PRCV 2021（北京→珠海）
第四届中国模式识别与计算机视觉大会

深度信息在显著性目标检测任务中的应用初探

主办单位：中国图象图形学学会、中国人工智能学会、中国计算机学会、中国自动化学会
承办单位：北京科技大学、北京交通大学、北京邮电大学 协办单位：中山大学、清华大学

汇报人：丛润民
北京交通大学数字媒体信息处理中心

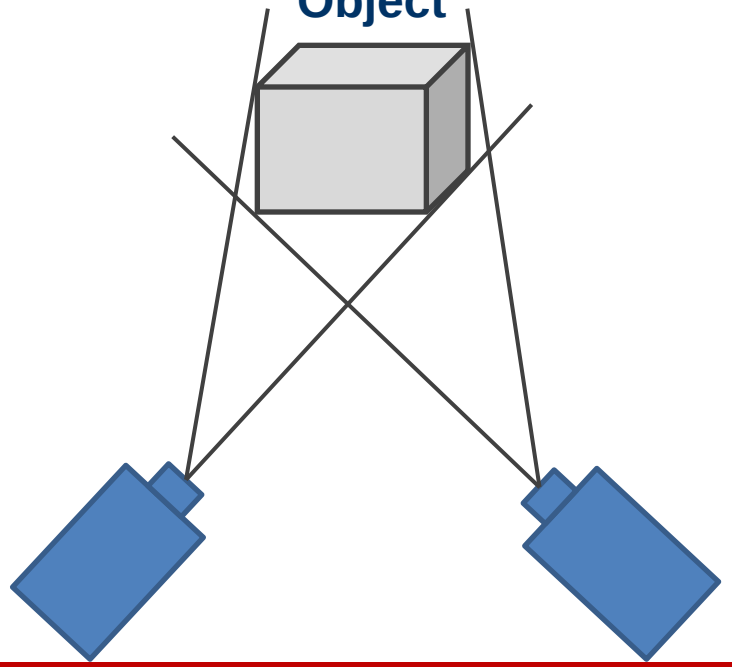
2021-12-21

Outline

- Introduction
- Cross-modality Discrepant Interaction Network for RGB-D Salient Object Detection, ACM MM 2021
- DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection, TIP 2021
- Future Work

Introduction

Object



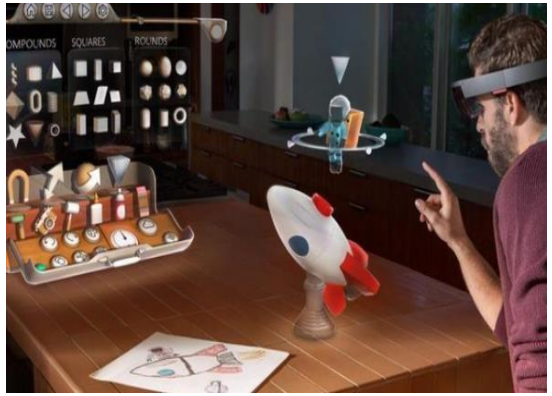
Depth Cameras

Kinect

ORBBEC

Photographing

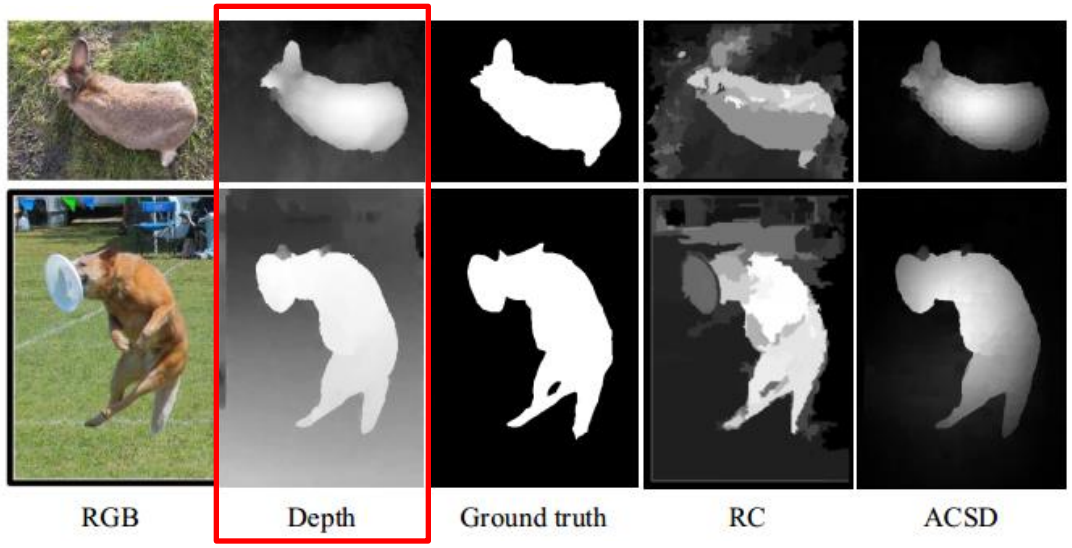
Driverless



CVPR 2022/ICCV 2011大会主席、香港科技大学权龙教授说过“真正意义上的计算机视觉要超越识别，感知三维场景。我们活在三维空间里，要做到交互和感知，就必须将世界恢复到三维。”

Introduction

RGB-D Salient Object Detection



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

depth quality perception

ACM multimedia
Chengdu, China OCT 20-24 2021

Cross-Modality Discrepant Interaction Network for RGB-D Salient Object Detection

Chen Zhang, Runmin Cong*, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, Sam Kwong

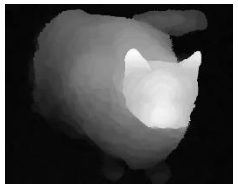
***ACM International Conference on Multimedia (ACM MM)*, 2021**

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

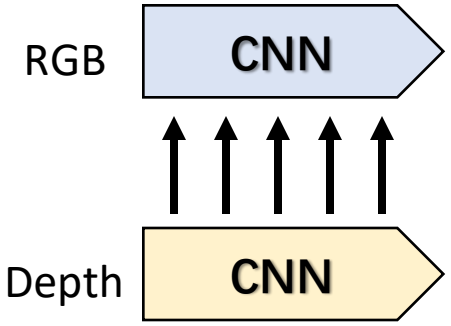
Motivation



How to effectively **integrate** the complementary information from RGB image and its corresponding depth map?

