053	0.783	0.750	0.107	0.795	0.759	0.130	0.796	0.741	0.120
EGNet (ICCV19)	0.913	0.892	0.033	0.704	0.707	0.135	0.845	0.838	0.087	0.867	0.856	0.070
BASNet (CVPR19)	0.916	0.894	0.030	0.842	0.851	0.061	0.862	0.834	0.084	0.890	0.878	0.054
PoolNet (CVPR19)	0.907	0.885	0.035	0.764	0.749	0.110	0.847	0.830	0.095	0.874	0.860	0.068
AFNet (CVPR19)	0.897	0.878	0.035	0.847	0.859	0.058	0.841	0.817	0.094	0.890	0.880	0.055
PiCAR (CVPR18)	0.907	0.890	0.036	0.864	0.871	0.055	0.849	0.834	0.104	0.887	0.882	0.060
R^3Net (IJCAI18)	0.857	0.845	0.045	0.711	0.672	0.144	0.843	0.818	0.089	0.805	0.771	0.105

Method	NLPR-test Dataset			STEREO797 Dataset			SIP Dataset			DUT Dataset		
	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$
DPANet (ours)	**0.924**	**0.927**	**0.025**	**0.919**	**0.915**	**0.039**	**0.906**	**0.883**	**0.052**	**0.918**	**0.904**	0.047
AF-Net (Arxiv19)	0.904	0.903	0.032	0.905	0.893	0.047	0.870	0.844	0.071	0.862	0.831	0.077
DMRA (ICCV19)	0.887	0.889	0.034	0.895	0.874	0.052	0.883	0.850	0.063	0.913	0.880	0.052
CPFP (CVPR19)	0.888	0.888	0.036	0.815	0.803	0.082	0.870	0.850	0.064	0.771	0.760	0.102
PCFN (CVPR18)	0.864	0.874	0.044	0.884	0.880	0.061	–	–	–	0.809	0.801	0.100
PDNet (ICME19)	0.905	0.902	0.042	0.908	0.896	0.062	0.863	0.843	0.091	0.879	0.859	0.085
TAN (TIP19)	0.877	0.886	0.041	0.886	0.877	0.059	–	–	–	0.824	0.808	0.093
MMCI (PR19)	0.841	0.856	0.059	0.861	0.856	0.080	–	–	–	0.804	0.791	0.113
CTMF (TC18)	0.841	0.860	0.056	0.827	0.829	0.102	–	–	–	0.842	0.831	0.097
RS (ICCV17)	0.900	0.864	0.039	0.857	0.804	0.088	–	–	–	0.807	0.797	0.111
EGNet (ICCV19)	0.845	0.863	0.050	0.872	0.853	0.067	0.846	0.825	0.083	0.888	0.867	0.064
BASNet (CVPR19)	0.882	0.894	0.035	0.914	0.900	0.041	0.894	0.872	0.055	0.912	0.902	**0.041**
PoolNet (CVPR19)	0.863	0.873	0.045	0.876	0.854	0.065	0.856	0.836	0.079	0.883	0.864	0.067
AFNet (CVPR19)	0.865	0.881	0.042	0.905	0.895	0.045	0.891	0.876	0.055	0.880	0.868	0.065
PiCAR (CVPR18)	0.872	0.882	0.048	0.906	0.903	0.051	0.890	0.878	0.060	0.903	0.892	0.062
R^3Net (IJCAI18)	0.832	0.846	0.049	0.811	0.754	0.107	0.641	0.624	0.158	0.841	0.812	0.079

TABLE III

COMPARISONS OF INFERENCE TIME OF DIFFERENT DEEP LEARNING BASED RGB-D SOD METHODS.

	CTMF	MMCI	TAN	PDNet	PCFN
Time (s)	0.63	0.05	0.07	0.07	0.06
	CPFP	AF-Net	DMRA	D^3Net	Ours
Time (s)	0.17	0.03	0.06	0.05	0.03

TABLE IV

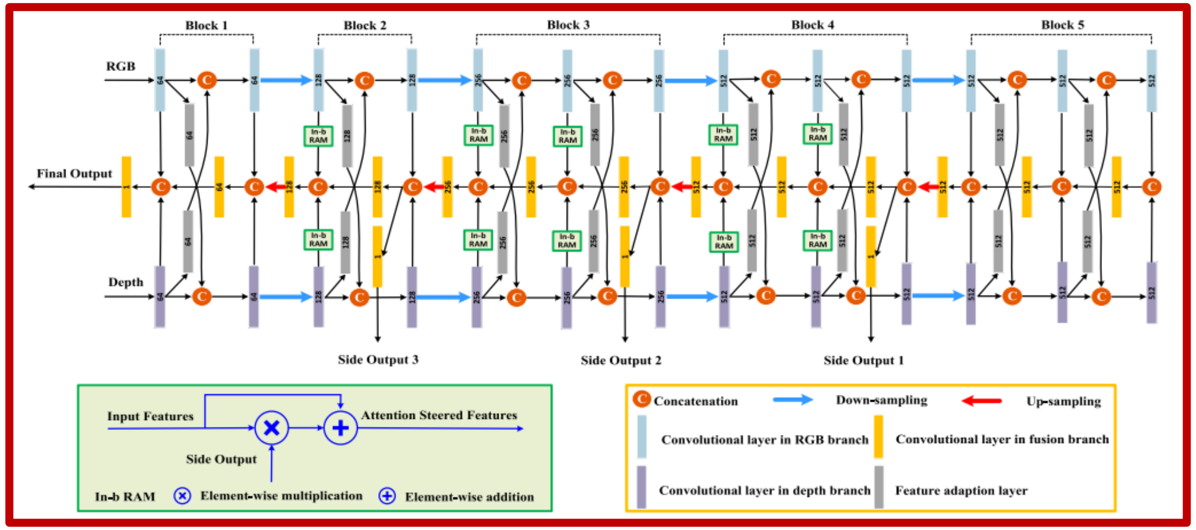
ABLATION STUDIES ON NJUD-TEST, SIP, AND STEREO797 DATASETS.

	NJUD-test Dataset			SIP Dataset			STEREO797 Dataset		
	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	MAE $\downarrow$
DPANet	0.930	0.921	0.035	0.904	0.883	0.051	0.915	0.911	0.041
concatenation	0.919	0.914	0.039	0.904	0.876	0.056	0.912	0.905	0.044
summation	0.923	0.915	0.038	0.906	0.881	0.054	0.910	0.904	0.045
hard manner	0.908	0.902	0.047	0.893	0.868	0.064	0.905	0.899	0.050
w/o depth	0.908	0.903	0.043	0.864	0.837	0.074	0.913	0.908	0.042

Conclusion

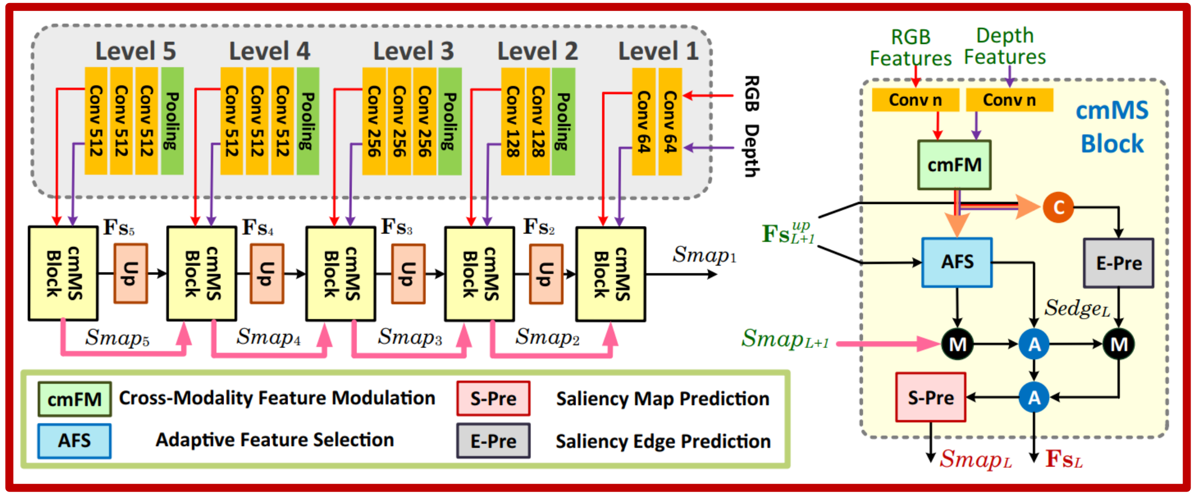
- We model a saliency-orientated depth potentiality perception module to evaluate the potentiality of the depth map and weaken the contamination.
- We propose a GMA module to highlight the saliency response and regulate the fusion rate of the cross-modal information.
- The multi-scale and multi-modality feature fusion are used to generate the discriminative RGB-D features and produce the saliency map.
- Experiments on eight RGB-D datasets demonstrate that the proposed network outperforms other 15 state-of-the-art methods under different evaluation metrics.

