Improved Saliency Detection in RGB-D Images Using Two-Phase Depth Estimation and Selective Deep Fusion

Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin

***Abstract*— To solve the saliency detection problem in RGB-D images, the depth information plays a critical role in distinguishing salient objects or foregrounds from cluttered backgrounds. As the complementary component to color information, the depth quality directly dictates the subsequent saliency detection performance. However, due to artifacts and the limitation of depth acquisition devices, the quality of the obtained depth varies tremendously across different scenarios. Consequently, conventional selective fusion-based RGB-D saliency detection methods may result in a degraded detection performance in cases containing salient objects with low color contrast coupled with a low depth quality. To solve this problem, we make our initial attempt to estimate additional high-quality depth information, which is denoted by Depth^{+}. Serving as a complement to the original depth, Depth^{+} will be fed into our newly designed selective fusion network to boost the detection performance. To achieve this aim, we first retrieve a small group of images that are similar to the given input, and then the inter-image, nonlocal correspondences are built accordingly. Thus, by using these inter-image correspondences, the overall depth can be coarsely estimated by utilizing our newly designed depth-transferring strategy. Next, we build fine-grained, object-level correspondences coupled with a saliency prior to further improve the depth quality of the previous estimation. Compared to the original depth, our newly estimated Depth^{+} is potentially more informative for detection improvement. Finally, we feed both the original depth and the newly estimated Depth^{+} into our selective deep fusion network, whose key novelty is to achieve an optimal complementary balance to make better decisions toward improving saliency boundaries.**

***Index Terms*— RGB-D saliency detection, inter-image correspondences, low-level saliency, selective deep fusion.**

I. Introduction and Motivation

IMAGE saliency detection aims to rapidly and accurately locate salient objects in a given scene, and currently, it frequently serves as a preprocessing tool in various applications including object tracking [1], video saliency detection [2]–[5], image quality assessment [6], background subtraction [7], [8], and object recognition [9].

Different from the conventional RGB image saliency [10]–[13], which has achieved remarkable progress in recent years, RGB-D saliency is a relatively new research topic. Compared to RGB image saliency [14], [15], which relies its saliency mechanism in RGB-spanned multiscale/multilevel contrasts that might encounter feature conflict, the depth information in RGB-D images provides a new venue to potentially alleviate this problem.

Generally, the current mainstream saliency-revealing solutions in the depth channel [16] usually follow the assumption that salient objects should be located at different depth layers to their nearby, non-salient surroundings. Thus, quite a few uniqueness-based depth-saliency-revealing strategies have been proposed, including the depth customized enclosure distribution [17], [18], the center-surround difference computation [19], and the multiscale uniqueness computation [20], [21]. As a result, the saliency detection performance can be significantly improved by simply fusing the depth saliency with the RGB saliency.

However, there still exists one persistent problem, which causes the current state-of-the-art RGB-D methods to reach a performance bottle-neck; i.e., the depth information itself may not always be trustworthy to separate a salient object from its non-salient surroundings (see the demonstrations in Fig. 1). Moreover, saliency clues revealed from those untrustworthy depths may result in an even worse fused saliency. Therefore, the main focus of this paper is to estimate an additional high-quality depth map to alleviate the abovementioned difficulty. The key rationale of our method is based on the phenomenon that images with similar scenes should exhibit a similar depth layout [22], [23]. Thus, we estimate the depth by transferring the depth information from other RGB-D images.

Our depth estimation method consists of two components: coarse-level depth estimation and saliency-aware depth enhancement. We utilize the former component (i.e., coarse-level depth estimation) to estimate the initial image-level depth; the depth-transferring procedure mainly considers both the mid-level and the object-level inter-image similarity. Next, we utilize the latter component (i.e., saliency-aware depth enhancement) to further improve the depth quality from the saliency perspective; the core rationale is to enlarge the

Manuscript received February 11, 2019; revised August 20, 2019; accepted January 1, 2020. Date of publication January 30, 2020; date of current version February 13, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61802215 and Grant 61806106, in part by the Natural Science Foundation of Shandong Province under Grant ZR2019BF011 and Grant ZR2019QF009, and in part by the National Science Foundation of USA under Grant IIS-1715985 and Grant IIS-1812606. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pavan Turaga. *(Corresponding author: Chong Peng.)*

Chenglizhao Chen, Jipeng Wei, Chong Peng, and Weizhong Zhang are with the College of Computer Science and Technology, Qingdao University, Qingdao 266071, China (e-mail: pchong1991@163.com).

Hong Qin is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA.

Digital Object Identifier 10.1109/TIP.2020.2968250

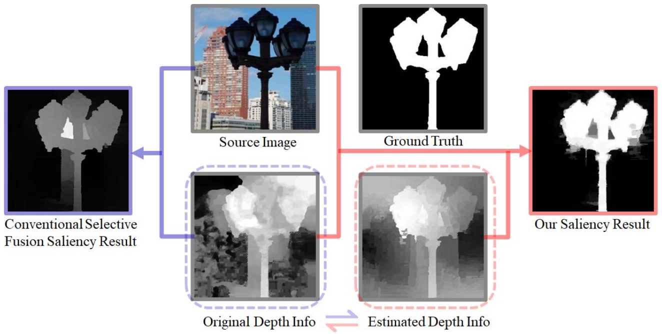

Fig. 1. The motivation behind our method. Conventional selective fusion-based RGB-D saliency methods (e.g., [20] and [17]) easily produce imperfect detection results for RGB-D images with low-quality depth. Thus, our method attempts to estimate a novel depth map, which can be regarded as as a complementary component to the original depth, to increase the detection performance.

difference in depths between those potentially salient regions from their non-salient surroundings, making the estimated depth map more suitable for the subsequent saliency detection. Then, both the original depth and the novel depth will be fed into our newly designed deep fully convolutional network to reveal image-level saliency clues. These revealed saliency clues will be further utilized to reduce the problem domain of our subsequent nonlocal selective deep fusion network, whose key focus is to ensure an effective complementary state between multichannel data while avoiding feature-conflict-induced learning ambiguity. Specifically, the major contributions of this paper can be summarized as follows:

- We propose a novel two-phase, exemplar-driven approach to the estimation of relatively trustworthy depth. Our newly estimated depth may potentially alleviate the learning ambiguity when both the RGB information and the depth information are incapable of separating the salient object from its non-salient surroundings.
- We design a novel deep selective saliency fusion network to compute an optimal complementary state between the RGB information, the original depth, and the newly estimated depth. Compared to the conventional end-to-end fusion schemes, our method makes full use of the image-level saliency assumptions to reduce the problem domain of nonlocal selective fusion, which is more suitable for robust RGB-D saliency detection.

II. Related Works

A. Saliency-Revealing Methods for the Depth Channel

RGB-D saliency detection is a rapidly growing field. In its early stages, studies consider depth as an additional information to formulate saliency clues to balance RGB saliency, which mainly focuses on the designation of the saliency-revealing strategy in the depth channel. Maki *et al.* [24] proposed to utilize an early computational model of depth-based attention by integrating stereo disparity, image flow, and motion. Zhang *et al.* [25] proposed to utilize both the depth contrast and the motion contrast to formulate stereoscopic visual attention for video data. Ju *et al.* [19] proposed to formulate depth saliency clues via the anisotropic center-surround difference to simultaneously enable the saliency-revealing scope covering both the fine-grained global structure and coarse-grained local detail. Then, Feng *et al.* [18] proposed to measure the background enclosure degree to represent the saliency of the target region in the depth channel. Additionally, to respect one important, common phenomenon of salient objects (i.e., salient regions are often characterized by an unusual surface orientation profile different from its nearby surroundings), Feng *et al.* [26] proposed to utilize the histogram of surface orientation to measure the corresponding saliency degree in the depth channel. Although many improvements have been made, the hidden core rationale of these methods [20], [27] still follow the general principle of conventional multiscale/multilevel/multidirectional depth contrast.

B. Handcrafted Fusion-Based Methods

Another branch of RGB-D saliency detection methods is fusion-based methods, and their rationale mainly focuses on pursuing an optimal complementary state between color and depth. Peng *et al.* [28] proposed to utilize the Bayesian integration scheme to fuse previously computed multistage saliency clues into one saliency prediction. Ren *et al.* [29] proposed to estimate multiple saliency priors from the depth channel, e.g., the background, depth, and orientation priors. Then, these priors are integrated with the color saliency to jointly boost the detection performance. More directly, Guo *et al.* [30] adopted the multiplicative-based fusion strategy to fuse RGB saliency clues with depth saliency clues. Although fusion-based RGB-D saliency detection methods can outperform conventional methods that rely on color information only, the possible performance improvement of these methods is limited because it is difficult for these handcrafted fusion strategies [31] to achieve an optimal balance between the RGB information and the depth information.

