Depth-oriented Enhancement and Application

Runmin Cong (丛润民)

第十六届图像图形技术与应用学术会议（IGTA2021）

2021-06-06@北京理工大学

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13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，国家级人才计划青年项目入选者1人，教育部新世纪优秀人才2人，中国科协青年人才托举工程入选者1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、北京市科协青年人才托举工程入选者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

多媒体与视觉处理实验室

Multimedia & Vision Processing Lab

多媒体与视觉处理实验室隶属于北京交通大学数字媒体信息处理研究中心，于2019年7月成立，负责人为丛润民副教授，主要从事多媒体理解、视觉处理等相关研究工作。

丛润民博士2019年从天津大学毕业后加入北京交通大学，先后主持了国家自然科学基金、北京市科技新星计划、中央高校基本科研业务费、中国博士后基金（特别资助、面上资助）等7项科研项目。在IEEE TIP、TCyb、TII、TMM、TCSVT、TGRS、NeurIPS、CVPR、ECCV、AAAI、IJCAI、ACM MM等国际学术期刊及会议上发表论文40余篇，其中CCF-A/IEEE Trans论文26篇，ESI高被引论文6篇，谷歌引用1400+；授权国家发明专利10项。担任SCI期刊Signal, Image and Video Processing 编委。

丛 润 民 副教授
中 国 科 协 青 托
人 社 部 香 江 学 者
北 京 市 科 技 新 星
北 京 市 科 协 青 托
CSIG优 博 获 得 者

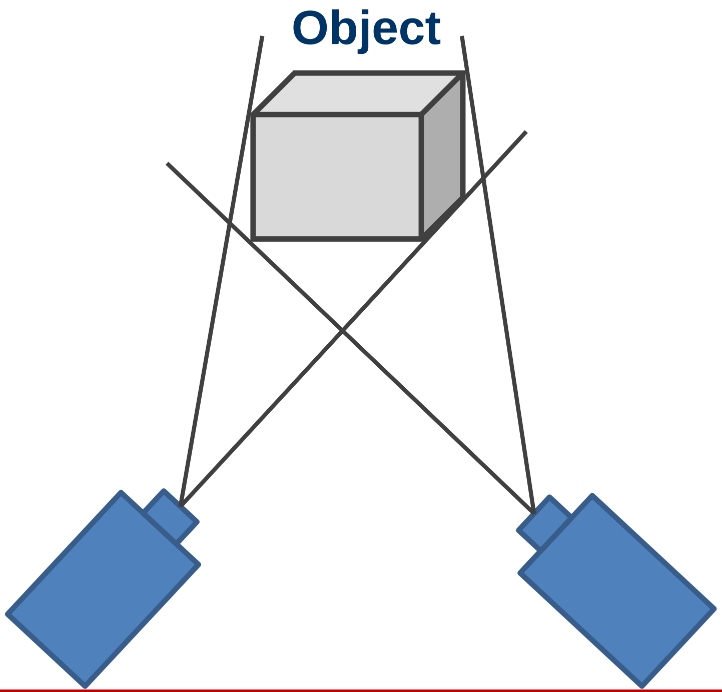
研究方向

- 注意力感知与显著性计算
- 遥感影像智能解译与分析
- 开放环境下视觉内容增强
- 多媒体大数据理解与应用

Outline

- Introduction
- Depth Map Super-Resolution
  - ➢ Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline, CVPR 2021
- RDG-D Salient Object Detection
  - ➢ DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection, TIP 2021
- Future Work

Introduction

Depth Cameras

Kinect

ORBBEC

photographing

Driverless

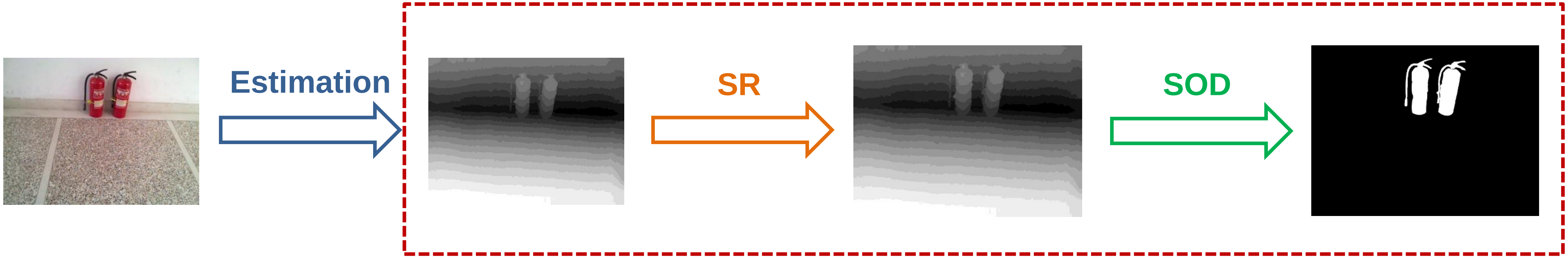
CVPR 2022/ICCV 2011大会主席、香港科技大学权龙教授说过“真正意义上的计算机视觉要超越识别，感知三维场景。我们活在三维空间里，要做到交互和感知，就必须将世界恢复到三维。”

Depth-oriented Perception: From Generation to Application

Generation — Monocular/Stereo Depth Estimation

Optimization — Depth Completion, Depth Super-resolution, Depth Hole Filling

Application — Segmentation, Detection, Recognition



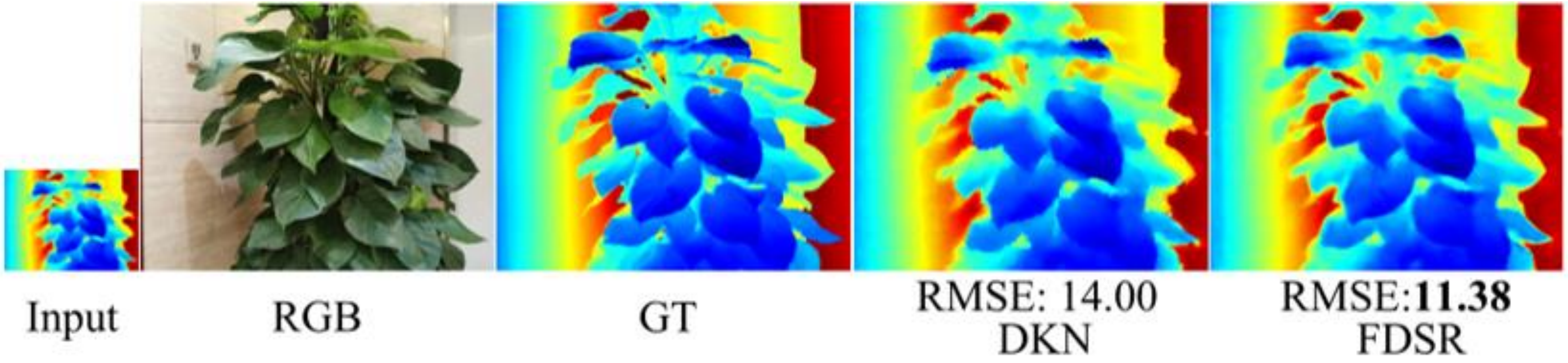
Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, Yao Zhao

***IEEE/CVF Conference on Computer Vision and Pattern Recongnition*, 2021**

https://github.com/lingzhi96/RGB-D-D-Dataset

Motivations



- The resolution of depth maps cannot match the resolution of RGB images. Depth map super-resolution (SR) is an effective solution.
- Limited by the lack of real-world paired LR and HR depth maps, most existing depth map SR methods use down-sampling to obtain paired training samples.
- The sharp boundaries and elaborate details in the depth map SR are hard to recover especially when the scaling factor is large. How to design a fast and accurate depth map SR model to generate HR depth maps.