Other Works



ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection, TCyb 2021

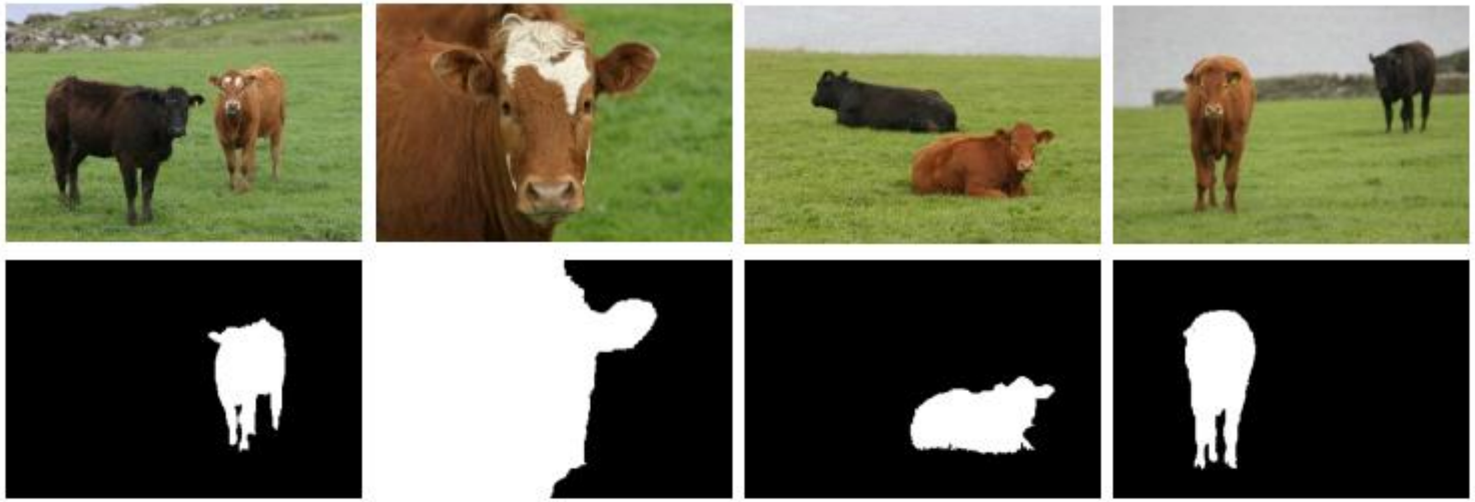
https://github.com/Li-Chongyi/ASIF-Net



RGB-D Salient Object Detection with Cross-Modality Modulation and Selection, ECCV 2020

https://li-chongyi.github.io/Proj_ECCV20

Co-salient Object Detection



Problems and Important Issues

how to explore and preserve inter-image correspondence among multiple images to constrain the common properties of salient object is a challenge.

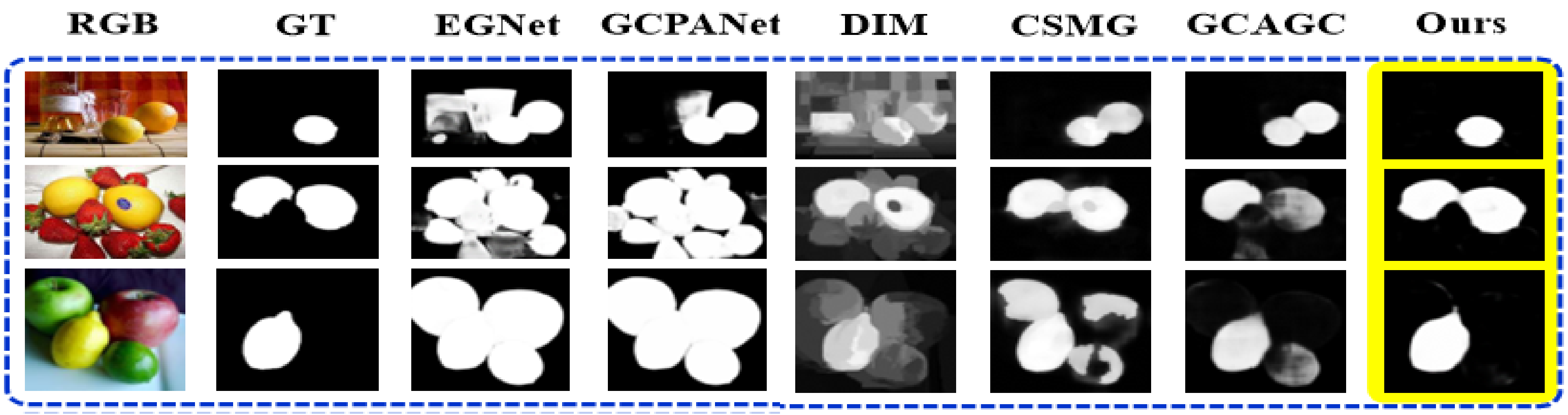
CoADNet: Collaborative Aggregation-and-Distribution Networks for Co-Salient Object Detection

Qijian Zhang Runmin Cong* Junhui Hou Chongyi Li Yao Zhao

***Conference on Neural Information Processing Systems (NeurIPS)*, 2020**

https://rmcong.github.io/proj_CoADNet.html

Motivations



- Co-Salient Object Detection (CoSOD) aims at discovering the salient objects that repeatedly appear in a query group containing two or more relevant images.
- One challenging issue is **how to effectively capture the co-saliency cues by modeling and exploiting the inter-image relationships**.

Motivations

- **Insufficient group-wise relationship modeling.** The learned group representations in the previous studies vary with different order of the input group images, leading to unstable training and vulnerable inference.
- **Competition between intra-image saliency and inter-image correspondence.** The learned group semantics in the previous studies were directly duplicated and concatenated with individual features. In fact, this operation implies that different individuals receive identical group semantics, which may propagate redundant and distracting information from the interactions among other images.
- **Weakened group consistency during feature decoding.** In the feature decoding of the CoSOD task, existing up-sampling or deconvolution based methods ignore the maintenance of inter-image consistency, which may lead to the inconsistency of co-salient objects among different images and introduce additional artifacts.

Contributions

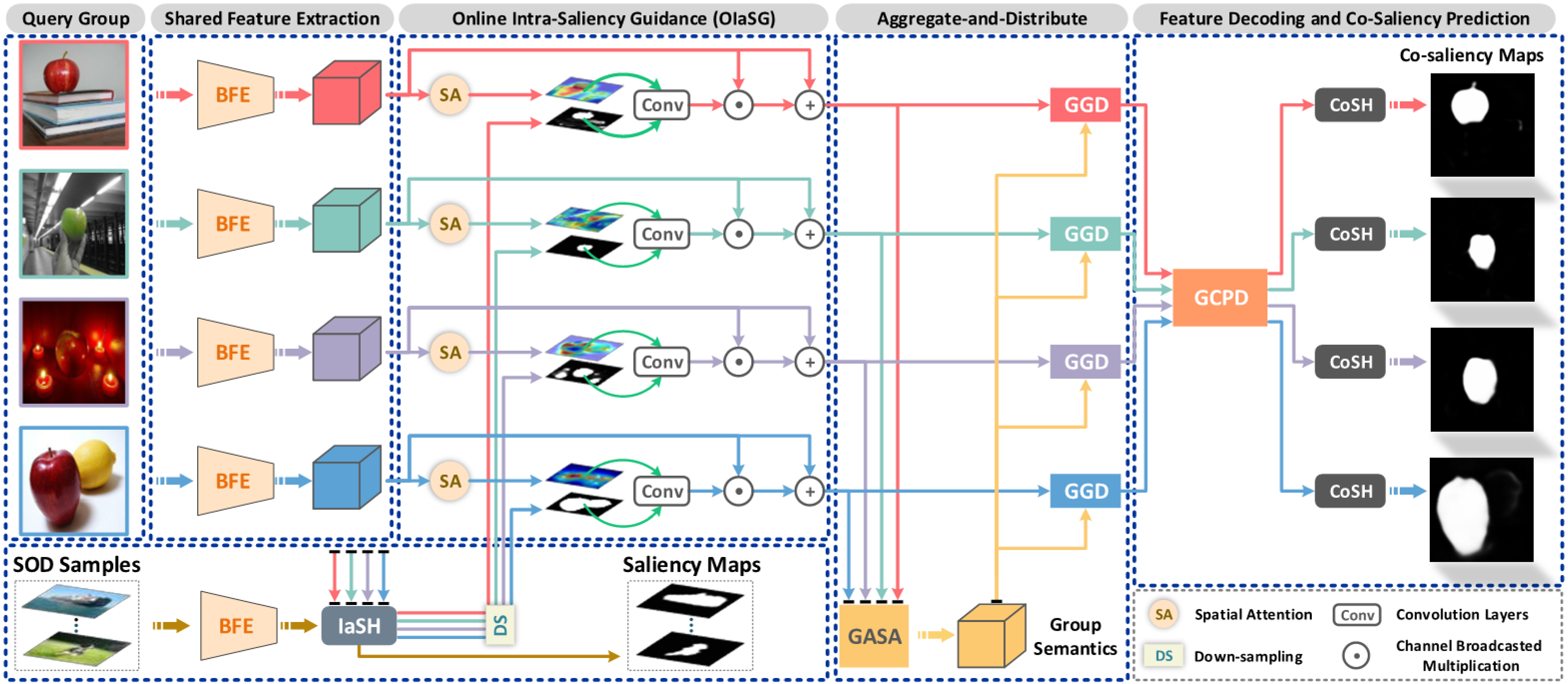
The proposed CoADNet provides some insights and improvements in terms of modeling and exploiting inter-image relationships in the CoSOD workflow, and produces more accurate and consistent co-saliency results on four prevailing co-saliency benchmark datasets.