C. Deep Fusion-Based Methods

Benefiting from the recently developed deep learning frameworks, the RGB-D saliency fusion problem can now be selectively balanced; i.e., the final fused saliency map may be automatically biased toward to the RGB channel or the depth channel. Qu *et al.* [20] adopted a CNN to selectively fuse multiple previously handcrafted low-level saliency clues to achieve high-performing RGB-D saliency prediction. Similarly, Shigematsu *et al.* [17] extracted multiple mid-level, handcrafted depth features as inputs for a network with two fully connected layers to make the saliency fusion process more robust. By using the newly designed cross-modal residual function, Chen and Li [32] further improved the complementary state between the RGB and depth features from a multilevel, deep fusion perspective. Although many improvements have been made; for example, the deep learning framework-based methods easily reach the performance bottleneck, which is mainly caused by limited depth quality (i.e., either induced by the hardware limitations or external disturbances). Thus, we propose to follow the exemplar-driven approach [33] to ensure an acceptable depth quality, and then we feed it into

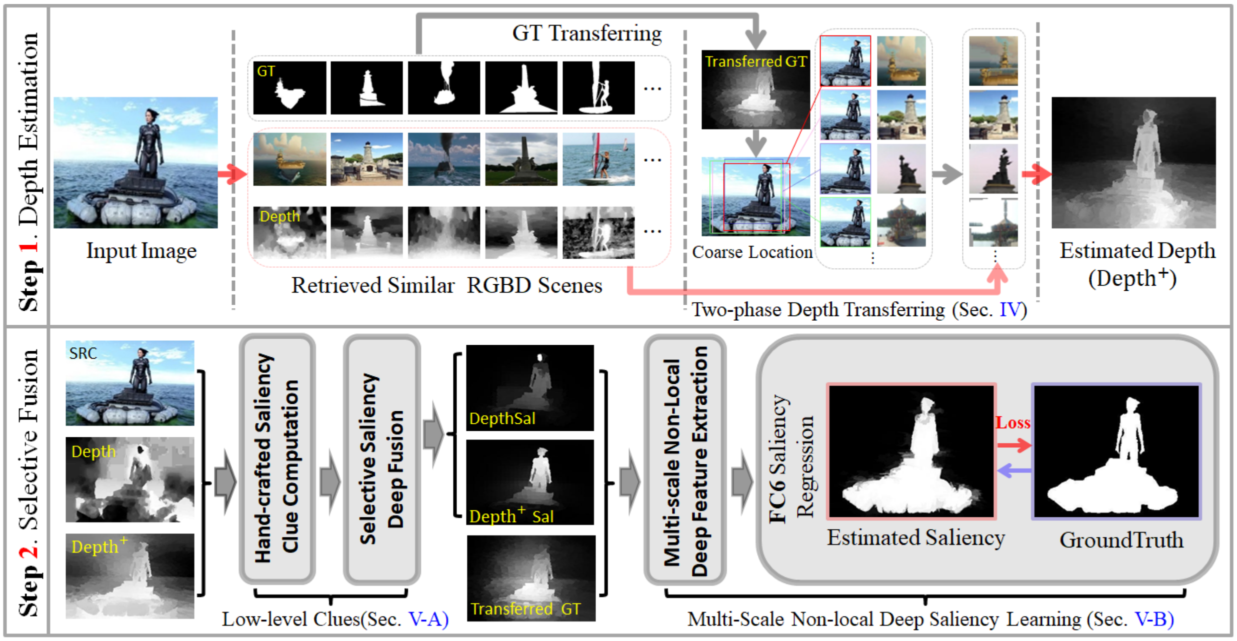
Step 1. Depth Estimation
Input Image
GT Transferring
GT
Depth
Retrieved Similar RGBD Scenes
Transferred GT
Coarse Location
Two-phase Depth Transferring (Sec. IV)
Estimated Depth (Depth+)
Step 2. Selective Fusion
SRC
Depth
Depth+
Hand-crafted Saliency Clue Computation
Selective Saliency Deep Fusion
DepthSal
Depth+ Sal
Transferred GT
Low-level Clues(Sec. V-A)
Multi-scale Non-Local Deep Feature Extraction
FC6 Saliency Regression
Loss
Estimated Saliency
GroundTruth
Multi-Scale Non-local Deep Saliency Learning (Sec. V-B)

Fig. 2. The functional pipeline of our method. Our method consists of two main steps. Step 1 estimates an additional depth map of the given input image; Step 2 utilizes our newly designed selective deep fusion network to achieve high-performance RGB-D saliency detection by obtaining an optimal fusion balance.

our newly designed selective deep fusion network to improve the RGB-D saliency detection performance. Moreover, due to overfitting, it is difficult for a deep learning-based depth estimation solution [34] to benefit RGB-D image saliency detection.

III. Method Overview

As shown in Fig. 2, our method mainly consists of two components: the depth estimation component and the selective deep fusion component. Given an input image, we utilize the common element to retrieve a subgroup of images sharing similar scenes to the given input. Then, for each image pair (i.e., the given input image and one of the retrieved subgroup images), we propose to coarsely build the inter-image regional correspondences by interactively considering the pixel-level similarity, the mid-level similarity and the object-level similarity. Once the coarse-level inter-image correspondences have been built, we use them in our coarse-level depth estimation (see Sec. IV-A). Based on the coarsely estimated depth, we further improve its quality by enlarging the depth differences between those potentially salient regions and its non-salient surroundings (see Sec. IV-B). Then, we utilize our newly designed selective deep saliency fusion network to achieve an optimal balance between the RGB information, the original depth information, and our newly estimated depth information, which will be further discussed in Sec. V.

IV. Two-Phase Depth Estimation

In [33], it has been proven that two images with similar scenes tend to have similar saliency layouts, and here, we assume that such rationale is also effective for the "depth channel", i.e., two images with similar RGB topology should have similar overall depth layouts. Thus, we adopt GIST features to retrieve subgroup RGB-D images with similar overall scene layouts (i.e., top-K images). That is, 1) we compute the GIST feature for each image in our database; and 2) we rank all the images in our database using the L2 distance in the GIST features between the input image and the other images in the database. Then, we determine dense inter-image correspondences between the given image and each its retrieved subgroup image (see Sec. IV-A) and use it to transfer depth information as the coarsely estimated depth layout of the given image.

Specifically, the depth value of salient objects between two aligned RGB-D images may be different, and the depth map produced using the depth-transferring strategy tends to obscure the area around the salient object (see Depth$^-$ in Fig. 3), and it is difficult to use the depth map to effectively benefit the salient object detection problem. Therefore, based on the previously estimated coarse depth layout, we further conduct another round of depth estimation (see Sec. IV-B) to pursue a distinct depth layer for the salient object compared to its non-salient nearby surroundings (see Depth$^+$ in Fig. 3).

A. Coarse-Level Depth Estimation

To transfer depth information from the retrieved subgroup of RGB-D images, we need to build the inter-image correspondences beforehand. Moreover, because regions belonging to an identical object tend to exhibit similar depth values, we simultaneously use both the object-level homogeneity and the non-local similarity to build the inter-image correspondences. Given an input RGB-D image, we initially adopt the classic SLIC [35] algorithm to conduct mid-level superpixel decomposition and use the EdgeBox [36] method to propose object-level rectangles. Then, the inter-image correspondences can be built by interactively conducting the binary alignments between superpixels and object proposals over the classic SIFT flow method [37], provided pixel-wise correspondences.

In general, the SIFT flow algorithm consists of matching densely sampled, pixel-wise SIFT features between two images while preserving spatial discontinuities. To achieve this goal, the SIFT flow algorithm converts the matching problem into a binary alignment problem by solving the energy minimization problem. Therefore, the dense binary matching from the SIFT flow algorithm may be slightly different from the conventional image matching problem, in which the former does not need to use a predefined hard threshold to produce the pixel-level correspondence. Meanwhile, because the SIFT flow part is slightly beyond the main focus of our work, we directly use the classic, simple SIFT flow algorithm [37] as the preprocessing tool to obtain the initial inter-image dense correspondences.

The pixel-level correspondences provided by the SIFT flow method can be represented by matrix $\mathbf{Q}_p \in \{0,1\}^{W\times H, W\times H}$, where W and H represent the image width and image height, respectively, and the non-zero elements in $\mathbf{Q}_p$ represent the existence of pixel-wise correspondence.

Thus, based on pixel-wise correspondences $\mathbf{Q}_p$, we adopt the majority voting strategy to measure the superpixel-level similarity $S \in [0,1]^{N\times N}$ and then use it to formulate the mid-level inter-image correspondences; i.e., $\mathbf{Q}_{sp} \in \{0,1\}^{N\times N}$, where N represents the superpixel number and $\mathbf{Q}_{sp}$ is the column-wise correspondence subject to $\mathbf{Q}_{sp} \times 1^{N\times 1} = 1^{N\times 1}$. Here, we formulate the superpixel-wise similarity (S) as Eq. 1.

$$S(sp_i^{\mathbf{I}}, sp_j^{\mathbf{IS}}) = \frac{num\{\xi(\mathbf{Q}_p, i) \cap \xi(\mathbf{Q}_p, j)\}}{num\{\xi(\mathbf{Q}_p, i) \cup \xi(\mathbf{Q}_p, j)\}}, \tag{1}$$

where the superscripts **I** and **IS** denote the input image, respectively, and one of its retrieved subgroup images and $sp_i^{\mathbf{I}}$ denotes the i-th superpixel in the input image; function $\xi(\mathbf{Q}_p, i)$ returns the pixel-wise correspondences of the i-th superpixel; function $num\{\cdot\}$ returns the total number of elements in its input.