Contributions

a) We build the first and large-scale depth map SR benchmark dataset named **RGB-D-D dataset**, towards the real scenes and real correspondences. This dataset bridges the gap between theoretical research and real-world applications, and also flourishes the depth related tasks in terms of benchmark dataset.

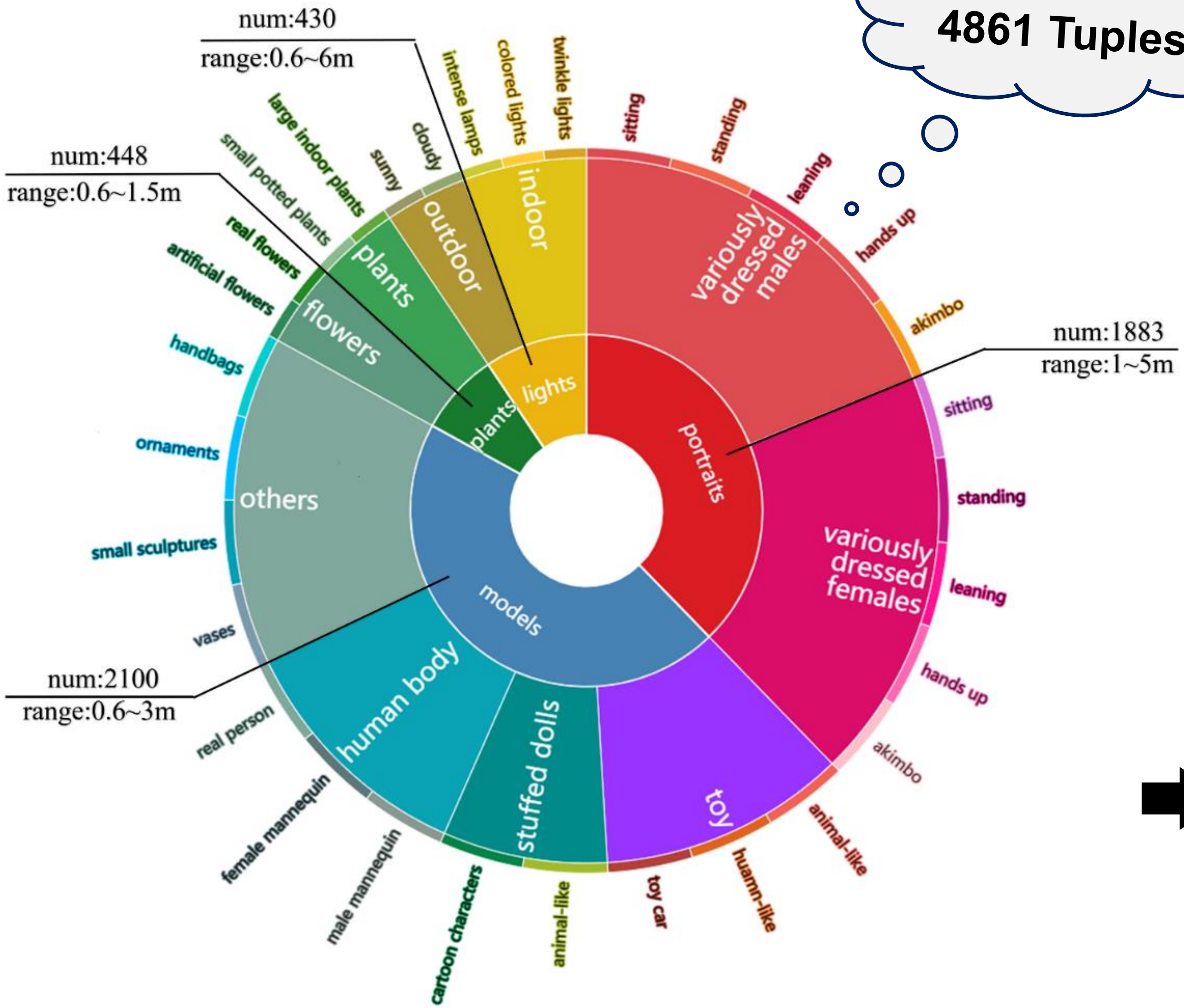
b) We design a fast depth map super-resolution (**FDSR**) baseline, in which a high-frequency guided multi-scale structure is introduced to provide the frequency guidance and exploit the contextual information. Such decomposition strategy can improve the **efficiency** while retaining the reconstruction **performance**.

c) Our network achieves the superior performance on the public datasets and our RGB-D-D benchmark dataset in terms of the speed and accuracy. Moreover, for the real-world depth map SR task, our algorithm can generate more accurate results with clearer boundaries and to some extent correct the value errors.

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

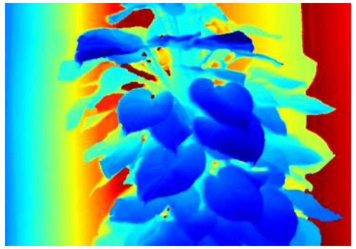
https://github.com/lingzhi96/RGB-D-D-Dataset