1. We design an **online intra-saliency guidance** module for supplying saliency prior knowledge, which is jointly optimized to generate trainable saliency guidance information.

2. We propose a **two-stage aggregate-and-distribute architecture** to learn group-wise correspondences and co-saliency features, including a group-attentional semantic aggregation and a gated group distribution module.

3. A **group consistency preserving decoder** is designed to exploit more sufficient inter-image constraints to generate full-resolution co-saliency maps while maintaining group-wise consistency.

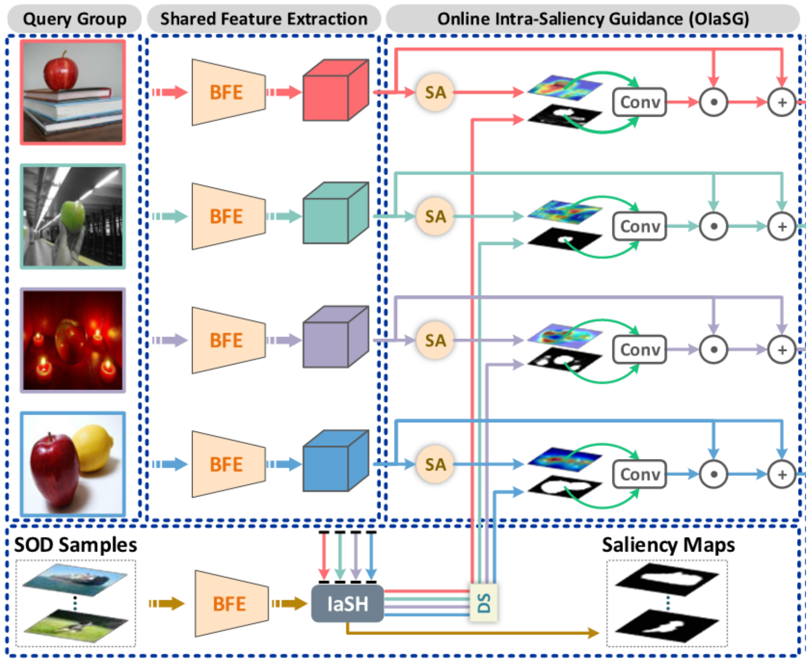
Our Method


Query Group
Shared Feature Extraction
Online Intra-Saliency Guidance (OIaSG)
Aggregate-and-Distribute
Feature Decoding and Co-Saliency Prediction
BFE
SA
Conv
GGD
GCPD
CoSH
Co-saliency Maps
SOD Samples
IaSH
DS
Saliency Maps
GASA
Group Semantics
SA Spatial Attention
Conv Convolution Layers
DS Down-sampling
Channel Broadcasted Multiplication

Online Intra-Saliency Guidance

The challenges of CoSOD are that 1) the salient objects within an individual image may not occur in all the other group images, and 2) the repetitive patterns are not necessarily visually attractive, making it difficult to learn a unified representation to combine these two factors. Thus, **we adopt a joint learning framework to provide trainable saliency priors as guidance information to suppress background redundancy**.

- Intra-saliency head (IaSH) to infer online saliency maps;
- Fuse online saliency priors with spatial feature in an attention way;
- In this way, we obtain a set of intra-saliency features (IaSFs) $\{U^{(n)}\}_{n=1}^{N}$ with suppressed background redundancy.



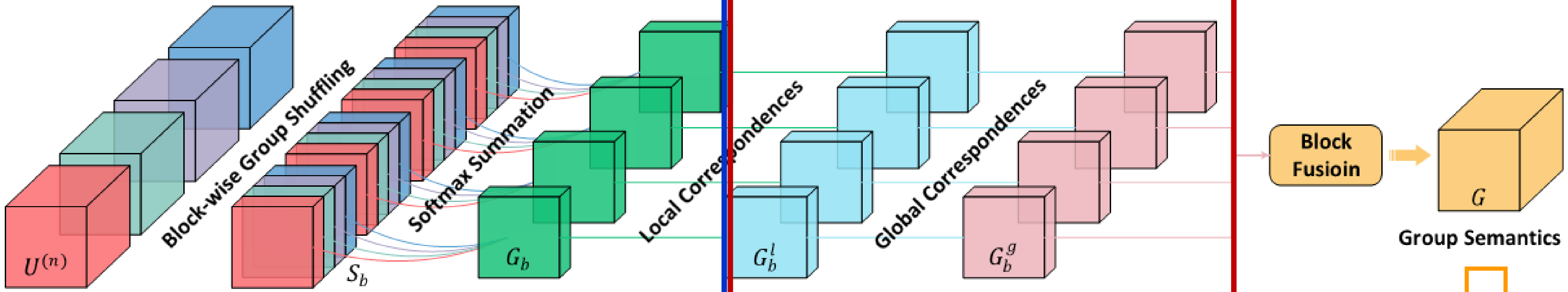
Group-Attentional Semantic Aggregation

To efficiently capture discriminative and robust group-wise relationships, we investigate three key criteria:

1) **Insensitivity to input order** means that the learned group representations should be insensitive to the input order of group images;
2) **Robustness to spatial variation** considers the fact that co-salient objects may be located at different positions across images;
3) **Computational efficiency** takes the computation burden into account especially when processing large query groups or high-dimensional features.

⇨ we propose a computation-efficient and order-insensitive **group-attentional semantic aggregation (GASA) module** which builds local and global associations of co-salient objects in group-wise semantic context.

Group-Attentional Semantic Aggregation



block-wise group shuffling that re-arranges the feature channels at the block level
channel-wise softmax + summation