Based on the superpixel-wise similarity (i.e., function S in Eq. 1), the superpixel-level correspondences $\mathbf{Q}_{sp}$ can be formulated as:

$$\mathbf{Q}_{sp}(i,j) \leftarrow \begin{cases} 1, & if \ \ S(sp_i^{\mathbf{I}}, sp_j^{\mathbf{IS}}) = \max \underbrace{\{S(sp_i^{\mathbf{I}}, sp_k^{\mathbf{IS}})\}}_{k\in\{1,2,\ldots,N\}} \\ 0, & otherwise. \end{cases} \tag{2}$$

Once the superpixel-level inter-image correspondences ($\mathbf{Q}_{sp}$ in Eq. 2) have been computed, we can easily build the object-level inter-image correspondences (Eq. 4) using the object-level similarity. Here, we represent the object-level inter-image correspondences as $\mathbf{Q}_{ob} \in \{0,1\}^{M\times M}$, where M represents the object proposal number (empirically set to 10) and $\mathbf{Q}_{ob}$ is also column-wisely subject to $\mathbf{Q}_{ob} \times 1^{M\times 1} = 1^{M\times 1}$. We formulate the object-level similarity $O \in [0,1]^{M\times M}$ as Eq. 3.

$$O(o_i^{\mathbf{I}}, o_j^{\mathbf{IS}}) = \frac{\zeta(\mathbf{Q}_{sp}, i) \cap \zeta(\mathbf{Q}_{sp}, j)}{\zeta(\mathbf{Q}_{sp}, i) \cup \zeta(\mathbf{Q}_{sp}, j)}, \tag{3}$$

where $o_i^{\mathbf{I}}$ denotes the i-th object proposal in the given image **I** and function $\zeta(\mathbf{Q}_{sp}, i)$ returns the previously established superpixel-wise inter-image correspondences of the i-th object proposal. Thus, we represent the object-level correspondence matrix $\mathbf{Q}_{ob}$ as follows:

$$\mathbf{Q}_{ob}(i,j) \leftarrow \begin{cases} 1, & if \ \ O(o_i^{\mathbf{I}}, o_j^{\mathbf{IS}}) = \max \underbrace{\{O(o_i^{\mathbf{I}}, o_k^{\mathbf{IS}})\}}_{k\in\{1,2,\ldots,M\}} \\ 0, & otherwise. \end{cases} \tag{4}$$

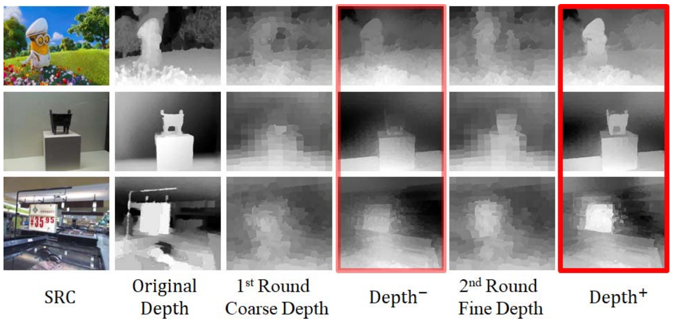

Fig. 3. Depth quality improvements achieved by our two-phase depth estimation.

Specifically, instead of directly using the pixel-wise correspondences $\mathbf{Q}_p$ provided by the SIFT flow method, we choose to use the mid-level superpixel correspondence matrix $\mathbf{Q}_{sp}$ because of the following reasons:

1) the mid-level decomposition (i.e., superpixels) can effectively preserve the boundary information to produce robust object-level correspondences;

2) it can effectively alleviate the computational burden.

Thus far, the object-level inter-image correspondences $\mathbf{Q}_{ob}$ are already capable of guiding the transference of depth values between two aligned object proposals in the object-level depth estimation. However, due to the absence of local details, it is difficult to use the object-level depth to improve the saliency detection performance. To improve it, for each pairs of already aligned object proposals, we recursively conduct the superpixel-level realignment to obtain accurate inter-image correspondences, which can be formulated as a minimization problem:

$$\begin{aligned} & if \ \mathbf{Q}_{ob}(i,j) = 1, \\ & \quad arg \min_{\mathbf{Q}_f} || \sum_{k\in o_i^{\mathbf{I}}} \sum_{l \in o_j^{\mathbf{IS}}} S(sp_k^{\mathbf{I}}, sp_l^{\mathbf{IS}}) \times \mathbf{Q}_f(k,l)||_2, \\ & \quad s.t., \ \mathbf{Q}_f \times 1^{d\times 1} = 1^{d\times 1}, \end{aligned} \tag{5}$$

where $\mathbf{Q}_f \in \{0,1\}^{d\times d}$ represents the final inter-image, non-local correspondences, $d = \max\{\text{size}(o_i^{\mathbf{I}}), \text{size}(o_j^{\mathbf{IS}})\}$, function size$(\cdot)$ returns the superpixel number of the corresponding object proposal and the similarity measurement $S(\cdot)$ is identical to that in Eq. 1. We also fill in the missing elements in $\mathbf{Q}_f$ with dummy nodes. In fact, Eq. 5 can be efficiently solved by the Hungarian algorithm [38] in polynomial time. To this end, we coarsely estimate the depth information (cD) by using the correspondences provided by matrix $\mathbf{Q}_f$, which can be formulated as:

$$\mathrm{cD}_u = \frac{1}{Z}\sum_{k=1}^{K}[\sum_i \sum_j \mathbf{Q}_{ob}^k(i,j) \cdot \underbrace{\{\sum_v \mathbf{Q}_f^k(u,v)\cdot \mathrm{D}_v^k\}}_{\text{only if } sp_u \in o_i^{\mathbf{I}} \text{ and } sp_v \in o_j^{\mathbf{IS}}}] \ , \quad (6)$$

where cD_u denotes the coarsely estimated depth information of the u-th superpixel, Z denotes the total transferred time, D_v^k denotes the original depth value of the v-th superpixel in **IS**, and K is the total number of images in **IS**.

We also show the qualitative demonstrations of the coarsely estimated depth in the 1st-round coarse depth column of Fig. 3. The postprocessed results can be found in the Depth^- column in Fig. 3, which is obtained by applying a spatial weighting scheme (we follow the suggestion of [27]) over the estimated coarse depth cD.

B. Second-Round Depth Estimation

In general, the coarsely estimated depth (cD, Eq. 6) is generally reasonable, yet the depth quality (Depth^-) may still be incapable of effectively benefiting the subsequent RGB-D saliency detection due to its obscured local details. Moreover, the quality of cD also heavily relies on the following 3 aspects:

1) the original depth quality of the retrieved sub-group RGB-D images (**IS**);

2) the similarity degree between the given input RGB-D image (**I**) and the retrieved subgroup of RGB-D images (**IS**);

3) the accuracy of the SIFT flow algorithm provided pixel-level inter-image correspondences ($\mathbf{Q}_p$).

Since the former 2 aspects are objectively determined by either the RGB-D image database or the depth-sensing equipment, here we mainly focus on the last aspect to further improve the quality of our estimated depth map.

In general, the SIFT flow method itself is heavily sensitive to both scale variations and movement (or view-angle-change)-induced displacements. Thus, the SIFT flow method provides pixel-level inter-image correspondences ($\mathbf{Q}_p$) that may become untrustworthy if the salient object undergoes fast movement coupling with extensive scale variation. Therefore, based on the previously estimated coarse depth map cD, we perform another round of depth transference, which is saliency-aware via enlarging the different in depth between those potentially salient regions and its non-salient nearby surroundings.

To locate potentially salient regions, here we formulate our coarse-level saliency (cS) via replacing the last element $\mathrm{D}_v^{\mathbf{IS}}$ (Eq. 6) with the corresponding saliency ground truth of **IS** (salGT$\in \{0, 1\}$); see Eq. 7, in which all other components are identical to those in Eq. 6. In fact, the computation of cS can be regarded as a byproduct of our coarse-level depth estimation, which has almost no additional computational costs.

$$\mathrm{cS}_u = \frac{1}{Z}\sum_{k=1}^{K}[\sum_i \sum_j \mathbf{Q}_{ob}^k(i,j) \cdot \underbrace{\{\sum_v \mathbf{Q}_f^k(u,v)\cdot \mathrm{salGT}_v^k\}}_{\text{only if } sp_u \in o_i^{\mathbf{I}} \text{ and } sp_v \in o_j^{\mathbf{IS}}}], \quad (7)$$

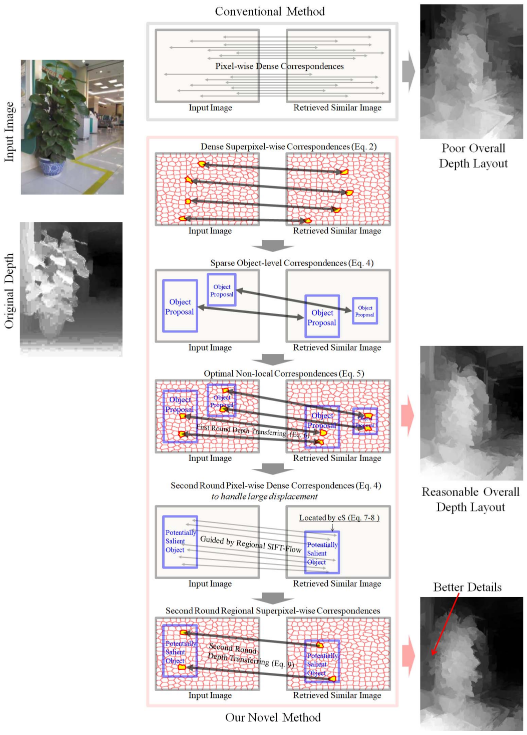
Conventional Method
Pixel-wise Dense Correspondences
Input Image
Retrieved Similar Image
Poor Overall Depth Layout
Input Image
Original Depth
Dense Superpixel-wise Correspondences (Eq. 2)
Input Image
Retrieved Similar Image
Sparse Object-level Correspondences (Eq. 4)
Object Proposal
Object Proposal
Object Proposal
Object Proposal
Input Image
Retrieved Similar Image
Optimal Non-local Correspondences (Eq. 5)
Object Proposal
Object Proposal
First Round Depth Transferring (Eq. 6)
Object Proposal
Object Proposal
Input Image
Retrieved Similar Image
Reasonable Overall Depth Layout
Second Round Pixel-wise Dense Correspondences (Eq. 4)
to handle large displacement
Located by cS (Eq. 7-8)
Potentially Salient Object
Guided by Regional SIFT-Flow
Potentially Salient Object
Input Image
Retrieved Similar Image
Better Details
Second Round Regional Superpixel-wise Correspondences
Potentially Salient Object
Second Round Depth Transferring (Eq. 9)
Potentially Salient Object
Input Image
Retrieved Similar Image
Our Novel Method

Fig. 4. Pictorial demonstration toward the main steps to build inter-image correspondences via interactively considering both the "object-level homogeneity" and the "non-local similarity".

