HSCS: Hierarchical Sparsity Based Co-saliency Detection for RGBD Images

Runmin Cong, Jianjun Lei, *Senior Member, IEEE*, Huazhu Fu, *Senior Member, IEEE*, Qingming Huang, *Fellow, IEEE*, Xiaochun Cao, *Senior Member, IEEE*, and Nam Ling, *Fellow, IEEE*

***Abstract*—Co-saliency detection aims to discover common and salient objects in an image group containing more than two relevant images. Moreover, depth information has been demonstrated to be effective for many computer vision tasks. In this paper, we propose a novel co-saliency detection method for RGBD images based on hierarchical sparsity reconstruction and energy function refinement. With the assistance of the intrasaliency map, the inter-image correspondence is formulated as a hierarchical sparsity reconstruction framework. The global sparsity reconstruction model with a ranking scheme focuses on capturing the global characteristics among the whole image group through a common foreground dictionary. The pairwise sparsity reconstruction model aims to explore the corresponding relationship between pairwise images through a set of pairwise dictionaries. In order to improve the intra-image smoothness and inter-image consistency, an energy function refinement model is proposed, which includes the unary data term, spatial smooth term, and holistic consistency term. Experiments on two RGBD co-saliency detection benchmarks demonstrate that the proposed method outperforms the state-of-the-art algorithms both qualitatively and quantitatively.**

***Index Terms*—Co-saliency detection, RGBD images, global sparsity reconstruction, pairwise sparsity reconstruction, energy function refinement.**

Manuscript received February 26, 2018; revised June 9, 2018, August 24, 2018, and October 22, 2018; accepted November 15, 2018. Date of publication December 3, 2018; date of current version June 21, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Crandall. This work was supported in part by the National Natural Science Foundation of China under Grants 61520106002, 61722112, 61731003, 61332016, 61620106009, U1636214, 61602345, in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences under Grant QYZDJ-SSW-SYS013, and in part by the Technology Research and Development Program of Tianjin under Grant 15ZXHLGX00130. *(Corresponding author: Jianjun Lei.)*

R. Cong and J. Lei are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: rmcong@tju.edu.cn; jjlei@tju.edu.cn).

H. Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: huazhufu@gmail.com).

Q. Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: qmhuang@ucas.ac.cn).

X. Cao is with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: caoxiaochun@iie.ac.cn).

N. Ling is with the Department of Computer Engineering, Santa Clara University, Santa Clara, CA 95053 USA (e-mail: nling@scu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2884481

I. Introduction

VISUAL attention mechanism enables people to quickly locate the most interesting parts or salient objects from a complex scene. As a branch of computer vision, saliency detection is devoted to enabling a computer to discover these salient regions automatically. This process has been applied in a large number of visual tasks, such as segmentation [1]–[3], retargeting [4], enhancement [5], foreground annotation [6], and blur detection [7]. The last decade has witnessed the vigorous development and qualitative leap in performance of image saliency detection [8]–[26].

With the recent explosive growth of data volume, people need to process multiple relevant images collaboratively. As an extension of traditional image saliency detection, co-saliency detection aims to discover common and salient objects in an image group containing multiple relevant images [27]. This has been successfully applied in co-segmentation [28]–[30], co-localization [31], and image matching [32]. Different from image saliency detection, the co-saliency detection model needs to consider the common attributes of salient objects in an image group through the inter-image constraint. In other words, the co-salient objects should not only be prominent with respect to the backgrounds in each individual image, but also should recur throughout the whole image group.

In addition to color appearance, humans can perceive the distance mapping of a scene, which is known as depth information. With the development of imaging devices and technologies, capturing the depth representation information becomes increasingly convenient. Moreover, depth information has been proven to be useful for many computer vision tasks, such as segmentation [33], enhancement [34], and saliency detection [35]–[42]. However, most of the existing methods focus on handling the RGBD images rather than the RGBD image group. In this paper, the depth feature is not only served as a constraint in the inter-image correspondence modelling, but also used as a color information supplement in the refinement component.

The corresponding relationship among multiple images plays an important role in co-saliency detection. In other words, in addition to the saliency attribute in an individual image, the repetitiveness constraint across the whole image group is also crucial to suppress the background and non-common salient regions. In existing methods, the inter-image correspondence is simulated as a matching process [43]–[48], clustering process [49]–[51], low-rank problem [52], [53], propagation pro-

cess [54]–[56], or learning process [57]–[61]. However, the matching- and propagation-based methods are often time consuming, while the clustering based methods are sensitive to the noise. To overcome these problems, the sparsity-based technique is a good choice and has demonstrated the potential to improve the performance of many tasks, including saliency detection [62]–[65]. For the sparsity-based saliency detection methods, the background or foreground dictionary is used to reconstruct each processing unit, and the saliency is measured by the reconstruction error. In addition to describing the saliency of an individual image, sparsity representation can be used to constrain the inter-image correspondence capturing and to achieve inter saliency detection. In this paper, a hierarchical sparsity reconstruction model is innovatively proposed to capture a more comprehensive inter-image relationship by considering the global and local inter-image information. The hierarchical sparsity property includes two complementary aspects, i.e., (1) The co-salient objects in the whole image group should belong to the same category and have similar appearance. Therefore, a global foreground dictionary with a ranking scheme is built to reconstruct each image and to capture the global inter-image correspondence, which is called global sparsity reconstruction. (2) The relationship among multiple images can be decomposed into a combination of multiple pairwise correspondences. Therefore, a set of foreground dictionaries constructed by other images are utilized to reconstruct the current image and obtain multiple pairwise inter saliency maps from the local perspective.

The co-salient objects in different images of the same group should be similar and consistent in appearance. Thus, a superior co-saliency detection model should guarantee the local smoothness in each individual image and global consistency in the whole image group. In this paper, we propose an energy function refinement model to attain a more consistent and accurate co-saliency result, which includes the unary data term, spatial smooth term, and holistic consistency term. The data term constrains the updating degree of the refinement algorithm, and the smooth term favors that all the spatially adjacent regions with similar appearance should be assigned to consistent saliency scores. In addition to these two traditional terms, a holistic consistency term is specifically designed for the co-saliency detection task, which imposes the appearances of co-salient objects to be consistent in the whole image group.

In this paper, we provide an effective and efficient co-saliency detection method for RGBD images based on hierarchical sparsity reconstruction and energy function refinement. The main contributions are summarized as follows:

1) A co-saliency detection method for RGBD images is proposed that integrates the intra saliency detection, hierarchical inter saliency detection based on global and pairwise sparsity reconstructions, and energy function refinement. The hierarchical sparsity representation is firstly used to capture the inter-image correspondence in co-saliency detection.
2) The global sparsity reconstruction is utilized to capture the global characteristic among the whole image group through a common foreground dictionary. Moreover, a ranking scheme is designed to guide the foreground seed selection and optimize the common foreground dictionary.
3) The inter-image relationship is simulated as a combination of multiple pairwise correspondences. The pairwise sparsity reconstruction model utilizes a set of foreground dictionaries produced by other images to explore local inter-image information.
4) To improve the intra-image smoothness and inter-image consistency, an energy function refinement model is proposed. A holistic consistency term is specifically designed for the co-saliency detection task, which constrains the appearances of co-salient objects to be consistent in the whole image group.

The rest of the paper is organized as follows. Section II reviews related works. Section III presents the details of the proposed RGBD co-saliency detection method. The experimental comparisons and discussions are presented in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

In this section, we briefly review the related works of image saliency detection and co-saliency detection.

