Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection

Jia-Xing Zhao[1,*] Yang Cao[1,*] Deng-Ping Fan[1,*] Ming-Ming Cheng[1] Xuan-Yi Li[1] Le Zhang [2]
[1] TKLNDST, CS, Nankai University [2] A*STAR
https://mmcheng.net/rgbdsalpyr/

Abstract

The large availability of depth sensors provides valuable complementary information for salient object detection (SOD) in RGBD images. However, due to the inherent difference between RGB and depth information, extracting features from the depth channel using ImageNet pre-trained backbone models and fusing them with RGB features directly are sub-optimal. In this paper, we utilize contrast prior, which used to be a dominant cue in none deep learning based SOD approaches, into CNNs-based architecture to enhance the depth information. The enhanced depth cues are further integrated with RGB features for SOD, using a novel fluid pyramid integration, which can make better use of multi-scale cross-modal features. Comprehensive experiments on 5 challenging benchmark datasets demonstrate the superiority of the architecture ***CPFP*** *over 9 state-of-the-art alternative methods.*

1. Introduction

Salient object detection (SOD) aims at distinguishing the most visually distinctive objects or regions in a scene. It has a wide range of applications, including video/image segmentation [17, 40], object recognition [46], visual tracking [3], foreground maps evaluation [14, 15], image retrieval [6, 16, 22, 38], content-aware image editing [8], information discovery [58], photo synthesis [5, 29], and weakly supervised semantic segmentation [52]. Recently, convolutional neural networks (CNNs) based methods [28, 36, 39] have become the main stream for SOD tasks, achieving promising results in challenging benchmarks [13]. However, existing CNNs-based SOD method mainly deal with RGB images, which may produce unsatisfying results when objects in the images share similar appearance with the background stuff.

Depth information from popular devices, *e.g.*, Kinect and iPhone X, provides important complementary information for identifying salient objects, as demonstrated in Fig. 1. Although several RGBD based SOD benchmarks [32, 42] and methods [4, 18, 20, 49] have been proposed in the last few years, how to effectively utilize depth information, especially in the context of deep neural networks [4], remains largely unexplored.

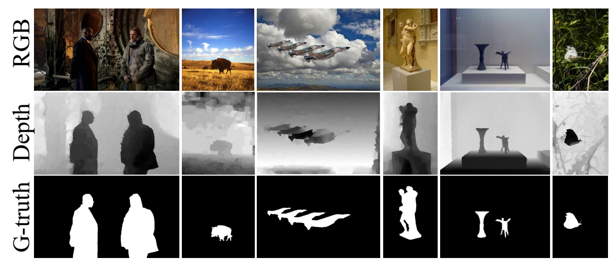

Figure 1. Samples from RGBD saliency datasets: NJU2000 [32], NLPR [42] and SSB [41]. Depth information plays an important complementary role for finding the salient objects.

Existing RGBD based SOD methods typically fuse RGB and depth input/features by simple concatenation, either via fusion at an early stage [42, 49], fusion at a late stage [18], or fusion at a middle stage [20], as shown in Fig. 2. We argue that direct cross-modal fusion via simple concatenation might be suboptimal due to two major challenges:

1) **Shortage of high-quality depth maps.**

Depth maps captured from state-of-the-art sensors are much noisier and textureless than RGB images, posting a challenge for the depth feature extraction. We lack well pre-trained backbone networks for extracting powerful features from depth maps, as an ImageNet [10] like large scale depth maps dataset is unavailable.

2) **Suboptimal multi-scale cross-modal fusion.**

The two modalities, *i.e.*, depth and RGB, have very different properties, making an effective multi-scale fusion of both modalities difficult. For instance, compared with the rest colors, 'green' color has a much stronger correlation with the 'plants' category. However, none depth value has such a correlation. The inherent difference between the two modalities may cause incompatibility problems when simple fusion strategies such as linear combination or concate-

*denotes joint first author. M.M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

nation are employed.

Instead of extracting features from depth maps using ImageNet pre-trained backbone networks, and then fusing the RGB and depth information as done in existing approaches [4, 18, 20, 49], we propose to enhance the depth information using the **contrast prior**. Then the enhanced depth map is used as an attention map to work with RGB features for high-quality SOD results. Before the popularity of CNNs, contrast prior used to be a dominant cue for discovering salient objects, not only in computer vision community [2, 7, 30, 43], but also in neuroscience [11] and cognitive psychology [50]. By re-employing the contrast prior with our *contrast-enhanced net*, we bridge the representative CNN features from the RGB channel and the powerful saliency prior from the depth channel. Specifically, we propose a *contrast loss* in the contrast-enhanced net by measuring the contrast between salient and non-salient regions as well as their coherence. Designed in a fully differentiable way, the contrast-enhanced net can be easily trained via back propagation and work with other CNN modules.

Effective multi-scale cross-modal feature fusion is desired for high-quality RGBD based SOD. Different from existing multi-scale feature fusion based CNN methods [4, 27, 28, 55], we need to additionally take care of the feature compatibility problem. We design **fluid pyramid integration** to fuse cross-modal (RGB and depth) information in a hierarchical manner. Inspired by Hou *et al*. [28] and Zhao *et al*. [55], our integration scheme contains a rich set of short-connections from higher CNN layers to lower CNN layers, while integrating features in a pyramid style. During the integration process, features from both modalities pass through several non-linear layers, enabling the back-propagation mechanism to adjust their representations for better compatibility.

We experimentally verify the effectiveness of our model designs via extensive ablation studies and comparisons. Even with the simple backbone network (VGG-16 [48]), our method demonstrates significant performance when compared with state-of-the-art RGBD-based SOD methods. In summary, our main contributions are three-fold.