The pictorial demonstrations toward the transferred saliency (cS) can be found in the transferred GT in Fig. 2. Then, based on cS, we perform object proposal ranking via Eq. 8 to select one rectangular box (so) to coarsely locate the salient object.

$$so^{\mathbf{I}} \leftarrow arg\max_{o_i} \frac{||\mathrm{G} \odot \mathrm{Rect}(\mathrm{cS}, i)||_1}{||\mathrm{G}||_1} + \lambda_1 \cdot \mathrm{Score}_i, \quad (8)$$

where o_i denotes the i-th EdgeBox-provided object proposal [36] in the given input RGB-D image, function $\mathrm{Rect}(cS, i)$ returns the transferred saliency value of the i-th object proposal, G represents the Gaussian function with $\sigma = 0.1$ of identical size to the given object proposal, Score_i represents the objectness score of the i-th object proposal, and the balance parameter λ_1 can be automatically selected via performing parameter optimization (method of least squares) over the dataset **IS**.

Due to the availability of the salient ground truths, for each RGB-D image in the retrieved subgroup **IS**, the salient objects can also be precisely located and then tightly warped by rectangular boxes, and we denote these boxes by $so^{\mathbf{IS}}$.

Once both the salient object in the input RGB-D image (**I**) and the salient ground truths in the retrieved subgroup images (**IS**) have been obtained, we will conduct second-round depth transference to enhance the estimated depth, which can be formulated by Eq. 9.

$$\mathrm{fD}(sp_u) = \frac{1}{Z}\sum_{k=1}^{K} \underbrace{\{\sum_v \mathbf{Q}_{sal}^k(u,v)\cdot \mathrm{D}_v^{\mathbf{IS}}\}}_{\text{only if } sp_u \in \mathrm{so}^{\mathbf{I}} \text{ and } sp_v \in \mathrm{so}_{\mathrm{i}}^{\mathbf{IS}}} \ , \quad (9)$$

where K is the total number of images in **IS**, $\mathrm{fD}(sp_u)$ denotes the transferred depth information of the u-th superpixel

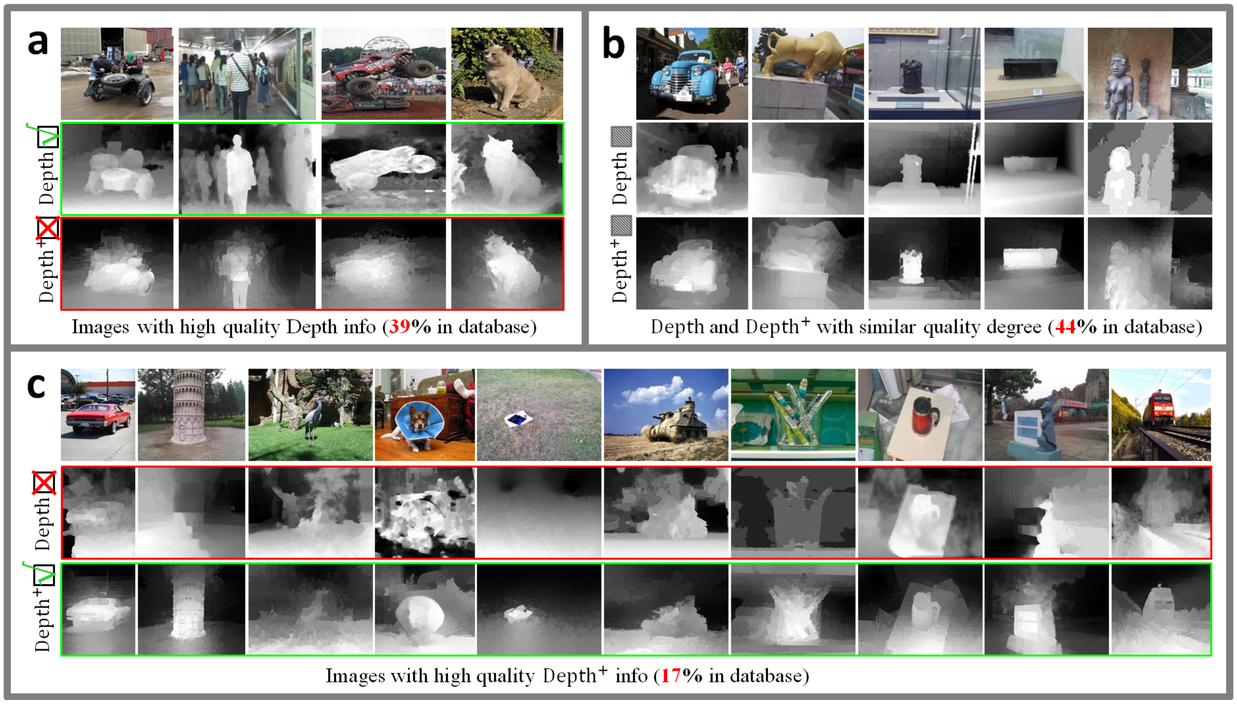
a
Depth
Depth+
Images with high quality Depth info (39% in database)
b
Depth
Depth+
Depth and Depth+ with similar quality degree (44% in database)
c
Depth
Depth+
Images with high quality Depth+ info (17% in database)

Fig. 5. Depth-quality comparisons between the original depth (Depth) and our newly estimated depth (Depth$^+$). By feeding Depth and Depth$^+$ to the simple RGB-D deep saliency model [17], we regard Depth quality>Depth$^+$ quality only if the Depth-corresponding maximum F-measure (saliency predictions) is larger than the Depth$^+$-corresponding maximum F-measure, and vice versa.

(belonging to the coarsely located salient object rectangle $so^{\mathbf{I}}$ from Eq. 8), and $\mathbf{Q}^k_{sal}$ denotes the superpixel-level correspondences between the target rectangle $so^{\mathbf{I}}$ and the k-th retrieved rectangles $so^{\mathbf{IS}}$. The detailed steps can be summarized as follows:

1) perform a new round of pixel-wise SIFT flow alignments between $so^{\mathbf{I}}$ and $so^{\mathbf{IS}}$ to obtain $\mathbf{Q}_p$;

2) based on $\mathbf{Q}_p$, we further compute the superpixel-level correspondences (similar to Eq. 2) to obtain $\mathbf{Q}_{sal}$ in Eq. 9.

To this end, we directly formulate the final estimated depth information finalD via simply applying additive fusion as follows:

$$\text{finalD} \leftarrow \text{cD(Eq. 6)} + \text{fD(Eq. 9)}. \tag{10}$$

Here, we will provide a pictorial demonstration (Fig. 4) to review the main steps raised above. As shown in the top row of Fig. 4, the pixel-wise inter-image correspondence-based depth transference may easily lead to an estimated depth map with poor overall layout, which is mainly induced by the lower consideration of the mid-level depth homogeneity. To address this problem, we build inter-image correspondences by interactively considering object-level homogeneity and nonlocal similarity. Thus, we can output a depth map with a much more reasonable overall depth layout (see the red arrow in the middle row of Fig. 4). Then, we use second-round depth transference to further improve the details of the estimated depth (see the pictorial demonstration in the bottom-left corner of Fig. 4).

C. Saliency-Aware Depth Enhancement

From the RGB-D saliency-revealing perspective, a desired finalD should simultaneously possess the following attributes:

1) the estimated depth value of the salient object should be different from its non-salient nearby surroundings;

2) regions inside the salient object should have similar depth values.

To ensure these attributes, we propose to take full advantage of the transferred saliency (cS, Eq. 7) to further enhance the depth quality. That is, we update finalD (Eq. 10) as follows:

$$\begin{aligned}\text{finalD}(sp_i) \leftarrow\ & 0.5 \cdot \text{finalD}(sp_i) \\ & + \frac{(1-0.5)\cdot sign(\text{D}_{\text{FG}}-\text{D}_{\text{BG}})}{1+exp(-abs(\text{D}_{\text{FG}}-\text{D}_{\text{BG}}))},\end{aligned} \tag{11}$$

where D_{FG} denotes the spatially weighed depth of the i-th superpixel sp_i (see Eq. 12) and D_{BG} denotes the transferred depth toward sp_i non-salient nearby surroundings (see Eq. 13).

$$\text{D}_{\text{FG}}(sp_i) = \sum_{j\in\theta} \frac{exp\{-\omega_1 \cdot \text{CDist}(sp_i, sp_j)\} \cdot \text{finalD}(sp_j)}{exp\{-\omega_1 \cdot \text{CDist}(sp_i, sp_j)\}}, \tag{12}$$

where θ determines the nonlocal nearby neighborhood (with a superpixel center coordinate L2 distance ≤ 45) of the i-th superpixel sp_i. Function $\text{CDist}(sp_i, sp_j)$ returns the L2 RGB distance between two given superpixels, and ω_1 is a weighting parameter.

$$\text{D}_{\text{BG}}(sp_i) = \sum_{j\in\theta} \frac{exp\{\omega_2 \cdot \text{SDist}(sp_i, sp_j\}) \cdot \text{finalD}(\text{sp}_\text{j})}{exp\{\omega_2 \cdot \text{SDist}(sp_i, sp_j)\}}, \tag{13}$$

where ω_2 is another weighting parameter and function $\text{SDist}(sp_i, sp_j)$ returns the L2 saliency distance (i.e., cS, Eq. 7) between two given superpixels.