local + global

model inter-image relationships with long-range semantic dependencies

high computational complexity

order-sensitivity

spatial variation of CoSal object

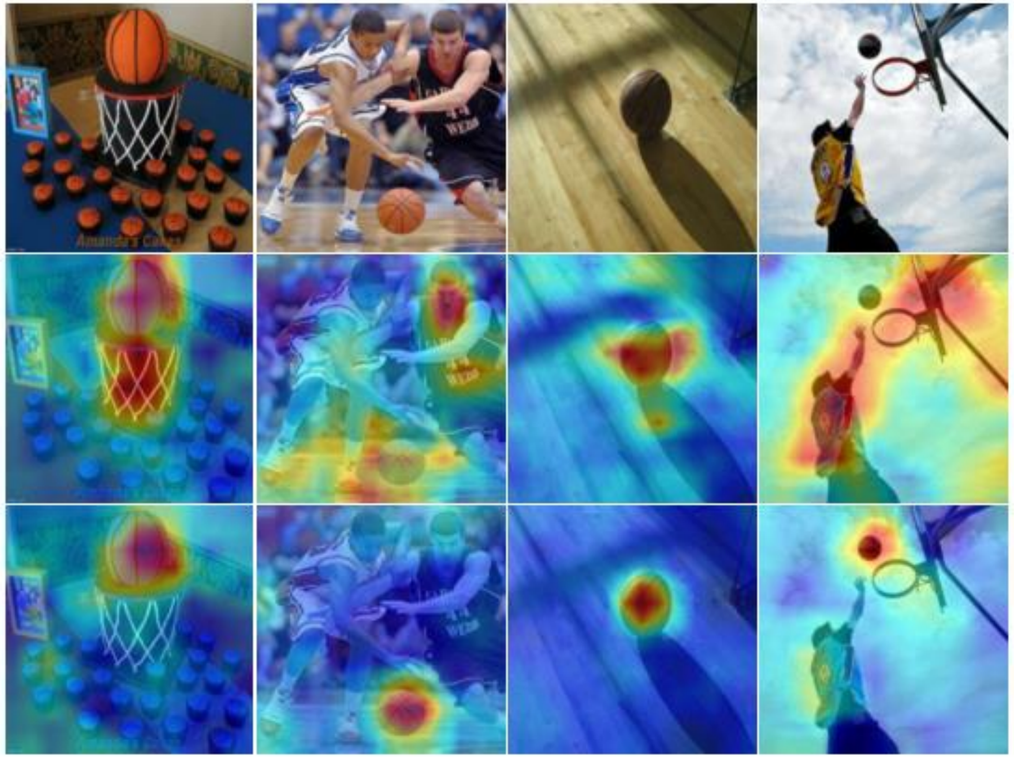
Gated Group Distribution

The group-wise semantics encode the relationships of all images, which may include some distracting information redundancy for co-saliency prediction of different images.

We propose a gated group distribution (GGD) module to adaptively distribute the most useful group-wise information to each individual. To achieve this, we construct a group importance estimator that learns dynamic weights to combine group semantics with different IaSFs through a gating mechanism.

$$X^{(n)} = P \otimes G + (1 - P) \otimes U^{(n)}$$

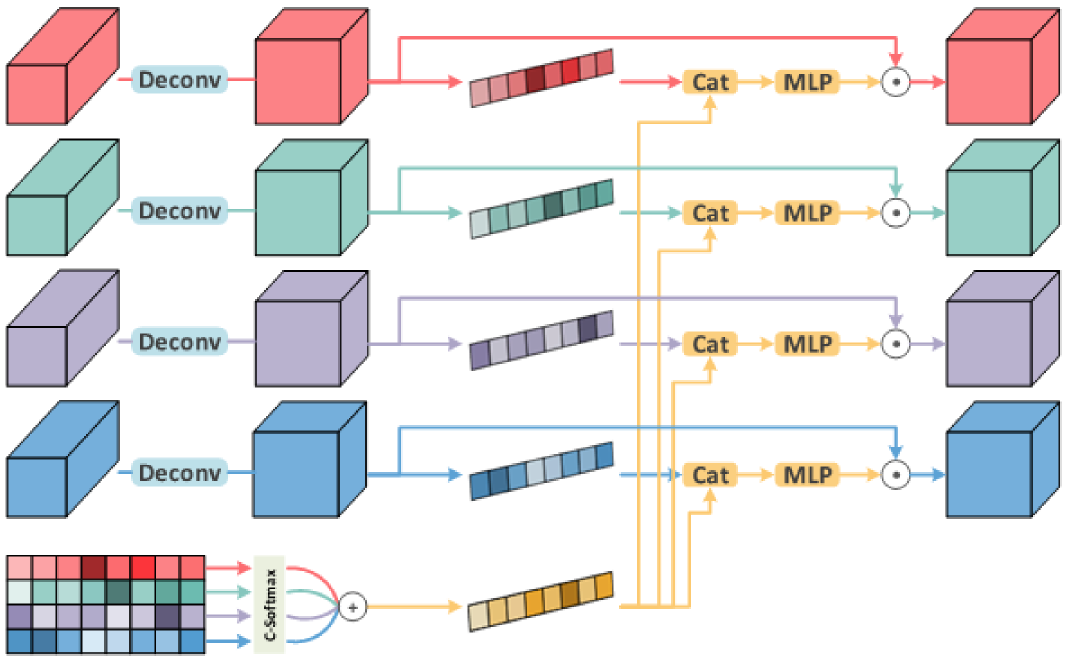
$$P = \sigma\left(f^{p}\left(SE\left(U_g^{(n)}\right)\right)\right)$$



Group Consistency Preserving Decoder

The most common up-sampling or deconvolution based feature decoders are not suitable for CoSOD tasks because they **ignore the inter-image constraints and may weaken the consistency between images during the prediction process**. Thus, **we propose a group consistency preserving decoder (GCPD) to consistently predict full-resolution co-saliency maps.**

- GCPD includes three cascaded feature decoding (FD) units;
- Learn a compact group feature vector y, and combine it with the vectorized deconvolution representations;
- The finest spatial resolution, which are further fed into a shared co-saliency head (CoSH) to generate full-resolution co-saliency maps;



Supervisions

We jointly optimize the co-saliency and single image saliency predictions in a multi-task learning framework.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_c + \beta \cdot \mathcal{L}_s$$

co-saliency loss:

$$\mathcal{L}_c = -\left(\sum_{n=1}^{N} \left(T_c^{(n)} \cdot \log\left(M^{(n)}\right) + \left(1 - T_c^{(n)}\right) \cdot \log\left(1 - M^{(n)}\right)\right)\right) / N$$

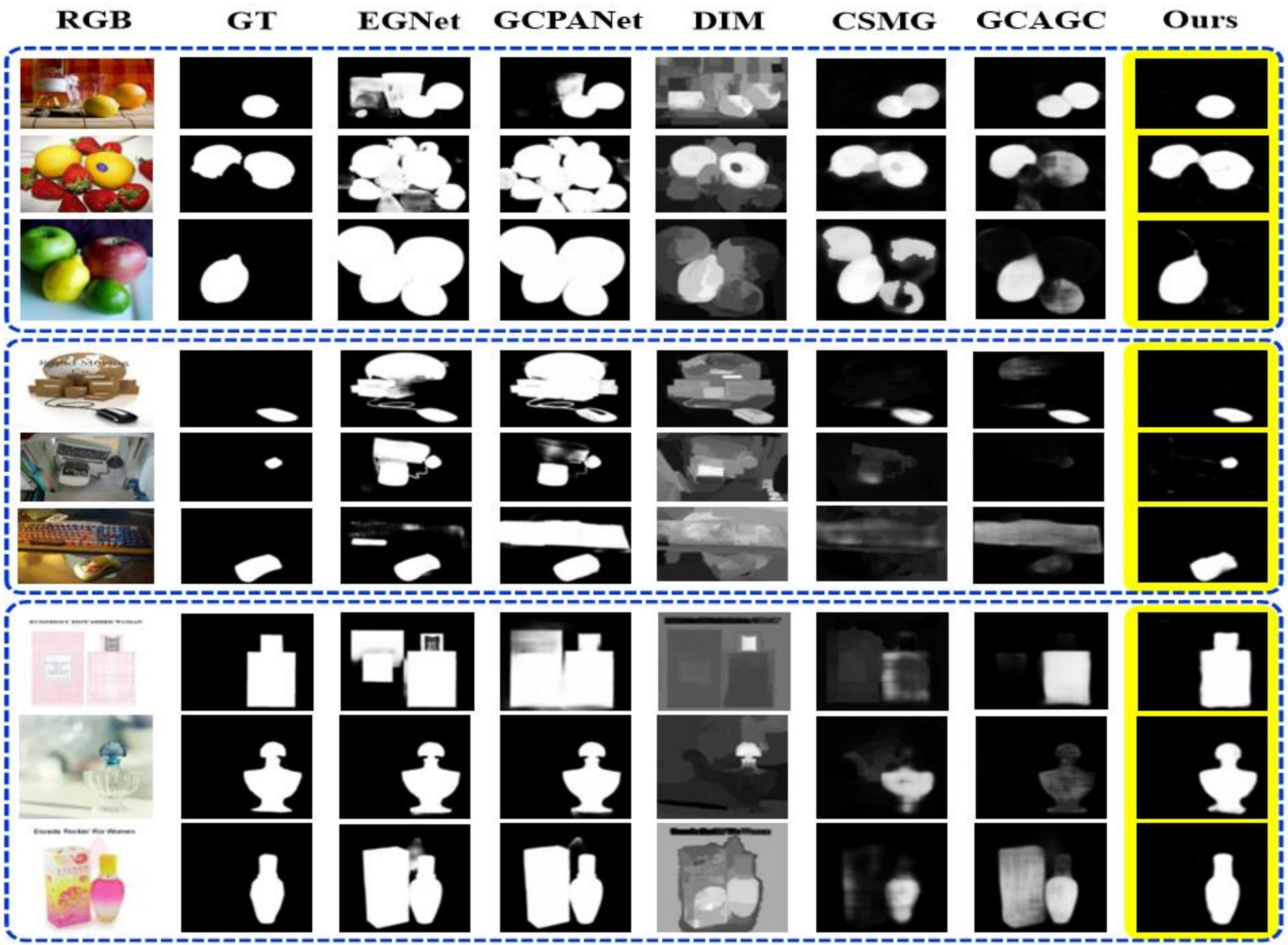
auxiliary saliency loss:

$$\mathcal{L}_s = -\left(\sum_{k=1}^{K} \left(T_s^{(k)} \cdot log\left(A^{(k)}\right) + (1 - T_s^{(k)}) \cdot \log\left(1 - A^{(k)}\right)\right)\right) / K$$