- We design a contrast loss to utilize the contrast prior, which has been widely used in non-deep learning based method, for depth map enhancement. Our RGBD based SOD model successfully utilize the strengths of both traditional contrast prior as well as the deep CNN features.
- We propose a fluid pyramid integration strategy to make better use of multi-scale cross-modal features, whose effectiveness has been experimentally verified.
- Without bells and whistles, *e.g*., HHA [24], superpixels [54] or CRF [33], our model outperforms 9 state-of-the-art alternatives with a large margin, over 5 widely used benchmark datasets.

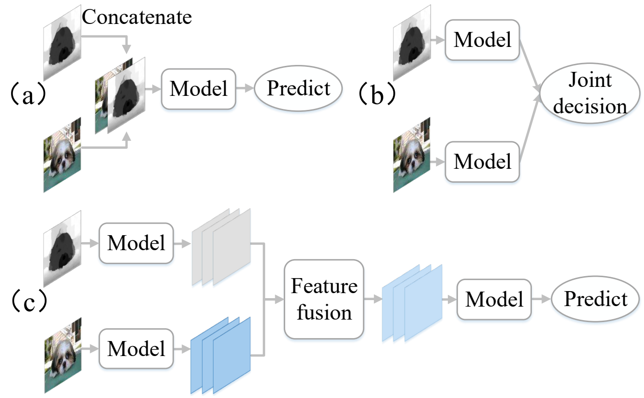

Figure 2. Three kinds of methods of using depth maps. (a) Early fusion (*e.g*. [42, 49]) (b) Late fusion (*e.g*. [18]) (c) Middle fusion (*e.g*. [20]) The details are introduced in Sec. 2.2.

2. Related Works

2.1. SOD

Earlier work for SOD relies on various hand-designed feature [7, 26, 37, 41]. Recently, learned representation is becoming the de-facto standard with much-improved performances. Li *et al*. [35] extracted multi-scale feature for each superpixel by the pre-trained deep convolutional network to derive the saliency map. The feature of three different scale bounding boxes surrounding each superpixel is combined into a feature vector to integrate the multi-scale information. In [56], Zhao *et al*. presented a multi-context deep learning framework for salient object detection in which two different CNNs are used to extract global and local context information, respectively. Lee *et al*. [34] considered both high-level feature extracted from CNNs and hand-crafted feature. The high-level feature and the hand-craft feature encoded using multiple 1×1 convolutional and ReLU layers are fused into a feature vector. Among the above-mentioned methods, the inputs are all superpixels so that the models have to be run many times to obtain the saliency object prediction results. Liu *et al*. [39] designed a two-stage network, in which a coarse downscaled prediction map is produced and refined in a hierarchical and progressive manner by another network. Li *et al*. [36] proposed a deep contrast network, which not only considers the pixel-wise information but also fuses the segment-level guidance into the network. A deep architecture with *short connection* is introduced in [28] which adds the connections from the high-level feature to the low-level feature based on the *HED* architecture [53].

2.2. RGBD based SOD

As shown in Fig. 2, existing RGBD saliency object detection approaches can be divided into three categories. The first scheme, as represented in Fig. 2(a), fuses the input in the earliest stage and regards the depth map as one channel of input directly [42, 49]. Fig. 2(b) stands for the second scheme which employs the "late fusion" strategy. More specifically, individual predictions from both RGB and depth are produced, and the results are integrated into a separate post-processing step such as pixel-wise summation and multiplication. For example, Fan *et al*. [18] used depth contrast and depth weighted color contrast to measure the saliency value of the regions. Fang *et al*. [19] used the depth extracted from DC-T coefficients to represent the energy for image patches. Cheng *et al*. [9] computed the saliency by laws of the visually salient stimuli in both color and depth spaces. Besides, Desingh *et al*. [12] leveraged nonlinear support vector regression to fuse these predicted maps. The third scheme, as shown in Fig. 2(c), combines the depth feature and RGB feature extracted from different networks. For instances, Feng *et al*. [20] proposed novel RGBD saliency feature to capture the spread of angular directions. Similarly, R.Shigematsu *et al*. [47] proposed to capture background enclosure, as well as low-level depth cues.

Recently, CNNs are adopted in RGBD saliency detection failed to obtain the more discriminative learning-based feature. CNNs-based methods almost belong to the third scheme as mentioned above. In [44], Qu *et al*. firstly generated RGB and depth feature vectors for each superpixel/patch, then fed these vectors into a CNN to derive the saliency confidence value, finally used a Laplacian propagation to obtain the final saliency map. Han *et al*. [25] proposed a two-view(RGB and depth) CNN to obtain the feature from RGB images and corresponding depth image, then simultaneously connected these feature with a new fully connected layer to get the final saliency map. Chen *et al*. [4] designed a progressive fusion method. For fusing the multi-scale information, it skip-connects the predictions from all the deeper layers to the shallower layers. While the information in different scales has been predicted as prediction map before fusing, that is to say, the cross-modal complementing for the feature is already finished before multi-scale fusing.

3. Proposed Method

The overall architecture CPFP is shown in Fig. 3. Feature-enhanced module(FEM) and fluid pyramid integration are applied in VGG-16. Based on contrast prior, FEM enhances RGB features at five stages of VGG-16. Details are introduced in Sec. 3.1. Then multi-scale cross-modal features are integrated by the fluid pyramid. Please see details in Sec. 3.2.