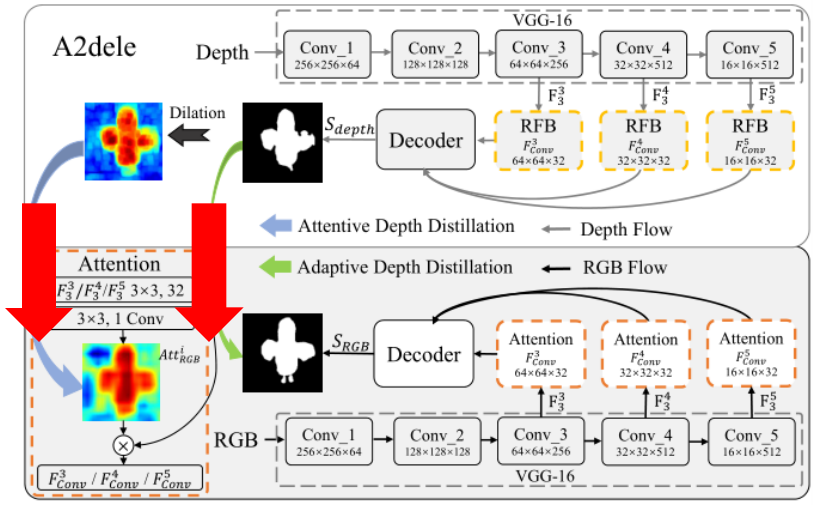




Motivation

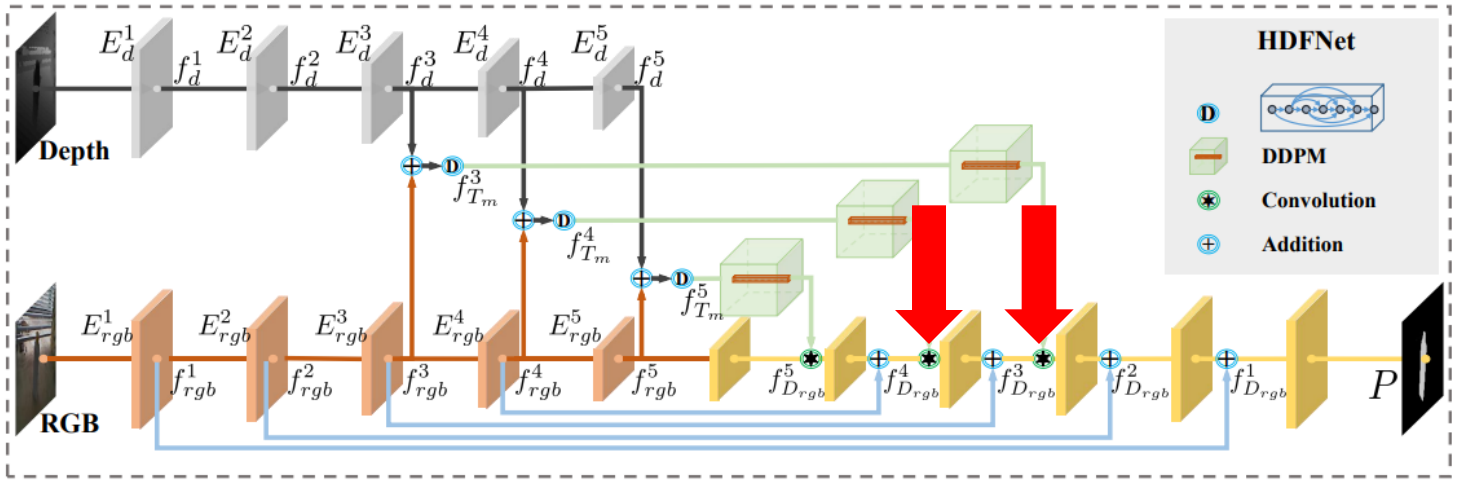


(a)

(a) **Unidirectional interaction mode**, which uses the depth cues as auxiliary information to supplement the RGB branch.

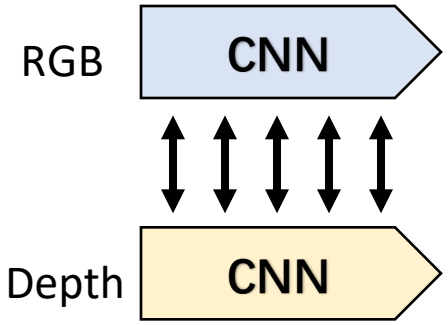


(CVPR 2020)



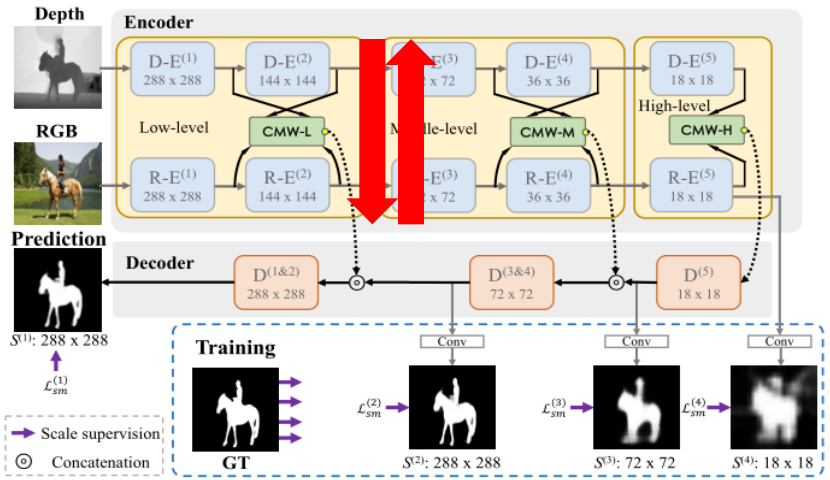
(ECCV 2020)

Motivation

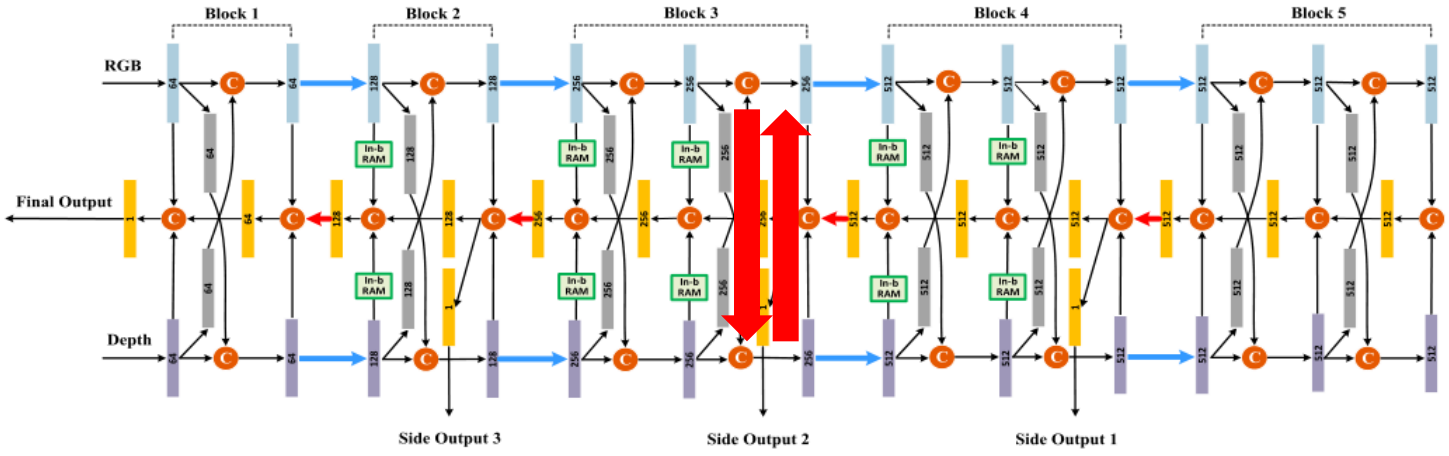


(b)

(b) **Bidirectional interaction mode**, which treats RGB and depth cues equally to achieve cross-modality interaction.



(ECCV 2020)



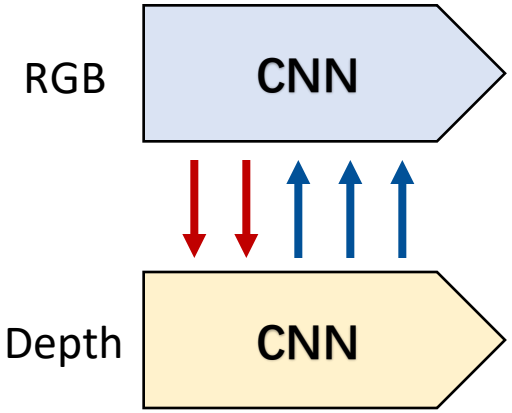
(TCyb 2021)

Motivation

How can we fully exploit the strengths of both modalities and provide clear guidance?