Experiments

- Benchmark Datasets: CoSOD3k, Cosal2015, MSRC, and iCoseg.
- Evaluation Metrics: Precision-Recall (P-R) curve, F-measure, MAE score, and S-measure
- Implementation Details: a sub-group containing 5 images are randomly selected from a certain query group. All input images are resized to 224 $\times$224. In each training iteration, 24 sub-groups from COCO-SEG and 64 samples from DUTS are simultaneously fed into the network for optimizing the objective function. In our experiment, we provide the results under two backbones including ResNet-50 and Dilated ResNet-50, and the training process converges until 50,000 iterations. The average inference time for a single image is 0.07 seconds.

Experiments



Experiments

	Cosal2015 Dataset			CoSOD3k Dataset			MSRC Dataset			iCoseg Dataset		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
CPD [49]	0.8228	0.0976	0.8168	0.7661	0.1068	0.7788	0.8250	0.1714	0.7184	0.8768	0.0579	0.8565
EGNet [58]	0.8281	0.0987	0.8206	0.7692	0.1061	0.7844	0.8101	0.1848	0.7056	0.8880	0.0601	0.8694
GCPANet [5]	0.8557	0.0813	0.8504	0.7808	0.1035	0.7954	0.8133	0.1487	0.7575	0.8924	0.0468	0.8811
UMLF [20]	0.7298	0.2691	0.6649	0.6895	0.2774	0.6414	0.8605	0.1815	0.8007	0.7623	0.2389	0.6828
CODW [53]	0.7252	0.2741	0.6501	–	–	–	0.8020	0.2645	0.7152	0.8271	0.1782	0.7510
DIM [25]	0.6363	0.3126	0.5943	0.5603	0.3267	0.5615	0.7419	0.3101	0.6579	0.8273	0.1739	0.7594
GoNet [23]	0.7818	0.1593	0.7543	–	–	–	0.8598	0.1779	0.7981	0.8653	0.1182	0.8221
CSMG [54]	0.8340	0.1309	0.7757	0.7641	0.1478	0.7272	0.8609	0.1892	0.7257	0.8660	0.1050	0.8122
RCGS [43]	0.8245	0.1004	0.7958	–	–	–	0.7692	0.2134	0.6717	0.8005	0.0976	0.7860
GCAGC [55]	0.8666	0.0791	0.8433	0.8066	0.0916	0.7983	0.7903	0.2072	0.6768	0.8823	0.0773	0.8606
CoADNet-V	0.8748	0.0644	0.8612	0.8249	0.0696	0.8368	0.8597	0.1139	0.8082	0.8940	0.0416	0.8839
CoADNet-R	0.8771	0.0609	0.8672	0.8204	**0.0643**	0.8402	**0.8710**	**0.1094**	**0.8269**	0.8997	0.0411	0.8863
CoADNet-DR	**0.8874**	**0.0599**	**0.8705**	**0.8308**	0.0652	**0.8416**	0.8618	0.1323	0.8103	**0.9225**	**0.0438**	**0.8942**

Experiments

Modules					Cosal2015 Dataset			CoSOD3k Dataset		
Baseline	OIaSG	GASA	GGD	GCPD	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
✓					0.7402	0.1406	0.7459	0.7099	0.1170	0.7320
✓	✓				0.8023	0.1161	0.7967	0.7489	0.1138	0.7721
✓	✓	✓			0.8465	0.0946	0.8209	0.8008	0.0915	0.8089
✓	✓	✓	✓		0.8682	0.0712	0.8534	0.8211	0.0815	0.8223
✓	✓	✓	✓	✓	**0.8874**	**0.0599**	**0.8705**	**0.8308**	**0.0652**	**0.8416**

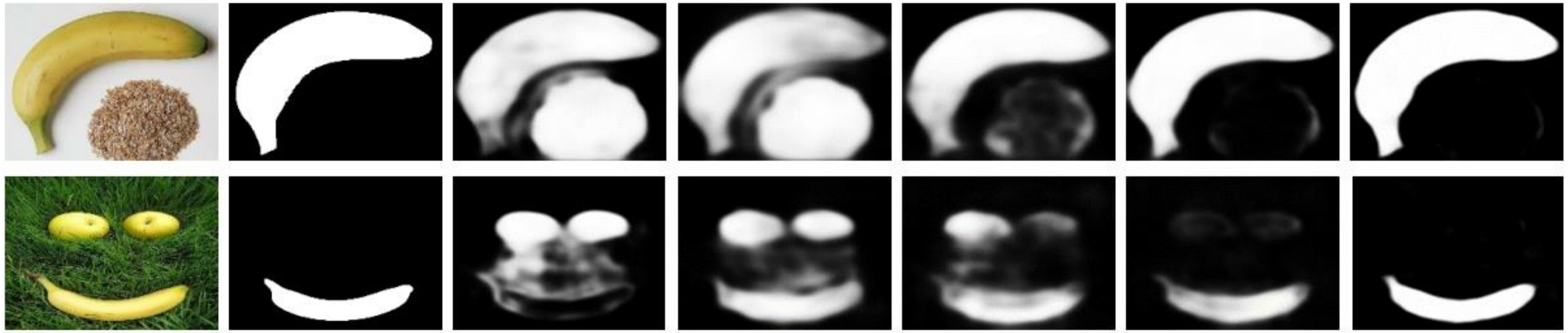


Figure 6: Visualization of different ablative results. From left to right: Input image group, Ground truth, Co-saliency maps produced by the Baseline, Baseline+OIaSG, Baseline+OIaSG+GASA, Baseline+OIaSG+GASA+GGD, and the full CoADNet.

Experiments

Table 3: Detection performance of our CoADNet-V using CoSOD3k as the training set.

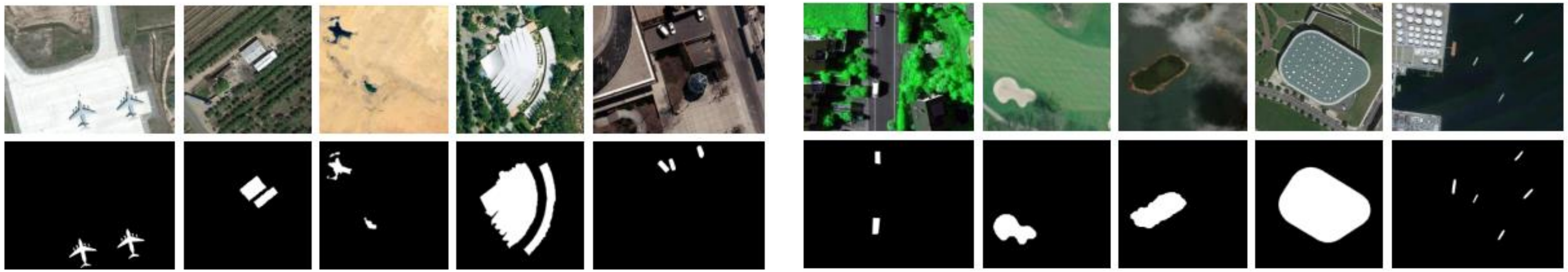
	Cosal2015 Dataset			MSRC Dataset			iCoseg Dataset		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
CoADNet-V	0.8592	0.0818	0.8454	0.8347	0.1558	0.7670	0.8784	0.0725	0.8569

We need an appropriate dataset to train our CoSOD network!!