RGB-D-D Dataset

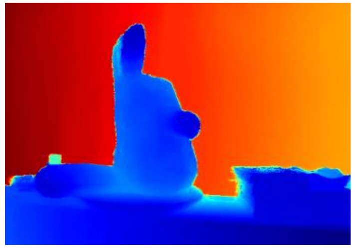


Total:
4861 Tuples

potraits



plants

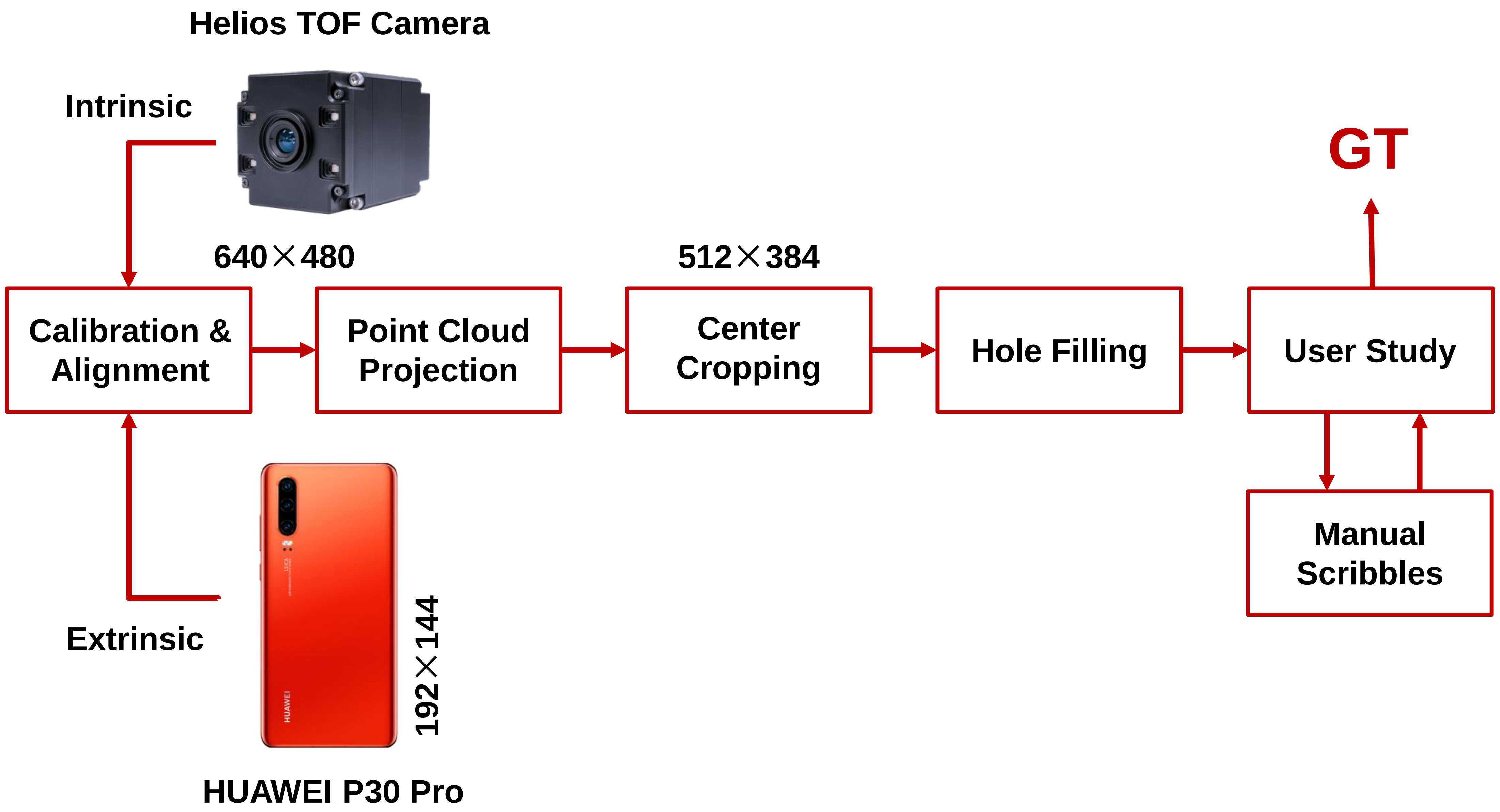


models

lights

- Bokeh Rendering
- Optimize the edge of objects
- Low-quality depth map
- Effect of complex illumination
- ……