3.1. Feature-enhanced Module(FEM)

We propose to enhance the feature from RGB input by modulating them with information from the depth map. However, simply modulating with depth map may degenerate the final performance as depth maps are usually noisy. Instead, we propose a novel **Feature-enhanced Module** consisting of a **Contrast Enhance Net** to learn an enhanced depth map and a Cross-Modal Fusion strategy for feature modulation. The feature-enhanced Module is independent of network backbone for RGB stream. Here we use the VGG-16 suggested in [4] for fair comparison and the last three layers are truncated. VGG-16 network includes five convolution blocks and the outputs of the blocks are [2, 4, 8, 16, 32] times down-sampled respectively. As shown in Fig. 3, we add a feature-enhanced module(**FEM**) at the end of each block to obtain enhanced feature. FEM contains contrast enhanced net and cross-modal fusion, which will be introduced in Sec. 3.1.1 and Sec. 3.1.2.

3.1.1 Contrast-enhanced Net(CEN)

Motivated by previous work [14], the contrast between foreground and background as well as uniform distribution in the foreground are dominant in SOD. To use this prior effectively, we design a *contrast loss* in our contrast-enhanced net. The structure of Contrast Enhance Net is illustrated in Fig. 3. To measure the effect of contrast loss scientifically, for the other parts in CEN we choose several common layers and simple structure, which will not dominate the performance. The parameter details are introduced in Sec. 4.1. Contrast loss contains three items: the foreground object distribution loss l_f, the background distribution loss l_b and the whole depth image distribution loss l_w. In our case, we simply regard the salient objects in an image as the foreground objects.

Firstly, the enhanced map should be coherent with the original depth map for both foreground and background objects. Therefore, for the generated enhanced map, the foreground object distribution loss l_f and the background distribution loss l_b could be represented as:

$$\begin{aligned} l_f &= -\log\big(1-4*\sum_{(i,j)\in F}\frac{(p_{i,j}-\hat{p_f})^2}{N_f}\big), \\ l_b &= -\log\big(1-4*\sum_{(i,j)\in B}\frac{(p_{i,j}-\hat{p_b})^2}{N_b}\big), \end{aligned} \tag{1}$$

F and B are the salient object area and background in the ground truth. N_f and N_b denote the number of pixels in salient object and background, respectively. Similarly, $\hat{p_f}$ and $\hat{p_b}$ represent the mean of values in the foreground and in the background of enhanced map, respectively.

$$\hat{p_f} = \sum_{(i,j)\in F}\frac{p_{i,j}}{N_f}, \hat{p_b} = \sum_{(i,j)\in B}\frac{p_{i,j}}{N_b}. \tag{2}$$

As defined in Eqn. 1, we model the internal variance of salient objects and background to promote consistency with the original depth map. A sigmoid layer is used to squash the outputs of the Contrast Enhance Net to [0, 1]. In this case, the maximum variance of the internal variance is 0.25,

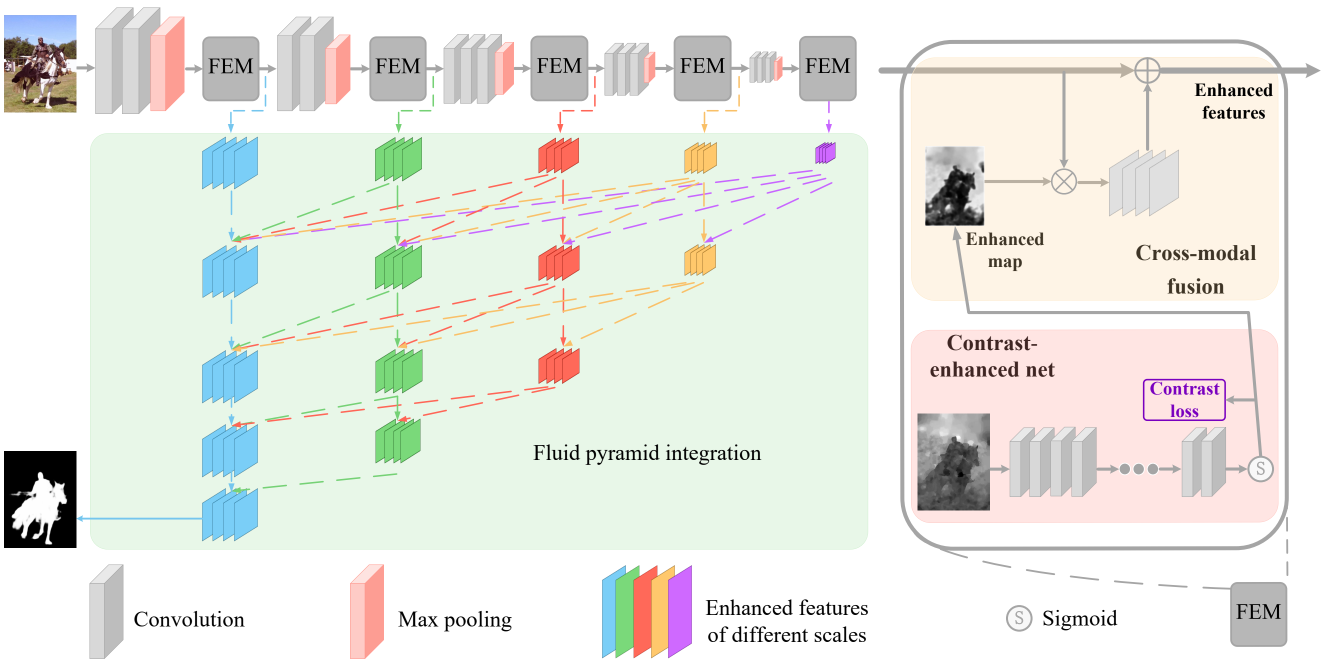

Figure 3. Architecture CPFP. The architecture contains two modules: feature-enhanced modules(FEM) and fluid pyramid integration module. FEM contains two submodules: Contrast-enhanced net and cross-modal fusion. In contrast-enhanced net, we utilize a novel contrast loss to leverage the contrast prior in the deep network to generate the enhanced map, and then get the enhanced features by the cross-modal fusion at all the 5 stages of VGG-16. The fluid pyramid integration method is designed to fuse the multi-scale cross-modal features. The details of our architecture are introduced in Sec. 3.

thus we multiply the variance by 4 to ensure the range of the log function is from 0 to 1.

