Going From RGB to RGBD Saliency: A Depth-Guided Transformation Model

Runmin Cong, *Member, IEEE*, Jianjun Lei, *Senior Member, IEEE*, Huazhu Fu, *Senior Member, IEEE*, Junhui Hou, *Member, IEEE*, Qingming Huang, *Fellow, IEEE*, and Sam Kwong, *Fellow, IEEE*

***Abstract*—Depth information has been demonstrated to be useful for saliency detection. However, the existing methods for RGBD saliency detection mainly focus on designing straightforward and comprehensive models, while ignoring the transferable ability of the existing RGB saliency detection models. In this article, we propose a novel depth-guided transformation model (DTM) going from RGB saliency to RGBD saliency. The proposed model includes three components, that is: 1) multilevel RGBD saliency initialization; 2) depth-guided saliency refinement; and 3) saliency optimization with depth constraints. The explicit depth feature is first utilized in the multilevel RGBD saliency model to initialize the RGBD saliency by combining the global compactness saliency cue and local geodesic saliency cue. The depth-guided saliency refinement is used to further highlight the salient objects and suppress the background regions by introducing the prior depth domain knowledge and prior refined depth shape. Benefiting from the consistency of the entire object in the depth map, we formulate an optimization model to attain more consistent and accurate saliency results via an energy function, which integrates the unary data term, color smooth term, and depth consistency term. Experiments on three public RGBD saliency detection benchmarks demonstrate the effectiveness and performance improvement of the proposed DTM from RGB to RGBD saliency.**

***Index Terms*—Depth cue, energy function optimization, refined depth shape prior (RDSP), RGBD images, saliency detection, transformation model.**

Manuscript received October 15, 2018; revised February 24, 2019 and May 7, 2019; accepted July 22, 2019. Date of publication August 20, 2019; date of current version July 10, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002900, in part by the Fundamental Research Funds for the Central Universities under Grant 2019RC039, in part by the National Natural Science Foundation of China under Grant 61722112, Grant 61520106002, Grant 61731003, Grant 61836002, Grant 61620106009, Grant U1636214, Grant 61873142, Grant 61772344, and Grant 61672443, in part by the Key Research Program of Frontier Sciences, CAS under Grant QYZDJ-SSW-SYS013, in part by Hong Kong Research Grants Council (RGC) General Research Funds under Grant 9042038 (CityU 11205314) and Grant 9042322 (CityU 11200116), and in part by Hong Kong RGC Early Career Schemes under Grant 9048123. This article was recommended by Associate Editor H. Lu. *(Corresponding author: Jianjun Lei.)*

R. Cong is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: rmcong@bjtu.edu.cn).

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

H. Fu is with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE (e-mail: hzfu@ieee.org).

J. Hou and S. Kwong are with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518000, China (e-mail: jh.hou@cityu.edu.hk; cssamk@cityu.edu.hk).

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

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

Digital Object Identifier 10.1109/TCYB.2019.2932005

I. Introduction

SIMULATING the human visual attention mechanism, salient object detection aims at locating and segmenting the interesting part or attractive object from a given scene [1], which has been widely applied to many vision tasks, such as retrieval [2], segmentation [3], enhancement [4], and quality assessment [5]. According to the different processing data, saliency detection tasks can be roughly divided into three categories, including: 1) image saliency detection for the individual image [6]–[20]; 2) co-saliency detection for the multiple images [21]–[24]; and 3) video saliency detection for the video sequences [25]–[28]. When faced with a scene, humans cannot only capture the appearance of the object through the visual system but also perceive the depth information of the scene [29], [30]. Benefiting from the recent development of 3-D sensing technology, depth representation of the scene can be captured more conveniently and accurately. The effectiveness of depth information has been demonstrated in many computer vision tasks, such as image segmentation [31]; super resolution [32], [33]; and saliency detection [34]–[49].

In the existing RGBD saliency detection methods, depth information is mainly used in two ways, that is, one directly and explicitly incorporates it into the feature pool as a supplement to the color feature, and the other is to capture the implicit attributes from the depth map through some designed depth descriptors. However, these methods mainly focus on designing a straightforward and comprehensive RGBD saliency detection model. In fact, for RGBD saliency detection, more efforts should be made to make full use of depth information with the assistance of the existing RGB saliency models. In this article, we propose a novel depth-guided transformation model (DTM), which effectively exploits any existing RGB saliency model to work well in RGBD saliency scenarios. In the proposed DTM, depth information is utilized in three aspects: 1) saliency initialization; 2) saliency refinement; and 3) saliency optimization.