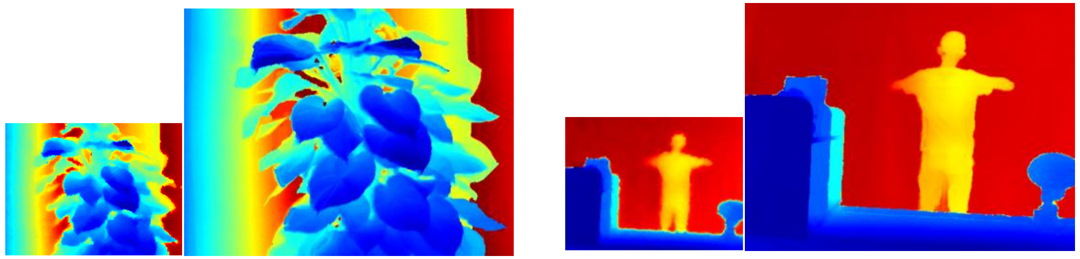
Dataset Processing



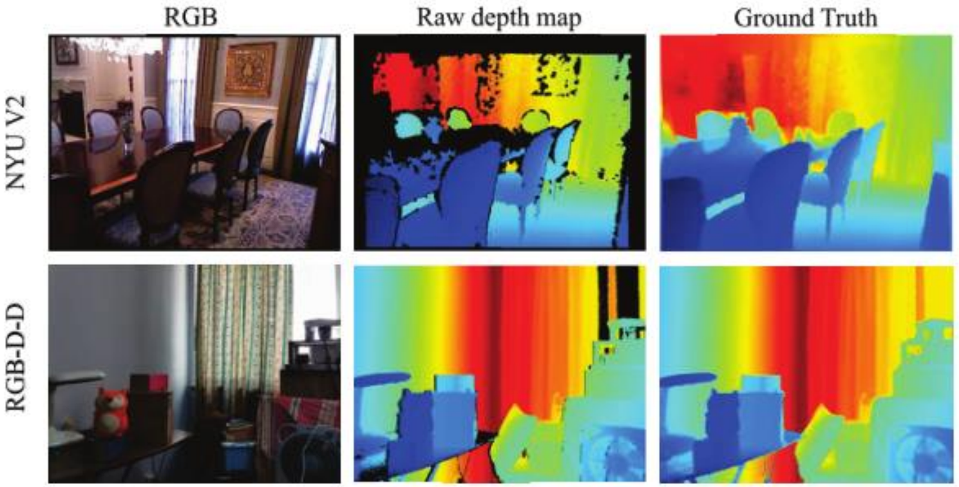
Dataset Statistic

- Real Scenes

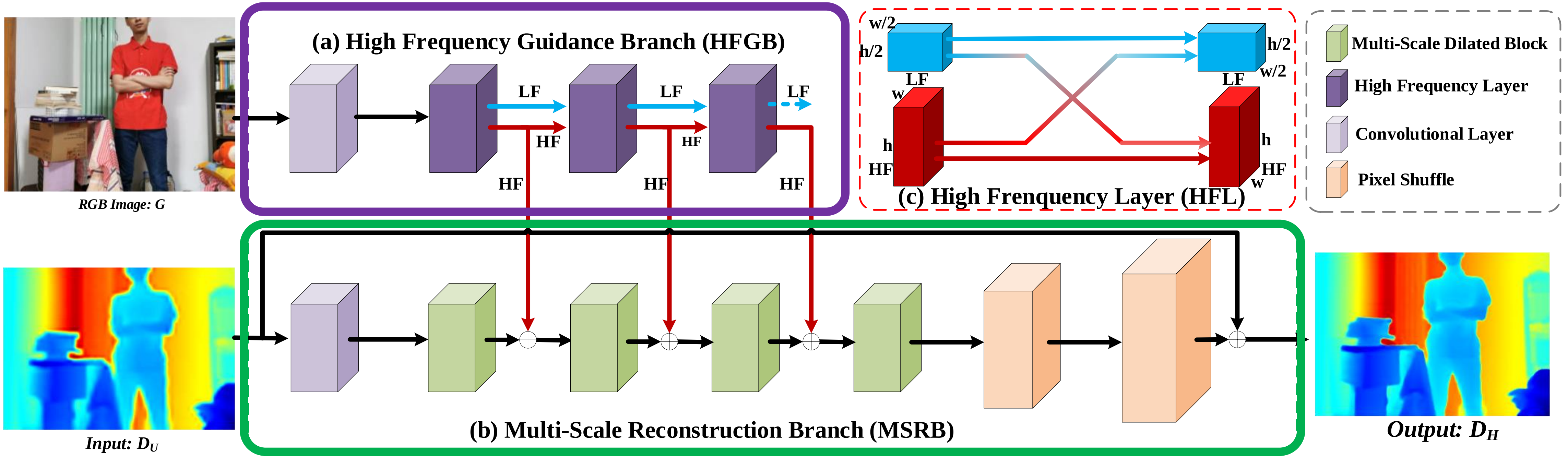
- Real Correspondences



- High Quality

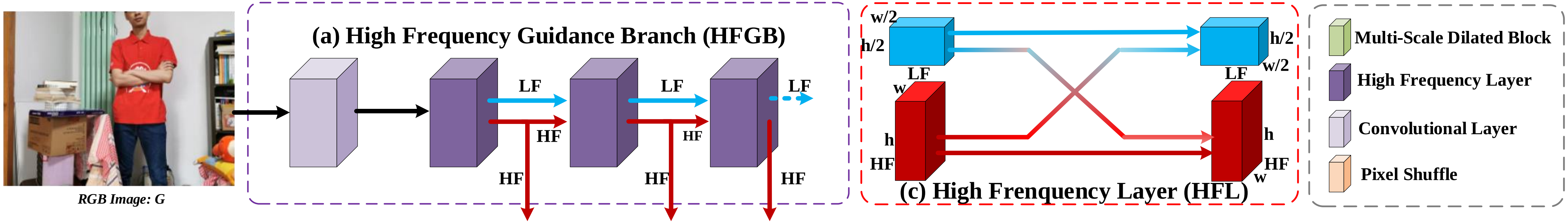


Our Method



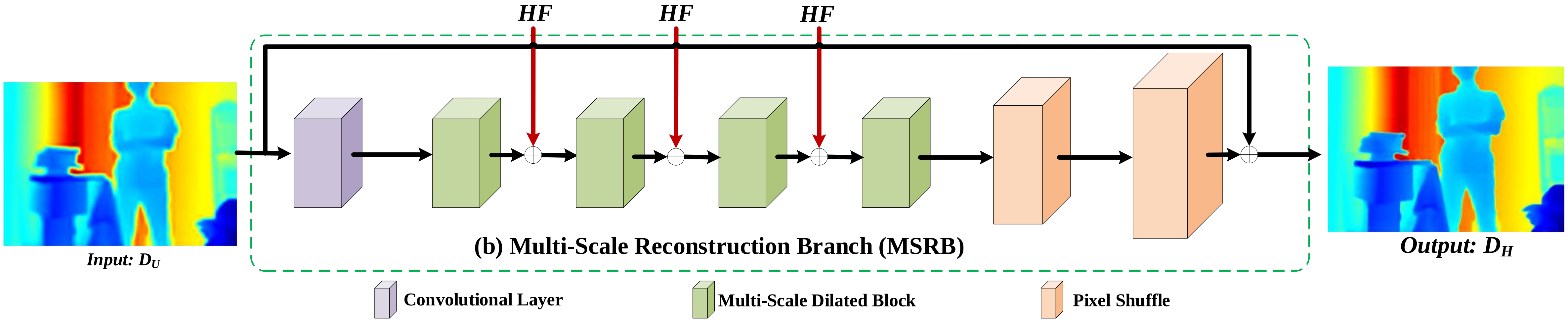
Our framework progressively equip with four **multi-scale reconstruction blocks** to exploit the contextual information under different receptive fields in MSRB, meanwhile, the **high-frequency guidance** extracted from the HFGB is integrated with the multiscale contextual information to enhance the ability of detail recovery for depth map SR.

High-Frequency Guidance Branch



- A direct high-frequency decomposition method is designed, where the octave convolution is utilized to decompose the RGB features into high- and low-frequency components.
- The high-frequency components are effectively used to guide depth map SR. Such design focuses on the useful high-frequency detail information to improve the performance, while it reduces the computation complexity due to the low-frequency components are not used in the MSRB.

Multi-Scale Reconstruction Branch



(b) Multi-Scale Reconstruction Branch (MSRB)

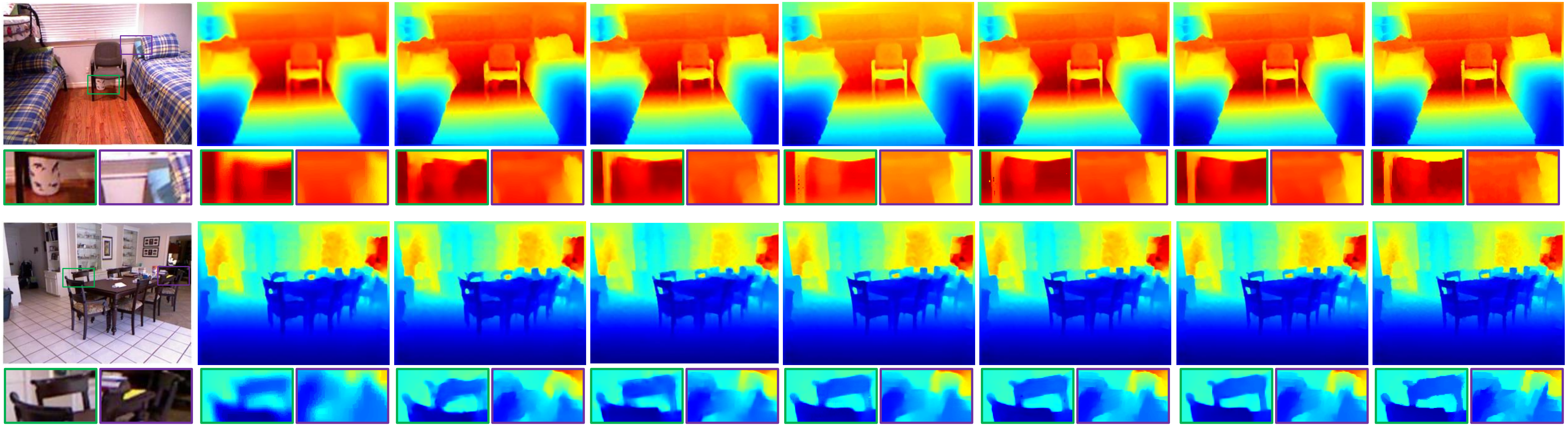
- This branch aims to progressively recover HR depth map through utilizing mulit-scale contextual information. We first use one convolution layer to initial feature extraction. Then, to exploit the contextual information under different receptive fields, we combine dilated convolutions with different dilated rates to form a multi-scale dilated block (MSDB), and one convolution layer is used to integrate the concatenated features.
- As for feature combination, three levels of high-frequency features extracted by HLFs are fused with different MSDBs respectively in the early stage of MSRB.

Experiments

- Benchmark Datasets: NYU v2 (1449 RGB-D images), RGB-D-D (4861 RGB-D images).
- We sample 1000 RGB-D image pairs from the NYU v2 dataset for training and the rest 449 image pairs for testing. As for RGB-D-D dataset, we randomly split 1586 portraits, 380 plants, 249 models for training and 297 portraits, 68 plants, 40 models for testing.
- Evaluation Metrics:

RMSE:	pixel wise depth map SR accuracy
Depth Value Errors:	global confidence depth map SR accuracy
Depth Edge Errors:	edge area depth map SR accuracy

Experiments on NYU v2



	(a) RGB	(b) SDF	(c) SVLRM	(d) DJFR	(e) FDKN	(f) DKN	(g) FDSR	(h) GT
Running Time (s)		25 (CPU)	0.10 (GPU)	0.01 (GPU)	0.01 (GPU)	0.17 (GPU)	0.01 (GPU)	