Conclusion

- We proposed an end-to-end CoSOD network by investigating how to model and utilize the inter-image correspondences.
- We first decoupled the single-image SOD from the CoSOD task and proposed an OIaSG module to provide learnable saliency prior guidance.
- Then, the GASA and GGD modules are integrated into a two-stage aggregate-and-distribute structure for effective extraction and adaptive distribution of group semantics.
- Finally, we designed a GCPD structure to strengthen inter-image constraints and predict full-resolution co-saliency maps.
- Experimental results and ablative studies demonstrated the superiority of the proposed CoADNet and the effectiveness of each component.

Salient Object Detection in Optical RSIs



Challenges

1

Optical RSI may include diversely scaled objects, various scenes and object types, cluttered backgrounds, and shadow noises.

2

Sometimes, there is even no salient region in a real outdoor scene, such as the desert, forest, and sea.

Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images

Qijian Zhang, Runmin Cong*, Chongyi Li, Ming-Ming Cheng, Yuming Fang, Xiaochun Cao, and Yao Zhao

IEEE Transaction on Image Processing, 2021

https://rmcong.github.io/proj_DAFNet.html

Challenges

a) First, salient objects are often corrupted by **background interference and redundancy**.

b) Second, salient objects in RSIs present much more **complex structure and topology** than the ones in NSIs, which poses new **challenges in capturing complete object regions**.

c) Third, for the optical RSI SOD task, there is **only one dataset** (i.e., ORSSD [6]) available for model training and performance evaluation, which contains 800 images totally. This dataset is pioneering, but **its size is still relatively small**.

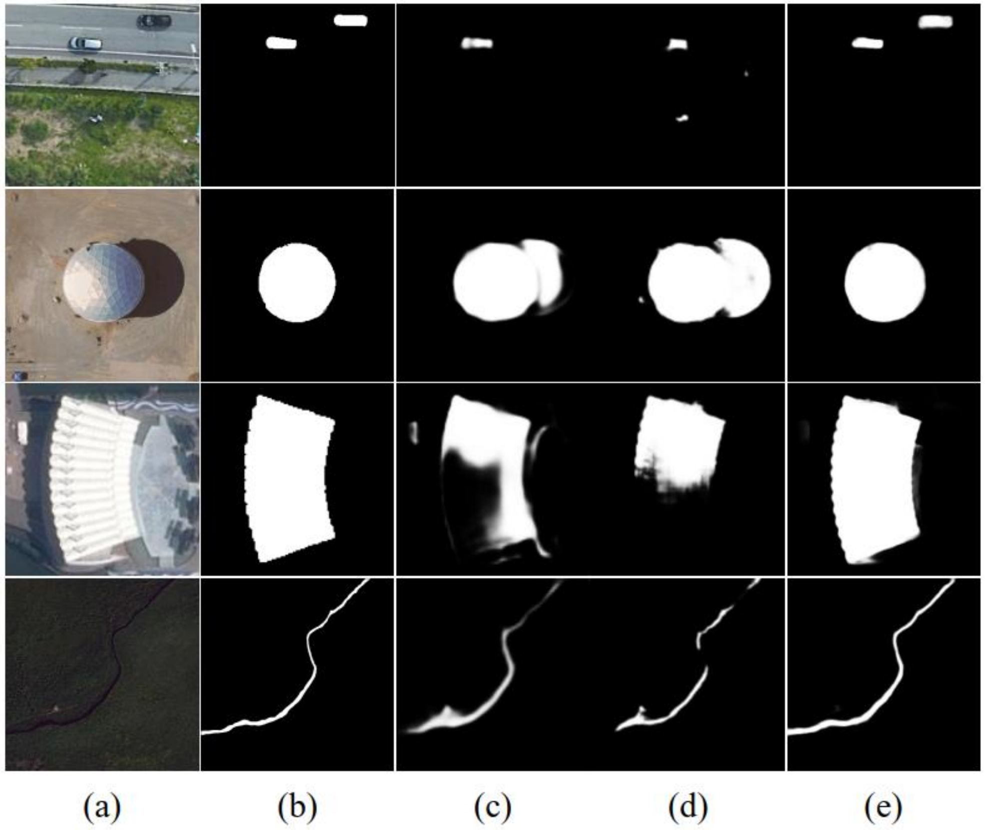


Fig. 1. Visual illustration of SOD results for optical RSIs by applying different methods. (a) Optical RSIs. (b) Ground truth. (c) PFAN [11]. (d) LVNet [6]. (e) Proposed DAFNet.

[6] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9156–9166, 2019

Contributions

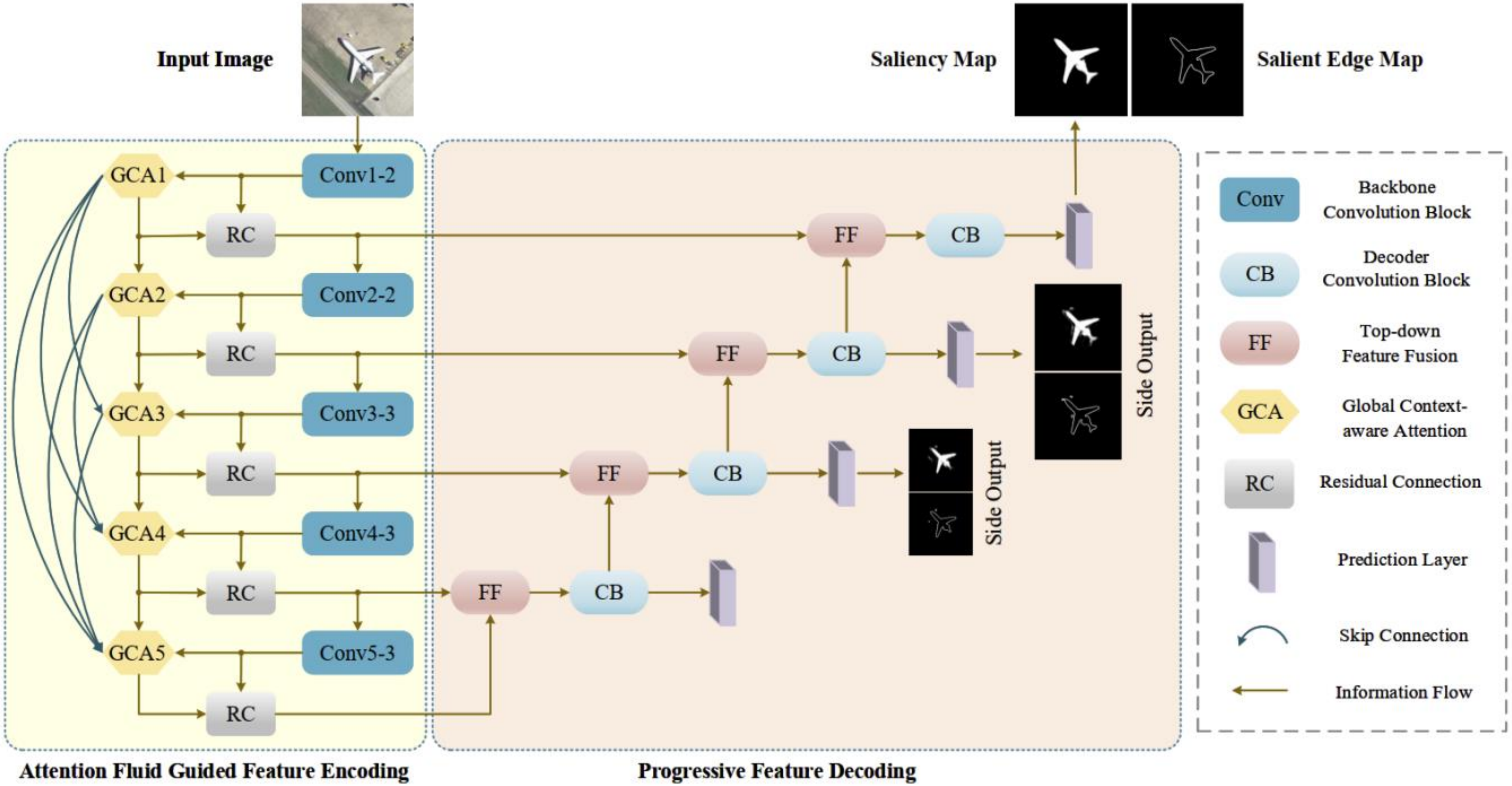
a) An end-to-end Dense Attention Fluid Network (DAFNet) is proposed to achieve SOD in optical RSIs, equipped with a **Dense Attention Fluid (DAF) structure** decoupled from the backbone feature extractor and a **Global Context-aware Attention (GCA) mechanism**.

b) The DAF structure is designed to **combine the multi-level attention cues**, where shallow-layer attention cues flow into the attention units of deeper layers so that low-level attention cues could **be propagated as guidance information to enhance the high-level attention**.

c) The GCA mechanism is proposed to **model the global context semantic relationships** by a global feature aggregation module, and **tackle the scale variation** by a cascaded pyramid attention module.

d) A **large-scale benchmark dataset** including 2, 000 images and corresponding pixel-wise annotations is constructed for SOD in optical RSIs. The proposed DAFNet **consistently outperforms 15 state-of-the-art competitors** in the experiments.