A. Image Saliency Detection

The last decade has witnessed the considerable development of saliency detection for RGB image, and a large number of methods have been presented [8]–[26]. Li *et al.* [9] used the reconstruction error to measure the saliency of a region, where the salient region corresponds to a larger reconstruction error. Zhu *et al.* [11] proposed a principled optimization framework integrating multiple low-level cues to achieve saliency detection with the help of a robust background measure. Peng *et al.* [15] proposed a structured matrix decomposition based saliency detection method guided by high-level priors and obtained competitive performance. Recently, deep learning has been demonstrated the power in saliency detection. Li and Yu [17] proposed an end-to-end deep contrast network for saliency detection, which integrates the multi-scale fully convolutional stream and the segment-wise spatial pooling stream. Hou *et al.* [19] introduced short connections into the skip-layer structures within the holisitcally-nested edge detector architecture to achieve image saliency detection. Zhang *et al.* [20] utilized an encoder Fully Convolutional Network (FCN) and a corresponding decoder FCN to detect the salient object, in which the reformulated dropout and hybrid up-sampling are designed. In addition, some new measurements for saliency evaluation are proposed, such as Structure-measure [25] and Enhanced-alignment measure [26].

In addition, the introduction of depth information makes RGBD images more in accordance with the human vision system, which could improve the performance of saliency detection [35]–[42]. To achieve RGBD saliency detection, some depth features and measures are utilized. Ju *et al.* [36] proposed an Anisotropic Center-Surround Difference (ACSD) measure to calculate the depth-aware saliency map. Considering the quality of depth map, Cong *et al.* [38] proposed an RGBD saliency

detection method via depth confidence analysis and multiple cues fusion, where the depth confidence measure works as a controller to constrain the introduction of depth information in the saliency model. In [39], depth information is used as a regional feature for low-level contrast-based saliency computation, and also worked as a weighting term for mid-level saliency evaluation. Qu *et al.* [40] designed a convolutional neural network to automatically learn the interaction between low-level cues and saliency result for RGBD saliency detection.

B. Co-saliency Detection

In order to achieve co-saliency detection, the inter-image correspondence can be captured by different techniques, such as similarity matching [43]–[48], clustering [49]–[51], low-rank decomposition [52], [53], propagation [54]–[56], and learning [57]–[61]. Liu *et al.* [46] proposed a novel co-saliency detection model integrating the global similarity on the fine segmentation level with the object prior on the coarse segmentation level, where the inter-image correspondence is formulated as global similarity of each region. Li *et al.* [47] proposed a two-stage saliency model guided co-saliency detection method, in which the first stage recovers the co-salient parts through the efficient manifold ranking, and the second stage captures the inter-image correspondence via a ranking scheme. In [49], an efficient cluster-based co-saliency detection algorithm for multiple images is proposed, which takes the cluster as the basic unit to represent the multi-image relationship by integrating the contrast, spatial, and corresponding cues. Cao *et al.* [52] proposed a saliency fusion framework for co-saliency detection based on the rank constraint, which is valid for multiple images and also works well on single image saliency detection. In [54], co-saliency detection is formulated as a two-stage saliency propagation problem, including the intra-saliency propagation stage and the inter-saliency propagation stage. With the training process, the learning based methods always achieve competitive performance. Zhang *et al.* [57] proposed a co-saliency detection model under the Bayesian framework with some higher-level features extracted by the convolutional neural network. Wei *et al.* [58] proposed an end-to-end group-wise deep co-saliency detection model, where the semantic block is utilized to obtain the basic feature representation, the group-wise and single feature representation blocks are used to capture the group-wise interaction information and individual image information, and the collaborative learning structure with convolution-deconvolution model aims to output the co-saliency map. Han *et al.* [61] introduced the metric learning into co-saliency detection to jointly learn the discriminative feature representation and co-salient object detector, which can handle the wide variation in image scene and achieve superior performance.

Furthermore, a few methods are proposed to achieve RGBD co-saliency detection by introducing the depth cue as an effective privileged information. Song *et al.* [66] utilized the bagging-based clustering method to detect the co-salient objects from RGBD images group, where the average depth value, depth range, and Histogram of Oriented Gradient (HOG) on depth map are extracted to represent the depth attributes. Cong *et al.* [67] proposed a co-saliency detection method for RGBD images by using the multi-constraint feature matching and cross label propagation, where the inter-image relationship is explored at the superpixel and image levels. In [68], an iterative co-saliency detection framework for RGBD images is proposed, which integrates the addition scheme, deletion scheme, and iterative scheme. The addition scheme aims at generating the RGBD saliency map by introducing the depth shape prior into the existing saliency detection model, the deletion scheme focuses on capturing the inter-image correspondence via a common probability function, and the iterative scheme is served as an optimization process through a refinement-cycle.

Compared to the traditional saliency detection, there are three challenges for RGBD co-saliency detection, i.e., (1) how to utilize the depth information to enhance the identification of salient objects, (2) how to capture the corresponding relationship among multiple images, and (3) how to guarantee the consistency and smoothness of the final co-saliency map. To address these challenges, we propose a hierarchical sparsity based co-saliency detection method for RGBD images. The depth feature is used as an additional constraint and supplement cue in the inter-image correspondence and refinement components. In addition, an effective inter saliency model is designed to capture the inter-image correspondence, which integrates the global sparsity reconstruction and pairwise sparsity reconstruction. Moreover, we also formulate an energy function refinement to achieve a superior and globally consistent co-saliency map.

III. Proposed Method

A. Framework

The flowchart of the proposed hierarchical sparsity based co-saliency detection method for RGBD images is shown in Fig. 1, which includes intra saliency calculation, hierarchical inter saliency detection based on global and pairwise sparsity reconstructions, and energy function refinement.

According to the definition of co-saliency detection, the co-salient objects should be prominent in an individual image. Therefore, the intra saliency map is firstly calculated for each individual image. We denote the input RGB images in a group as $\{I^i\}_{i=1}^N$, and the corresponding depth maps as $\{D^i\}_{i=1}^N$, where N is the number of images in the group. For computational efficiency and structural representation, each RGB image I^i is abstracted into some superpixels $\mathbf{R}^i = \{r_m^i\}_{m=1}^{N^i}$ through the SLIC algorithm [69], where N^i represents the number of superpixels in image I^i. In light of the effectiveness and robustness of the DCMC method [38],[1] we chose it as the basic method for intra saliency detection, and the intra saliency value of superpixel r_m^i is denoted as $S_a(r_m^i)$.

The background of each image may be diverse within the same image group, while the co-salient objects tend to have a similar appearance in all images. Therefore, the co-salient regions can be reconstructed better than the background regions through a sparsity framework with the foreground dictionary. In

[1] https://rmcong.github.io/proj_RGBD_sal.html

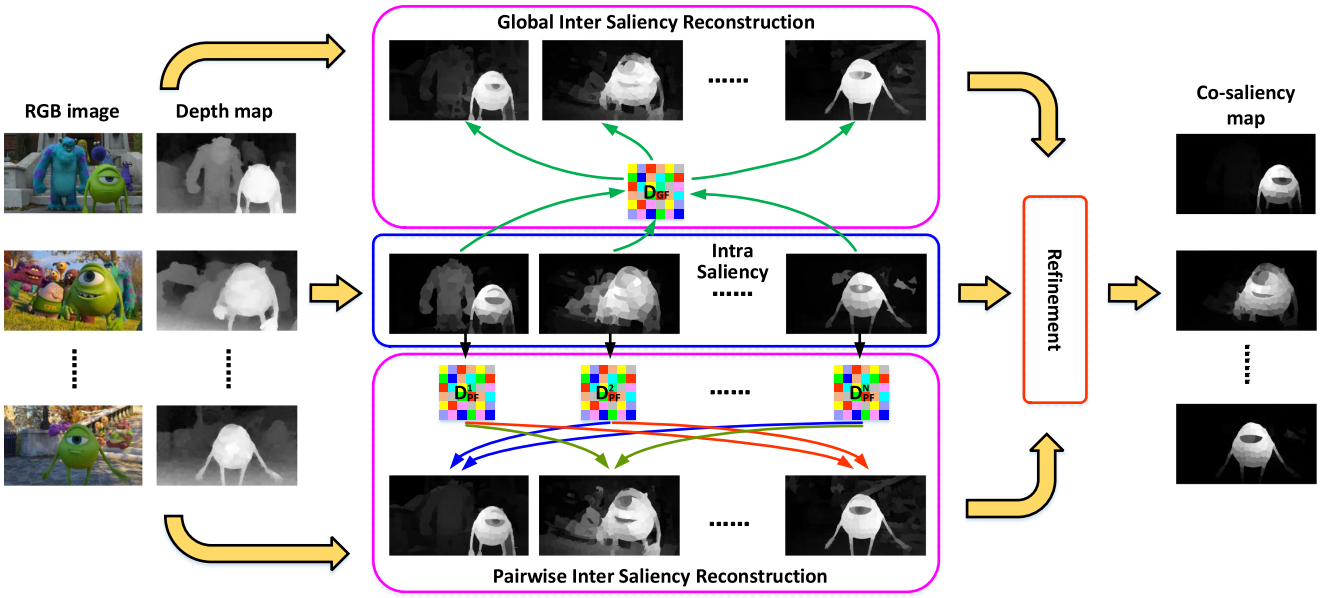


Fig. 1. The flowchart of the proposed RGBD co-saliency detection method.