Experiments on NYU v2

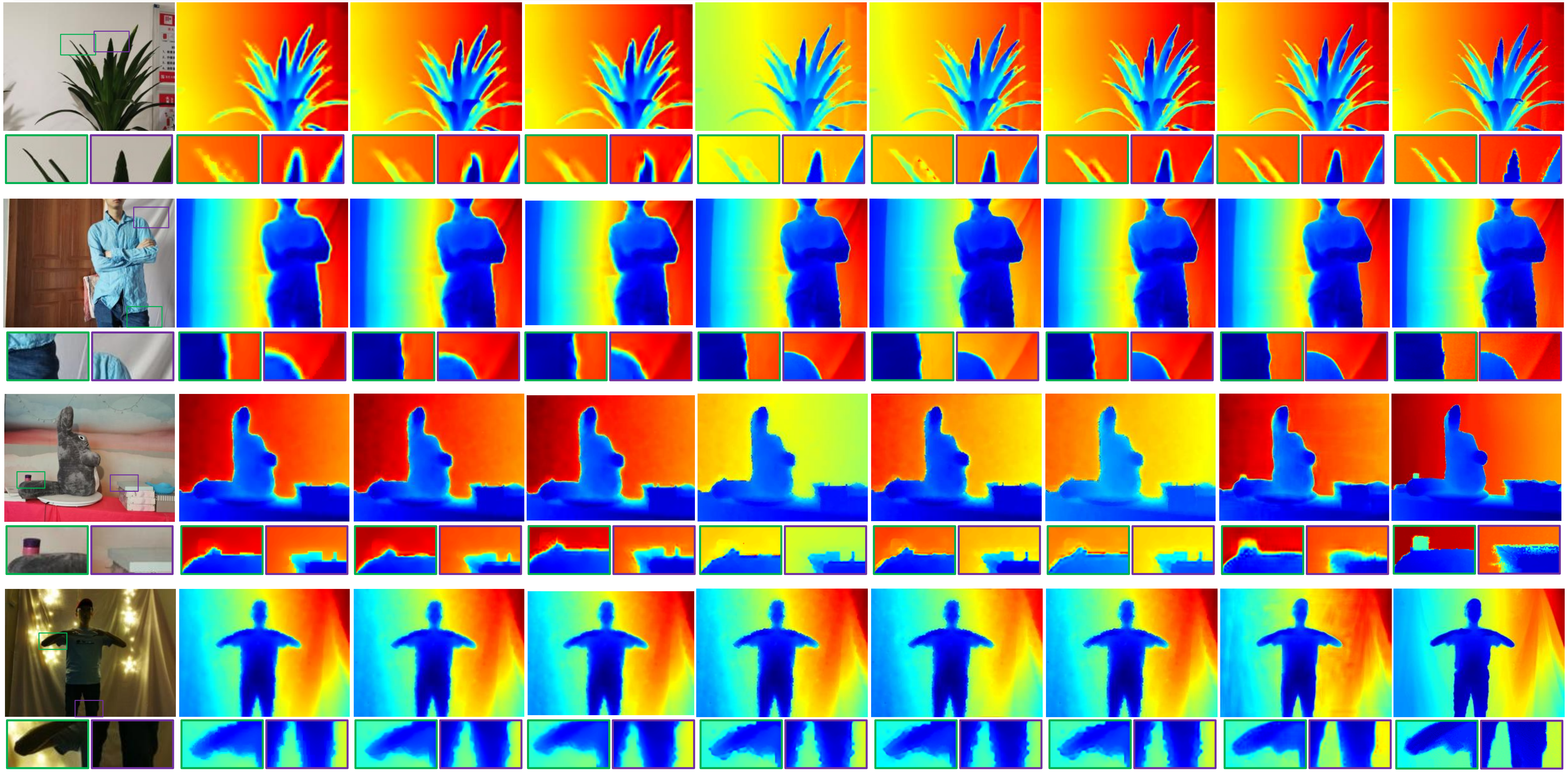
RMSE	Bicubic	MRF [7]	GF [12]	JBU [18]	TGV [8]	Park [31]	SDF [22]	FBS [4]	DMSG [14]	PAC [38]	DJF [22]	DJFR [23]	DKN [16]	FDKN [16]	FDSR
$\times 4$	8.16	7.84	7.32	4.07	4.98	5.21	5.27	4.29	3.02	2.39	3.54	3.38	1.62	1.86	**1.61**
$\times 8$	14.22	13.98	13.62	8.29	11.23	9.56	12.31	8.94	5.38	4.59	6.2	5.86	3.26	3.58	**3.18**
$\times 16$	22.32	22.2	22.03	13.35	28.13	18.1	19.24	14.59	9.17	8.09	10.21	10.11	6.51	6.96	**5.86**

Table 1. Comparisons with the state-of-the-art methods in terms of RMSE on NYU v2 [28]. The depth values are measured in centimeter.

Percentage	Value Errors (in 10 m)			Edge Errors		
	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$
SDF [22]	0.42	1.28	3.52	4.20	10.19	25.06
SVLRM [30]	1.08	2.56	5.76	6.04	24.28	49.26
DJF [22]	1.05	2.74	6.25	9.87	30.38	55.35
DJFR [23]	1.04	2.72	6.25	6.78	25.01	53.98
FDKN [16]	0.04	0.24	1.00	0.83	3.27	13.03
DKN [16]	0.05	0.20	1.10	0.95	2.95	13.78
FDSR	**0.04**	**0.18**	**0.69**	**0.78**	**2.60**	**9.44**

Table 2. Value errors and edge errors on NYU v2 [28].

Experiments on RGB-D-D



(a) RGB (b) SDF (c) SVLRM (d) DJFR (e) FDKN (f) DKN (g) FDSR (h) FDSR+ / FDSR++ (i) GT

Experiments on RGB-D-D

RMSE	SDF [22]	SVLRM [30]	DJF [22]	DJFR [23]	FDKN [16]	DKN [16]	FDSR	FDSR$^+$
$\times 4$	2.00	3.39	3.41	3.35	1.18	1.30	1.16	**1.11**
$\times 8$	3.23	5.59	5.57	5.57	1.91	1.96	1.82	**1.71**
$\times 16$	5.16	8.28	8.15	7.99	3.41	3.42	3.06	**3.01**

Table 3. Quantitative depth map SR results on RGB-D-D. FDSR$^+$ is trained in downsampling manner on RGB-D-D)

	SDF [22]	SVLRM [30]	DJF [22]	DJFR [23]	FDKN [16]	DKN [16]	FDSR	FDSR^{++}
RMSE	7.16	8.05	7.90	8.01	7.50	7.38	7.50	**5.49**
Value Errors	2.86	3.62	3.62	3.67	2.85	2.83	2.90	**1.71**
Edge Errors	52.78	51.87	50.56	52.28	51.73	51.90	51.89	**42.89**

Table 4. RMSE, value errors and edge errors of depth SR results. FDSR^{++} is trained on RGB-D-D in real-world training manner.

Percentage	Value Errors (in 3 m)			Edge Errors		
	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$
SDF [22]	0.33	0.90	2.37	3.22	8.74	20.71
SVLRM [30]	0.80	2.11	4.58	5.08	15.18	34.30
DJF [22]	0.82	2.19	4.89	5.65	17.07	35.32
DJFR [23]	0.79	2.15	4.78	5.26	15.66	34.54
FDKN [16]	0.11	0.28	0.94	1.39	3.41	11.73
DKN [16]	0.14	0.33	1.54	2.11	3.55	12.93
FDSR	0.10	0.26	0.76	1.38	3.09	12.47
FDSR$^+$	**0.09**	**0.21**	**0.67**	**1.15**	**2.79**	**11.68**

Table 5. Value errors and edge errors of depth SR results on RGB-D-D. FDSR$^+$ is trained in downsampling training manner.