Secondly, the contrast between the foreground and background objects should be enhanced. Hence we define the whole depth image distribution loss l_w as:

$$l_w = -\log\left(\hat{p_f} - \hat{p_b}\right)^2. \tag{3}$$

We ensure the contrast between foreground objects and background as large as possible by modeling the mean difference. The $\hat{p_f}$ and $\hat{p_b}$ are between 0 and 1, thus the value of the parameter in the log function range from 0 to 1.

Finally, the contrast loss l_c can be represented as:

$$l_c = \alpha_1 l_f + \alpha_2 l_b + \alpha_3 l_w, \tag{4}$$

where α_1 and α_2 and α_3 are pre-defined parameters. We suggest setting them to 5, 5 and 1 respectively.

As shown in Fig. 4, the enhanced depth maps have higher contrasts compared with the original depth maps. Besides, The distributions in foreground and background are more uniform.

3.1.2 Cross-modal Fusion

Cross-modal fusion is a sub-module of the feature-enhanced module which aims at modulating RGB feature with the enhanced depth map. The role of the one-channel enhanced map is similar to the attention map [21, 51]. To be specific, we multiply the RGB feature maps from each block by the enhanced depth map to enhance the contrast of feature between salient and non-salient regions. A residual connection is further added to preserve the original RGB feature. We call these feature maps **enhanced feature** $\tilde{F}$, which is computed as:

$$\tilde{F} = F + F \otimes D_E, \tag{5}$$

F is the original RGB feature and D_E denotes the enhanced map generated by the proposed contrast enhanced net. $\otimes$ denotes the pixel-wise multiplication.

As shown in Fig. 3, by plugging the feature-enhanced module into the end of each block, we obtain enhanced features of five different scales, $\tilde{F}_1$, $\tilde{F}_2$, $\tilde{F}_3$, $\tilde{F}_4$, $\tilde{F}_5$, respectively.

3.2. Fluid Pyramid Integration(FPI)

When dealing with cross-modal information, feature compatibility is the key point. Motivated by the recent success in multi-scale feature fusion, we design a fluid pyramid architecture as shown in Fig. 3. The fluid pyramid can make fuller use of cross-modal feature in the multi-scale level, which helps to ensure feature compatibility.

Concretely, our pyramid has 5 tiers. The first tier is composed of five nodes and each node is a set of enhanced features of different scales. Then, we construct the first node of the second tier by up-sampling $\tilde{F}_2$, $\tilde{F}_3$, $\tilde{F}_4$, $\tilde{F}_5$ to the same size as $\tilde{F}_1$ and adding these up-sampled features. Similarly, we up-sample $\tilde{F}_3$, $\tilde{F}_4$, $\tilde{F}_5$ to the same size as $\tilde{F}_2$ and adding them to construct the second node of the second tier. In this way, for the $nth(n \in \{1,2,3,4,5\})$ tier of the pyramid, there are n nodes in total and each node is integrated with all the higher-level information from the $(n-1)th$ tier of the pyramid ($0th$ in this case turns back to the modified VGG-16 backbone). Followed by a transition convolution layer and a sigmoid layer, we obtain the final saliency map P. Compared with [4] which concatenates the predicted saliency maps, the proposed integration approach works on the feature maps. While feature reserves richer cross-modal information before fused in multi-scale level. That is to say, fluid pyramid integrates information in both multi-scale level and cross-modal level. Compared with [55] which fuses features in the traditional pyramid way, FPI leads all the high-level features into low-level features for every node at each tier of the pyramid by richer connection, which called *fluid connection*. Fluid connection provides more interactions for cross-modal features in different scales, which helps feature compatibility in multi-scale level.

Inspired by [53], we add deep supervisions to the enhanced depth map of each scale. Therefore, the total loss L could be represented as:

$$L = l_s + \sum_{i=1}^{5} l_{c_i}, \tag{6}$$

where l_s represents the cross-entropy loss between the predicted map and saliency ground truth. l_{c_i} represent the **contrast loss** in the ith feature enhance module. Contrast loss has been mentioned above and cross-entropy loss could be computed as:

$$l_f = Y \log P + (1-Y)\log(1-P), \tag{7}$$

where P and Y denote the predicted map and saliency ground-truth map, respectively.

4. Experiments

4.1. Implementation Details

The proposed idea is generally independent of the network backbone. In this work, we choose VGG-16 [48] for a fair comparison. The proposed network is implemented using the Caffe library [31]. Following [4], we randomly select 1400 samples from the NJU2000 [32] and 650 samples from the NLPR [42] for training. We also sample 100 images from NJU2000 and 50 from NLPR as the validation set. The rest images are for testing. We randomly flip the images in the training set for data augmentation.

Parameter details in Contrast-enhanced Net. We simply use two convolutional layers followed by ReLU layers, repeatedly to ensure that the enhanced maps have the same size as the original feature map. In the first convolutional layer, kernel size, number of channel and stride are set to be (4, 32, 2). In the second convolutional layer, kernel size, number of channel and stride are set to be (3, 32, 1). After repeating this two-layer block until feature maps hold the same size as RGB feature in fusion position. Then two more convolutional layers are followed. Their kernel size, channel number and stride are (3, 32, 1) and (3, 1, 1) respectively. After that, the output is thrown into a sigmoid layer to generate the final enhanced map. A sigmoid layer is adopted to ensure that the values of enhanced map fall in the range [0, 1].