Our Method


Input Image
Saliency Map
Salient Edge Map
Conv1-2
Conv2-2
Conv3-3
Conv4-3
Conv5-3
GCA1
GCA2
GCA3
GCA4
GCA5
RC
FF
CB
Side Output
Side Output
Conv
Backbone Convolution Block
CB
Decoder Convolution Block
FF
Top-down Feature Fusion
GCA
Global Context-aware Attention
RC
Residual Connection
Prediction Layer
Skip Connection
Information Flow
Attention Fluid Guided Feature Encoding
Progressive Feature Decoding

Attention Fluid Guided Feature Encoding

- The attention fluid guided feature encoding consists of:
  - a feature fluid that generates hierarchical feature representations with stronger discriminative ability by incorporating attention cues mined from the corresponding global context-aware attention modules.
  - an attention fluid where low-level attention maps flow into deeper layers to guide the generation of high-level attentions .

Global Context-aware Attention Mechanism

- We investigate a novel **global context-aware attention (GCA) mechanism** that explicitly captures the long-range semantic dependencies among all spatial locations in an attention manner. The GCA module consists of two main functional components：
  - The **global feature aggregation (GFA)** module consumes raw side features generated from the backbone convolutional block and produces aggregated features that encode global contextual information.
  - The **cascaded pyramid attention (CPA)** module is used to address the scale variation of objects in optical RSIs, which takes the aggregated features from GFA as input and produces a progressively refined attention map under a cascaded pyramid framework.

Global Context-aware Attention Mechanism

Global Feature Aggregation

- The GFA module aims to achieve feature alignment and mutual reinforcement between saliency patterns by aggregating global semantic relationships among pixel pairs, which is beneficial to generate intact and uniform saliency map.

- Aggregated feature map F^s with global contextual dependencies:

$$F^s = f^s + \delta \cdot (f^s \odot G^s)$$

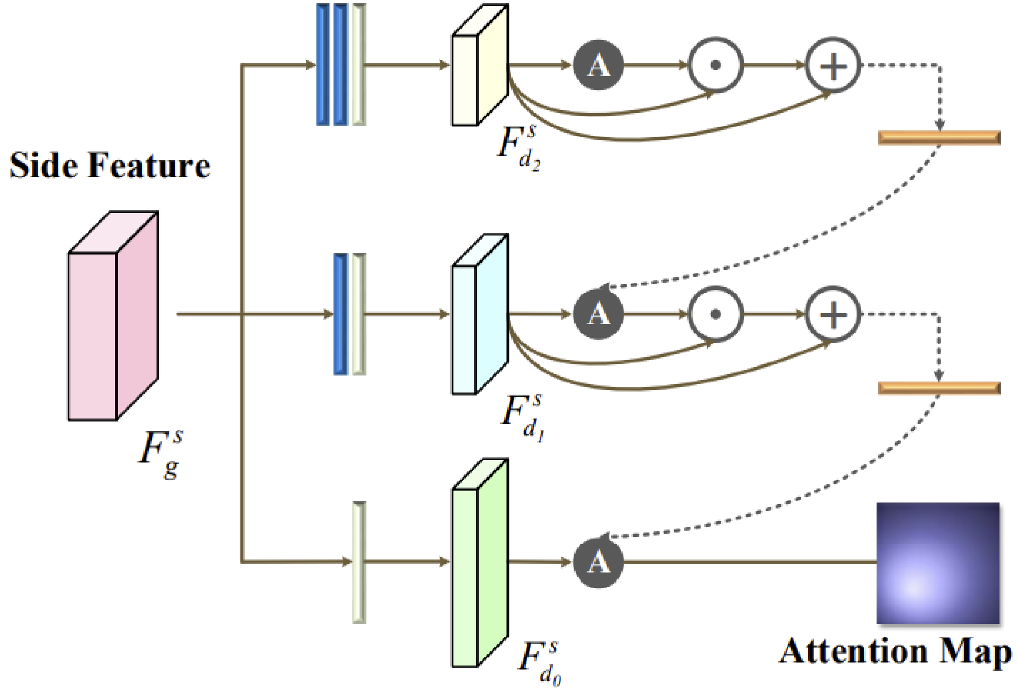
- Refined feature map F_g^s with more compact channel information:

$$F_g^s = F^s \odot \Gamma^s$$

Global Context-aware Attention Mechanism

● Cascaded Pyramid Attention

- We design a cascaded pyramid attention to progressively refine both features and attentive cues from coarse to fine.
- The CPA module produces a full-resolution attention map $\hat{A}^s$ at the original feature scale, which can be formulated as:



$$\hat{A}^s = Att(concat(F^s_{d_0}, (F^s_{d_1} \odot \boxed{A^s_{d_1}} + F^s_{d_1}) \uparrow))$$

$$A^s = Att(F^s_g) = \sigma\left(conv\left(concat\left(avepool(F^s_g), maxpool(F^s_g)\right); \hat{\theta}\right)\right)$$

Dense Attention Fluid Structure

- Each GCA module consumes a raw side feature map f^s, and produces an attention map $\hat{A}^s$.
- First, we build sequential connections among the attention maps generated from hierarchical feature representations. Moreover, considering the hierarchical attention interaction among different levels, we add feed-forward skip connections to form the attention fluid. Formally, the above updating process is denoted as:

$$\hat{A}^s \leftarrow \sigma(conv(concat((\hat{A}^1)\downarrow, \ldots, (\hat{A}^{s-1})\downarrow, \hat{A}^s)))$$

- With the updated attention map, the final feature map at the s^{th} convolution stage F_c^s can be generated via the residual connection:

$$F_c^s = concat(F_{d_0}^s, (F_{out_1}^s)\uparrow) \odot (\hat{A}^s + O^s)$$

Progressive Feature Decoding

- Each decoding stage consists of three procedures.
- First, we employ top-down feature fusion (FF) to align the spatial resolution and number of channels between adjacent side feature maps via up-sampling and 1×1 convolution, and then perform pointwise summation.
- Second, a bottleneck convolutional block (CB) is deployed to further integrate semantic information from fusion features.
- Third, we deploy a mask prediction layer and an edge prediction layer for the decoded features, and use a Sigmoid layer to map the range of saliency scores into [0, 1].
- The final output of our DAFNet is derived from the predicted saliency map at the top decoding level.

Loss Function

- To accelerate network convergence and yield more robust saliency feature representations, we formulate a hierarchical optimization objective by applying deep supervisions to the side outputs at different convolution stages. We further introduce edge supervisions to capture fine-grained saliency patterns and enhance the depiction of object contours.

$$\ell = \sum_{s=1}^{3} (\omega_m^s \cdot \boxed{\ell_m^s} + \omega_e^s \cdot \boxed{\ell_e^s})$$

class-balanced binary cross-entropy loss function for saliency supervision

class-balanced binary cross-entropy loss function for salient edge supervision

EORSSD Dataset

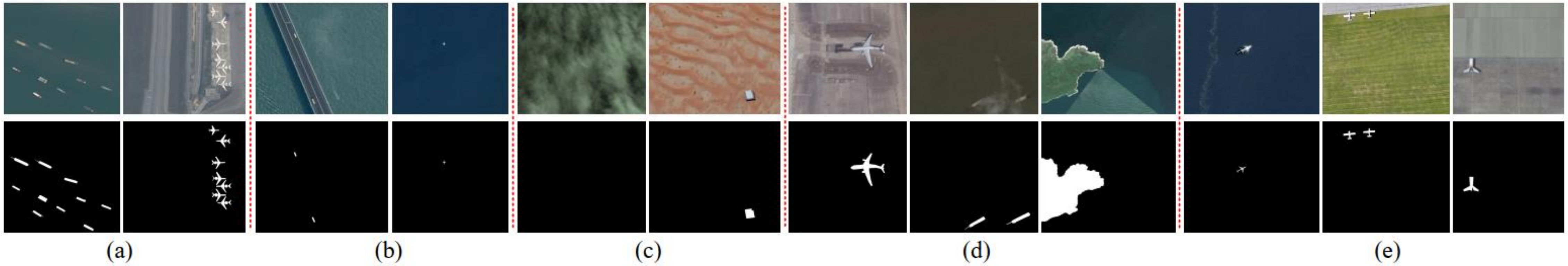


Fig. 4. Visualization of the more challenging EORSSD dataset. The first row shows the optical RSI, and the second row exhibits the corresponding ground truth. (a) Challenge in the number of salient objects. (b) Challenge in small salient objects. (c) Challenge in new scenarios. (d) Challenge in interferences from imaging. (e) Challenge in specific circumstances.