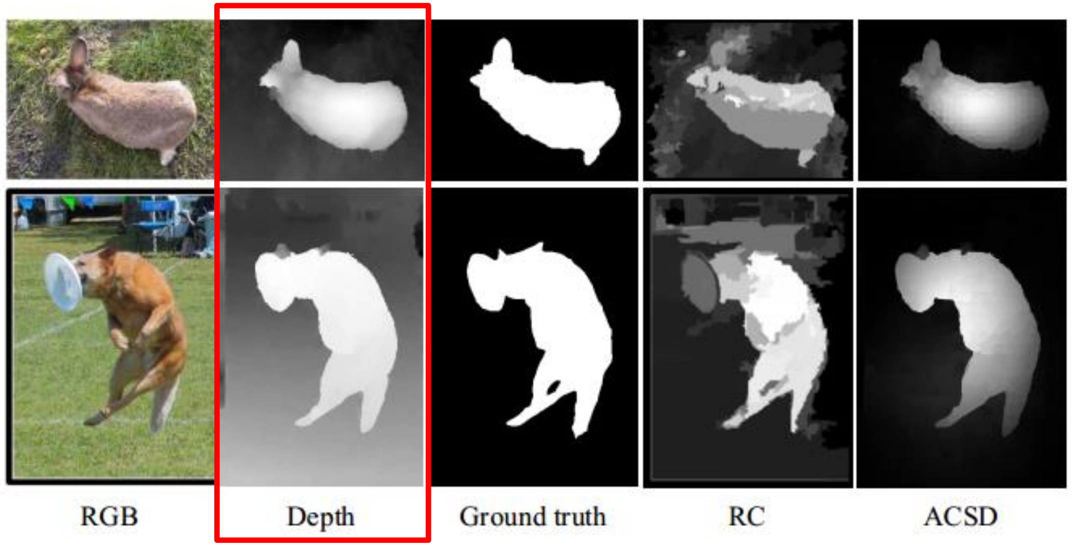
Methods	NYU v2 [28]			RGB-D-D		
	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$
w/o HFGB	2.02	3.90	7.58	1.16	1.88	3.47
w/o HFL	1.68	3.21	5.89	1.13	1.85	3.20
FDSR	**1.61**	**3.18**	**5.86**	**1.11**	**1.71**	**3.01**

Table 6. RMSE evaluation of HFL and HFGB.

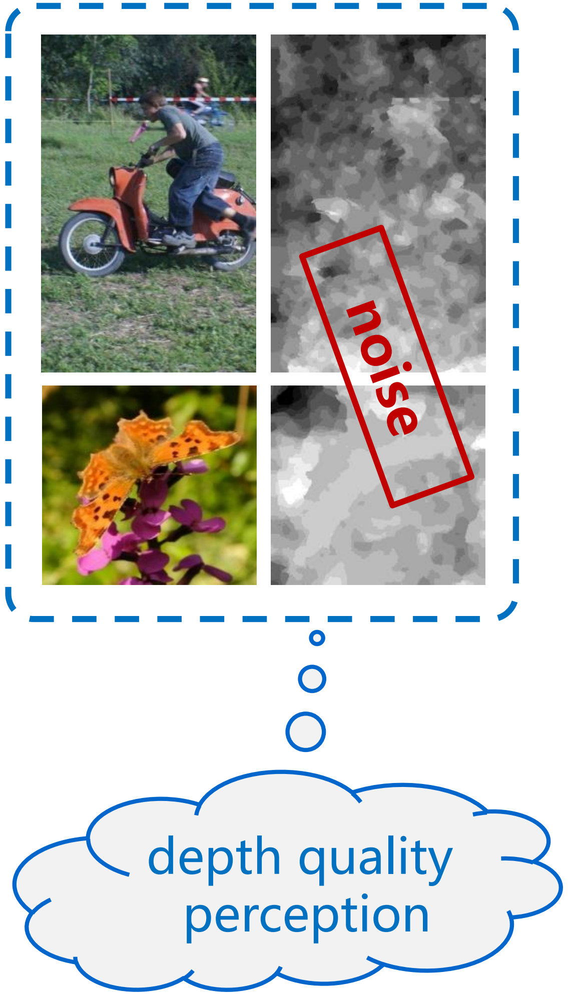
Conclusion

- We build the first benchmark dataset which satisfy both real scene and real correspondence. The dataset contains paired LR and HR depth maps in multiple scenarios, and contributes the completely new benchmark dataset for real-world depth map SR research.
- Furthermore, the “RGB-D-D” triples not only can complete the traditional depth-related tasks, such as depth estimation, depth completion, etc. but also have significant potential to promote the application of depth maps on portable intelligent electronics.
- We also provide a fast and accurate depth map SR baseline adaptively focusing on the high-frequency components of the guidance and suppress the low-frequency components and achieve the competitive performance on public datasets and our proposed dataset. What’s more, it has an ability to cope with the task of real-world depth map SR.

RGB-D Salient Object Detection



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



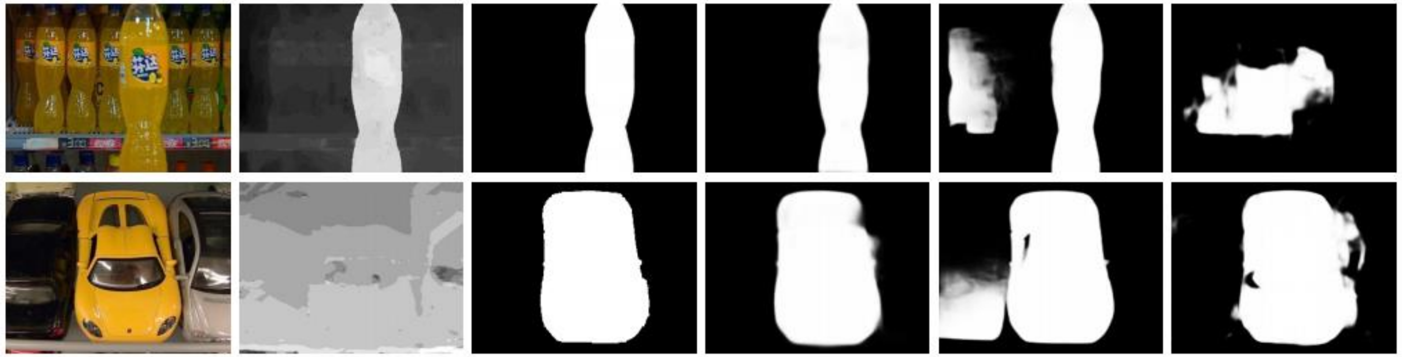
DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection

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

IEEE Transactions on Image Processing, 2021

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

Motivations



(a) (b) (c) (d) (e) (f)