We demonstrate the updated finalD in the second-round fine depth column of Fig. 3. We also adopt the commonly used

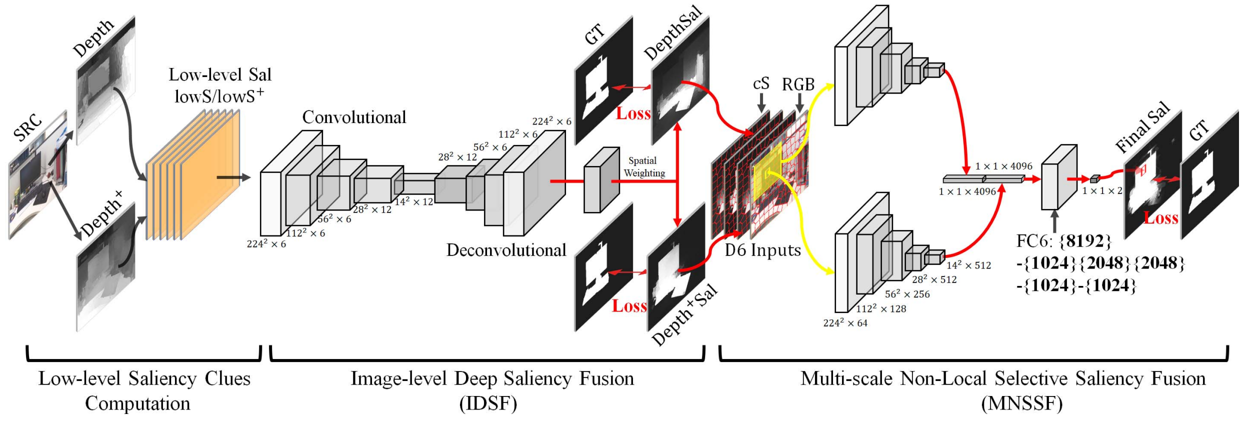

Fig. 6. The architecture of our newly designed network, which mainly consists of two subnetworks: the **IDSF** network and **MNSSF** network.

spatial weighting scheme [12], [39] to postprocess the newly estimated finalD, and the depth quality differences between the unsmoothed depth maps (i.e., the first-round coarse depth vs the second-round fine depth) and the smoothed depth maps (i.e., Depth^{-} and Depth^{+}) can be found in Fig. 3.

Specifically, we summarize the rationale of our saliency-driven second-round depth enhancement as follows:

1) we utilize the objectness prior to handling the common scale/displacement-sensitive limitation of the adopted SIFT flow method;

2) we further enlarge the depth difference between the color-similarity-indicated regions (Eq. 12) and the cS (Eq. 7)-determined non-salient nearby surroundings (Eq. 13).

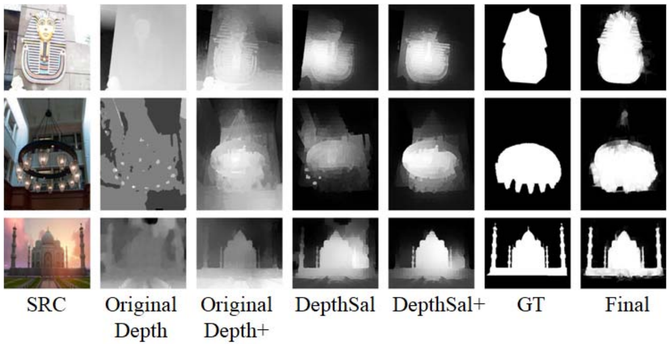

Fig. 7. Pictorial demonstration of our lowS (Eq. 14) and LowS^{+} (Eq. 15) clues; the final saliency predictions can be found in the last column.

V. Computation of Deep Saliency and Selective Fusion

Thus far, we have already obtained the estimated depth map (i.e., Depth^{+}). As we mentioned before, the newly estimated Depth^{+} is informatively complementary to the original depth information (denoted by Depth). Here, we demonstrate the quality quantitative comparisons between the Depth and the Depth^{+} in Fig. 5. As shown in Fig. 5, almost 17% of the time, Depth^{+} outperforms Depth in RGB-D saliency detection. To achieve an optimal complementary state between Depth and Depth^{+}, we propose to reveal the low-level saliency clues from Depth and Depth^{+} first. These newly revealed low-level saliency clues will jointly improve the RGB-D saliency detection via our newly designed selective fusion network.

A. Image-Level, Low-Level Saliency Fusion

Here, we will introduce the low-level saliency computation to formulate the saliency clues either from the RGB or the Depth/Depth^{+} features. To achieve this computation, we propose to adopt three commonly used handcrafted features, including the local/global/boundary contrast computation [27], to reveal the low-level saliency clues. We represent these computed low-level saliency clues (i.e., lowS and lowS^{+}) in Eq. 14 and Eq. 15.

$$\begin{aligned}\text{lowS} = \{&\text{loc(RGB)},\ \text{glo(RGB)},\ \text{bou(RGB)},\\ &\text{loc(Depth)},\ \text{glo(Depth)},\ \text{bou(Depth)}\},\end{aligned} \tag{14}$$

$$\begin{aligned}\text{lowS}^{+} = \{&\text{loc(RGB)},\ \text{glo(RGB)},\ \text{bou(RGB)},\\ &\text{loc(Depth}^{+}),\ \text{glo(Depth}^{+}),\ \text{bou(Depth}^{+})\},\end{aligned} \tag{15}$$

where loc(·) denotes the local contrast computation with a contrast computation range of 50, glo(·) denotes the global contrast computation, and bou(·) represents the boundary contrast [40]. It should also be noted that we conduct all the abovementioned low-level saliency computations over the mid-level superpixels, and the component qualitative demonstrations can be found in Fig. 7.

Since these low-level saliency clues are complementary to each other, we propose to fuse lowS and lowS^{+} into single saliency maps (i.e., DepthSal and Depth^{+}Sal) using our **IDSF** (image-level deep saliency fusion) network (see the network structure in Fig. 6). Our **IDSF** network receives 6-channel, low-level saliency clues as inputs. All these inputs are fed into 8 fully convoluted layers to automatically conduct selective saliency fusion by regressing the hidden parameters toward the given saliency ground truth:

$$\text{DepthSal} = SW\{dec[con(lowS, \Theta_c), \Theta_d]\}, \tag{16}$$

$$\text{Depth}^{+}\text{Sal} = SW\{dec[con(lowS^{+}, \Theta_c), \Theta_d]\}, \tag{17}$$

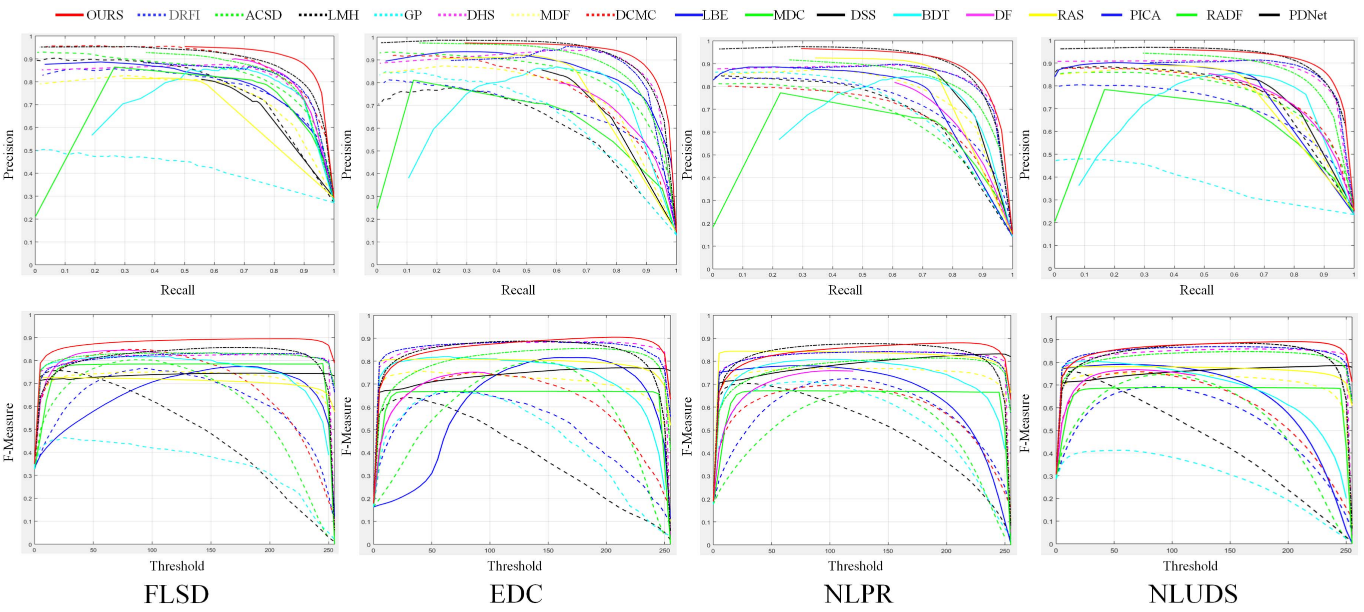

Fig. 8. The quantitative comparisons between our method and 16 state-of-the-art methods including ACSD14 [19], LMH14 [28], GP15 [29], DHS16 [41], DCMC16 [21], LBE16 [18], MDF16 [11], DRFI17 [42], MDC17 [43], DSS18 [10], BDT17 [17], DF17 [20], RAS18 [44], PICA18 [45], PDNet18 [46], and RADF18 [47] in the NLPR [28], NJUD [19], DEC [48] and FLSD [49] datasets. The first row shows the PR curves, the second row demonstrates the F-measure curves, and the third row demonstrates the corresponding precision, recall, and F-measure after using an adaptive binary threshold.

where $SW\{\cdot\}$ denotes the common thread superpixel-wise spatial weighting operation, which is identical to the one adopted in Sec. IV-B, aiming to sharpen the depth boundary. dec and con represent the deconvolution layers and the convolutional layers, respectively. All the parameters of the convolution (Θ_c) and deconvolution (Θ_d) layers are learnable.