Saliency Initialization: For different RGB saliency detection methods, their performances vary greatly with respective superiority and drawback, especially for some challenging

cases, such as the complex scenes and various target types. Given all that, with the help of the explicit depth cue, a multilevel RGBD saliency method integrating the global compactness saliency cue and local geodesic saliency cue are proposed to generate RGBD saliency initialization of the proposed transformation model. The global compactness saliency cue combines color compactness and depth compactness into a formulation as a robust global saliency representation. The local geodesic saliency cue introduces the novel depth weight and graph relationship to measure the saliency from the local perspective. Finally, the original RGB saliency is combined with the global and local saliency cues to achieve RGBD saliency initialization, where the depth cue works as an explicit supplement to the color feature.

Saliency Refinement: In addition to the explicit depth feature information, some implicit attributes captured from the depth map are useful to refine the saliency map, such as the depth domain knowledge and depth shape cue. Therefore, we propose a depth-guided saliency refinement model to further highlight the salient objects and suppress the background regions by introducing the depth domain knowledge prior and refined depth shape prior (RDSP). In general, the photographer usually places the salient object closer to the camera. Thus, the depth value of the salient object is different from the distant background region. Based on this prior knowledge, the prior depth domain knowledge is used to describe the depth distance mapping information. In addition, an RDSP is studied to capture the shape information from the depth map, which introduces the color consistency constraint and refines optimal seeds selection. The RDSP can obtain more accurate and complete shape attributes from the depth map.

Saliency Optimization: From the depth map, the interior of the entire object usually has a consistent depth distribution. In other words, the depth information is beneficial to improve the consistency and smoothness of the acquired saliency map. Thus, we formulate an energy function-based optimization model to attain more consistent and accurate saliency results, which integrates the unary data term, color smooth term, and depth consistency term. The data term controls the updating degree of the optimization algorithm, the color smooth term restricts the spatially adjacent regions with similar color appearance and should be assigned to approximate saliency scores, and the depth consistency term enforces that adjacent regions with similar depth distribution should be assigned to consistent saliency scores.

The main idea of our method is to adapt any existing RGB saliency model to RGBD images, which could inherit the performance of RGB image saliency and utilize the depth information to enhance performance. The contributions of this article are summarized as follows.

1) The biggest advantage of our method is to fully exploit the depth cue and provide a general transformation model going from RGB saliency to RGBD saliency. The proposed model enables any existing RGB saliency model to work well in RGBD saliency scenarios with significant performance improvement.
2) Multilevel RGBD saliency initialization is proposed to integrate the global compactness and the local geodesic saliency cues, where the depth feature is used as a supplement to the color information.
3) To capture more accurate and complete shape information from the depth map, an RDSP is proposed, which considers the color consistency constraint and updates the optimal seeds selection.
4) To improve the accuracy and consistency, an optimization strategy with depth constraints is designed, which introduces the depth consistency relationship as an additional term in the energy optimization function.

The remainder of this article is organized as follows. The related works on the RGB saliency detection and RGBD saliency detection are introduced in Section II. Section III presents the details of the proposed depth-guided transformation framework. The experimental comparisons and analyses are discussed in Section IV. Finally, the conclusion is drawn in Section V.

II. Related Work

A. RGB Image Saliency Detection

The past few decades have witnessed the considerable technology development and encouraging performance improvement of saliency detection for RGB image, and numerous bottom-up and top-down models have been presented [6]–[20].