(1) Depth map has relatively distinct details for describing the salient objects, but can not distinguish different object instances at the same depth level.

(2) RGB image contains more affluent semantic information, but the complex background interference may cause the salient objects to be flawed.



(c)

Mode (c), a discrepant interaction mode is proposed!

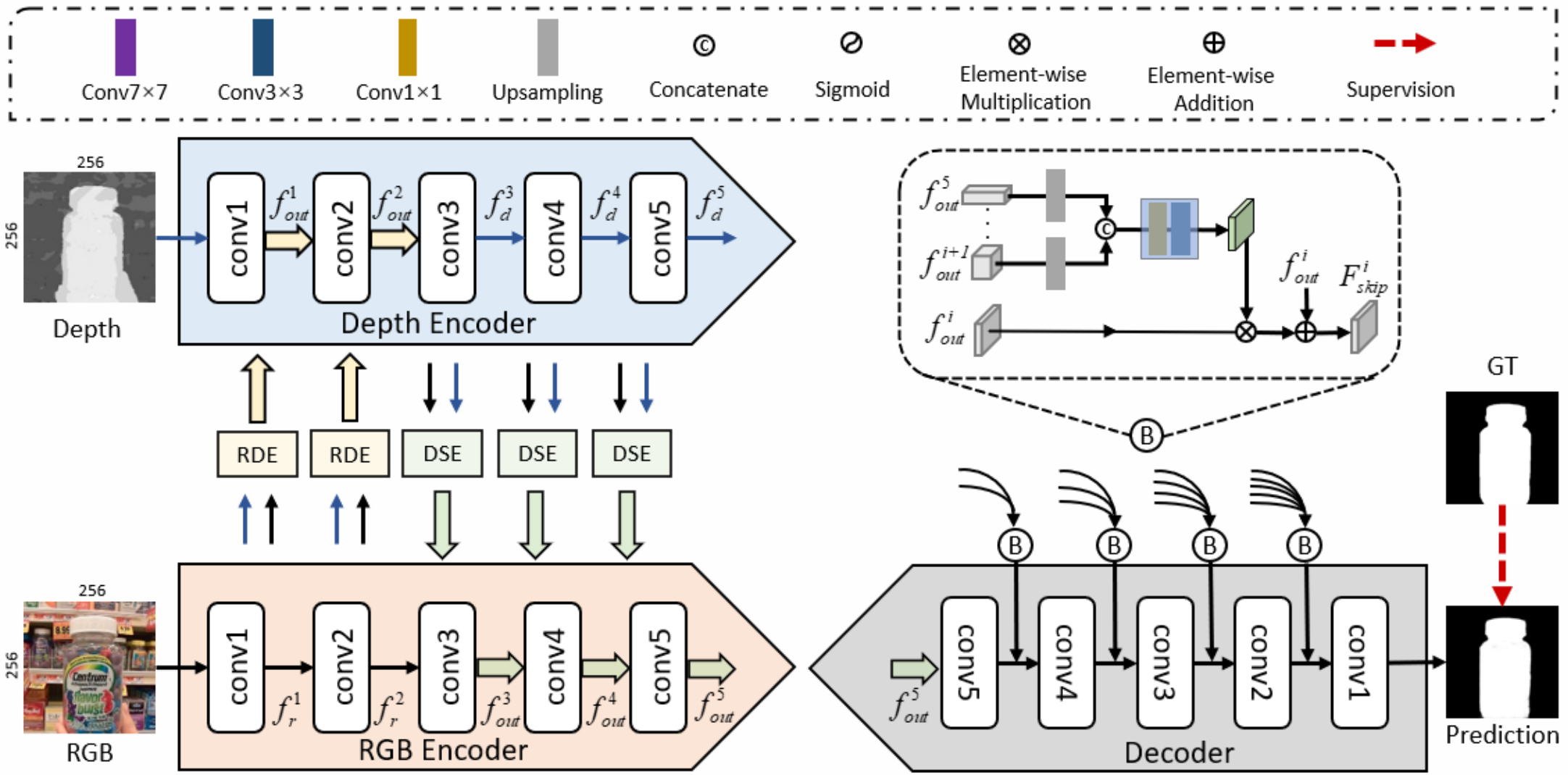
Contributions

- We propose a **Cross-modality Discrepant Interaction Network (CDINet)**, which differentially models the dependence of two modalities according to the feature representations of different layers.

- We design an **RGB-induced Detail Enhancement (RDE)** module in low-level stage and a **Depth-induced Semantic Enhancement (DSE)** module in high-level stage to enhance self-modal feature by utilizing complementary modality.

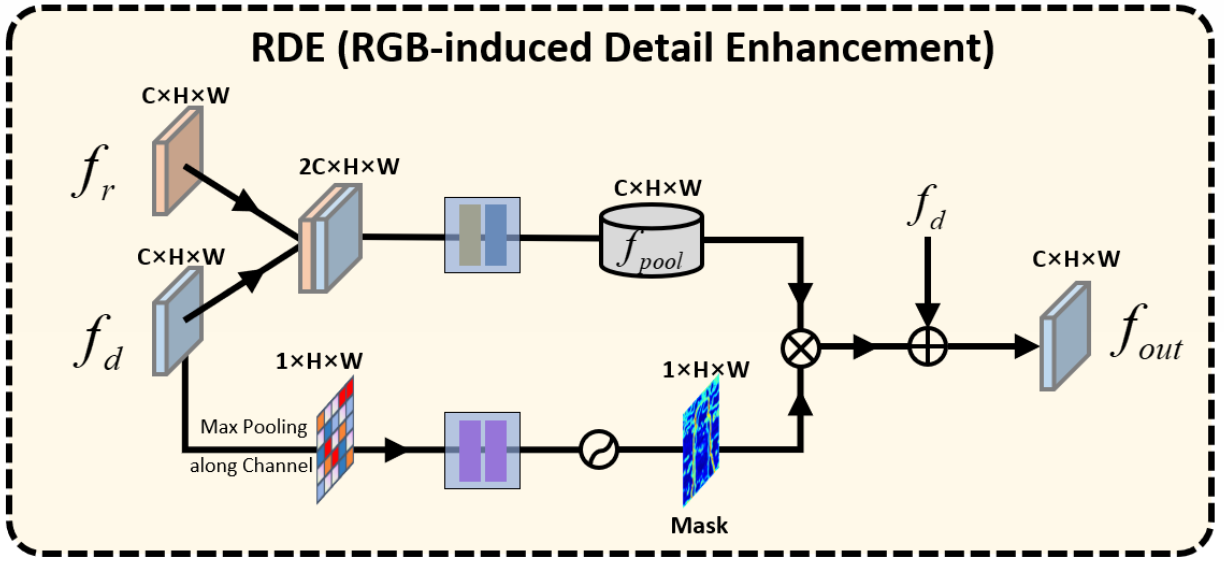
- We design a **Dense Decoding Reconstruction (DDR)** structure, which generates a semantic block by leveraging multiple high-level encoder features to upgrade the skip connection in the feature decoding.