Thus far, we have already obtained the Depth-based saliency assumption DepthSal and the Depth$^+$-based saliency assumption Depth$^+$Sal, and we propose to further perform nonlocal selective saliency fusion in a multiscale manner to jointly improve the RGB-D saliency detection performance via simultaneously feeding both DepthSal and Depth$^+$Sal into our newly designed deep network **MNSSF** (multiscale, nonlocal selective saliency fusion); see the demonstrations in the middle-bottom row of Fig. 2.

B. Multiscale, Nonlocal Deep Saliency Network

Since our newly estimated depth information, Depth$^+$, is complementary to the original depth information, Depth, there also exists a similar complementary relationship between the newly computed image-level saliency, i.e., DepthSal and Depth$^+$Sal. Moreover, since the previously obtained cS (Eq. 7) can coarsely locate the salient object, we utilize it to reduce the RGB-D saliency revealing problem domain. Here, we propose to utilize multiscale, nonlocal deep features coupled with one strong regressor to enlarge the feature margin between the salient foregrounds and its non-salient nearby surroundings.

We formulate the inputs of our deep network **MNSSF** as 6-channel data (D6, Eq. 18), i.e., DepthSal, Depth$^+$Sal, cS, and RGB. We follow the [11]-proposed multiscale (local and nonlocal deep feature concatenation, see the yellow rectangles in Fig. 6), superpixel-wise VGG16 deep features (with dimensions of 4096+4096) to represent the non-local contrast, see Eq. 18.

$$\mathrm{L_i} \rightleftharpoons \mathrm{FC6(F_i)},\ \mathrm{F_i} \leftarrow cat\{\Theta_{\mathrm{loc}}(\mathrm{sp_i}, \mathrm{D6}), \Theta_{\mathrm{non}}(\mathrm{sp_i}, \mathrm{D6})\}, \quad (18)$$

where L represents the binary saliency ground truths, FC6 denotes the 6-layer fully connected network, $cat\{\cdot, \cdot\}$ is the concatenation operation, $\Theta_{loc/non}$ denotes the local/nonlocal VGG16 deep features, and sp_i and D6 denote the i-th superpixel corresponded multi-scale topology and the 6 channel inputs, respectively. To achieve multiscale detections, we vary the superpixel decomposition size in both the training and testing stages to further increase the detection result robustness. The detailed network architecture can be found in Fig. 6, and more implementation details can be found in the next section.

VI. Experiments and Discussion

A. Implementation Details

We implement our method using MATLAB 2016b with CAFFE. We oversegment all the input RGB-D images into two scales; i.e., we initially assign the number of superpixels $\{300, 500\}$ to handle the scale problem. We empirically assign the number of retrieved subgroup RGB images K (Sec. IV), the weighing parameter ω_1 (Eq. 12) and ω_2 (Eq. 13) as 10, 0.01, and 0.1, respectively. All of these parameters are fixed throughout all the experiments. All the evaluations were conducted on a workstation with an NVIDIA GTX 1080Ti GPU and an Intel Xeon W-2133 CPU (6 cores with 12 threads) and 32 GB RAM. Meanwhile, we randomly select 1500 RGB-D images from the adopted datasets for training and use the remaining data for testing. We trained the network for 50k iterations using stochastic gradient descent (SGD) with a

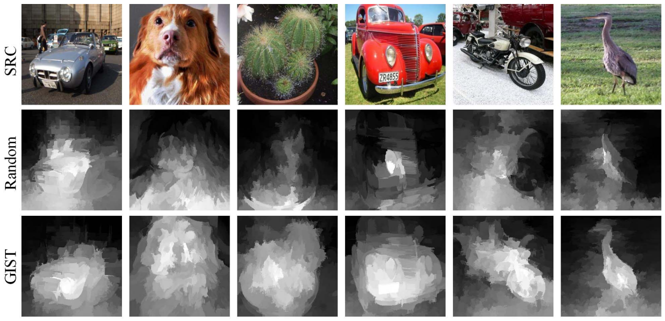

Fig. 9. Qualitative comparisons between the GIST-based image retrieval strategy and the “randomly picking one” scheme.

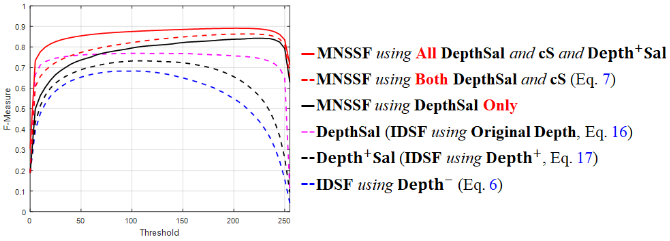

Fig. 10. Component evaluation results (i.e., the F-measure curve) in the NJUD [19] dataset.

TABLE I

QUANTITATIVE COMPARISONS BETWEEN THE GIST-BASED IMAGE RETRIEVAL STRATEGY AND THE “RANDOMLY PICKING ONE” SCHEME IN THE 4 PUBLICLY AVAILABLE DATASETS. THE ADOPTED EVALUATION METRICS ARE FROM [50]

	Metric	maxF	S-measure	MAE	adpEm	meanEm	maxEm	adpFm	meanFm
FLSD	**GIST**	0.889	0.887	0.057	0.904	0.910	0.928	0.864	0.857
	Rand	0.662	0.553	0.333	0.405	0.518	0.757	0.627	0.499
DEC	**GIST**	0.897	0.900	0.032	0.917	0.928	0.971	0.794	0.826
	Rand	0.735	0.665	0.174	0.647	0.663	0.897	0.531	0.510
NLPR	**GIST**	0.877	0.881	0.038	0.883	0.906	0.952	0.758	0.800
	Rand	0.667	0.592	0.232	0.551	0.577	0.833	0.479	0.440
NJUD	**GIST**	0.888	0.888	0.050	0.888	0.912	0.940	0.841	0.845
	Rand	0.615	0.531	0.346	0.422	0.505	0.733	0.570	0.467

TABLE II

QUANTITATIVE PROOFS OF THE PERFORMANCE DEGRADATION AFTER REPLACING THE ESTIMATED DEPTH MAP Depth$^+$ BY A COARSELY ESTIMATED SALIENCY MAP. THE ADOPTED EVALUATION METRICS ARE FROM [50]

	Metric	maxF	S-measure	MAE	adpEm	meanEm	maxEm	adpFm	meanFm
FLSD	**Depth+**	0.889	0.887	0.057	0.904	0.910	0.928	0.864	0.857
	cS	0.825	0.829	0.088	0.873	0.857	0.878	0.821	0.806
DEC	**Depth+**	0.897	0.900	0.032	0.917	0.928	0.971	0.794	0.826
	cS	0.880	0.894	0.037	0.919	0.922	0.959	0.797	0.822
NLPR	**Depth+**	0.877	0.881	0.038	0.883	0.906	0.952	0.758	0.800
	cS	0.873	0.881	0.038	0.886	0.906	0.950	0.761	0.801
NJUD	**Depth+**	0.888	0.888	0.050	0.888	0.912	0.940	0.841	0.845
	cS	0.811	0.825	0.089	0.849	0.851	0.877	0.790	0.783

moment of 0.9, a weight decay of 0.005, and a learning rate of 0.001.

B. Datasets

We conducted many quantitative experiments to validate the effectiveness of our method. We also compared our method to 16 state-of-the-art methods in 4 publicly available RGB-D datasets (NLPR [28], NJUD [19], DEC [48] and FLSD [49]) to demonstrate the advantages of our method. The NLPR dataset contains 1000 RGB-D images captured by a Microsoft Kinect device in both indoor and outdoor scenes. The NJUDS dataset contains 2000 RGB-D images. The depth maps are synthesized using an optical flow method. The LFSD dataset contains 100 RGB-D images with depth information captured by a Lytro light field camera. The DEC dataset contains 135 RGB-D indoor images sensed by a Microsoft Kinect device.

C. Component Evaluation

To validate the effectiveness of our method, we perform the component evaluation via F-measure curves over the entire testing dataset (see the details in Fig. 10). As shown in Fig. 10, the combination directly feeding Depth$^-$ into the **IDSF** (image-level deep saliency fusion) network exhibits the worst performance. Then, by replacing Depth$^-$ with Depth$^+$, the saliency detection performance can be remarkably improved because Depth$^+$ can effectively enlarge the depth difference between those potentially salient regions and their non-salient nearby surroundings. It should also be noted that the original Depth can generally outperform Depth$^+$, which is consistent with our previous quantitative comparison of the depth quality between Depth and Depth$^+$ (Fig. 5). Specifically, a significant performance improvement can be easily observed after introducing the **MNSSF** network (using DepthSal only) to conduct nonlocal selective saliency fusion (see the black solid line in Fig. 10). Additionally, by feeding the previously transferred saliency map cS (Eq. 7) into the **MNSSF** network, the performance can be further improved slightly (see the red dashed line in Fig. 10). Moreover, the overall performance can be further improved by simultaneously utilizing DepthSal, Depth$^+$Sal and cS to achieve an optimal complementary state.