In [7], saliency detection is modeled as the dense and sparse reconstruction process, and the reconstruction error is used to measure the saliency of a region. Zhu *et al.* [8] proposed a principled optimization framework to achieve saliency detection by using a robust boundary connectivity measure. In [11], saliency detection is formulated as a structured matrix decomposition problem guided by high-level priors. Yuan *et al.* [12] proposed a novel saliency detection method with reversion correction and regularized random walk ranking, and obtained competitive performance. Recently, deep learning has demonstrated the superior performance in saliency detection. Han *et al.* [13] proposed a bottom-up salient object detection framework based on the background prior, where more powerful representations are learned from the stacked denoising autoencoder, and the separation of salient objects from backgrounds is formulated as a problem of measuring reconstruction residuals of deep autoencoders. Li and Yu [14] proposed an end-to-end deep contrast network for saliency detection, including the multiscale fully convolutional stream and the segment-wise spatial pooling stream. Zhang *et al.* [15] proposed an encoder–decoder fully convolutional network with reformulated dropout and hybrid upsampling strategies to detect the salient object. In [16], short connections are introduced into the skip-layer structures within the holistically nested edge detector architecture to achieve saliency detection. Deng *et al.* [18] proposed a recurrent residual refinement network (R^3Net) for saliency detection, where residual refinement blocks are utilized to recurrently learn the difference between the coarse saliency map and the ground truth by alternatively harnessing the low-level and high-level features. Moreover, some works focus on the performance evaluation for the saliency detection task [19] or exploit some new data sources [20].

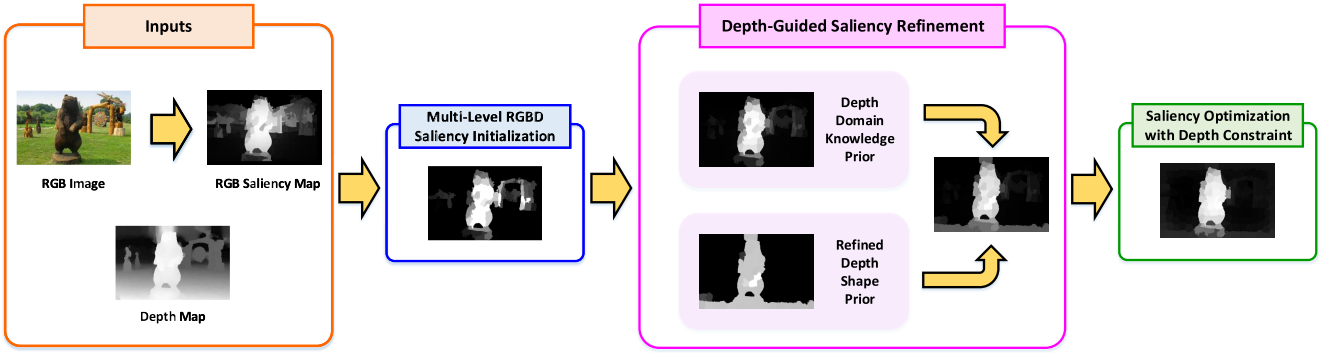

Fig. 1. Flowchart of the proposed DTM.

B. RGBD Image Saliency Detection

The depth information is introduced as an additional feature or a novel depth measure to enhance the identification of the salient object from the RGBD images, and lots of methods have been proposed [34]–[49].

Fang *et al.* [35] combined the color, luminance, texture, and depth features to calculate the feature contrast and produce the stereoscopic saliency map. In [36], an anisotropic center-surround difference (ACSD) measure is proposed to measure the depth-aware saliency map. Feng *et al.* [38] proposed a local background enclosure (LBE) measure to capture the salient structure from the depth map. Cong *et al.* [39] proposed a saliency detection method for RGBD images based on the depth confidence analysis and multiple cues fusion, where the depth confidence measure is used to evaluate the quality of the depth map. Wang and Wang [41] proposed an RGBD saliency detection method by using the minimum barrier distance (MBD) transform and multilayer cellular automata-based saliency fusion. Peng *et al.* [44] calculated the depth saliency through a multicontextual contrast model considering the contrast prior, global distinctiveness, and background cue of the depth map. Moreover, a multistage RGBD saliency model was proposed by combining the low-level feature contrast, mid-level region grouping, and high-level prior enhancement. The main difference between our method and the work [44] lies in how to utilize the depth information. In [44], the depth information is used as a feature to calculate the local, global, and background contexts. In contrast, the use of the depth information in our method is more comprehensive and thorough. First, the depth cue works as an explicit supplement to the color feature in the saliency initialization model by combining the global compactness and local geodesic saliency cues. Second, some implicit attributes captured from the depth map are used to further refine the saliency map, such as the depth domain knowledge prior and RDSP. Third, inspired by the observation that the interior of the entire object usually has a consistent depth distribution, the depth cue is utilized to improve the consistency and smoothness of the acquired saliency map through an energy function-based optimization model.