Method


Conv7×7
Conv3×3
Conv1×1
Upsampling
Concatenate
Sigmoid
Element-wise Multiplication
Element-wise Addition
Supervision
256
256
Depth
conv1
conv2
conv3
conv4
conv5
Depth Encoder
RDE
RDE
DSE
DSE
DSE
256
256
RGB
RGB Encoder
Decoder
B
GT
Prediction

RGB-induced Detail Enhancement Module

The RGB-induced Detail Enhancement (RDE) module can transfer **detail supplement information** from RGB modality to depth modality in low-level encoder stage.



We first adopt two cascaded convolutional layers to fuse the underlying visual features of two modalities.

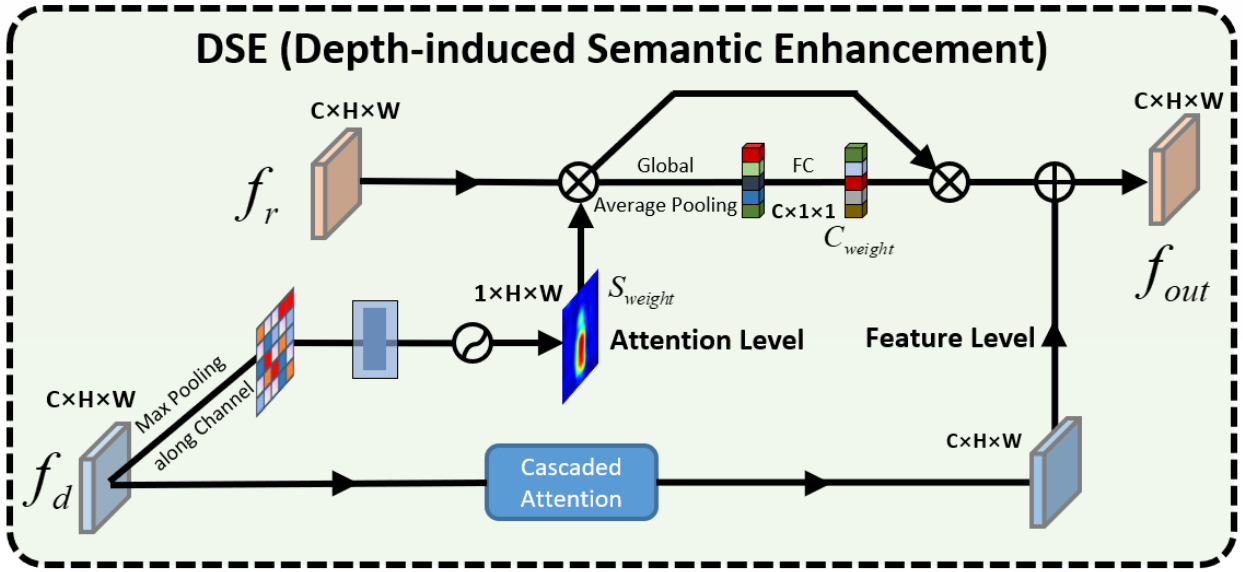
$$f_{pool}^{i} = conv_3(conv_1([f_r^i, f_d^i])),$$

The depth features generate a spatial attention mask, and obtain the required supplement information from the perspective of depth modality.

$$f_{out}^{i} = \sigma(conv_7(conv_7(maxpool(f_d^i)))) \odot f_{pool}^{i} + f_d^i,$$

Depth-induced Semantic Enhancement Module

The Depth-induced Semantic Enhancement (DSE) module can assist RGB branch in capturing clearer and fine-grained semantic attributes by utilizing the **positioning accuracy** and **internal consistency** of high-level depth features.



First, we learn an attention vector from the depth features to guide RGB modality to focus on the region of interest:

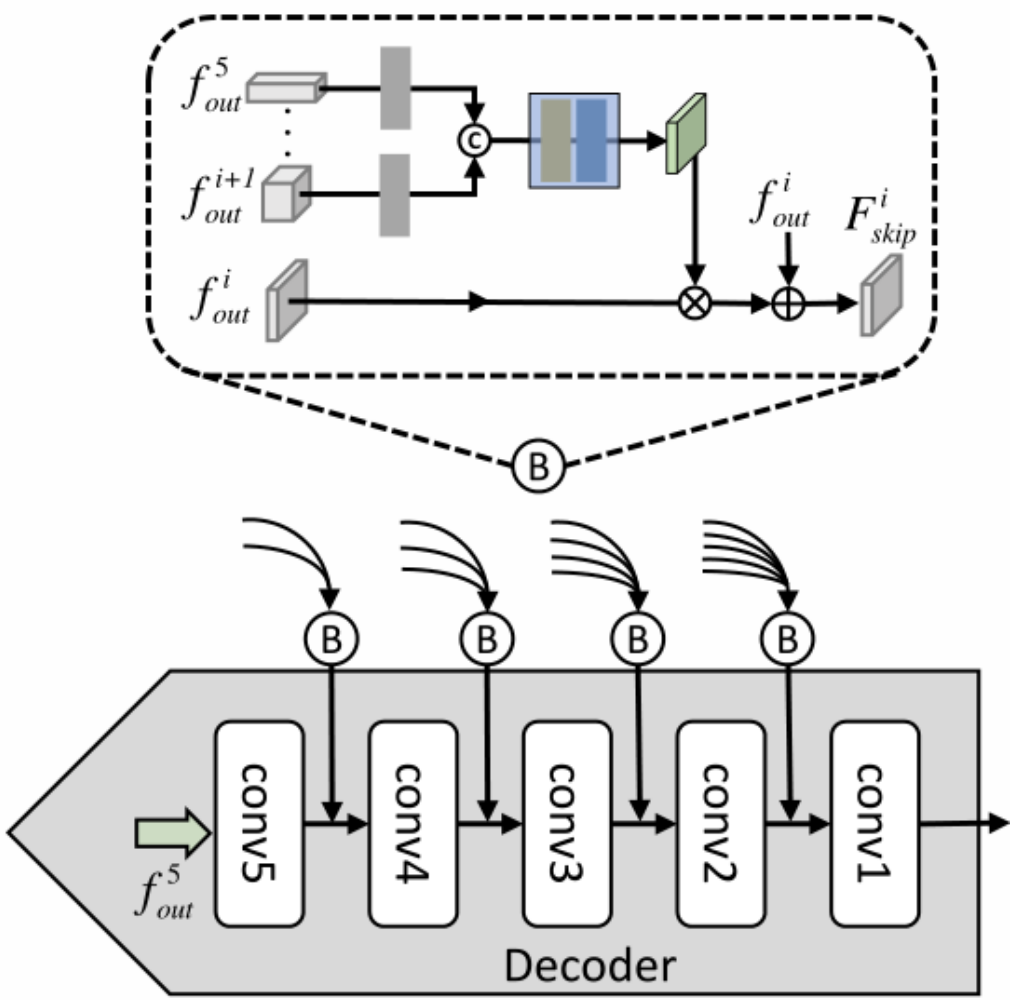
$$S_{weight} = \sigma(conv_3(maxpool(f_d^i)), \quad f_{rs}^i = S_{weight} \odot f_r^i,$$

$$C_{weight} = \sigma(FC(GAP(f_{rs}^i))), \quad D_{att}^i = C_{weight} \odot f_{rs}^i,$$

Then we use cascaded attention to enhance the depth features, the features that eventually flow into the next layer of the RGB branch can be expressed as:

$$f_{out}^i = D_{att}^i + D_{add}^i + f_r^i.$$

Dense Decoding Reconstruction Structure



Existing question

While the traditional skip connection introduces supplementary information, it also introduces additional interference information.