Training. During the training phase, we train our network for 10,000 iterations. The initial learning rate is set to 1e-7 and divided by 10 after 7,000 iterations. Weight decay and momentum are set to 0.0005 and 0.9, respectively. We train our network on a single NVIDIA TITAN X GPU. The batch size and iter size are set to 1 and 10, respectively. The parameters of newly added convolutional layers are all initialized with Gaussian kernels. For image whose length or width is larger than 400, we resize it to new length and width, in which the maximum value is 400 while keeping the length-width ratio unchanged.

Inference. During the inference phase, we resize the predicted saliency maps to keep the same resolution as original RGB images.

4.2. Datasets and Evaluation Metrics

Datasets. We conduct our experiments on 5 following widely used RGBD datasets. **NJU2000** [32] contains 2003 stereo image pairs with diverse objects and complex, challenging scenarios, along with ground-truth map. The stereo images are gathered from 3D movies, the Internet, and photographs taken by a Fuji W3 stereo camera. **NLPR** [42] is also called **RGBD1000** dataset which including 1,000 images. There may exist multiple salient objects in each image. The structured light depth images are obtained by the Microsoft Kinect under different illumination conditions. **SSB** [41] is also called **STEREO** dataset, which consists of 1000 pairs of binocular images. **LFSD** [37] is a small dataset which contains 100 images with depth information and human labeled ground truths. The depth information was obtained via the Lytro light field camera. **RGBD135** [9] is also named **DES** which consists of seven indoor scenes and contains 135 indoor images collected by Microsoft Kinect.

Evaluation Metrics. We adopt 4 commonly used metrics, namely S-measure, mean F-measure, max F-measure and mean absolute error (MAE), and the recently released structure measure (S-measure [14]) to evaluate the performance of different methods [2].

The F-measure is a harmonic mean of average precision and average recall, formulated as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{8}$$

we set $\beta^2 = 0.3$ to weigh precision more than recall as

Dataset	Metric	LHM [42]	GP [45]	LBE [20]	SE [23]	CDCP [57]	DF [44]	MDSF [49]	CTMF [25]	PCF [4]	Our CPFP
SSB1000	S-measure ↑	0.562	0.588	0.660	0.708	0.713	0.757	0.728	**0.848**	**0.875**	**0.879**
	meanF ↑	0.378	0.405	0.501	0.610	0.643	0.616	0.527	**0.758**	**0.818**	**0.842**
	maxF ↑	0.683	0.671	0.633	0.755	0.668	0.756	0.719	**0.831**	**0.860**	**0.873**
[41]	MAE ↓	0.172	0.182	0.250	0.143	0.149	0.141	0.176	**0.086**	**0.064**	**0.051**
NJU2000	S-measure ↑	0.514	0.527	0.695	0.664	0.669	0.763	0.748	**0.849**	**0.877**	**0.878**
	meanF ↑	0.328	0.357	0.606	0.583	0.594	0.663	0.628	**0.779**	**0.840**	**0.850**
	maxF ↑	0.632	0.647	0.748	0.747	0.621	0.815	0.775	**0.845**	**0.872**	**0.877**
[32]	MAE ↓	0.205	0.211	0.153	0.169	0.180	0.136	0.157	**0.085**	**0.059**	**0.053**
LFSD	S-measure ↑	0.557	0.640	0.736	0.698	0.717	0.791	0.700	**0.796**	**0.794**	**0.828**
	meanF ↑	0.396	0.519	0.611	0.640	0.680	0.679	0.521	**0.756**	**0.761**	**0.811**
	maxF ↑	0.712	0.787	0.726	0.791	0.703	**0.817**	0.783	**0.791**	0.779	**0.826**
[37]	MAE ↓	0.211	0.183	0.208	0.167	0.167	0.138	0.190	**0.119**	**0.112**	**0.088**
RGBD135	S-measure ↑	0.578	0.636	0.703	0.741	0.709	0.752	0.741	**0.863**	**0.842**	**0.872**
	meanF ↑	0.345	0.411	0.576	0.619	0.585	0.604	0.523	**0.756**	**0.765**	**0.815**
	maxF ↑	0.511	0.600	0.788	0.745	0.631	0.766	0.746	**0.844**	**0.804**	**0.838**
[9]	MAE ↓	0.114	0.168	0.208	0.089	0.115	0.093	0.122	**0.055**	**0.049**	**0.037**
NLPR	S-measure↑	0.630	0.654	0.762	0.756	0.727	0.802	0.805	**0.860**	**0.874**	**0.888**
	meanF ↑	0.427	0.443	0.626	0.624	0.621	0.684	0.649	**0.753**	**0.809**	**0.840**
	maxF ↑	0.622	0.603	0.745	0.720	0.655	0.792	0.793	**0.834**	**0.847**	**0.869**
[42]	MAE ↓	0.108	0.155	0.081	0.099	0.117	0.078	0.095	**0.063**	**0.052**	**0.036**

Table 1. Quantitative comparison results including S-measure, mean F-measure, maximum F-measure and MAE on 5 popular datasets. ↑ & ↓ denote larger and smaller is better, respectively. Top three scores in each row are marked in red, blue, and green, respectively.

suggested in [1]. Following [2], we provide the mean F-measure, max F-measure using different thresholds(0-255).

Let P and Y denote the saliency map and the ground truth that is normalized to [0, 1]. For fair comparison on non-salient regions [2], we compute the MAE score by:

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - Y(x,y)|, \tag{9}$$

where W and H are the width and height of the saliency map.

Both MAE and F-measure metrics ignore the structure similarity assessment, however, behavioral vision studies have shown that the human visual system is highly sensitive to structures in scenes [14]. Thus, we additionally introduce the S-measure [14] for a more comprehensive evaluation. The S-measure combines the region-aware (S_r) and object-aware (S_o) structural similarity as their final structure metric:

$$S - measure = \alpha * S_o + (1 - \alpha) * S_r, \tag{10}$$

where $\alpha \in [0, 1]$ is the balance parameter and set 0.5.