The deep-learning technique is also introduced into RGBD saliency detection and achieves competitive performance. Qu *et al.* [45] designed a convolutional neural network to learn the interaction between the low-level cues and saliency result for RGBD saliency detection, where the raw saliency feature vectors are taken as the input. Han *et al.* [46] proposed a saliency detection method for RGBD images based on the convolutional neural network, which transfers the structure of the color deep network to be applicable for depth view and fuses both views automatically to obtain the final saliency map. Chen and Li [47] proposed an end-to-end RGBD salient object detection network, which fuses both cross-modal and cross-level features complementarily. Chen *et al.* [48] presented a multiscale multipath fusion network with cross-modal interactions for the RGBD saliency detection, which advances the traditional two-stream fusion architecture with a single-fusion path by diversifying the fusion paths and introducing the cross-modal interactions in multiple layers. Chen and Li [49] proposed a three-stream attention-aware multimodal fusion network for the RGBD saliency detection, where the cross-modal distillation stream is used to augment the RGB-D representation capacity in the bottom-up path, and the channel-wise attention mechanism is introduced to adaptively select the complementary feature maps in the top-down inference path.

Most of the above-mentioned RGBD saliency detection methods are mainly devoted to designing a new model, while ignoring the transfer ability and superior performance of the existing RGB saliency detection models. Therefore, in this article, we propose a DTM, which transfers the existing RGB saliency model to RGBD saliency scenarios.

III. Proposed Method

As shown in Fig. 1, we make full use of the depth information to enhance the saliency performance and propose a transformation model from RGB to RGBD saliency. There are three main steps in the proposed framework. 1) The multilevel RGBD saliency initialization integrates the global compactness and the local geodesic saliency cues to generate the stereoscopic saliency initialization, where the depth feature is explicitly used as a supplement to the color information. 2) The depth-guided saliency refinement focuses on further highlighting the salient objects and suppressing the background regions. It is a propagation procedure to refine and update the stereoscopic saliency initialization by exploring

the implicit depth information, that is, depth domain knowledge and depth shape constraint. 3) The saliency optimization with depth consistency constraints is designed to improve the accuracy and consistency through an energy function optimization considering depth consistency. It is called the optimization model because we design a holistic energy function to obtain the optimized saliency result by solving the optimization problem. The details will be introduced in the following sections.

A. Multilevel RGBD Saliency Initialization

Due to the lack of depth information, the original RGB saliency detection algorithm may fail to accurately highlight the salient objects and effectively suppress the background regions. In order to exploit the depth feature and guarantee the basic performance of the proposed transformation model, a multilevel RGBD saliency model is proposed to generate the RGBD saliency initialization, where the global compactness saliency cue is worked on as a robust global representation combining the color compactness and depth compactness, and the local geodesic saliency cue is utilized to measure the saliency from the local perspective with the novel depth weight and graph relationship.

1) Global Compactness Saliency Cue: Inspired by the fact that the salient region has a compact spread in the spatial domain, while the background region owns a larger spread over the entire image, compactness prior is thus widely used to distinguish the salient object and background region in the color space. In fact, there is a similar spatial distribution characteristic in the depth domain, that is, the salient region has a centralized depth distribution near the image center. The compactness cue is a global descriptor that does not rely on any assumptions and describes the spatial distribution of the entire image in the given domain. Motivated by this, we integrate the color compactness and depth compactness into a formulation to define global saliency.

Following the existing works [44], the input RGB image I is first abstracted into some compact and homogenous superpixels $\mathbf{R} = \{r_m\}_{m=1}^{N}$ by the SLIC algorithm [50], where N is the number of superpixels. Here, the superpixel segmentation algorithm considering the depth information also can be applied, which may further improve saliency performance. Then, the similarity between two superpixels in the Lab color space and depth space are, respectively, defined as

$$\begin{cases} a_{ij}^{c} = \exp\left(-||\boldsymbol{c}_i - \boldsymbol{c}_j||_2/\sigma^2\right) \\ a_{ij}^{d} = \exp\left(-\lambda_d \cdot |d_i - d_j|/\sigma^2\right) \end{cases} \tag{1}$$

where $\boldsymbol{c}_i$ represents the mean color value of superpixel r_i in Lab color space, d_i is the mean depth value of superpixel r_i, $\sigma^2 = 0.1$ is a constant to control the strength of the similarity, $\lambda_d = \exp((1 - m_d) \cdot CV \cdot H) - 1$ is the depth confidence measure [39], m_d is the mean value of the depth map, CV denotes the coefficient of variation, and H represents the depth frequency entropy. The better the quality the depth map is, the higher the value λ_d is.