this paper, the corresponding relationship among multiple images is simulated as a hierarchical sparsity framework considering the global and pairwise sparsity reconstructions. The global inter saliency reconstruction model describes the inter-image correspondence from the perspective of the whole image group via a common reconstruction dictionary, while the pairwise inter saliency reconstruction model utilizes a set of foreground dictionaries produced by other images to capture local inter-image information.

Finally, an energy function refinement model, including the unary data term, spatial smooth term, and holistic consistency term, is proposed to improve the intra-image smoothness and inter-image consistency and to generate the final co-saliency map. The spatial smooth term is used to optimize the intra-image smoothness, and the holistic consistency term is specifically designed for co-saliency detection task to update the inter-image consistency. Hierarchical inter saliency detection based on global and pairwise sparsity reconstructions, as well as the energy function refinement, are detailed in the following sections.

B. Global Inter Saliency Reconstruction

The co-salient objects in a whole image group should belong to the same category and have a similar appearance. Therefore, a global foreground dictionary is built to reconstruct each image and capture the global inter-image correspondence. First, some initial foreground seeds are selected based on all the intra saliency maps in the image group. Then, a ranking filter is designed to eliminate the interference seeds and to determine the optimal foreground seeds. Next, the feature vectors of foreground seeds are extracted to construct the global foreground dictionary. Finally, the reconstruction error produced by the sparsity framework is utilized to measure the global inter saliency.

1) Initial Foreground Seeds Selection: The intra saliency map provides effective single image saliency description. We assume that most of the co-salient objects can be included in these saliency maps. Thus, the top K superpixels in image I^i with larger intra saliency values are selected as the foreground seeds. Then, all these seeds from different images are combined into an initial foreground seed set $\Phi_{init} = \Phi^1_{init} \cup \Phi^2_{init} \cup \ldots \Phi^N_{init}$, where Φ^i_{init} denotes the foreground seed set of image I^i, and N is the number of images.

2) Ranking Based Seeds Filtering: Since the intra saliency result is not completely accurate, some disturbances may be wrongly included in the initial foreground seed set, such as the backgrounds and non-common salient regions, which may degenerate the reconstruction accuracy. Therefore, we designed a ranking scheme to filter the interferences and refine the foreground seeds.

In general, the co-salient objects satisfy three constraints, i.e., (a) the category should be same, (b) the color appearance should be similar, and (c) the depth distribution should be approximate. Combining these three constraints, a novel measure is designed to evaluate the local consistency of superpixels belonging to the initial foreground seed set. First, all the initial seed superpixels are grouped into five clusters by using the K-means++ clustering [70], and each superpixel is assigned to a corresponding cluster center $\{\boldsymbol{c}_i\}_{i=1}^{N\cdot K}$, respectively. Then, introducing the clustering, color, depth, and saliency constraints, the consistency measure is defined as follows:

$$mc(r_m) = \left[\sum_{\substack{n=1 \\ n\neq m}}^{N\cdot K} (1 - \|\boldsymbol{c}_m - \boldsymbol{c}_n\|_2) \cdot \omega_{mn}\right] \cdot S_a(r_m) \quad (1)$$

and

$$\omega_{mn} = \exp\left(-\frac{\chi^2(\boldsymbol{h}_m, \boldsymbol{h}_n) + \lambda_{\min} \cdot |d_m - d_n|}{\sigma^2}\right) \quad (2)$$

Fig. 2. Some examples of ranking scheme for foreground seeds selection. The first second rows are the RGB images and depth maps, the third row shows the initial foreground seeds marked in red, and last row presents the final foreground seeds marked in yellow after ranking scheme.

where $r_m, r_n \in \Phi_{init}$; $\boldsymbol{c}_m$ is the cluster center of superpixel r_m; ω_{mn} represents the feature similarity between superpixel r_m and superpixel r_n; $S_a(r_m)$ is the intra saliency value of superpixel r_m; $N \cdot K$ is the total number of initial foreground seeds; $\|\cdot\|_2$ is the l_2-norm function; and σ^2 is a parameter to control strength of the similarity, which is set to 0.1 in all experiments following [38]. $\boldsymbol{h}_m$ denotes the color histogram of superpixel r_m in the Lab color space; $\chi^2(\cdot)$ represents the Chi-square distance; and d_m is the mean depth value of superpixel r_m. $\lambda_{\min} = \min(\lambda_m, \lambda_n)$ denotes the minimum depth confidence measure of two input depth maps, where $\lambda_m = \exp((1 - m_m) \cdot CV_m \cdot H_m) - 1$ is the depth confidence measure of the input depth map D^m; m_m denotes the mean value of the whole depth image; CV_m represents the coefficient of variation; and H_m is the depth frequency entropy. More details can be found in [38]. A larger mc corresponds to higher consistency with respect to other foreground seeds. In other words, the larger the consistency measure is, the higher the probability of the superpixel being the foreground seed. Finally, the top 80% of initial seeds with larger consistency measure values are reserved as the final foreground seeds, which is denoted as Φ_{fin}.

Some illustrations of the foreground seeds are shown in Fig. 2, where the third row presents the visualization of the initial foreground seeds marked in red, and the final foreground seeds marked in yellow after ranking scheme are shown in the last row. As can be seen, some backgrounds (e.g., the lawns located by the blue arrows) in the second and third images are wrongly selected as the foregrounds in the initial seeds set. With the ranking scheme, the correct foreground seeds (i.e., the dark bird) are successfully reserved, while the backgrounds are effectively eliminated.

3) Sparsity-Based Global Reconstruction: Four types of low-level features, including color components, depth attribute, spatial location, and texture distribution, are utilized to describe each superpixel r_m^i as $\boldsymbol{f}_m^i = [\boldsymbol{l}_m^i \ d_m^i \ \boldsymbol{p}_m^i \ \boldsymbol{t}_m^i]^T$, where $\boldsymbol{l}$ is the 9-dimensional color components in the RGB, Lab, and HSV color spaces; d denotes the depth value; $\boldsymbol{p}$ corresponds to the 2-dimensional spatial coordinates; and $\boldsymbol{t}$ represents the 15-dimensional texton histogram [71]. The feature representations of the stacking superpixels in the final foreground seeds set Φ_{fin} are constructed as the global foreground dictionary, which is denoted as $\mathbf{D}_{GF}$.

Under the same reconstruction dictionary, the reconstruction error between foreground and background regions should be different. Thus, the image saliency can be measured by the reconstruction error [9]. We compute the reconstruction error by the sparsity representation with a global foreground dictionary, and each superpixel r_m^i is encoded by:

$$\alpha_m^{i*} = \arg\min_{\alpha_m^i} \|\boldsymbol{f}_m^i - \mathbf{D}_{GF} \cdot \alpha_m^i\|_2^2 + \xi \cdot \|\alpha_m^i\|_1 \quad (3)$$

where α_m^{i*} is the optimal sparse coefficient for superpixel r_m^i; $[\mathbf{D}_{GF}]_{L \times |\Phi_{fin}|}$ denotes the global foreground dictionary; $|\Phi_{fin}|$ represents the number of final foreground seeds; $L = 27$ is the feature dimension of each superpixel in our work; $\boldsymbol{f}_m^i$ is the feature representation of superpixel r_m^i; $\|\cdot\|_1$ is the l_1-norm function; and ξ is set to 0.01 as suggested in [9].