Solution

We propose a dense decoding reconstruction (DDR) structure, which generates a semantic block by densely connecting the higher-level encoding features to provide more comprehensive semantic guidance:

$$B^i = conv_3(conv_1([up(f_{skip}^{i+1}), ..., up(f_{skip}^5)])),$$

$$F_{skip}^i = B^i \odot f_{skip}^i + f_{skip}^i,$$

Experiments

- Benchmark Datasets: NJUD (1985 RGB-D images), NLPR (1000 RGB-D images), STEREO (797 RGB-D images), LFSD (100 RGB-D images), and DUT (1200 RGB-D images).
- Evaluation Metrics:

$$F_\beta = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \quad MAE = \frac{1}{H \times W} \sum_{y=1}^{H} \sum_{x=1}^{W} |S(x,y) - G(x,y)|, \quad S = \alpha * S_o + (1 - \alpha) * S_r,$$

- Implementation Details: All the training and testing images are resized to 256 × 256, and the depth map is simply copied to three channels as input. Then, to avoid overfitting, we use random flipping and rotating to augment the training samples. Moreover, we apply the usual binary cross-entropy loss function to optimize the proposed network, and the Adam algorithm is used to optimize our network with the batch size of 4 and the initial learning rate of 1e-4 which is divided by 5 every 40 epochs.

Experiments

		MMCI [3]	TAN [2]	CPFP [37]	DMRA [26]	FRDT [35]	SSF [34]	S2MA [22]	A2dele [27]	JL-DCF [13]	PGAR [4]	DANet [38]	cmMS [20]	BiANet [36]	D3Net [10]	ASIFNet [19]	CDINet Ours
		2019	2019	2019	2019	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2021	-
		PR	TIP	CVPR	ICCV	ACM MM	CVPR	CVPR	CVPR	CVPR	ECCV	ECCV	ECCV	TIP	TNNLS	TCyb	-
NLPR	$F_\beta \uparrow$	.8149	.8631	.8675	.8749	.8976	.8986	.9017	.8815	.8915	.9153	.9013	.9031	.8764	.8969	.8907	**.9162**
	$S_\alpha \uparrow$	.8557	.8861	.8884	.8892	.9129	.9141	.9155	.8979	.9097	**.9297**	.9152	.9176	.9000	.9117	.9079	.9273
	$MAE \downarrow$	.0591	.0410	.0359	.0339	.0290	.0259	.0298	.0285	.0295	.0245	.0283	.0277	.0325	.0296	.0295	**.0240**
NJUD	$F_\beta \uparrow$	.8526	.8741	.7661	.8883	.8982	.9000	.8888	.8733	.9042	.9068	.8927	.9034	.9121	.8996	.8886	**.9215**
	$S_\alpha \uparrow$	.8588	.8785	.7984	.8804	.8992	.9002	.8943	.8704	.9022	.9089	.8971	.9051	.9119	.9002	.8902	**.9188**
	$MAE \downarrow$	.0789	.0605	.0794	.0521	.0467	.0422	.0532	.0510	.0413	.0422	.0463	.0432	.0399	.0465	.0472	**.0354**
DUT	$F_\beta \uparrow$	.7671	.7903	.7180	.8975	.9263	.9242	.8997	.8923	.8612	.9171	.8954	.9090	.8156	.7855	.8245	**.9372**
	$S_\alpha \uparrow$	.7913	.8083	.7490	.8879	.9159	.9157	.9031	.8864	.8758	.9136	.8894	.9070	.8368	.8152	.8396	**.9274**
	$MAE \downarrow$	.1126	.0926	.0955	.0477	.0362	.0340	.0440	.0426	.0556	.0372	.0465	.0405	.0745	.0848	.0724	**.0302**
STEREO	$F_\beta \uparrow$	.8425	.8705	.8601	.8861	.8987	.8903	.8158	.8864	.8740	.9008	.8199	.8971	.8844	.8495	.8800	**.9033**
	$S_\alpha \uparrow$	.8559	.8775	.8714	.8858	.9004	.8920	.8424	.8868	.8855	.9054	.8410	.8999	.8882	.8687	.8820	**.9055**
	$MAE \downarrow$	.0796	.0591	.0537	.0474	.0428	.0449	.0746	.0431	.0509	.0422	.0712	.0429	.0497	.0578	.0485	**.0410**
LFSD	$F_\beta \uparrow$	-	-	.8214	.8523	.8555	.8626	.8310	.8280	.8217	.8390	.8417	.8623	.7287	.8062	.8602	**.8746**
	$S_\alpha \uparrow$	-	-	.8199	.8393	.8498	.8495	.8292	.8258	.8171	.8444	.8375	.8491	.7422	.8167	.8520	**.8703**
	$MAE \downarrow$	-	-	.0953	.0830	.0809	.0751	.1018	.0839	.1031	.0818	.1031	.0792	.1340	.1023	.0809	**.0631**

Experiments

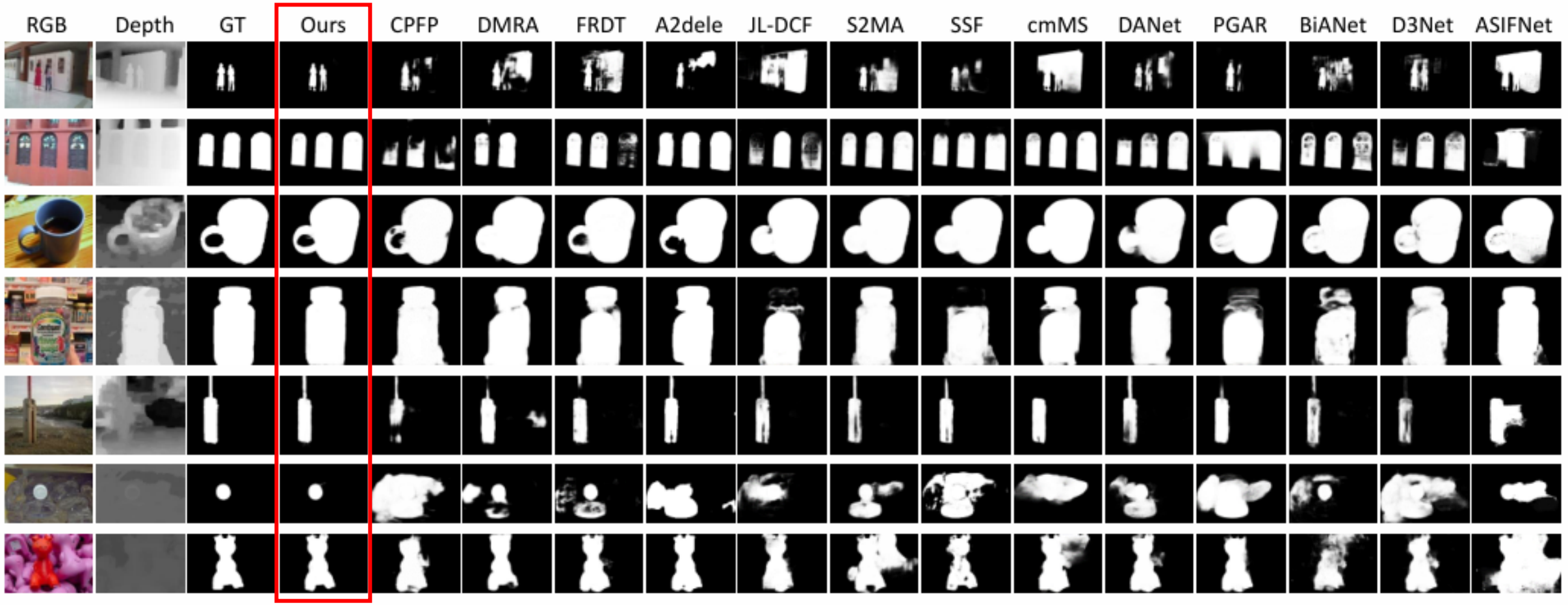


Figure 3: Visual comparisons with other state-of-the-art RGB-D methods in some representative scenes.