4.3. Ablation Experiments and Analyses

In this section, we explore the effect of different components in the proposed method on the NJU2000 dataset.

Feature-enhanced Module. To prove the effectiveness of the proposed contrast-enhanced net. We compare the results of using the backbone(denoted by B) with the results adding FEM in the backbone(denoted by B + C). As shown in Tab. 2, comparing the 1st and 3rd rows, we could see

Figure 4. The visual comparison between depth images and their enhanced maps. The contrast between salient and non-salient regions is promoted and meanwhile the values in these regions become more consistent.

that the proposed FEM brings obvious improvement. In addition, we show some visual comparisons between depth images and their enhanced maps in Fig. 4. Obviously, compared with the original depth images, the contrast between salient and non-salient regions is promoted and meanwhile the value within both regions become more consistent. Besides, we also evaluate the results directly using the original depth maps as enhanced maps(denoted by B + D, the 2nd row in Tab. 2), which shows that B + D have negative effects. It is reasonable. From the original depth map shown in Fig. 4, we can see that contrast between salient and no-salient regions is not obvious enough and there are more noises within salient region and background. While B + C makes a difference, the visual instances are shown in Fig. 6. Comparing the results generated by the backbone(B, the 3rd column in Fig. 6) and backbone fusing original depth map (B + D, the 4th column Fig. 6), we could see that origi-

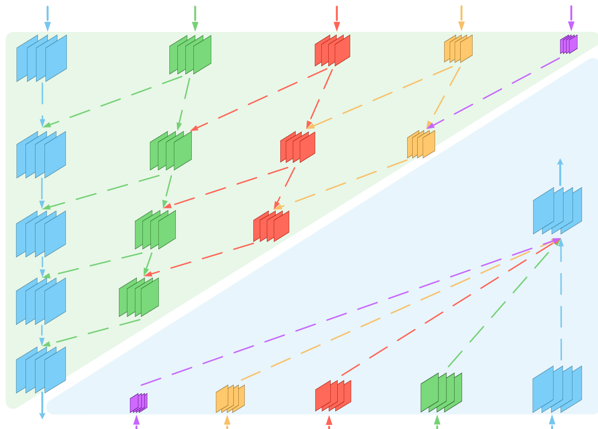

Figure 5. The different fusion method. The one located at the upper left is the pyramid fusion architecture [55] (P in Tab. 2). The other one located at the lower right is the simple multi-scale fusion architecture [36] (M in Tab. 2).

nal depth map does not work well. When we add our proposed feature-enhanced module into the backbone to fuse the cross-modal information, the results are shown in the 5th column of Fig. 6 (B + C). Regions that are mistaken for the salient object in the backbone are successfully removed with the help of depth information. It shows that after enhancing depth maps with contrast prior, depth information helps a lot when detection from RGB features meets difficulties. For example, some regions in RGB maps are noisy(because of color, texture, brightness, *et al.*) are in trivial distribution in depth level.

Fluid Pyramid Integration. Compared with some traditional multi-scale methods [36, 55], the proposed integration can utilize information more fully, which helps cross-modal feature compatibility in multi-scale levels. In Tab. 2, the 3rd and last rows show the performance before adding FPI(B + C) and after adding FPI(B + C + FP). Numerically, the pyramid integration strategy is very effective and contributes by nearly ten percentage points. To illustrate the role of pyramid architecture, we firstly adopt the simple fusion method in which we up-sample the multi-scale features to the same size and concatenate them directly [36] as shown in the lower right of Fig. 5. We denote this method as B + C + M and show the performance in the 4th row in Tab. 2. The results show that the help of this multi-scale fusion method is very limited. Then we use a pyramid architecture to fuse these feature hierarchically [55] as shown in the upper left of Fig. 5, which is denoted as B + C + P in the 5th row of Tab. 2. Numerically, the pyramid fusion is much more effective than the direct fusion method and contributes improvement by nearly four points. Then we add the fluid connection on the pyramid, the result is further improved as shown in the 6th row. Visually, as shown in Fig. 6, compared the results between the 5th (B + C) and the 6th (B + C + M) column. It could be seen that after fusing

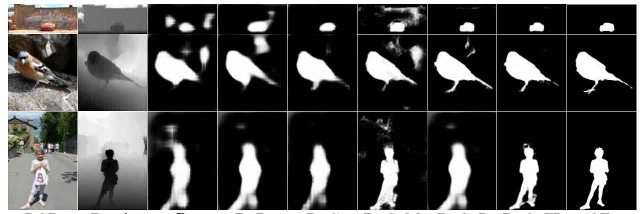

Figure 6. Visual comparison with different modules. The meaning of indexes could be seen in the caption of Tab. 2

Model	meanF↑	maxF↑	MAE↓
B [40]	0.714	0.791	0.115
B + D	0.708	0.788	0.121
B + C	0.756	0.806	0.094
B + FP	0.758	0.814	0.092
B + C + M	0.748	0.824	0.105
B + C + P	0.789	0.844	0.078
B + D + FP	0.783	0.842	0.081
B + C + FP	0.851	0.877	0.053

Table 2. Ablation studies of different modules. B denotes the base model (VGG). D denotes the depth map. B + D represents that we directly use the original depth map as an enhanced map. C denotes the contrast-enhanced net and M denotes simple multi-scale fusion as shown in the bottom right of Fig. 5. P denotes the pyramid fusion as shown in the top left of Fig. 5 and FP denotes the proposed fluid pyramid integration method. The details are introduced in Sec. 4.3.

the multi-scale information, the edge details have been improved. But the non-salient region which has been shielded by contrast prior(5th column) comes out again. The reason behind this phenomenon is that the cross-modal information fusing meets feature compatibility problem in multi-scale level. Then we leverage the pyramid architecture (B + C + P) to fuse the multi-scale information more fully. Non-salient region becomes smaller because features complement better. After we add the fluid connection (B + C + FP), fusing the high-level features into the low-level features at each tier of the pyramid, the location of the salient object becomes much better. Feature complementing achieves the best performance.

