Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection

Jia-Xing Zhao¹,* Yang Cao¹,* Deng-Ping Fan¹,* Ming-Ming Cheng¹ Xuan-Yi Li¹  Le Zhang²
¹ TKLNDST, CS, Nankai University ² A*STAR

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 mainstream 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.

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 pretrained 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](image)

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 $1 \times 1$ 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 super-pixel/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 $l_p$. 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:

$$l_f = -\log (1 - 4 \times \sum_{(i,j) \in F} \frac{(p_{i,j} - \hat{p}_f)^2}{N_f}),$$
$$l_b = -\log (1 - 4 \times \sum_{(i,j) \in B} \frac{(p_{i,j} - \hat{p}_b)^2}{N_b}),$$

\(F\) 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.

$$\hat{p}_f = \frac{\sum_{(i,j) \in F} p_{i,j}}{N_f}, \quad \hat{p}_b = \frac{\sum_{(i,j) \in B} p_{i,j}}{N_b}$$

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.
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:

\[
l_w = -\log(p_f - \hat{p}_b)^2.
\]

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:

\[
l_c = \alpha_1 l_f + \alpha_2 l_b + \alpha_3 l_w,
\]

where \( \alpha_1 \) and \( \alpha_2 \) and \( \alpha_3 \) 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:

\[
\tilde{F} = F + F \odot D_E,
\]

\( F \) is the original RGB feature and \( D_E \) denotes the enhanced map generated by the proposed contrast enhanced net. \( \odot \) 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 \( n \) nodes in total and each node is integrated with all the higher-level information from the \( (n-1) \)th 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. 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:

\[
L = l_s + \sum_{i=1}^{5} l_{c_i},
\]

(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 \( i \)th feature enhance module. Contrast loss has been mentioned above and cross-entropy loss could be computed as:

\[
l_f = Y \log P + (1-Y) \log(1-P),
\]

(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:

\[
F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall},
\]

(8)

we set \( \beta^2 = 0.3 \) to weigh precision more than recall as.
Our results of using the backbone (denoted by B) with the results of the proposed contrast-enhanced net. We compare the evaluation. The S-measure combines the region-aware ($S$) and object-aware ($S_o$) structural similarity as their final structure metric:

$$S - \text{measure} = \alpha \ast S_o + (1 - \alpha) \ast S_r,$$

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 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 values in these regions become more consistent.
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.) 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 FP(B + C) and after adding FP(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 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], CDTP [57], and PCF [3]. 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-
previous 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 (17JCYJJC43700, 18ZXZNGX00110).
References


[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
[54] Jiaxing Zhao, Bo Ren, Qibin Hou, and Ming-Ming Cheng. Flic: Fast linear iterative clustering with active search. In AAAI, 2018. 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
[61] Jiaxing Zhao, Bo Ren, Qibin Hou, and Ming-Ming Cheng. Flic: Fast linear iterative clustering with active search. In AAAI, 2018. 2
[64] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In ICCVW, 2017. 6, 7