Experiments

Table 2: Ablation analyses of different components on the NLPR and DUT datasets.

models	NLPR			DUT		
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$
CDINet	**.9162**	**.9273**	**.0240**	**.9372**	**.9274**	**.0302**
w/o RDE	.9153	.9261	.0251	.9327	.9226	.0338
w/o DSE	.9062	.9219	.0253	.9222	.9184	.0369
w/o DDR	.9154	.9258	.0248	.9296	.9238	.0334

Table 3: The effectiveness analyses of discrepant interaction structure on the NLPR and DUT datasets.

Number	NLPR			DUT		
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$
No.1	**.9162**	.9273	**.0240**	**.9372**	**.9274**	**.0302**
No.2	.9153	.9261	.0251	.9295	.9217	.0345
No.3	.9160	**.9298**	.0242	.9328	.9246	.0327

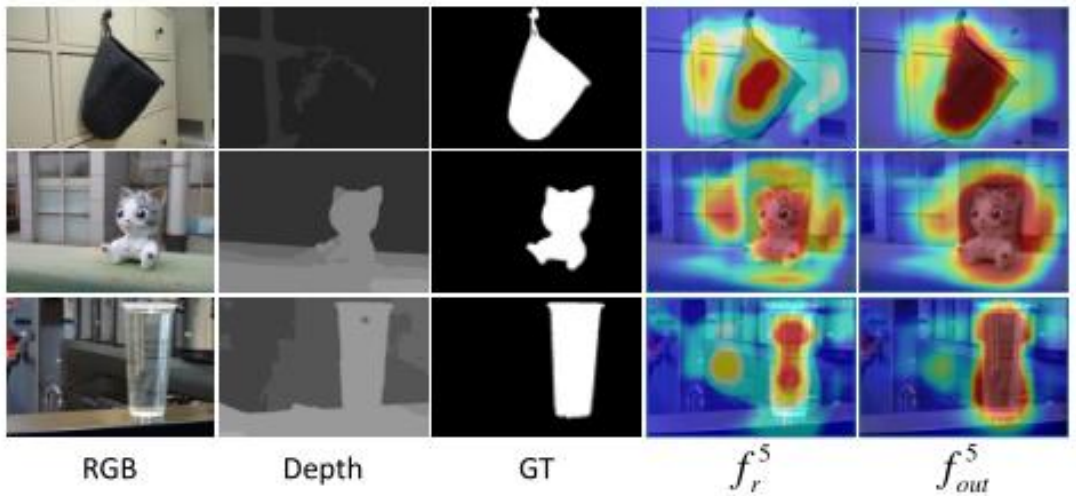


Figure 4: Feature visualization of the DSE module in the last layer of backbone.

Conclusion

- In this paper, we explore a novel cross-modality interaction mode and propose a cross-modality discrepant interaction network, which **explicitly models the dependence of two modalities in different convolutional layers**.
- To this end, two components (i.e., **RDE module and DSE module**) are designed to achieve differentiated cross-modality guidance. Furthermore, we also put forward a **DDR structure**, which generates a semantic block by leveraging multiple high-level features to upgrade the skip connection.
- The comprehensive experiments demonstrate that our network achieves competitive performance against state-of-the-art methods on five benchmark datasets, and our inference speed reaches the real-time level (i.e., **42 FPS**).

Highly Cited Paper

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

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

***IEEE Transactions on Image Processing*, 2021**

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

Motivations

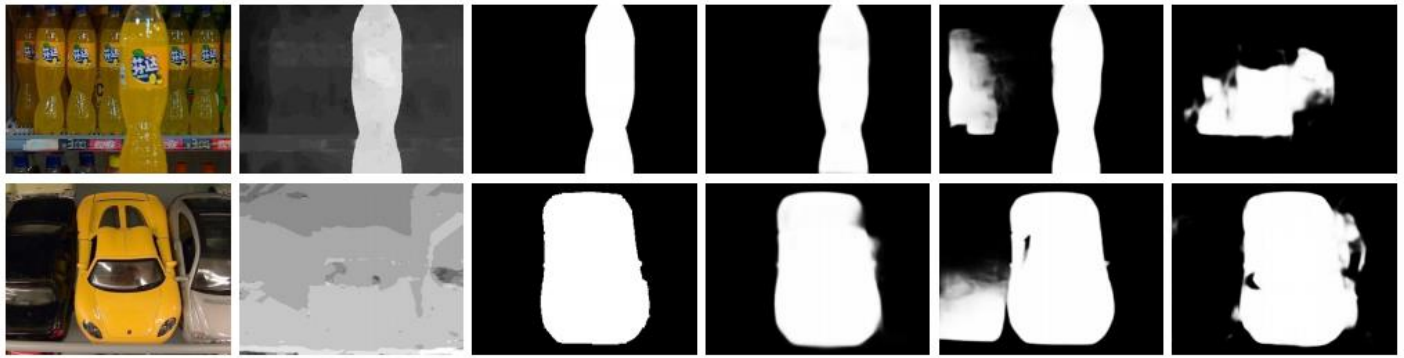


Fig. 1. Sample results of our method compared with others. RGB-D methods are marked in **boldface**. (a) RGB image; (b) Depth map; (c) Ground truth; (d) **Ours**; (e) BASNet [14]; (f) **CPFP** [33].