The foreground dictionary is used to achieve global reconstruction, thus, the superpixel with the smaller reconstruction error should be assigned to a greater saliency value and vice versa. The global inter saliency of superpixel r_m^i is defined as:

$$S_{gr}(r_m^i) = \exp(-\varepsilon_m^i/\sigma^2) = \exp(-\|\boldsymbol{f}_m^i - \mathbf{D}_{GF} \cdot \alpha_m^{i*}\|_2^2/\sigma^2) \quad (4)$$

where $S_{gr}(r_m^i)$ is the inter saliency of superpixel r_m^i through the global reconstruction; ε_m^i denotes the reconstruction error of superpixel r_m^i; and σ^2 is a weighted constant.

C. Pairwise Inter Saliency Reconstruction

The global reconstruction aims to describe the inter-image correspondence from the perspective of the whole image group. In fact, the relationship among multiple images can be decomposed into a combination of multiple pairwise correspondences, which benefits capturing the local inter-image information. In order to deeply explore a more comprehensive inter-image corresponding relationship, a sparsity-based pairwise reconstruction method is proposed to calculate the pairwise inter saliency. First, we construct a foreground dictionary for each image based on the corresponding intra saliency map, respectively. In this way, the N foreground dictionaries in an image group are obtained, where N denotes the number of images in the group. Then, each image is reconstructed by the $N-1$ foreground dictionaries derived from other images in the group, respectively. Finally, these $N-1$ reconstructed results are fused to generate the pairwise inter saliency map.

For each image I^k, the top K superpixels with larger intra saliency values are selected as the foreground seeds. Similar

to the sparsity-based global reconstruction, a 27-dimensional feature vector is used to represent each superpixel. Then, the feature representations of the stacking foreground superpixels in each image are constructed as the pairwise foreground dictionary, which is denoted as $\mathbf{D}_{PF}^{k}$. As mentioned earlier, the foreground pairwise dictionaries generated by other images can be utilized to reconstruct the current image and capture the local inter-image relationship. Using the pairwise foreground dictionary $\mathbf{D}_{PF}^{k}$ produced by the image I^k, the image I^i can be constructed and the saliency is measured as:

$$S_{pr}^{k}(r_m^i) = \exp(-\varepsilon_m^{k,i}/\sigma^2) = \exp(-\|\boldsymbol{f}_m^i - \mathbf{D}_{PF}^{k} \cdot \alpha_m^{k,i*}\|_2^2/\sigma^2) \tag{5}$$

where $S_{ir}^{k}(r_m^i)$ is the inter saliency through the pairwise reconstruction using the dictionary $[\mathbf{D}_{PF}^{k}]_{L\times K}$; $\varepsilon_m^{k,i}$ denotes the reconstruction error of superpixel r_m^i; $\alpha_m^{k,i*}$ is the optimal sparse coefficient of superpixel r_m^i; $k \in [1, 2, \ldots, N], k \neq i$ represents the index of pairwise foreground dictionary; and σ^2 is a weighted parameter. Therefore, we obtain $N-1$ saliency maps for each image through different pairwise dictionaries. At last, all these maps are fused to generate the final pairwise inter saliency map by:

$$S_{pr}(r_m^i) = \frac{1}{N-1} \cdot \sum_{\substack{k=1 \\ k\neq i}}^{N} S_{pr}^{k}(r_m^i). \tag{6}$$

The global inter saliency map describes the global inter-image correspondence from the whole image group, while the pairwise inter saliency map captures the local relationship from the pairwise images. Finally, these two inter saliency maps are combined as the hierarchical sparsity based inter saliency:

$$S_r(r_m^i) = \frac{1}{2} \cdot (S_{gr}(r_m^i) + S_{pr}(r_m^i)) \tag{7}$$

where $S_r(r_m^i)$ denotes the hierarchical sparsity based inter saliency of superpixel r_m^i.

D. Energy Function Refinement

In order to achieve a superior and globally consistent saliency map, a refinement model with an energy function is designed in our work. Three terms are included in the energy function: the unary data term T_u constrains the similarity between the final saliency map and initial saliency map; the spatial smooth term T_s favors that all the similar and spatially adjacent superpixels in an individual image should be assigned to consistent saliency scores; and the holistic consistency term T_h enforces that the appearance of the salient objects should be consistent within the whole image group. Therefore, the energy function is defined as:

$$E = T_u + T_s + T_h = \sum_m (\overline{s}_m - s_m)^2 + \sum_{(m,n)\in\Omega} \omega_{mn} \cdot (\overline{s}_m - \overline{s}_n)^2 + \sum_m g_m \cdot \overline{s}_m^2 \tag{8}$$

where $\overline{s}_m$ denotes the refined saliency value of superpixel r_m; $s_m = S_a(r_m) \cdot S_r(r_m)$ is the initial saliency value of superpixel r_m by combining the intra and inter saliencies; Ω represents the spatially adjacent set in an individual image; ω_{mn} denotes the similarity between two superpixels, which is defined in the same way as Eq. (2); and $g_m = \chi^2(\boldsymbol{h}_m, \boldsymbol{h}_g)$ is the color difference between the superpixel r_m and global foreground model via the chi-square distance of Lab color histograms. The top 20 superpixels with larger initial saliency value in each image are regarded as the foreground samples to represent the global foreground distribution.

Let $\boldsymbol{s} = [s_m]_{\aleph\times 1}$, and $\overline{\boldsymbol{s}} = [\overline{s}_m]_{\aleph\times 1}$, where $\aleph = \sum_{i=1}^{N} N_i$ is the total number of superpixels in the whole image group. Then, the energy function is rewritten in the matrix forms as:

$$\mathbf{E} = (\overline{\boldsymbol{s}} - \boldsymbol{s})^T \cdot (\overline{\boldsymbol{s}} - \boldsymbol{s}) + \overline{\boldsymbol{s}}^T \cdot (\mathbf{D} - \mathbf{W}) \cdot \overline{\boldsymbol{s}} + \overline{\boldsymbol{s}}^T \cdot \mathbf{G} \cdot \overline{\boldsymbol{s}} \tag{9}$$

where $\mathbf{W} = [\omega_{mn}]_{\aleph\times\aleph}^{(m,n)\in\Omega}$ is the spatial color similarity matrix; $\mathbf{D} = \text{diag}\,(d_1, d_2, \ldots, d_\aleph)$ represents the degree matrix; $d_i = \sum_{j=1,(i,j)\in\Omega}^{\aleph} \omega_{ij}$; and $\mathbf{G} = \text{diag}\,(g_1, g_2, \ldots, g_\aleph)$ is the difference matrix between the superpixels and global foreground model.

The minimization of the above energy function can be solved by setting the derivative with respect to $\overline{\boldsymbol{s}}$ to be 0, which is represented as:

$$\frac{\partial \mathbf{E}}{\partial \overline{\boldsymbol{s}}} = 2(\overline{\boldsymbol{s}} - \boldsymbol{s}) + 2(\mathbf{D} - \mathbf{W}) \cdot \overline{\boldsymbol{s}} + 2\mathbf{G} \cdot \overline{\boldsymbol{s}} = 0 \tag{10}$$

Combining the like terms, the solution is given by:

$$\overline{\boldsymbol{s}} = [\mathbf{I} + (\mathbf{D} - \mathbf{W}) + \mathbf{G}]^{-1} \cdot \boldsymbol{s} \tag{11}$$

where $\mathbf{I}$ is an identity matrix with the size of $\aleph \times \aleph$.

IV. Experiments

In this section, we evaluate the proposed RGBD co-saliency detection method on the RGBD CoSal150 dataset and RGBD CoSeg183 dataset. The qualitative and quantitative comparisons with some state-of-the-art methods are presented, and some discussions and analyses are conducted.