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

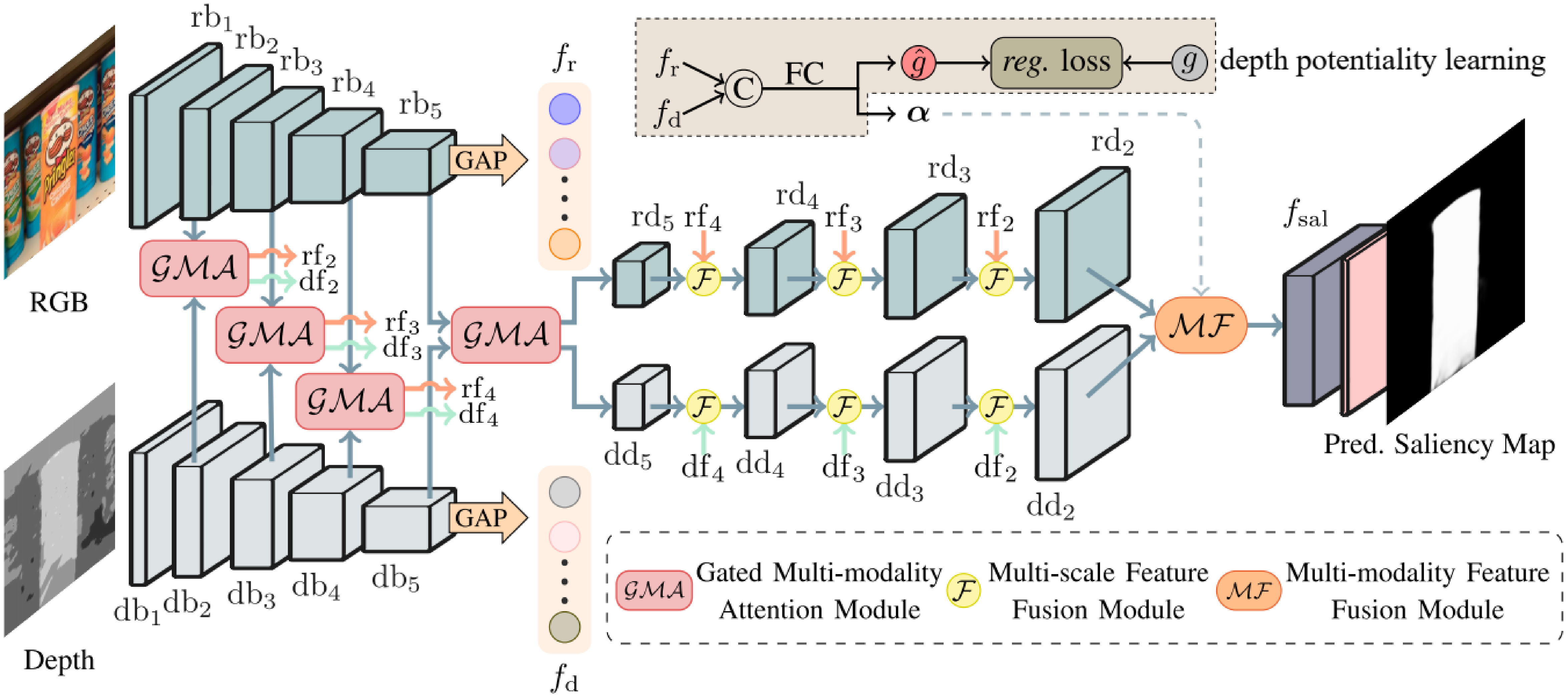
a) **For the first time, we address the unreliable depth map in the RGB-D SOD network in an end-to-end formulation**, and propose the DPANet by incorporating the depth potentiality perception into the cross-modality integration pipeline.

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.

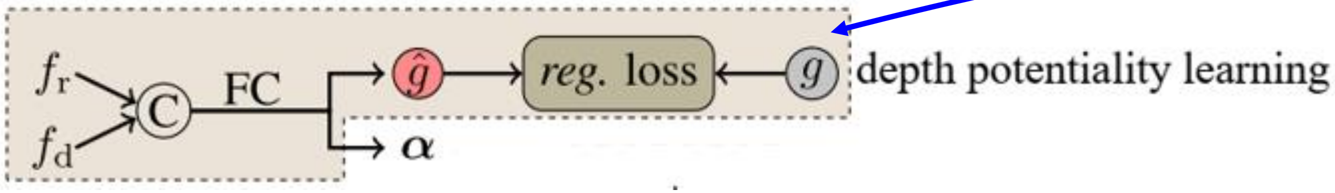
Our Method



rb1 rb2 rb3 rb4 rb5
GAP
f_r
RGB
GMA
rf2 df2
rf3 df3
rf4 df4
Depth
db1 db2 db3 db4 db5
f_d
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
Gated Multi-modality Attention Module
Multi-scale Feature Fusion Module
Multi-modality Feature Fusion Module

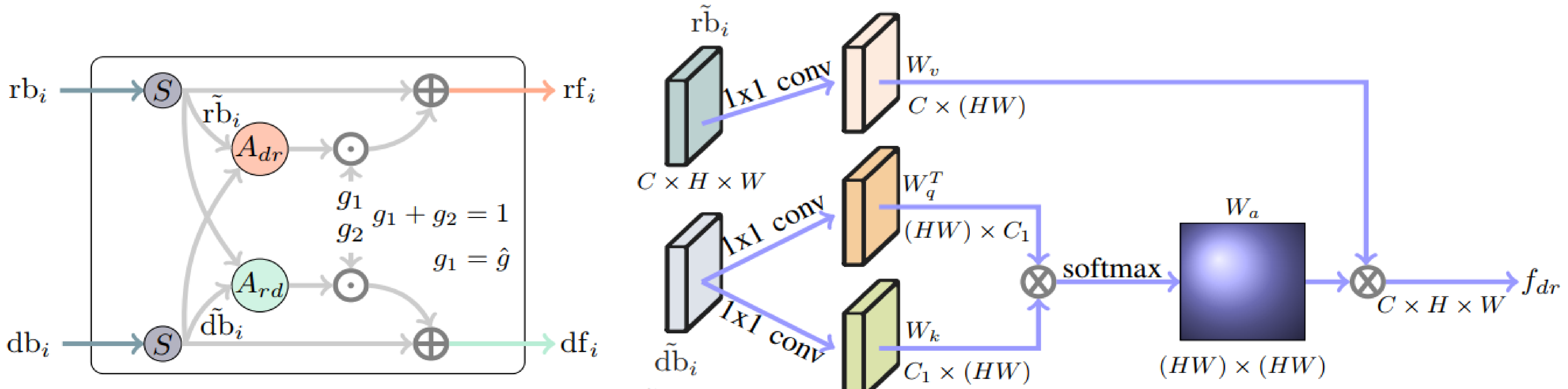
Depth Potentiality Perception

- 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.