- 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

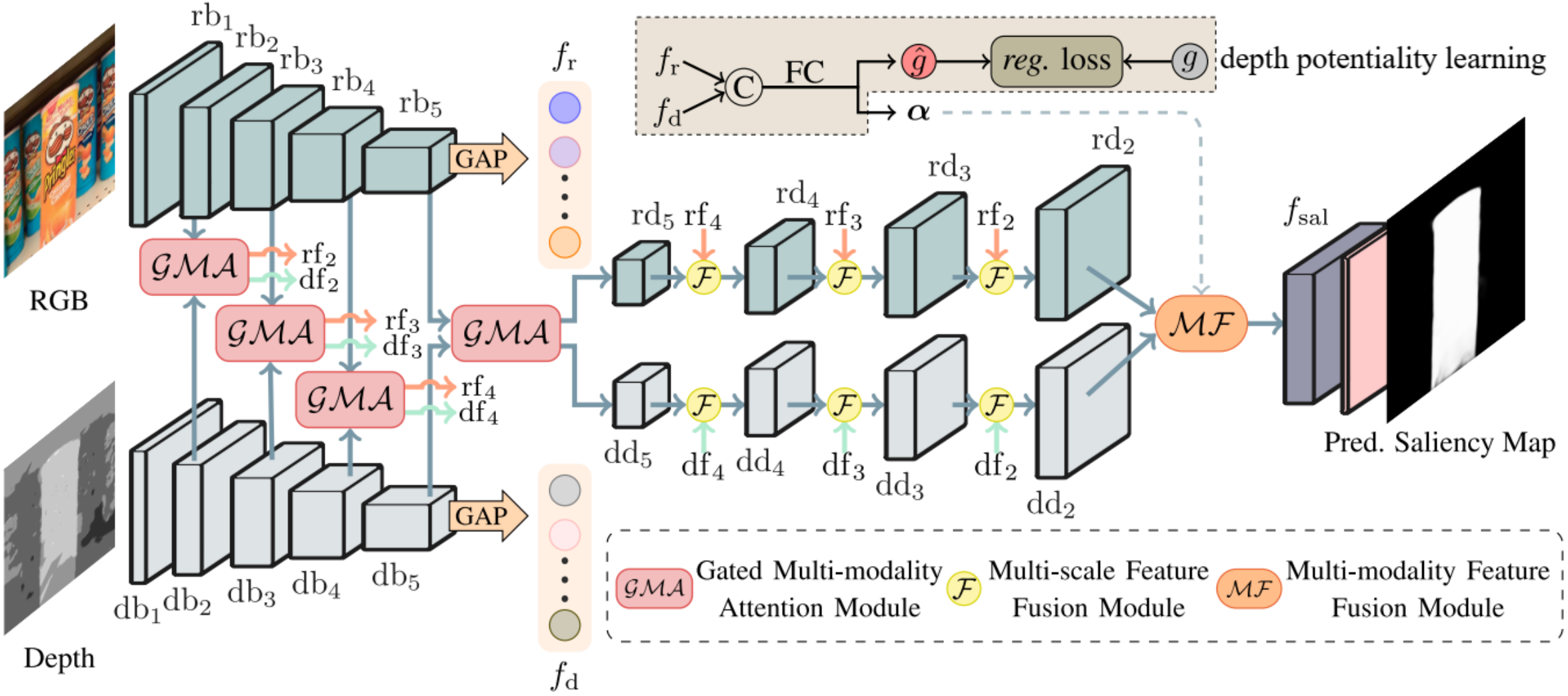
a) **For the first time, we address the unreliable depth map in the RGB-D SOD network in an end-to-end formulation**, and propose the DPANet by incorporating the depth potentiality perception into the cross-modality integration pipeline.

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.

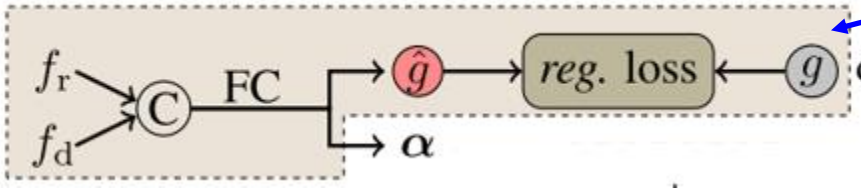
d) Without any pre-processing or post-processing techniques, the proposed network **outperforms 16 state-of-the-art methods on 8 RGB-D SOD datasets** in quantitative and qualitative evaluations.

Our Method


RGB
Depth
rb1 rb2 rb3 rb4 rb5
db1 db2 db3 db4 db5
GAP
GAP
GMA
GMA
GMA
GMA
rf2 df2
rf3 df3
rf4 df4
fr
fd
fr
fd
C
FC
ĝ
reg. loss
g
depth potentiality learning
α
rd5 rf4
rd4 rf3
rd3 rf2
rd2
dd5 df4
dd4 df3
dd3 df2
dd2
F
MF
fsal
Pred. Saliency Map
GMA Gated Multi-modality Attention Module
F Multi-scale Feature Fusion Module
MF Multi-modality Feature Fusion Module

Depth Potentiality Perception

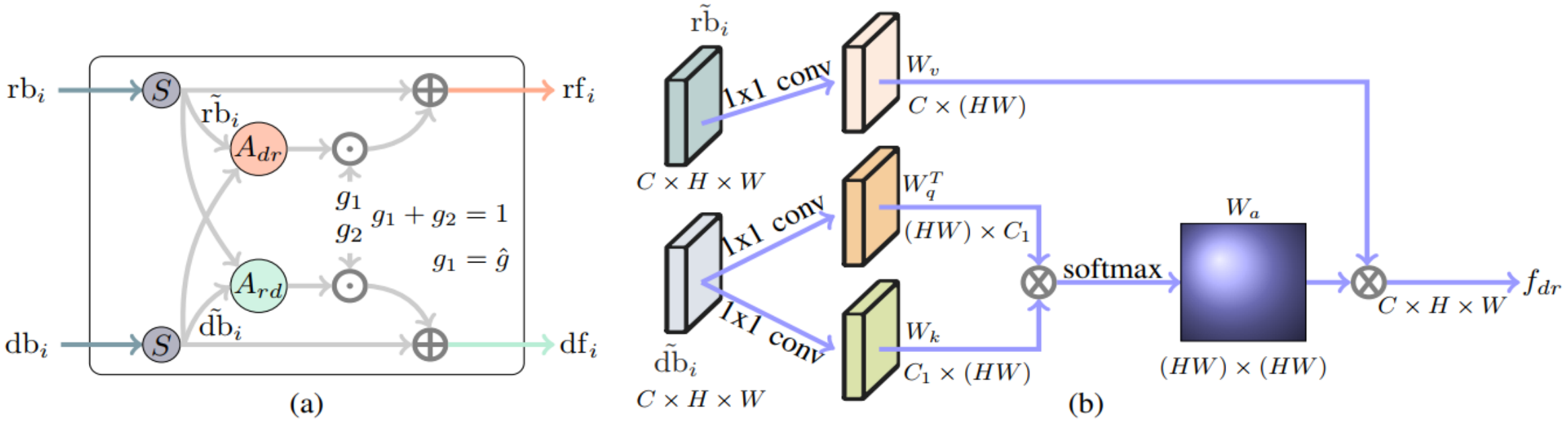
- Most previous works generally integrate the multi-modal features from RGB and corresponding depth information indiscriminately. However, **there exist some contaminations when depth maps are unreliable**.
- Since we do not hold any labels for depth map quality assessment, **we model the depth potentiality 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**.



depth potentiality learning

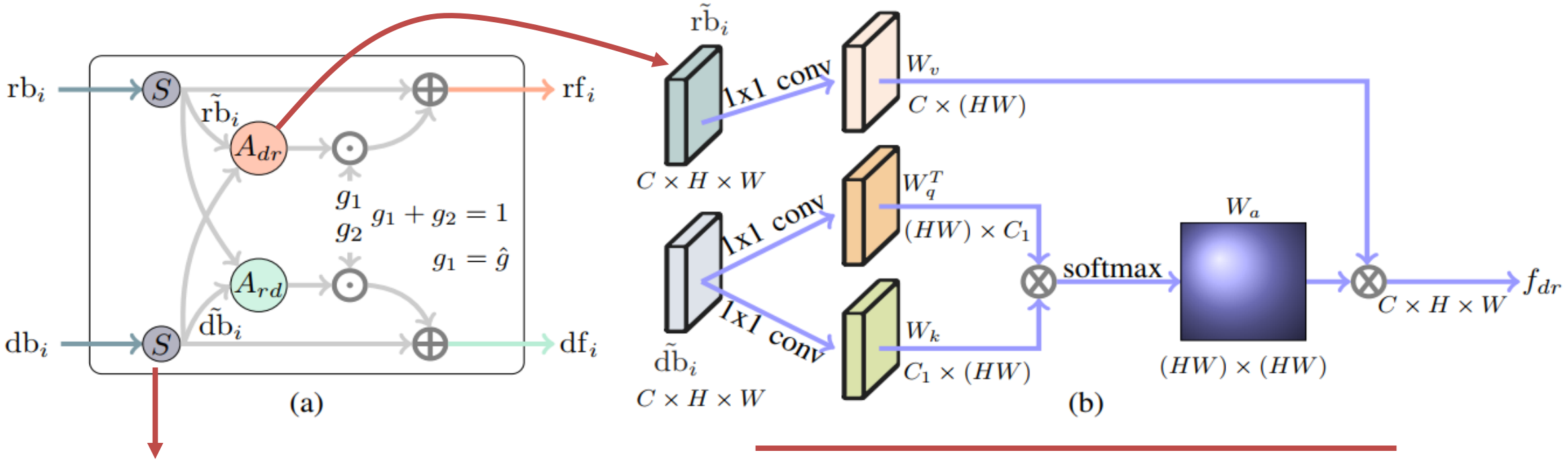
$$D(\tilde{I}, G) = \frac{(1+\gamma) \cdot D_{iou} \cdot D_{cov}}{D_{iou} + \gamma \cdot D_{cov}}$$

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

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

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.