A. Experimental Settings

In experiments, two public RGBD co-saliency detection datasets, named RGBD CoSal150 dataset[2] and RGBD CoSeg183 dataset[3], are used to evaluate the effectiveness of the proposed method. The RGBD CoSal150 dataset [67] consists of 150 RGBD images that are distributed in 21 indoor and outdoor scenes, and the pixel-level ground truth for each image is provided. The RGBD CoSeg183 dataset [28] is a challenging dataset due to the cluttered backgrounds, complex foreground patterns, and multiple objects, containing 183 RGBD images with pixel-level ground truth that are distributed in 16 indoor scenes. In our work, the number of superpixels for each image is set to 400, and the number of initial foreground seeds is set to $K = 40$. The project is available on our website.[4]

For quantitative evaluation, three criteria including the Precision-Recall (PR) curve, F-measure, and Mean Absolute

[2]https://rmcong.github.io/proj_RGBD_cosal.html
[3]http://hzfu.github.io/proj_rgbdseg.html
[4]https://rmcong.github.io/proj_RGBD_cosal_HSCS_tmm.html

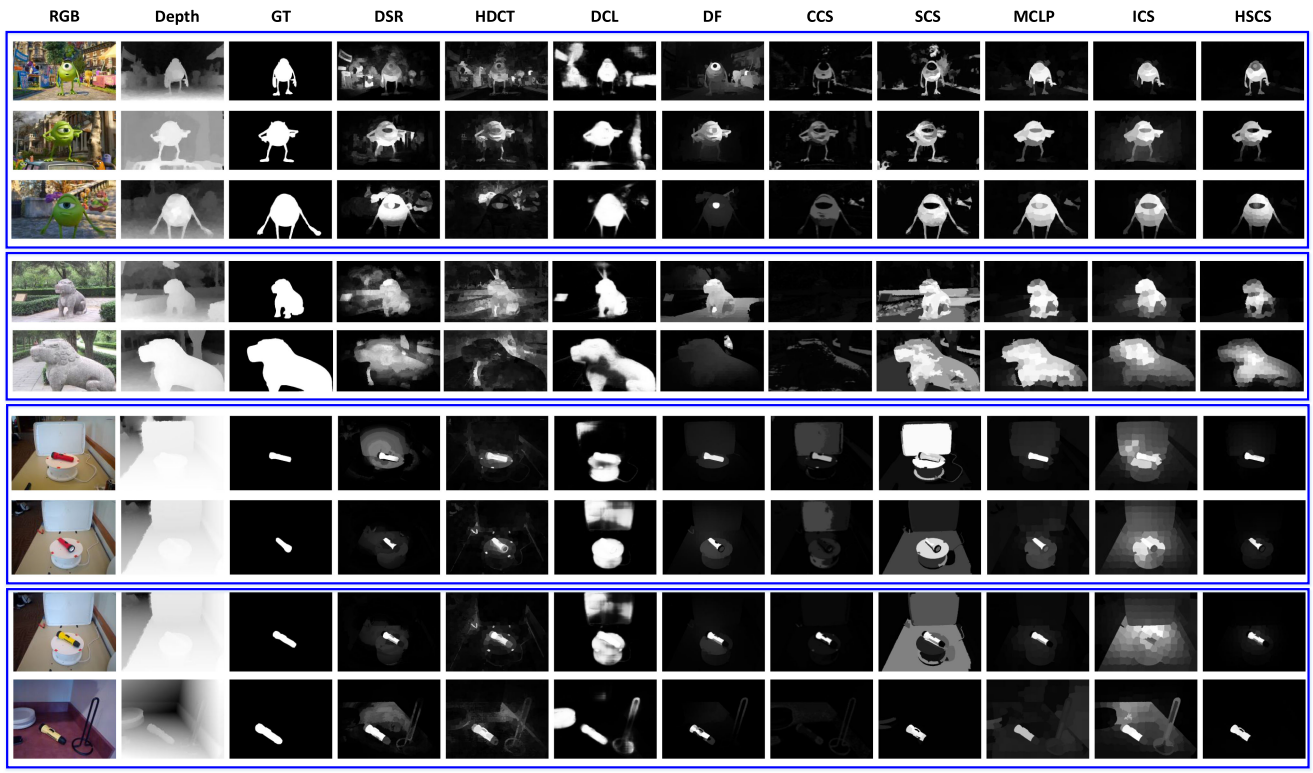


Fig. 3. Some visual examples of different methods.

Error (MAE) are introduced. The precision and recall scores are computed by comparing the binary saliency map against the ground truth, where the precision score represents the percentage of salient pixels which are allocated correctly in the obtained saliency map, and the recall value corresponds to the ratio of detected salient pixels with respect to the salient pixels in the ground truth. Thus, the PR curve can be drawn using the pairwise precision and recall scores. F-measure [72], [73] is an overall performance measurement, which is defined as the weighted mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2)\,\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{12}$$

where β^2 is set to 0.3 for emphasizing the precision as suggested in [74].

Mean Absolute Error (MAE) is calculated as the average pixel-wise difference between the obtained saliency map S and ground truth G [75], [76]:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(i,j) - G(i,j)| \tag{13}$$

where W and H are the width and height of the image, respectively.

B. Comparison With State-of-the-art Methods

We compared the proposed HSCS method with 17 state-of-the-art methods, including DSR [9], BSCA [10], DCLC [12], HDCT [14], SMD [15], DCL [17], DSS [19], R3Net [22], ACSD [36], DF [40], CTMF [41], PCFN [42], SCS [47], CCS [49], LRMF [53], ICS [67], and MCLP [68], where DCL, DSS, R3Net, DF, CTMF, and PCFN are the deep learning based methods. The visual comparisons are shown in Fig. 3, and the quantitative evaluations are reported in Fig. 4 and Table I.

In Fig. 3, four image groups, including the green cartoon in the virtual scene, sculpture in the outdoor scene, and the red and yellow flashlights in the indoor scene, are illustrated for visual comparison. Due to the lack of a high-level feature description and inter-image constraints, the unsupervised single image saliency detection methods (e.g., DSR [9], HDCT [14]) only roughly highlight the salient regions , while the background regions cannot be suppressed effectively (such as the street in the green cartoon group and the trees in the sculpture group). Benefitting from the strong learning ability of deep learning, the DCL [17] method achieves better performance with more consistent salient regions. However, there are still some wrongly detected backgrounds, such as the white object in the second image of the last group. Combining the depth cue and deep learning, the DF [40] method suppresses the background effectively, but it ignores the completeness of salient objects, such as the third image in the green cartoon group. For the RGB co-saliency detection methods (CCS [49] and SCS [47]), some foregrounds (such as the third image in the green cartoon group) are wrongly suppressed by the CCS method, and some backgrounds (such as the white board in the red flashlight group) are also inaccurately highlighted by the SCS method. Compared with the above methods, RGBD co-saliency detection methods (ICS [67] and MCLP [68]) achieve relatively superior performance with tangible salient objects. However, they still fail to suppress some common backgrounds, such as the ground in the sculpture group