Moreover, we quantitatively tested the performance of the “randomly picking one” strategy (see the details in Fig. 9). In fact, the “randomly picking one” strategy significantly degrades the performance of our method. Additionally, the quantitative proofs can be found in Table. I.

Additionally, we conducted a quantitative evaluation by replacing the estimated Depth$^+$ with the coarse saliency map (cS) as the input of our network, and the detailed result can be found in Table. II.

As shown in Table. II, the overall performance using Depth$^+$ significantly outperforms the coarse saliency alternation.

Thus, rather than directly feeding the coarse saliency map into our IDSF network, we take only the coarse saliency to facilitate the depth quality estimation, for example, to enable the second-round regional SIFT flow alignment and then highlight the salient object in the estimated depth map.

D. Quantitative Comparisons

We compare the proposed algorithm with 16 of the most representative state-of-the-art methods, including deep learning-based methods as well as other non-deep competitors: ACSD14 [19], LMH14 [28], GP15 [29], DHS16 [41], DCMC16 [21], LBE16 [18], MDF16 [11], DRFI17 [42], MDC17 [43], DSS18 [10], BDT17 [17], DF17 [20],

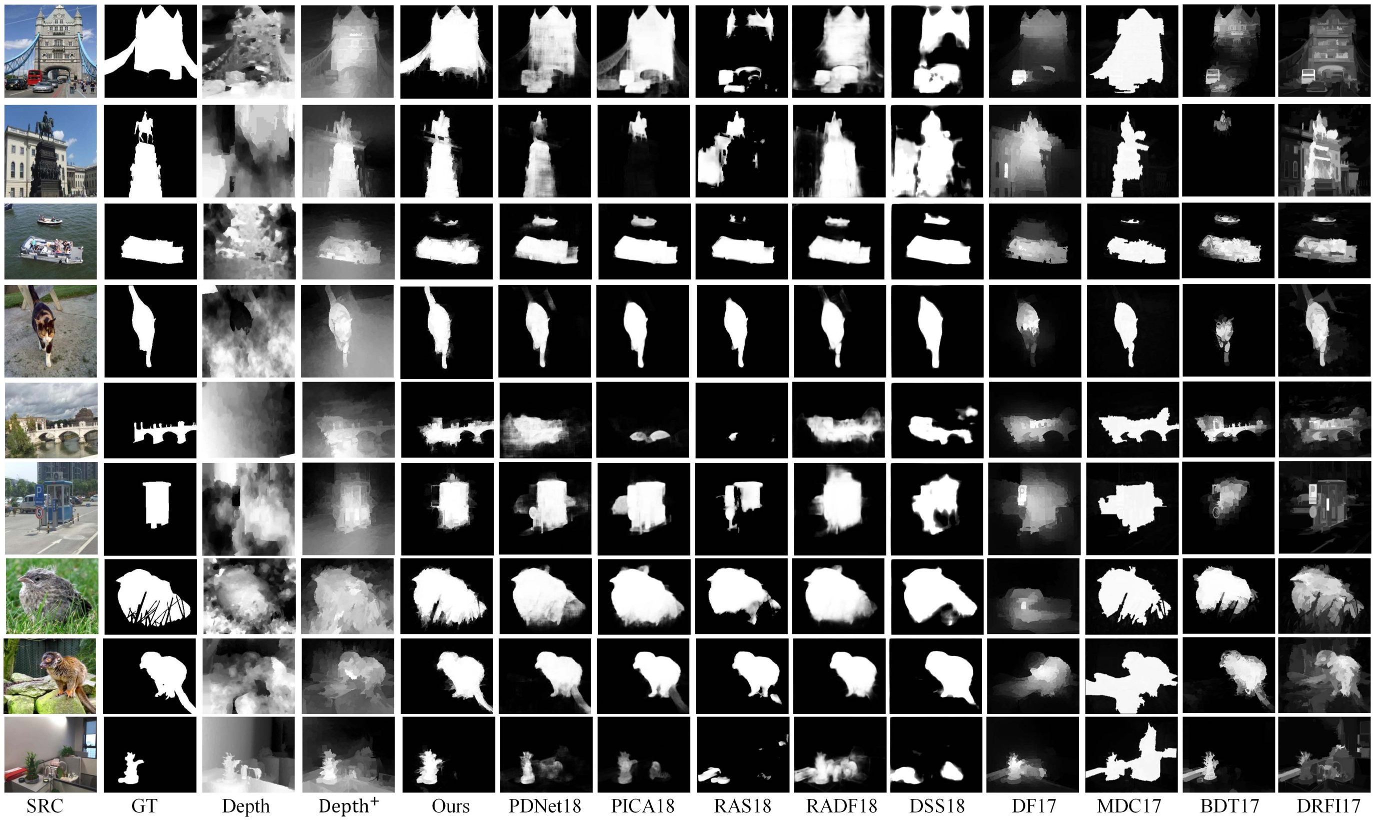

Fig. 11. Qualitative comparisons between our method and the state-of-the-art methods including PDNet18 [46], PICA18 [45],RAS18 [44], RADF18 [47], DSS18 [10], DF17 [20], MDC17 [43], BDT17 [17], and DRFI17 [42].

TABLE III

COMPARISON OF THE QUANTITATIVE RESULTS INCLUDING THE MAXIMUM F-MEASURE (A LARGER VALUE IS BETTER), THE MAE (A SMALLER VALUE IS BETTER) AND THE AUC (A LARGER VALUE IS BETTER). THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY

DATASET	**FLSD** [49]			**DEC** [48]			**NLPR** [28]			**NJUDS** [19]		
Metric	**maxF**	**AUC**	**MAE**	**maxF**	**AUC**	**MAE**	**maxF**	**AUC**	**MAE**	**maxF**	**AUC**	**MAE**
Ours	**0.895**	**0.984**	**0.057**	**0.905**	**0.992**	**0.032**	**0.880**	**0.990**	**0.038**	**0.891**	**0.985**	**0.050**
PDNet18 [46]	**0.857**	**0.973**	0.109	**0.888**	**0.991**	0.045	**0.876**	**0.989**	0.048	**0.884**	**0.983**	**0.072**
RADF18 [47]	0.833	0.944	0.107	0.855	0.979	0.057	0.814	0.979	0.066	0.847	0.963	0.869
RAS18 [44]	0.726	0.785	0.164	0.809	0.855	0.060	**0.843**	0.899	0.053	0.781	0.839	0.121
PICA18 [45]	0.834	**0.966**	**0.106**	**0.887**	0.973	**0.037**	0.842	**0.980**	**0.051**	**0.869**	**0.974**	**0.064**
DSS18 [10]	0.745	0.837	0.159	0.771	0.882	0.075	0.832	0.931	0.066	0.786	0.893	0.122
MDC17 [43]	0.787	0.927	0.141	0.668	0.917	0.104	0.671	0.923	0.121	0.689	0.882	0.171
BDT17 [17]	0.817	0.939	0.125	0.820	0.936	0.055	0.809	0.969	0.064	0.806	0.928	0.127
DF17 [20]	0.845	0.962	0.146	0.752	0.953	0.107	0.759	0.947	0.106	0.768	0.929	0.168
DRFI17 [42]	0.766	0.930	0.204	0.667	0.934	0.122	0.724	0.949	0.131	0.692	0.904	0.203
DHS16 [41]	0.825	0.938	**0.107**	0.885	0.969	**0.036**	0.841	0.971	**0.048**	0.859	0.956	0.076
MDF16 [11]	0.751	0.905	0.167	0.757	0.917	0.083	0.769	0.946	0.086	0.750	0.904	0.140
DCMC16 [21]	**0.849**	0.965	0.155	0.750	0.948	0.108	0.697	0.932	0.117	0.759	0.936	0.172
LBE16 [18]	0.774	0.940	0.209	0.816	**0.980**	0.208	0.780	0.895	0.082	0.791	0.936	0.153
GP15 [29]	0.464	0.627	0.282	0.672	0.884	0.166	0.711	0.906	0.139	0.412	0.617	0.266
ACSD14 [19]	0.803	0.951	0.183	0.798	0.976	0.153	0.670	0.926	0.164	0.746	0.929	0.194
LMH14 [28]	0.757	0.904	0.212	0.641	0.893	0.117	0.703	0.905	0.109	0.760	0.855	0.207

RAS18 [44], PICA18 [45], PDNet18 [46], and RADF18 [47]. For the objective comparisons, all the quantitative evaluations are conducted using the source codes provided by the authors and without changing the parameters.