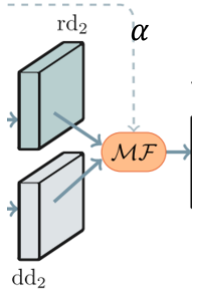
$$f_1 = \delta\left(up\left(conv_3(\mathrm{rd}_5)\right) \odot \mathrm{rf}_4\right)$$

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

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

- 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.



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

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

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

$\boldsymbol{\alpha}$ 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

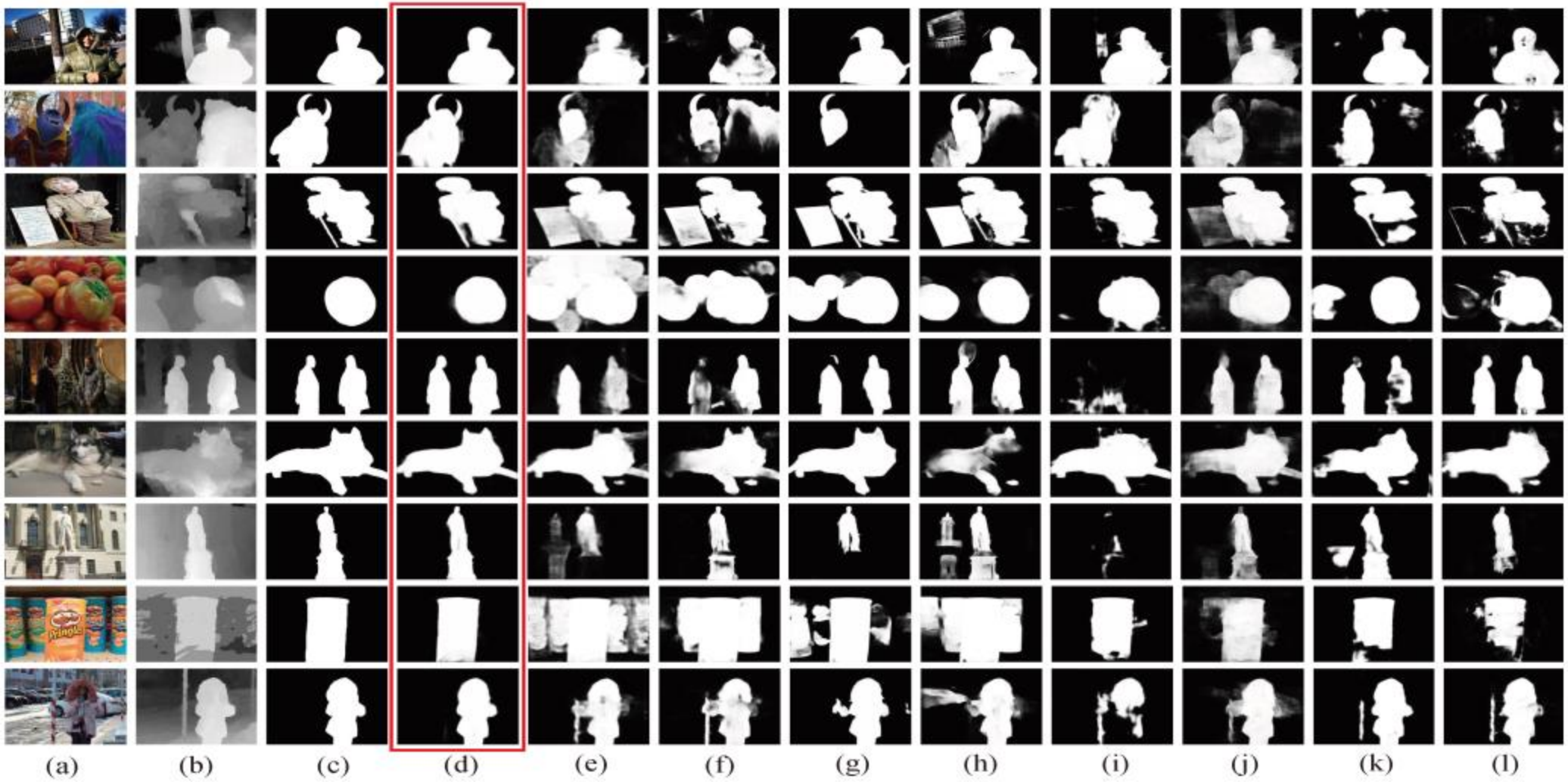


Fig. 4. Qualitative comparison of the proposed approach with some state-of-the-art RGB and RGB-D SOD methods, in which our results are highlighted by a red box. (a) RGB image. (b) Depth map. (c) GT. (d) DPANet. (e) PiCAR. (f) PoolNet. (g) BASNet. (h) EGNet. (i) CPFP. (j) PDNet. (k) DMRA. (l) AF-Net.

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.

Our work in RGB-D SOD

1. Runmin Cong, Jianjun Lei, et.al, Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion, IEEE Signal Processing Letters (SPL), vol. 23, no. 6, pp. 819-823, 2016.
2. Runmin Cong, Jianjun Lei, et.al, Going from RGB to RGBD saliency: A depth-guided transformation model, IEEE Transactions on Cybernetics (TCyb), vol. 50, no. 8, pp. 3627-3639, 2020. Highly Cited Paper
3. Chongyi Li, Runmin Cong*, et.al, ASIF-Net: Attention steered interweave fusion network for RGBD salient object detection, IEEE Transactions on Cybernetics (TCyb), vol. 50, no. 1, pp. 88-100, 2021. Highly Cited Paper
4. Zuyao Chen‡, Runmin Cong‡, et.al, DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection, IEEE Transactions on Image Processing (TIP), vol. 30, pp. 7012-7024, 2021. Highly Cited Paper
5. Chen Zhang, Runmin Cong*, et.al, Cross-modality discrepant interaction network for RGB-D salient object detection, ACM International Conference on Multimedia (ACM MM), pp. 2094-2102, 2021.
6. Chongyi Li, Runmin Cong*, et.al, RGB-D salient object detection with cross-modality modulation and selection, European Conference on Computer Vision (ECCV), pp. 225-241, 2020.
7. Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, et.al, Dynamic selective network for RGB-D salient object detection, IEEE Transactions on Image Processing (TIP), vol. 30, pp. 9179-9192, 2021.
8. Yudong Mao, Qiuping Jiang, Runmin Cong, et.al, Cross-modality fusion and progressive integration network for saliency prediction on stereoscopic 3D images, IEEE Transactions on Multimedia (TMM), 2021.

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 FOR WATCHING

Runmin Cong (丛润民)

Beijing Jiaotong University