Combining the color and depth attributes, the global compactness saliency cue is defined as

$$S_c(r_i) = 1 - \frac{\sum_{j=1}^{N} n_j \cdot \left(a_{ij}^{c} \cdot ||\boldsymbol{b}_j - \boldsymbol{u}_i||_2 + a_{ij}^{d} \cdot ||\boldsymbol{b}_j - \boldsymbol{p}_0||_2\right)}{\sum_{j=1}^{N} n_j \cdot \left(a_{ij}^{c} + a_{ij}^{d}\right)} \tag{2}$$

where a_{ij}^{c} and a_{ij}^{d} denote the color and depth similarities between superpixels r_i and r_j, respectively; n_j is the size of superpixel r_j; $\boldsymbol{b}_j = [b_j^x, b_j^y]$ is the centroid coordinate of superpixel r_j; $\boldsymbol{p}_0$ represents the coordinate of the image center; and $\boldsymbol{u}_i = [u_i^x, u_i^y] = [([\sum_{j=1}^{N} a_{ij}^{c} \cdot n_j \cdot b_j^x]/[\sum_{j=1}^{N} a_{ij}^{c} \cdot n_j]), ([\sum_{j=1}^{N} a_{ij}^{c} \cdot n_j \cdot b_j^y]/[\sum_{j=1}^{N} a_{ij}^{c} \cdot n_j])]$ is the color spatial mean. The compactness cue describes the global saliency by considering the color and depth attributes, and a higher value indicates the larger saliency probability.

2) Local Geodesic Saliency Cue: In an image, background regions are more easily connected to the image boundaries than the foreground regions. According to this observation, the saliency of a region can be calculated as the length of its shortest path to the background nodes, which is called the geodesic saliency measure [6]. In formulation, a virtual background node connected to all boundary regions is added to compute the saliency of the boundary regions, and the geodesic saliency of a region is defined as the accumulated edge weights along the shortest path from the region to the virtual background node on the graph. In this article, we calculate the geodesic saliency with the assistance of novel depth weight and optimized graph relationship from the local perspective.

First, an undirected weighted graph $G = (\upsilon, \varepsilon)$ is constructed, where υ denotes the set of nodes, including all of the superpixels $\{r_m\}_{m=1}^{N}$ plus a virtual background node B, and ε represents the set of edge link relationships between superpixels. In our model, three types of edges are defined: 1) the neighbor edge that connects the adjacent superpixels; 2) the boundary edge that connects the superpixel near the image boundary to the virtual background node; and 3) the background edge that connects the given background superpixels to the virtual background node, which is denoted as

$$\begin{aligned} \varepsilon = &\left\{(r_i, r_j)|\ r_i \text{ is adjacent to } r_j\right\} \\ &\cup \{(r_i, BV)|r_i \text{ is on image boundary}\} \\ &\cup \{(r_i, BG)|r_i \text{ is given background}\}. \end{aligned} \tag{3}$$

In the graph, the background edge is introduced to further constrain the link relationship generation. Considering the RGB saliency, depth cue, and background connectivity probability, a measurement is designed to evaluate the probability that a region belongs to the background, which is denoted as

$$P_b(r_i) = P_c(r_i) \cdot \exp\left(\frac{S_{\mathrm{RGB}}(r_i) + \lambda_d \cdot d_i}{\sigma^2}\right) \tag{4}$$

where $P_c(r_i)$ is the background connectivity probability of superpixel r_i defined as [8], $S_{\mathrm{RGB}}(r_i)$ is the input RGB saliency value of superpixel r_i, and d_i is the mean depth value of superpixel r_i. The smaller the RGB saliency and depth values, the higher the background connectivity probability, and the larger the P_b value is, indicating the superpixel is more likely to be a

background region. Then, the top 20% superpixels with larger P_b values will be selected as the background seeds to generate the boundary edges.

The edge weight between two superpixels is represented as

$$w_{ij} = \begin{cases} a_{ij}^{c} \cdot a_{ij}^{d}, & \text{if } (r_i, r_j) \in \varepsilon \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $(r_i, r_j) \in \varepsilon$ denotes the connected superpixels r_i and r_j on the graph.