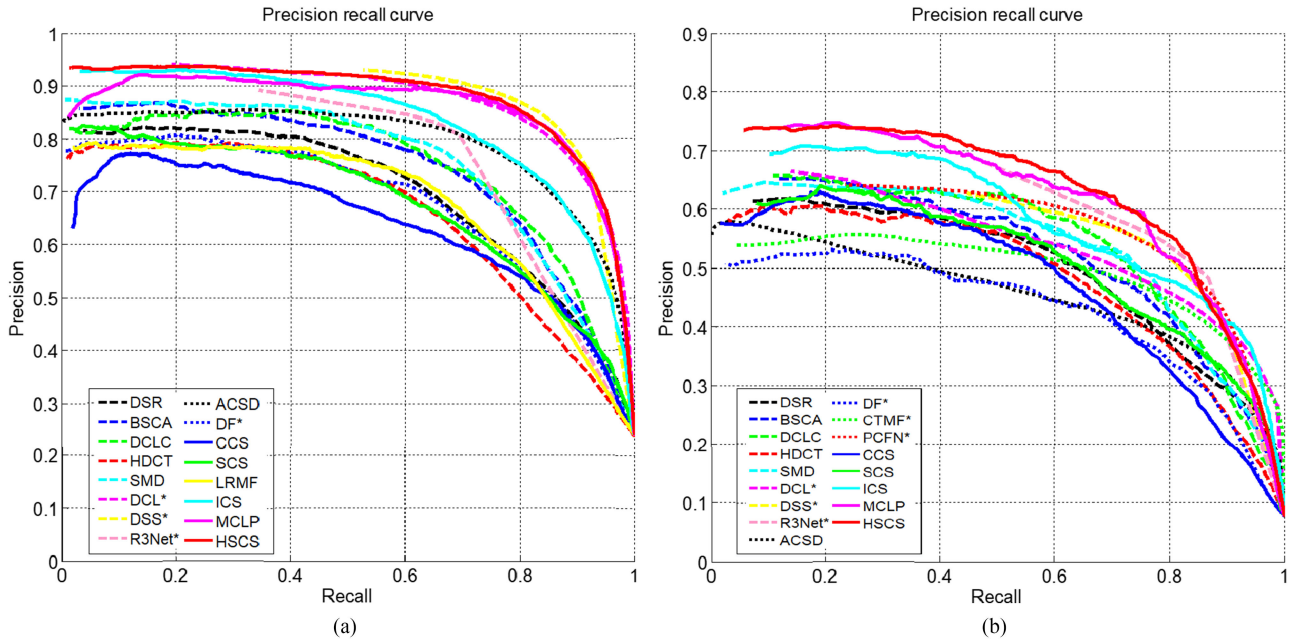


Fig. 4. PR curves of different methods on two RGBD co-saliency detection datasets, where "*" denotes the deep learning based methods. (a) RGBD CoSal150 dataset. (b) RGBD CoSeg183 dataset.

TABLE I
QUANTITATIVE COMPARISONS WITH DIFFERENT METHODS ON TWO DATASETS, WHERE "*" DENOTES THE DEEP LEARNING BASED METHODS

	RGBD CoSal150 Dataset		RGBD CoSeg183 Dataset	
	F-measure	MAE	F-measure	MAE
DSR [9]	0.6956	0.1867	0.5496	0.1092
BSCA [10]	0.7318	0.1925	0.5678	0.1877
DCLC [12]	0.7385	0.1728	0.5994	0.1097
HDCT [14]	0.6753	0.2146	0.5447	0.1307
SMD [15]	0.7494	0.1774	0.5760	0.1229
DCL* [17]	0.8345	0.1056	0.5531	0.0967
DSS* [19]	0.8540	0.0869	0.5972	0.0782
R3Net* [22]	0.7812	0.1296	0.6190	0.0678
ACSD [36]	0.7788	0.1806	0.4787	0.1940
DF* [40]	0.6844	0.1945	0.4840	0.1077
CTMF* [41]	–	–	0.5316	0.1259
PCFN* [42]	–	–	0.6049	0.0782
CCS [49]	0.6311	0.2138	0.5383	0.1210
SCS [47]	0.6724	0.1966	0.5553	0.1616
LRMF [53]	0.6995	0.1813	–	–
ICS [67]	0.7915	0.1790	0.6011	0.1544
MCLP [68]	0.8403	0.1370	0.6365	0.0979
HSCS	0.8500	0.1030	0.6466	0.0787

and the white board in the red flashlight group. By contrast, benefitting from the hierarchical reconstruction and global refinement, the proposed method can consistently highlight the salient objects and effectively suppress the backgrounds.

PR curves of different methods on two datasets are shown in Fig. 4. As can be seen, the proposed HSCS method reaches a higher precision on the whole PR curves. Moreover, the proposed method is even superior to some deep learning based methods (e.g., DCL [17], R3Net [22], DF [40], CTMF [41], and PCFN [42]). The quantitative measurements, including F-measure and MAE score, are reported in Table I. From the table, we can see that the proposed method achieves the competitive performance compared with 17 other state-of-the-art methods. On the RGBD CoSal150 dataset, the F-measure of the proposed method reaches 0.8500, and the maximum percentage gain reaches 34.7% compared with other methods. Especially, the proposed HSCS method also achieves the percentage gain of 8.8% compared with the deep learning based method (e.g., R3Net [22]). On the RGBD CoSeg183 dataset, the proposed method achieves the best performance in terms of F-measure, and the performance gains against others are more remarkable. The maximum percentage gain of the proposed method also reaches 35.1% in terms of F-measure. All these visual examples and quantitative measures demonstrate the effectiveness of the proposed method.

C. Discussions

In this section, we conduct some discussions, including the module analysis, depth and ranking scheme evaluation, parameter discussion, and running time.

1) Module Analysis: The key points of the proposed hierarchical sparsity-based co-saliency detection method for RGBD images include a hierarchical sparsity based inter saliency model and an energy function refinement model. For hierarchical sparsity based inter saliency generation, the global and pairwise sparsity reconstructions are used to capture the inter-image constraints from two aspects. We comprehensively evaluate each module on the RGBD CoSal150 dataset, and the F-measures are presented in Table II. The global inter saliency

TABLE II
F-MEASURE OF THE MAIN MODULES ON THE RGBD COSAL150 DATASET

Modules	F-measure
intra saliency detection	0.8348
global reconstruction	0.8145
pairwise reconstruction	0.7628
hierarchical inter saliency	0.8198
energy function refinement	0.8500

TABLE III
EVALUATION OF DEPTH AND RANKING SCHEME ON THE RGBD COSAL150 DATASET, WHERE "W/O" MEANS "WITHOUT" AND "W/" CORRESPONDS TO "WITH"

	F-measure
w/o depth and w/ ranking	0.7839
w/ depth and w/o ranking	0.8439
w/ depth and w/ ranking	0.8500

reconstruction captures the global corresponding relationship throughout the whole image group and achieves the F-measure of 0.8145. As a supplement, the multiple images relationship is formulated as pairwise correspondences by using the pairwise reconstruction model with a set of pairwise dictionaries, and the F-measure reaches 0.7628. Combining these two aspects, the hierarchical inter saliency structure can explore a more comprehensive inter-image relationship, and reaches 0.8198 in terms of F-measure, which is superior to most of the existing (co-)saliency detection methods (e.g., DSR [9], SMD [15], DF [40], SCS [47], LRMF [53], and ICS [67]). Finally, the co-saliency detection with energy function refinement achieves the best performance, and the percentage gain reaches 3.7% compared with the inter saliency models.

2) Depth and Ranking Scheme Evaluation: In this paper, the depth cue is not only served as a constraint in the inter-image correspondence modelling, but also used as a supplement of color information in the refinement component. In order to attain more robust and accurate foreground seeds for global dictionary construction, a ranking scheme is designed to filter the interferences and to obtain optimal foreground seeds. We conduct some experiments on the RGBD CoSal150 dataset to evaluate the influence of these two constraints, and the F-measures are reported in Table III. Compared with the first and the third rows, introducing the depth cue into the model, the performance is obviously improved with a percentage gain of 8.4%. Shown in the second and third rows, the performance with the ranking scheme is better than the model without the ranking scheme. In addition, some illustrations are shown in Fig. 2. As can be seen, with the ranking scheme, the correct foreground seeds are successfully reserved, while the backgrounds (such as the lawn in the second and third images) are effectively eliminated. All these data demonstrate the effectiveness of the depth information and ranking scheme.