4.4. Compare with the State-of-the-art

We compare our model with 9 RGBD based salient object detection models including LHM [42], GP [45], LBE [20], SE [23], CTMF [25], DF [44], MDSF [49], CDCP [57], and PCF [4]. Note that all the saliency maps of the above methods are produced by running source codes or pre-computed by the authors. For all the compared methods, we use the default settings suggested by the paper. For works which do not release the code currently, we appreciate the author helping to run the results.

As shown in Tab. 1, our method outperforms the state-of-the-art methods on most evaluation metrics contain Max F-measure, Mean F-measure and MAE. Compared with recently proposed CNNs-based methods, our method has ob-

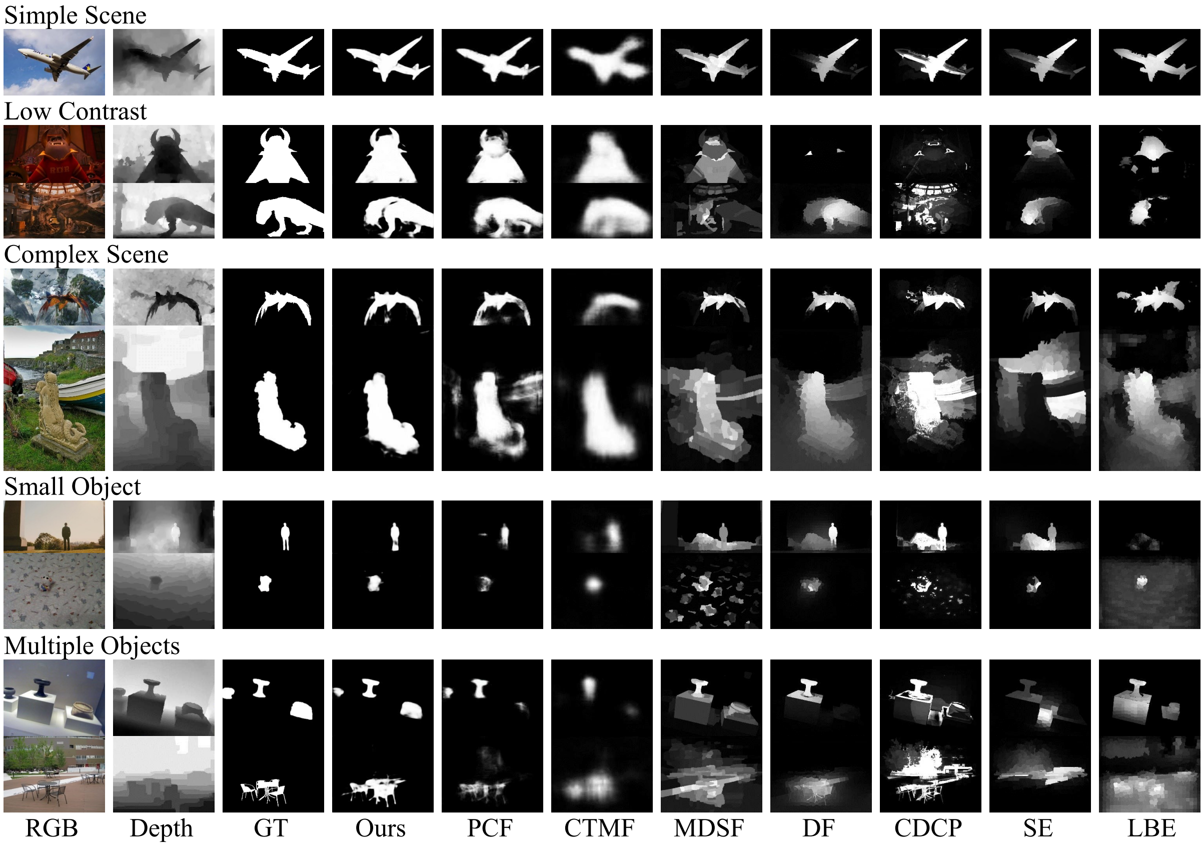

Figure 7. The visualization results from **SSB1000**, **NJU2000**, **LFSD**, **RGBD135** and **NLPR**.

vious advantages on commonly used datasets.

In Fig. 7, we show some visualization results. Especially, we summarize several challenging situations in salient object detection: low contrast, complex scene, small object and multiple objects. As shown in Fig. 7, we show a simple example in the 1st row and almost methods perform well. In 2nd-3rd rows, we show some low contrast images in which the color differences between the salient object and background are not obvious. However, if their depth difference is obvious as the showed samples, we could leverage these depth information to help the model to detect the salient objects. Compared with the early methods(right), our results are more complete. Compared to the learning-based methods such as PCF [4] and CTMF [25], the details are much better. Besides, we also sample some images (4th-5th rows) whose scene is complex. In these images, other methods mistake the background for the salient object due to the complexity of scene. However, our model performs very well. These two types of images further illustrate that the proposed way of using the depth information is reasonable. Then, we show other two challenging situations, small object and multiple objects. In these challenging cases, it can be seen that our model not only locates the salient object well through high-level information but also segment objects well through low-level information.