We evaluate our model using two widely adopted metrics: the precision-recall (PR) curve and the F-measure. Given a predicted saliency map, we perform binary segmentation with a hard threshold T. If the obtained foreground is consistent with the ground truth mask, it is considered a successful detection, and the final precision-recall curves are obtained by varying T from 0 to 255. As the recall rate is inversely proportional to the precision, the tendency of the trade-off between the precision and recall can truly indicate the overall video saliency detection performance. The F-measure is an important performance indicator when the precision rate conflicts with the recall rate and can be computed by

$$\mathrm{F-measure} = \frac{(1+\beta^2) \times \mathrm{Pre} \times \mathrm{Rec}}{\beta^2 \times \mathrm{Pre} + \mathrm{Rec}}, \tag{19}$$

where Pre and Rec denote the corresponding precision rate and recall rate, respectively, and we assign $\beta^2 = 0.3$ to be biased toward the precision rate. We also report the maximum F-measure, AUC, MAE in Table. I. Although commonly used, PR curves have limited value because they fail to consider truly negative pixels. For a more balanced comparison, we adopt the MAE as another evaluation metric. The MAE measures the numerical distance between the ground truth and the estimated saliency map and is more meaningful in evaluating the applicability of a saliency model in a task such as object segmentation. It is defined as the average pixel-wise absolute difference between the binary ground truth (GT) and the saliency map (SAL). The MAE evaluates the saliency detection accuracy by Eq. 20.

$$\mathrm{MAE} = \frac{1}{\mathrm{W} \times \mathrm{H}} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathrm{SAL(x, y)} - \mathrm{GT(x, y)}|\,, \tag{20}$$

where W and H represent the width and height of the given image, respectively, SAL(x, y) denotes the saliency value of the pixel with coordinates (x, y), and GT denotes the corresponding binary ground truth. The ROC curve can be conveniently generated according to the true positive rates and false positive rates obtained during the calculation of the PR curve. The AUC is calculated as the area under the ROC curve. A perfect model will achieve an AUC of 1, while random guessing will achieve an AUC of approximately 0.5.

$$\mathrm{AUC} = \frac{\sum_{\mathrm{i}\in \mathrm{Pos}} \mathrm{Rank_i} - \frac{\mathrm{PosNum}\times(\mathrm{PosNum}+1)}{2}}{\mathrm{PosNum}\times \mathrm{NegNum}}, \tag{21}$$

where $\mathrm{Rank_i}$ represents the serial order of the i-th element, $PosNum$ is the number of positive samples and $NegNum$ is the number of negative samples.

As shown in Fig. 11, the RGB information-based saliency detection methods (e.g., DSS18, RADF18, DHS16, etc.) frequently demonstrate a poor detection performance for low color contrast salient foregrounds while assigning large saliency values to non-salient, complex backgrounds. Benefiting from the newly available depth information, the state-of-the-art RGB-D methods achieve a much better detection performance. That is, both the low color contrast-induced hollow effects can be potentially alleviated by the depth continuity toward those inner regions of the salient object. The high color contrast-induced false-alarm detections may also be solved by the large depth differences between those salient regions and their non-salient nearby surroundings. However, because the current state-of-the-art RGB-D methods overly rely their saliency detection on the depth component, it is difficult to achieve a balance between the RGB information and the depth information, making the detection sensitive to the depth quality. By using the newly estimated Depth$^+$, our method can potentially alleviate the abovementioned problems.

Additionally, from the perspective of the depth-sensing equipment, the depth quality of the Lytro Light Field camera is significantly inferior to that of the Microsoft Kinect device. Thus, our method achieves a significant performance improvement in the LFSD dataset because all depth information in the LFSD dataset is sensed by the Lytro Light Field camera. However, our method can only slightly outperform the other methods in the NJUD dataset because the depth information in the NJUD dataset is formulated based on the optical flow method, which frequently exhibits the highest depth quality. Moreover, the F-measure curves also demonstrate the superiority of our method; i.e., our method outperforms the state-of-the-art methods by simply using a hard threshold ranging from 200 to 230 (see the middle row in Fig. 8).

E. Limitations

Our method has two limitations. First, our depth estimation procedure is time consuming. On a desktop computer with an i7-6700k 4.00 GHz CPU, a GTX 1080Ti GPU, and 32 GB RAM, it takes approximately 0.43 s to obtain the coarse depth (cD, Eq. 6) and another 0.31 s to perform the saliency-driven depth transference to obtain the fine depth (Depth$^+$). Additionally, since our deep learning network (**MNSSF**) follows a superpixel-wise method, it takes approximately 2 s to output the final saliency map.

Second, the performance of our method relies on the quality of the estimated Depth$^+$. Thus, our method may also produce failure detections when the quality of both Depth and Depth$^+$ are poor. Moreover, since a large dataset can ensure a strong similarity between the input image and its retrieved subgroup images, the quality of our newly estimated Depth$^+$ is positively related to the given dataset size.

VII. Conclusion and Future Works

In this paper, we have proposed a novel method to perform high-performance RGB-D saliency detection. Our method consists of two components: two-phase depth estimation and selective deep saliency fusion network. First-round depth transfer simultaneously considers both the mid-level and object-level inter-image similarity to coarsely estimate the depth information. Then, from a saliency perspective, second-round depth transfer mainly attempts to enlarge the depth difference between those potentially salient regions and their non-salient nearby surroundings. Our selective deep saliency fusion network consists of two subnetworks: the **IDSF** network and the **MNSSF** network. The **IDSF** network performs image-level selective saliency fusion to complement the RGB information with the depth information to reduce the subsequent problem domain. Then, the **MNSSF** network mainly attempts to compute an optimal nonlocal complimentary state between the original depth and the newly estimated depth. Both quantitative and qualitative comparisons indicated that our method outperformed all the current state-of-the-art methods.

As for our future works, we are particularly interested in developing an explicit solution to make blind depth quality assessments. Thus, the selective saliency fusion performance may be further improved if we can adaptively control its fusion balance using the previously assessed depth quality.

References

[1] C. Chen, S. Li, H. Qin, and A. Hao, “Real-time and robust object tracking in video via low-rank coherency analysis in feature space,” *Pattern Recognit.*, vol. 48, no. 9, pp. 2885–2905, 2015.

[2] C. Chen, S. Li, H. Qin, Z. Pan, and G. Yang, “Bilevel feature learning for video saliency detection,” *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3324–3336, Dec. 2018.

[3] C. Chen, Y. Li, S. Li, H. Qin, and A. Hao, “A novel bottom-up saliency detection method for video with dynamic background,” *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 154–158, Feb. 2018.

[4] C. Chen, G. Wang, C. Peng, X. Zhang, and H. Qin, “Improved robust video saliency detection based on long-term spatial-temporal information,” *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 1090–1100, Aug. 2020.

[5] Y. Li, S. Li, C. Chen, H. Qin, and A. Hao, “Accurate and robust video saliency detection via self-paced diffusion,” *IEEE Trans. Multimedia*, early access, 2019, doi: 10.1109/TMM.2019.2940851.

[6] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, “No-reference quality metric of contrast-distorted images based on information maximization,” *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4559–4565, Dec. 2017.

[7] C. Chen, S. Li, H. Qin, and A. Hao, “Robust salient motion detection in non-stationary videos via novel integrated strategies of spatio-temporal coherency clues and low-rank analysis,” *Pattern Recognit.*, vol. 52, pp. 410–432, Apr. 2016.

[8] C. Peng, C. Chen, Z. Kang, J. Li, and Q. Cheng, “Res-PCA: A scalable approach to recovering low-rank matrices,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7317–7325.

[9] J. Wu, Y. Zhang, and W. Lin, “Good practices for learning to recognize actions using FV and VLAD,” *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2978–2990, Dec. 2016.

[10] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[11] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[12] C. Chen, S. Li, H. Qin, and A. Hao, "Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2303–2316, Aug. 2015.

[13] G. Ma, C. Chen, S. Li, C. Peng, H. Qin, and A. Hao, "Salient object detection via multiple instance joint re-learning," *IEEE Trans. Multimedia*, to be published.

[14] M. Liang and X. Hu, "Feature selection in supervised saliency prediction," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 914–926, May 2015.

[15] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via SDAEs," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 487–498, Feb. 2016.

[16] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[17] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2749–2757.

[18] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2343–2350.

[19] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 1115–1119.

[20] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[21] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.

[22] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[23] R. Cong *et al.*, "An iterative co-saliency framework for RGBD images," *IEEE Trans. Cybern.*, vol. 49, no. 1, pp. 233–246, Jan. 2019.

[24] A. Maki, P. Nordlund, and J. Eklundh, "A computational model of depth-based attention," *Pattern Recognit.*, vol. 4, pp. 734–739, 1996.

[25] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3d video," in *Proc. Adv. Multimedia Modeling*, 2010, pp. 314–324.

[26] D. Feng, N. Barnes, and S. You, "HOSO: Histogram of surface orientation for RGB-D salient object detection," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl.*, 2017, pp. 1–8.

[27] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Aug. 2015.

[28] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[29] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 25–32.

[30] J. Guo, T. Ren, J. Bei, and Y. Zhu, "Salient object detection in RGB-D image based on saliency fusion and propagation," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2015, pp. 59–63.

[31] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663–667, May 2017.

[32] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3061.

[33] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.

[34] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3917–3925.

[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[36] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[37] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, Aug. 2011.

[38] J. Munkres, "Algorithms for the assignment and transportation problems," *J. Soc. Ind. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.

[39] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao, "Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3156–3170, Jul. 2017.

[40] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[41] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 678–686.

[42] J. Wang, H. Jiang, Z. Yuan, M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, pp. 251–268, Jun. 2017.

[43] X. Huang and Y. Zhang, "300-FPS salient object detection via minimum directional contrast," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4243–4254, Sep. 2017.

[44] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.

[45] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[46] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 199–204.

[47] X. Hu, L. Zhu, J. Qin, C. Fu, and P. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6943–6950.

[48] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. ACM Int. Conf. Internet Multimedia Comput. Service*, 2014, pp. 23–27.

[49] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2806–2813.

[50] D. Fan, W. Wang, M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8554–8564.