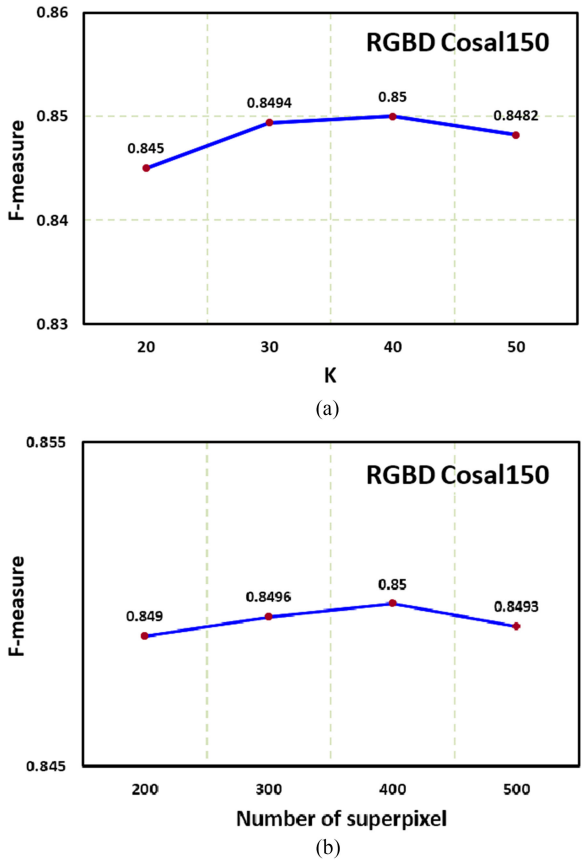


Fig. 5. The F-measure of different parameters on RGBD Cosal150 dataset. (a) The influence of different number of initial foreground seeds. (b) The influence of different number of superpixels.

3) Parameter Discussion: In this section, we mainly discuss the influence of different numbers of initial foreground seeds and superpixels. In the experiment, we evaluate one parameter by fixing the other parameters. The tendency chart of the F-measure is shown in Fig. 5. From Fig. 5(a), selecting 20 initial foreground seeds for each image is not enough to represent the common saliency attributes completely and degenerates the inter reconstruction result. As the seed number increase, the performance improves and reaches the optimum when K is set to 40. When K reaches 50, the performance of the algorithm begins to drop. The main reason for the drop after 50 is that too many seeds contain background regions and decrease the reconstruction accuracy. As mentioned above, the performance is not highly sensitive to the parameter K, and we set it to 40 in all experiments. In addition to the number of initial foreground seeds, we further discuss the influence of different numbers of superpixels in the experiments. From the curve shown in Fig. 5(b), when the number of superpixels is set to 400, the result achieves the best performance. In fact, the performance in different numbers of superpixels are similar, indicating that the proposed algorithm is insensitive to the number of superpixels.

4) Running Time: We compare the running time of the proposed method with others on a Quad Core 3.7 GHz workstation with 16GB RAM and implemented using MATLAB 2014a. The average running time is listed in Table IV. In general, compared

TABLE IV
COMPARISONS OF THE AVERAGE RUNNING TIME (SECONDS PER IMAGE) ON THE RGBD COSAL150 DATASET

Method	DCLC	SMD	DF	CCS	SCS	MCLP	ICS	HSCS
Time	1.96	7.49	12.95	2.65	2.94	41.03	42.67	8.29

with the image saliency detection method, co-saliency detection algorithm often requires more computation time, especially for the matching based methods (such as MCLP [68], ICS [67]). For the three RGBD co-saliency detection methods, under the same conditions, the MCLP method takes 41.03 seconds for one image, the ICS method takes 42.67 seconds, and the proposed HSCS method takes an average of 8.29 seconds to process one image. Since the commonly used superpixel-level matching process is replaced by the hierarchical sparsity based reconstruction to capture the inter-image correspondence, the computational efficiency of the proposed algorithm is clearly improved.

V. CONCLUSION

In this paper, a novel co-saliency detection method for RGBD images based on hierarchical sparsity reconstruction and energy function refinement is proposed. The major contribution lies in the hierarchical sparsity based inter saliency modelling, where the global inter-image model with a ranking scheme is used to capture the global characteristic among the whole image group through a common foreground dictionary, and the pairwise inter-image model is devoted to exploring the local corresponding relationship through a set of pairwise foreground dictionaries. In addition, an energy function refinement model is proposed to further improve the intra-image smoothness and inter-image consistency. The comprehensive comparisons and discussions on two RGBD co-saliency detection datasets have demonstrated that the proposed method outperforms other state-of-the-art methods qualitatively and quantitatively.

REFERENCES

[1] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Image cosegmentation via saliency-guided constraint clustering with cosine similarity," in *Proc. Assoc. Advancement Artif. Intell.*, 2017, pp. 4285–4291.

[2] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3395–3402.

[3] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[4] J. Lei *et al.*, "Depth-preserving stereo image retargeting based on pixel fusion," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1442–1453, Jul. 2017.

[5] C. Li, J. Guo, R. Cong, Y. Pang, and B. Wang, "Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5664–5677, Dec. 2016.

[6] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, "Saliency-aware nonparametric foreground annotation based on weakly labeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1253–1265, Jun. 2016.

[7] Y. Pang, H. Zhu, X. Li, and J. Pan, "Motion blur detection with an indicator function for surveillance robots," *IEEE Trans. Ind. Electron.*, vol. 63, no. 9, pp. 5592–5601, Sep. 2016.

[8] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 409–416.

[9] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. Int. Conf. Comput. Vision*, 2013, pp. 2976–2983.

[10] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 110–119.

[11] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 2814–2821.

[12] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, "Salient region detection via integrating diffusion-based compactness and local contrast," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3308–3320, Nov. 2015.

[13] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.

[14] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 9–23, Jan. 2016.

[15] H. Peng *et al.*, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[16] D.-P. Fan *et al.*, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. European Conference on Computer Vision*, 2018.

[17] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 478–487.

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

[19] Q. Hou *et al.*, "Deeply supervised salient object detection with short connections," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 5300–5309.

[20] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 212–221.

[21] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 6943–6950.

[22] Z. Deng *et al.*, "R^3 Net: Recurrent residual refinement network for saliency detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 684–690.

[23] L. Ye *et al.*, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, Aug. 2017.

[24] J. Ren, Z. Liu, X. Zhou, G. Sun, and C. Bai, "Saliency integration driven by similar images," *J. Vis. Commun. Image Representation*, vol. 50, pp. 227–236, Jan. 2018.

[25] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 4548–4557.

[26] D.-P. Fan *et al.*, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.

[27] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 1–31, Feb. 2018.

[28] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 4428–4436.

[29] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, "Object-based multiple foreground video co-segmentation via multi-state selection graph," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3415–3424, Nov. 2015.

[30] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.

[31] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1464–1471.

[32] A. Toshev, J. Shi, and K. Daniilidis, "Image matching via saliency region correspondences," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2007, pp. 1–8.

[33] J. Yang, Z. Gan, K. Li, and C. Hou, "Graph-based segmentation for RGB-D data using 3-D geometry enhanced superpixels," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 913–926, May 2015.

[34] M. Ni *et al.*, "Color-guided depth map super resolution using convolutional neural network," *IEEE Access*, vol. 2, pp. 26666–26672, 2017.