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

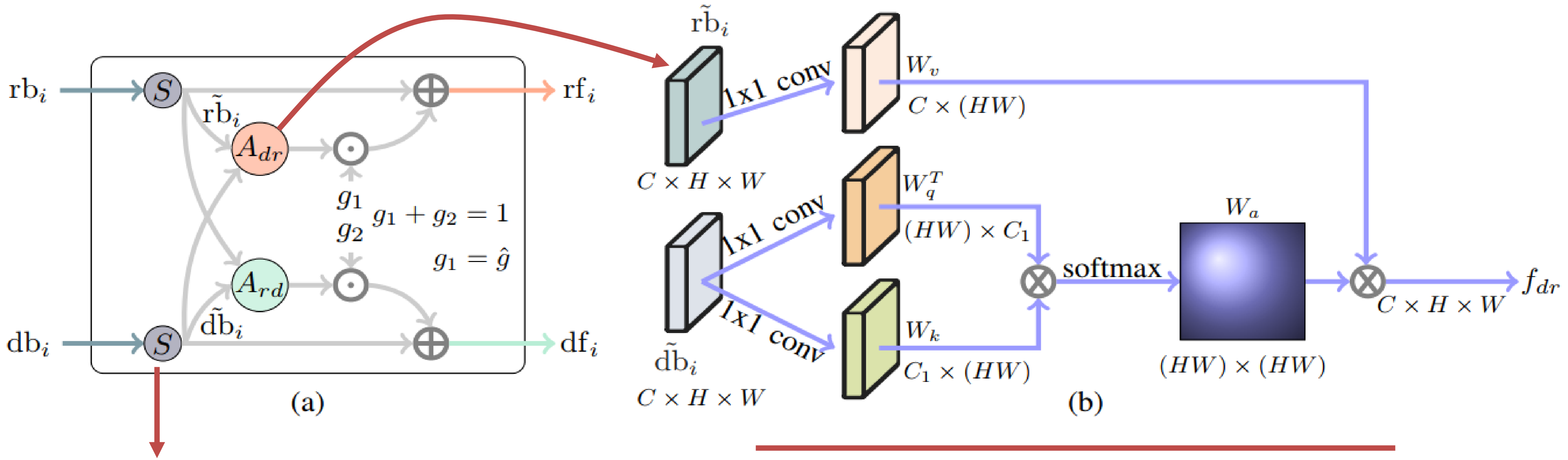
Gated Multi-modality Attention Module



(a) (b)

- 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

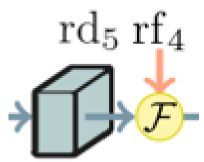
$$\mathrm{rf}_i = \widetilde{\mathrm{rb}}_i + g_1 \cdot f_{dr} \qquad g_1 = \hat{g}$$

$$\mathrm{df}_i = \widetilde{\mathrm{db}}_i + g_2 \cdot f_{rd} \qquad g_1 + g_2 = 1$$

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.



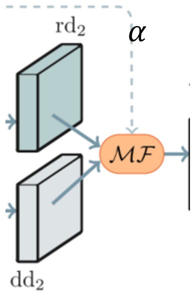
$$f_1 = \delta\big(up\big(conv_3(\mathrm{rd}_5)\big) \odot \mathrm{rf}_4\big)$$

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

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

• 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 :

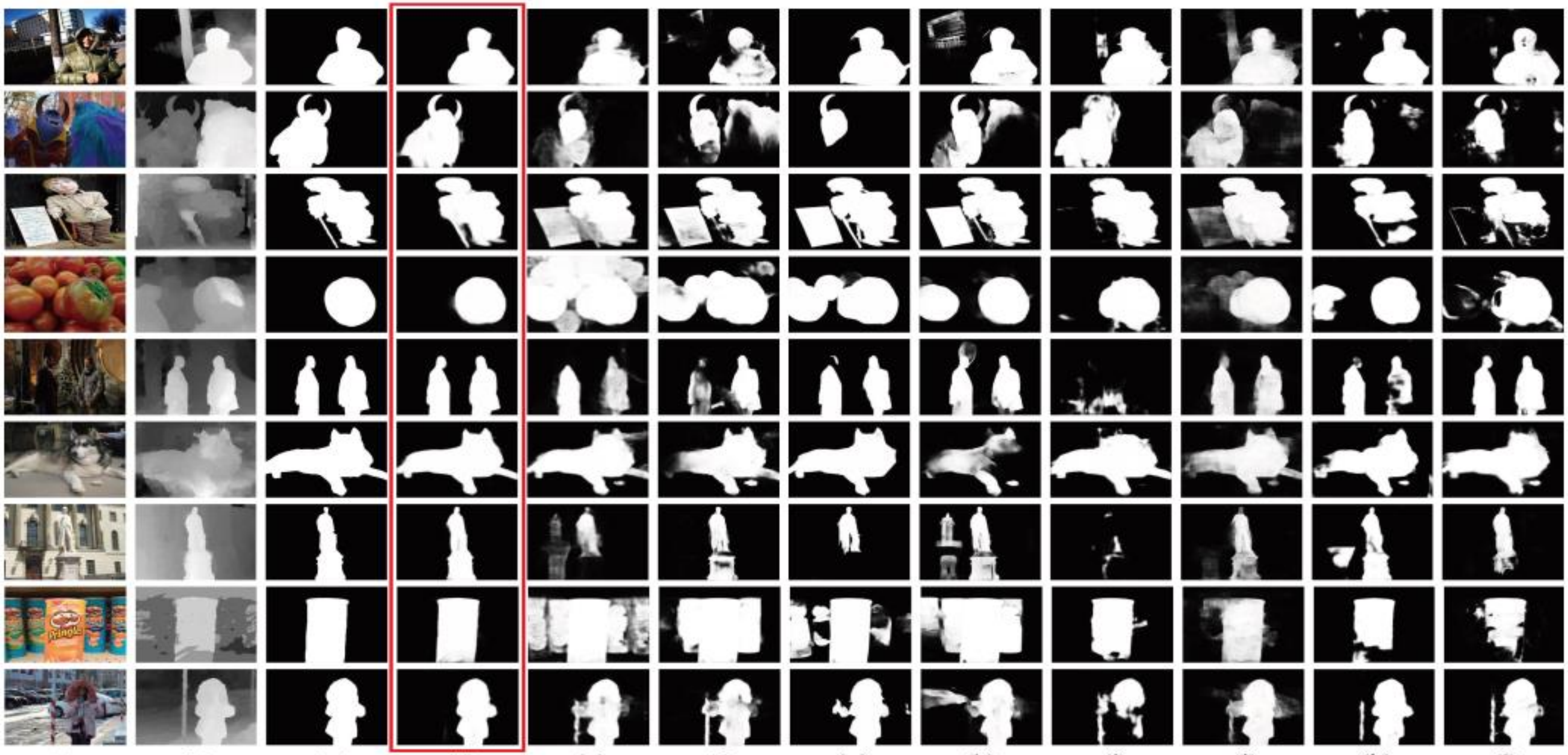
$$\mathcal{L}_{reg} = \begin{cases} 0.5(g-\hat{g})^2, & if\ |g-\hat{g}| < 1 \\ |g-\hat{g}| - 0.5, & otherwise \end{cases}$$

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



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

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

- Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong, Going from RGB to RGBD saliency: A depth-guided transformation model, IEEE Transactions on Cybernetics (TCyb), vol. 50, no. 8, pp. 3627-3639, 2020. **ESI Highly Cited Paper**
- Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang, 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. **ESI Highly Cited Paper**
- Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang, DPANet: Depth potentiality-aware gated Attention network for RGB-D salient object detection, IEEE Transactions on Image Processing (TIP), 2021.
- Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy, RGB-D salient object detection with cross-modality modulation and selection, European Conference on Computer Vision (ECCV), pp. 225-241, 2020.
- Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou, 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.

https://rmcong.github.io/   rmcong@bjtu.edu.cn

Future work

1

The whole process research of depth image processing tasks, such as depth estimation, hole filling, RGB-D semantic segmentation, etc.

2

For the depth SR task, explore more efficient RGB information guidance modes, such as multi-task learning. And explore solutions to the inconsistency of cross-modal information interaction.

3

For the SOD task, we can further extend new task with different data sources, try new learning based methods (such as weakly supervised learning), and find new ideas and solutions (such as instance-level SOD, saliency improvement and refinement).

Thanks

Runmin Cong (丛润民)
Beijing Jiaotong University

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