Finally, the local geodesic saliency cue of superpixel r is calculated as the accumulated edge weights along the shortest path from r to a virtual background node B on the graph

$$S_G(r) = \min_{r_1=r, r_2, \ldots, r_N=B} \sum_{i=1}^{N} w_{i,i+1}, \ (r_i, r_{i+1}) \in \varepsilon. \tag{6}$$

3) RGBD Saliency Initialization: Considering that compactness saliency is a global measure that does not rely on any assumptions and original saliency result, we first fuse it with the intersection between the global compactness saliency and local geodesic saliency. Then, combined with the RGB saliency result to generate the RGBD saliency initialization as

$$S_{ML}(r_i) = \frac{1}{2}(S_{\mathrm{RGB}}(r_i) + N[S_C(r_i) + S_C(r_i) \cdot S_G(r_i)]) \tag{7}$$

where $S_{\mathrm{RGB}}(r_i)$, $S_C(r_i)$, and $S_G(r_i)$ denote the input RGB saliency, global compactness saliency, and local geodesic distance saliency of superpixel r_i, respectively, and $N[\cdot]$ is a min–max normalization function.

B. Depth-Guided Saliency Refinement

The multilevel RGBD saliency model provides an initialization of the transformation framework by using the explicit depth feature. In the depth-guided saliency refinement, we exploit the implicit depth information to refine the saliency map. Generally, the salient object is placed near the camera by a photographer when taking a picture. Thus, the object with a large depth magnitude tends to be salient. Moreover, the depth distribution between the foreground and background regions is different. Therefore, the depth domain knowledge prior, including the depth distance and depth contrast, is proposed to refine the saliency map. In addition, although the depth map does not provide rich texture information as a color image, it provides effective shape attribute representation. Based on this, the RDSP refinement is proposed to capture the shape constraint from the depth map and refine the salient region.

1) Depth Domain Knowledge Prior: From a depth map, we can observe that: a) the salient object tends to be close the camera with a large depth magnitude and b) the salient object could be identified by the depth contrast compared with the background regions. However, limited by the depth-sensing technology, the poor quality of the depth map may degenerate the saliency performance. Thus, according to the quality of the depth map, the depth contrast and depth weighting are used as the prior depth domain knowledge to refine the RGBD saliency initialization. When the quality of the depth map is reliable (i.e., $\lambda_d \geq \tau_1$), the depth contrast could better describe the depth saliency characteristic, which is directly used to refine the initial saliency map. When the quality of the depth map is tolerable (i.e., $\tau_2 \leq \lambda_d < \tau_1$), the depth distance relationship is used to weigh the initial saliency map. When the depth map quality is poor (i.e., $\lambda_d < \tau_2$), which is unable to provide enough effective and accurate auxiliary information for saliency detection, we only retain the initial saliency result. Therefore, the saliency model with depth domain knowledge prior refinement is defined as

$$S_{DDK}(r_i) = \begin{cases} \frac{1}{2}(S_{ML}(r_i) + S_{DC}(r_i)), & \lambda_d \geq \tau_1 \\ S_{ML}(r_i) \cdot d_i, & \tau_2 \leq \lambda_d < \tau_1 \\ S_{ML}(r_i), & \text{otherwise} \end{cases} \tag{8}$$

where $S_{ML}(r_i)$ is the multilevel RGBD saliency initialization of superpixel r_i; d_i is the mean depth value of superpixel r_i; τ_1 and τ_2 are set to 0.8 and 0.3 in the experiments, respectively; and $S_{DC}(r_i) = \sum_{j=1, j\neq i}^{N} |d_i - d_j| \cdot \exp(-||\boldsymbol{b}_i - \boldsymbol{b}_j||_2/\sigma^2)$ is the depth contrast of superpixel r_i.

2) Refined Depth Shape Prior: From a depth map, some implicit attributes can be used to refine the saliency result, such as the shape and contour. In [21], depth shaper prior (DSP) was proposed to capture the depth shape attribute, in which some salient seeds are selected based on the given saliency map, and then propagated to generate the depth shape prior. However, there are two key issues that need to be further addressed.

(a) The depth map can better depict the shape and contour information of the object, while lacking an effective description of the textures and details because it is a gray image. Moreover, in the depth map, the ground area near the camera usually has a larger depth value, and even its depth value is almost the same as the salient object. The original DSP algorithm only depends on the depth value during propagation, which may induce some good-quality depth data, which fails to obtain a clear shape description. Fortunately, this problem can be easily solved in the color image. Therefore, the color consistency constraint is introduced to enhance the completeness of the entire object.