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

[36] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *Signal Process., Image Commun.*, vol. 38, pp. 115–126, Oct. 2015.
[37] J. Guo, T. Ren, and J. Bei, "Salient object detection in RGB-D image via saliency evolution," in *Proc. Int. Conf. Multimedia Expo*, 2016, pp. 1–6.
[38] R. Cong *et al.*, "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.
[39] H. Song *et al.*, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.
[40] L. Qu *et al.*, "RGBD salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.
[41] 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.
[42] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3051–3060.
[43] H. Li and K. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.
[44] Z. Tan, L. Wan, W. Feng, and C.-M. Pun, "Image co-saliency detection by propagating superpixel affinities," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 2114–2118.
[45] H. Li, F. Meng, and K. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1869–1909, Dec. 2013.
[46] Z. Liu, W. Zou, L. Li, L. Shen, and O. L. Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 88–92, Jan. 2014.
[47] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 588–592, May 2015.
[48] L. Ye, Z. Liu, J. Li, W.-L. Zhao, and L. Shen, "Co-saliency detection via co-salient object discovery and recovery," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2073–2077, Nov. 2015.
[49] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.
[50] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3196–3209, Jul. 2017.
[51] L. Wu, Z. Liu, H. Song, and O. Meur, "RGBD co-saliency detection via multiple kernel boosting and fusion," *Multimedia Tools Appl.*, no. 5, pp. 21185–21199, Aug. 2018.
[52] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.
[53] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *Proc. Int. Conf. Multimedia Expo*, 2015, pp. 1–6.
[54] C. Ge, K. Fu, F. Liu, L. Bai, and J. Yang, "Co-saliency detection via inter and intra saliency propagation," *Signal Process., Image Commun.*, vol. 44, pp. 69–83, May 2016.
[55] R. Huang, W. Feng, and J. Sun, "Color feature reinforcement for cosaliency detection without single saliency residuals," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 569–573, May 2017.
[56] C.-C. Tsai, X. Qian, and Y.-Y. Lin, "Segmentation guided local proposal fusion for co-saliency detection," in *Proc. Int. Conf. Multimedia Expo*, 2017, pp. 523–528.
[57] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2994–3002.
[58] L. Wei, S. Zhao, O. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3041–3047.
[59] D. Zhang *et al.*, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 594–602.
[60] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.
[61] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based for co-saliency detection framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2018.
[62] J. Yang, Y. Zhu, K. Li, J. Yang, and C. Hou, "Tensor completion from structurally-missing entries by low-tt-rankness and fiber-wise sparsity," *IEEE J. Sel. Topics Signal Process.*, pp. 1–15, 2017, doi: 10.1109/JSTSP.2018.2873990.
[63] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 5216–5223.
[64] H. Lu, X. Li, L. Zhang, X. Ruan, and M.-H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1592–1603, Apr. 2016.
[65] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Dense and sparse labeling with multi-dimensional features for saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1130–1143, May 2018.
[66] H. Song, Z. Liu, Y. Xie, L. Wu, and M. Huang, "RGBD co-saliency detection via bagging-based clustering," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1722–1726, Dec. 2016.
[67] R. Cong *et al.*, "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.
[68] R. Cong *et al.*, "An iterative co-saliency framework for RGBD images," *IEEE Trans. Cybern.*, pp. 1–14, 2017, doi: 10.1109/TCYB.2017.2771488.
[69] R. Achanta *et al.*, "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.
[70] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.
[71] T. Leung and J. Malik, "Recognizing surface using three-dimensional textons," in *Proc. Int. Conf. Comput. Vision*, 1999, pp. 1010–1017.
[72] R. Cong *et al.*, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–19, 2018, doi: 10.1109/TCSVT.2018.2870832.
[73] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.
[74] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[75] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.
[76] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.

Runmin Cong received the M.S. degree in communication and information system from the Civil Aviation University of China, Tianjin, China, in 2014. He is currently working toward the Ph.D. degree in information and communication engineering at Tianjin University, Tianjin, China. From December 2016 to February 2017, he was a visiting student with Nanyang Technological University, Singapore.

Since May 2018, he has been a Research Associate with the Department of Computer Science, City University of Hong Kong, Hong Kong. His research interests include computer vision, image processing, saliency detection, and 3-D imaging.

Mr. Cong is a Reviewer for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, etc.

Jianjun Lei (M'11–SM'17) received the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007.

From August 2012 to August 2013, he was a visiting Researcher with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA. He is currently a Professor with Tianjin University, Tianjin, China. His research interests include 3-D video processing, virtual reality, and artificial intelligence.

Dr. Lei is on the editorial board of *Neurocomputing and China Communications*.

Huazhu Fu (SM'18) received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2013.

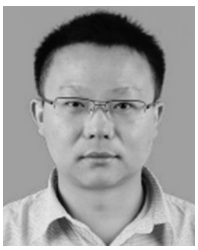
From 2013 to 2015, he was a Research Fellow with Nanyang Technological University, Singapore. From 2015 to 2018, he was a Research Scientist with the Institute for Infocomm Research, Agency for Science, Technology, and Research, Singapore. He is currently a Senior Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, image processing, and medical image analysis.

Dr. Fu is an Associate Editor for IEEE ACCESS and *BMC Medical Imaging*.

Qingming Huang (SM'08–F'18) received the bachelor's degree in computer science and the Ph.D. degree in computer engineering from the Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has published more than 400 academic papers in prestigious international journals including IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, etc., and top-level conferences such as ACM Multimedia, International Conference on Computer Vision, Conference on Computer Vision and Pattern Recognition, International Joint Conference on Artificial Intelligence, Very Large Data Bases Conference, etc. His research interests include multimedia video analysis, image processing, computer vision, and pattern recognition.

Dr. Huang is an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Acta Automatica Sinica*, and a Reviewer for various international journals including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, etc. He has been a General Chair, Program Chair, Track Chair, and Technical Program Committee member for various conferences, including ACM Multimedia, Conference on Computer Vision and Pattern Recognition, International Conference on Computer Vision, International Conference on Multimedia and Expo, Pacific-Rim Conference on Multimedia, Pacific-Rim Symposium on Image and Video Technology, etc.



Xiaochun Cao (SM'14) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA.

After graduation, he spent about three years as a Research Scientist with ObjectVideo Inc., Tysons, VA, USA. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 100 journal and conference papers.

Dr. Cao is a Fellow of the IET. He is on the editorial boards of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His dissertation was nominated for the University of Central Florida's University-Level Outstanding Dissertation Award. In 2004 and 2010, he was the recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

Nam Ling (S'88–M'90–SM'99–F'08) received the B.Eng. degree from the National University of Singapore, Singapore, in 1981, and the M.S. and Ph.D. degrees from the University of Louisiana, Lafayette, LA, USA, in 1985 and 1989, respectively.

From 2002 to 2010, he was an Associate Dean with the School of Engineering, Santa Clara University, Santa Clara, CA, USA. He is currently the Sanfilippo Family Chair Professor and the Chair for the Department of Computer Engineering, Santa Clara University. He is also a Consulting Professor with the National University of Singapore, a Guest Professor for Tianjin University, Tianjin, China, a Guest Professor for Shanghai Jiao Tong University, Shanghai, China, a Cuiying Chair Professor for Lanzhou University, Gansu, China, and a Distinguished Professor for Xian University of Posts and Telecommunications, Shaanxi, China. He has authored or coauthored more than 180 publications and standard contributions, including two books in the fields of video coding and systolic arrays. He has filed/granted over 15 U.S. patents.

Dr. Ling is an IET Fellow. He was named as an IEEE Distinguished Lecturer twice and was also an APSIPA Distinguished Lecturer. He received the IEEE ICCE Best Paper Award (First Place). He was a recipient of six awards from Santa Clara University, four at the University level (Outstanding Achievement, Recent Achievement in Scholarship, Presidents Recognition, and Sustained Excellence in Scholarship) and two at the School/College level (Researcher of the Year and Teaching Excellence). He was a Keynote Speaker for the IEEE Asia Pacific Conference on Circuits and Systems, the Visual Perception and Visual Computing Conference (twice), Joint Conferences on Pervasive Computing, the IEEE International Conference on Adaptive Science and Technology, teh IEEE Conference on Industrial Electronics and Applications, the IET FC Umedia, the IEEE Umedia, and the Workshop at the Xian University of Posts and Telecommunications, as well as a Distinguished Speaker for IEEE Conference on Industrial Electronics and Applications. He was a General Chair/Co-Chair for IEEE Hot Chips, Visual Perception and Visual Computing (twice), the IEEE International Conference on Multimedia and Expo, IEEE Umedia (thrice), and the IEEE Workshop on Signal Processing Systems. He was also a Technical Program Co-Chair for the IEEE International Symposium on Circuits and Systems, the APSIPA Annual Summit and Conference , the IEEE Asia Pacific Conference on Circuits and Systems, the IEEE Workshop on Signal Processing Systems (twice), the International Workshop on Digital and Computational Video, and the IEEE International Conference on Visual Communications and Image Processing. He was a Technical Committee Chair for the IEEE Circuits and Systems and Communications Technical Committee and the IEEE Technical Committee on Microprocessors and Microcomputers. He was a Guest Editor or Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I: REGULAR PAPERS, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, Springer *Journal of Signal Processing Systems*, Springer *Multidimensional Systems and Signal Processing*, etc.