5. Conclusion

In this paper, we develop a contrast-enhanced net supervised by a novel contrast loss for depth images. The proposed net enhances depth maps explicitly, based on contrast prior. Then enhanced map works with RGB features, to enhance the contrast between the salient and non-salient regions, and meanwhile guarantee the coherence within these regions. Besides, we design a fluid pyramid integration method to make better use of the multi-scale cross-modal features. Compared with multi-scale fusing strategies for single-modal features, fluid pyramid integration is designed fuller for cross-modal fusing in multi-scale level, to deal with feature compatibility better. Our approach significantly advances the state-of-the-art over the widely used datasets and is capable of capturing salient regions under challenging situations.

Acknowledgements. We would like to thank the anonymous reviewers for their useful feedback. This research was supported by NSFC (61572264), the national youth talent support program, the Fundamental Research Funds for the Central Universities (Nankai University, NO. 63191501) and Tianjin Natural Science Foundation (17JCJQJC43700, 18ZXZNGX00110).

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 6

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient Object Detection: A Benchmark. *IEEE TIP*, 24(12):5706–5722, 2015. 2, 5, 6

[3] Ali Borji, Simone Frintrop, Dicky N Sihite, and Laurent Itti. Adaptive object tracking by learning background context. In *CVPRW*, pages 23–30. IEEE, 2012. 1

[4] Hao Chen and Youfu Li. Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection. In *CVPR*, pages 3051–3060, 2018. 1, 2, 3, 5, 6, 7, 8

[5] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM TOG*, 28(5):124, 2009. 1

[6] Ming-Ming Cheng, Qi-Bin Hou, Song-Hai Zhang, and Paul L Rosin. Intelligent visual media processing: When graphics meets vision. *JCST*, 32(1):110–121, 2017. 1

[7] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015. 2

[8] Ming-Ming Cheng, Fang-Lue Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Repfinder: finding approximately repeated scene elements for image editing. *ACM TOG*, 29(4):83, 2010. 1

[9] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *ICIMCS*, page 23. ACM, 2014. 3, 5, 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[11] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 2

[12] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013. 3

[13] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*. Springer, 2018. 1

[14] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017. 1, 3, 5, 6

[15] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, pages 698–704, 2018. 1

[16] Deng Ping Fan, Juan Wang, and Xue Mei Liang. Improving image retrieval using the context-aware saliency areas. In *Applied Mechanics and Materials*, volume 734, pages 596–599. Trans Tech Publ, 2015. 1

[17] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019. 1

[18] Xingxing Fan, Zhi Liu, and Guangling Sun. Salient region detection for stereoscopic images. In *DSP*, pages 454–458, 2014. 1, 2, 3

[19] Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin. Saliency detection for stereoscopic images. *IEEE TIP*, 23(6):2625–2636, 2014. 3

[20] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In *CVPR*, pages 2343–2350, 2016. 1, 2, 3, 6, 7

[21] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, volume 2, page 3, 2017. 4

[22] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE TIP*, 21(9):4290–4303, 2012. 1

[23] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for rgb-d image via saliency evolution. In *IEEE ICME*, pages 1–6. IEEE, 2016. 6, 7

[24] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, pages 345–360. Springer, 2014. 2

[25] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 2017. 3, 6, 7, 8

[26] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2007. 2

[27] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, pages 447–456, 2015. 2

[28] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 5300–5309. IEEE, 2017. 1, 2

[29] Shi-Min Hu, Tao Chen, Kun Xu, Ming-Ming Cheng, and Ralph R Martin. Internet visual media processing: a survey with graphics and vision applications. *The Visual Computer*, 29(5):393–405, 2013. 1

[30] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 2

[31] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678. ACM, 2014. 5

[32] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE ICIP*, pages 1115–1119, 2014. 1, 5, 6

[33] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 2

[34] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *ICCV*, pages 660–668, 2016. 2

[35] Guanbin Li and Yizhou Yu. Visual saliency based on multi-scale deep features. In *ICCV*, pages 5455–5463, 2015. 2

[36] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *ICCV*, pages 478–487, 2016. 1, 2, 7

[37] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014. 2, 5, 6

[38] Guanghai Liu and Dengping Fan. A model of visual attention for natural image retrieval. In *IEEE ISCC-C*, pages 728–733, 2013. 1

[39] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *ICCV*, pages 678–686, 2016. 1, 2

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *ICCV*, pages 3431–3440, 2015. 1, 7

[41] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012. 1, 2, 5, 6

[42] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: a benchmark and algorithms. In *ECCV*, pages 92–109. Springer, 2014. 1, 2, 5, 6, 7

[43] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*. IEEE, 2012. 2

[44] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE TIP*, 26(5):2274–2285, 2017. 3, 6, 7

[45] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting Global Priors for RGB-D Saliency Detection. In *CVPRW*, pages 25–32, 2015. 6, 7

[46] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE TCSVT*, 24(5):769–779, 2014. 1

[47] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning RGB-D Salient Object Detection using background enclosure, depth contrast, and top-down features. In *ICCVW*, 2017. 3

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 5

[49] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE TIP*, 26(9):4204–4216, 2017. 1, 2, 6, 7

[50] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2

[51] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017. 4

[52] Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 39(11):2314–2320, 2017. 1

[53] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, pages 1395–1403, 2015. 2, 5

[54] Jiaxing Zhao, Bo Ren, Qibin Hou, and Ming-Ming Cheng. Flic: Fast linear iterative clustering with active search. In *AAAI*, 2018. 2

[55] Kai Zhao, Wei Shen, Shanghua Gao, Dandan Li, and Ming-Ming Cheng. Hi-fi: Hierarchical feature integration for skeleton detection. In *IJCAI*, 2018. 2, 5, 7

[56] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015. 2

[57] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *ICCVW*, 2017. 6, 7

[58] Jun-Yan Zhu, Jiajun Wu, Yan Xu, Eric Chang, and Zhuowen Tu. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE TPAMI*, 37(4):862–875, 2015. 1