(b) In order to bridge the relationship between the salient object and the depth shape, some salient regions are selected as root seeds in the DSP algorithm. However, the saliency value that is used as the sole selection criterion seems too one-sided, which may introduce some unexpected regions, such as the regions located on the image boundary, and degenerate the accuracy of the depth shape capturing. Therefore, the location constraint is introduced to further filter the initial salient seeds and achieve more robust propagation seeds.

To this end, an upgrade version called RDSP is proposed, which introduces the color constraint and refines the propagation seed selection. First, the top-K superpixels with highest saliency values after the depth domain knowledge prior refinement are selected as the initial salient seeds. Then, half of the superpixels closer to the image center in the initial seed set are determined as the root propagation seeds.

Similar to the DSP algorithm, depth propagation is applied to each root propagation seeds based on the smoothness and consistency decisions to obtain the depth shape result. Different from the DSP method, our RDSP method introduces

the color constraint in the process of determining neighborhood child nodes. In other words, the selected child nodes need to satisfy two constraints: a) the depth values between the root node and the parent node should be approximated and b) the color distribution with the parent node should be similar.

In the l-loop propagation, the superpixels directly adjacent to the $l-1$-loop child nodes that satisfy the smoothness and consistency decisions are selected as the l-loop child nodes.

a) *Smoothness Decision:* The depth difference with the color constraint between the neighbor superpixel and $l-1$-loop child seeds should be less than a given threshold T_1, as $N[|d_{nq} - d_{c_{l-1}}| \cdot (1 - a^c_{nq,c_{l-1}})] \leq T_1$, where d_{nq} is the depth value of the neighbor superpixel r_{nq}, $d_{c_{l-1}}$ is the average depth value of $l-1$-loop child seeds, $a^c_{nq,c_{l-1}}$ is the color similarity between superpixels r_{nq} and $l-1$-loop child seeds, and T_1 is set to 0.1 as suggested in [21].

b) *Consistency Decision:* The depth difference with the color constraint between the neighbor superpixel and root seed should be smaller than a given threshold T_2, as $N[|d_{nq} - d_{rk}| \cdot (1 - a^c_{nq,rk})] \leq T_2$, where d_{rk} is the depth value of the root seed r_{rk}, $a^c_{nq,rk}$ is the color similarity between superpixels r_{nq} and root seed r_{rk}, and T_2 is set to 0.2 as suggested in [21].

The RSDP value of the child node r_{cp} in the l-loop from the root seed r_k is defined as

$$\mathrm{RDSP}_k(r_{cp}) = 1 - \min\left(|d_{cp_l} - d_{c_{l-1}}|, |d_{cp_l} - d_{rk}|\right) \tag{9}$$

where d_{cp_l} denotes the depth value of the child node r_{cp} in the l-loop, $d_{c_{l-1}}$ is the average depth value of all child seeds in the $l-1$-loop, and d_{rk} represents the depth value of the root seed r_{rk}. The loop propagation will be continued until there is no neighboring superpixel that satisfies these two decisions.

Finally, the depth domain knowledge prior and the RDSP are combined to generate the depth-guided saliency refinement result as

$$S_{\mathrm{DR}}(r_i) = N[S_{\mathrm{DDK}}(r_i) + \mathrm{RDSP}(r_i)] \tag{10}$$

where $S_{\mathrm{DDK}}(r_i)$ denotes the saliency value of superpixel r_i with depth domain knowledge prior refinement, $\mathrm{RDSP}(r_i) = (\sum_{k=1}^{K} \mathrm{RDSP}_k(r_i))/K$ is the average RDSP value of superpixel r_i derived from all root seeds, and $N[\cdot]$ is a min–max normalization function.

The visual comparisons between the DSP and RDSP are presented in Fig. 2, where the third and last columns show the DSP and RDSP maps, respectively. As we can see, the DSP algorithm cannot capture the depth shape effectively with clear backgrounds so that the quality of the input depth map is degraded. Benefiting from the depth decisions with the color constraint and root seeds filtering, our RDSP descriptor accurately captures the shape of the salient object from the depth map and effectively suppresses the background interference. Even for the depth map with low contrast between the foreground and background, such as the third image, the RDSP method can still extract the shape completely.