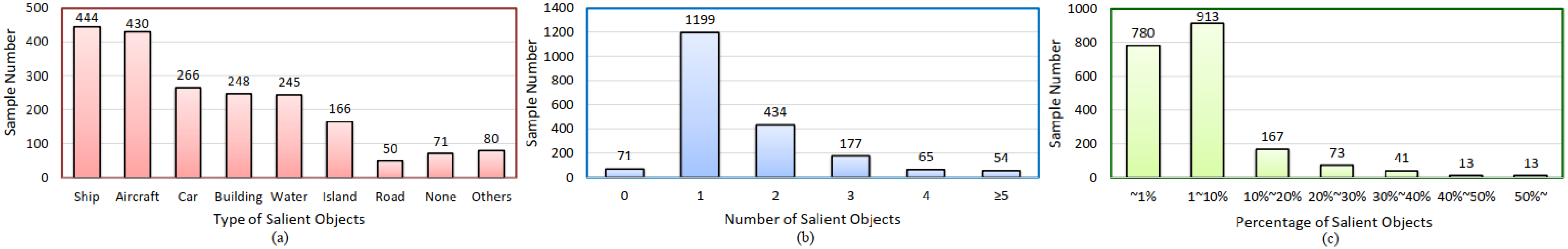


Fig. 5. Statistical analysis of EORSSD dataset. (a) Type analysis of salient object. (b) Number analysis of salient object. (c) Size analysis of salient object.

Download: https://github.com/rmcong/EORSSD-dataset

Experiments

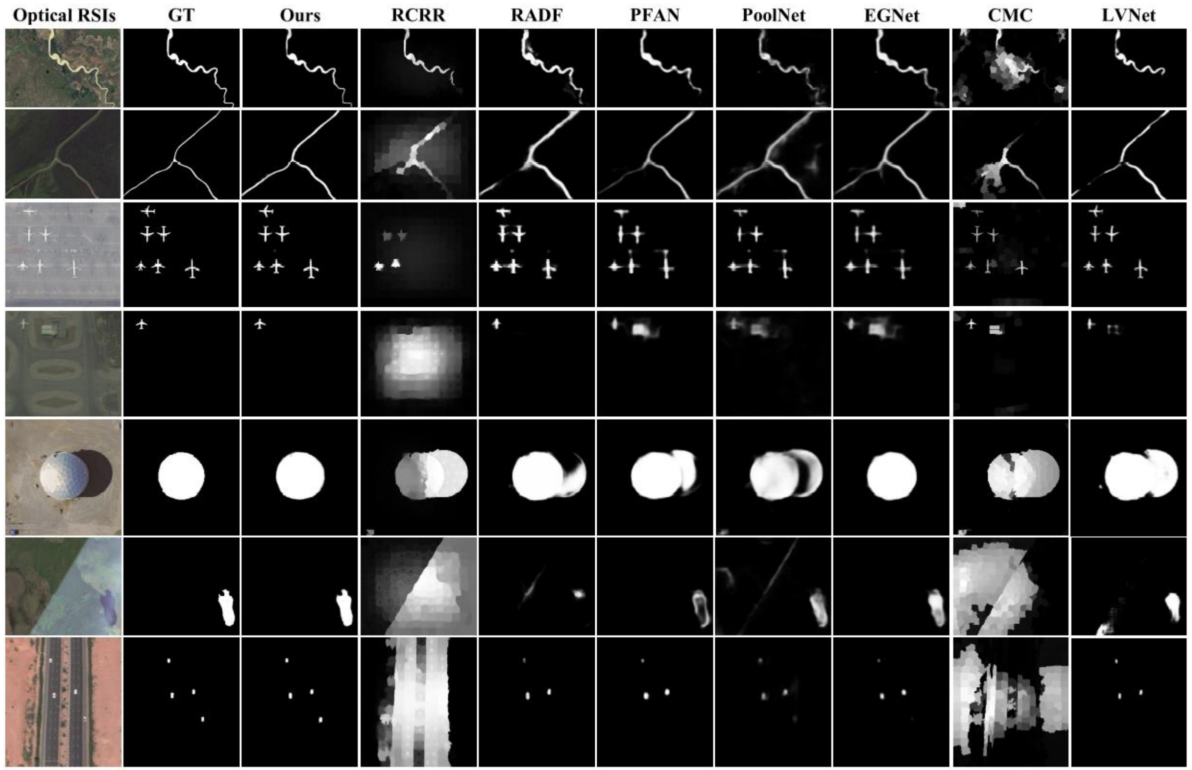


TABLE I
QUANTITATIVE COMPARISONS WITH DIFFERENT METHODS ON THE TESTING SUBSET OF THE ORSSD AND EORSSD DATASETS. TOP THREE RESULTS ARE MARKED IN RED, BLUE, AND GREEN RESPECTIVELY.

	ORSSD Dataset			EORSSD Dataset		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
DSG [26]	0.6630	0.1041	0.7195	0.5837	0.1246	0.6428
RRWR [25]	0.5950	0.1324	0.6835	0.4495	0.1677	0.5997
HDCT [22]	0.5775	0.1309	0.6197	0.5992	0.1087	0.5976
SMD [23]	0.7075	0.0715	0.7640	0.6468	0.0770	0.7112
RCRR [24]	0.5944	0.1277	0.6849	0.4495	0.1644	0.6013
DSS [28]	0.7838	0.0363	0.8262	0.7158	0.0186	0.7874
R3Net [27]	0.7998	0.0399	0.8141	0.7709	0.0171	0.8193
RADF [29]	0.7881	0.0382	0.8259	0.7810	0.0168	0.8189
PFAN [11]	0.8344	0.0543	0.8613	0.7740	0.0159	0.8361
PoolNet [39]	0.7911	0.0358	0.8403	0.7812	0.0209	0.8218
EGNet [16]	0.8438	0.0216	0.8721	0.8060	0.0109	0.8602
CMC [46]	0.4214	0.1267	0.6033	0.3663	0.1057	0.5800
VOS [45]	0.4168	0.2151	0.5366	0.3599	0.2096	0.5083
SMFF [41]	0.4864	0.1854	0.5312	0.5738	0.1434	0.5405
LVNet [6]	0.8414	0.0207	0.8815	0.8051	0.0145	0.8645
DAFNet-V	0.9174	0.0125	0.9191	0.8922	0.0060	0.9167
DAFNet-R	0.9235	0.0106	0.9188	0.9060	0.0053	0.9185

TABLE V
QUANTITATIVE EVALUATION OF ABLATION STUDIES ON THE TESTING SUBSET OF EORSSD DATASET.

Baseline	GFA	CPA	DAF	F_β	MAE	S_m
✓				0.8391	0.0125	0.8432
✓	✓			0.8504	0.0098	0.8661
✓	✓	✓		0.8742	0.0083	0.8760
✓	✓	✓	✓	0.8922	0.0060	0.9167

Conclusion

- This paper focuses on salient object detection in optical remote sensing images and proposes an end-to-end encoder-decoder framework dubbed as DAFNet, in which attention mechanism is incorporated to guide the feature learning.
- Benefiting from the attention fluid structure, our DAFNet learns to integrate low-level attention cues into the generation of high-level attention maps in deeper layers. Moreover, we investigate the global context-aware attention mechanism to encode long-range pixel dependencies and explicitly exploit global contextual information. In addition, we construct a new large-scale optical RSI benchmark dataset for SOD with pixel-wise saliency annotations.
- Extensive experiments and ablation studies demonstrate the effectiveness of the proposed DAFNet architecture.

Future work

1

New attempts in learning based saliency detection methods, such as small samples training, weakly supervised learning, and cross-domain learning.

2

Extending the saliency detection task in different data sources, such as light filed image, RGB-D video, and remote sensing image.

3

New ideas and solutions in saliency detection task, such as instance-level saliency detection and segmentation, saliency improvement and refinement.

Thanks

Runmin Cong (丛润民)
Beijing Jiaotong University

Homepage: https://rmcong.github.io/
E-mail: rmcong@bjtu.edu.cn