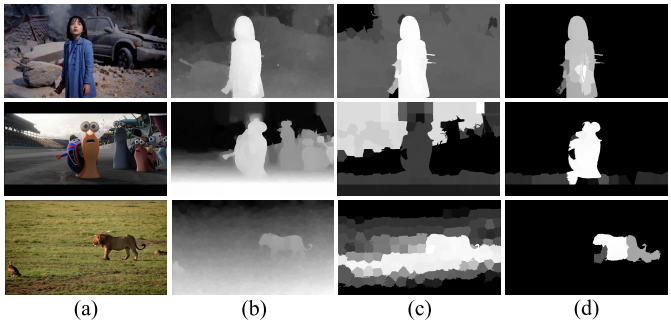

Fig. 2. Visual illustration of the proposed RDSP descriptor. (a) RGB image. (b) Depth map. (c) DSP map. (d) RDSP map.

C. Saliency Optimization With Depth Constraints

From the depth map, in addition to providing an effective shape description, the entire object usually has high consistency in the depth map. Therefore, depth information can be used to improve the consistency and smoothness of the acquired saliency map. In this article, a saliency optimization strategy with the depth constraint is formulated to attain more consistent and accurate saliency results, where the depth consistency relationship is introduced as an additional term in the energy function. The energy function integrates the unary data term, color smooth term, and depth consistency term. The data term E_u controls the updating degree between the final saliency map and initial saliency map. The color smooth term E_s constrains the spatially adjacent regions, which guarantees that similar color appearance should be assigned to similar saliency scores. The depth consistency term E_c imposes that the adjacent regions with similar depth distribution should be assigned to consistent saliency scores. The energy function is defined as

$$\begin{aligned} E = E_u + E_s + E_c = &\sum_i \left(s_i^* - s_i\right)^2 \\ &+ \sum_{(i,j)\in\Omega_s} \omega_{ij}^c \cdot \left(s_i^* - s_j^*\right)^2 + \sum_{(i,j)\in\Omega_s} \omega_{ij}^d \cdot \left(s_i^* - s_j^*\right)^2 \end{aligned} \tag{11}$$

where $s_i = S_{DR}(r_i)$ is the saliency value of superpixel r_i before optimization, s_i^* is the optimized saliency value of superpixel r_i, Ω_s represents the spatially adjacent set, and $\omega_{ij}^\star$ is the color ($\star = c$) or depth ($\star = d$) similarity between two adjacent superpixels, which is represented as

$$\omega_{ij}^\star = \begin{cases} a_{ij}^\star, & \text{if } (r_i, r_j) \in \Omega_s \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Then, the energy function defined in (11) can be rewritten as a matrix form

$$\begin{aligned} \mathbf{E} = \mathbf{E}_u + \mathbf{E}_s + \mathbf{E}_c = &\left(\boldsymbol{s}^* - \boldsymbol{s}\right)^T \cdot \left(\boldsymbol{s}^* - \boldsymbol{s}\right) \\ &+ \boldsymbol{s}^{*T} \cdot (\mathbf{D}_c - \mathbf{W}_c) \cdot \boldsymbol{s}^* + \boldsymbol{s}^{*T} \cdot (\mathbf{D}_d - \mathbf{W}_d) \cdot \boldsymbol{s}^* \end{aligned} \tag{13}$$

where $\boldsymbol{s} = [s_1, s_2, \ldots, s_N]^T$ denotes the saliency vector of all superpixels before optimization; $\boldsymbol{s}^* = [s_1^*, s_2^*, \ldots, s_N^*]^T$ corresponds to the optimized saliency vector; $\mathbf{W}_c = [\omega_{ij}^c]_{N\times N}$ and $\mathbf{W}_d = [\omega_{ij}^d]_{N\times N}$ are the color and depth affinity matrices, respectively; $\mathbf{D}_c = \mathrm{diag}(d_1^c, d_2^c, \ldots, d_N^c)$ and $\mathbf{D}_d =$

Fig. 3. Some visual examples of the RGB saliency and the corresponding DTM saliency maps. The first column shows the RGB images and the corresponding depth maps. From the second column to the seventh column in an image group, the first row is the different RGB saliency maps, and the second row presents the corresponding updated saliency maps through the proposed DTM.

$\text{diag}(d_1^d, d_2^d, \ldots, d_N^d)$ represent the color and depth degree matrices, respectively; and $d_i^\star = \sum_{j=1}^{N} \omega_{ij}^\star$.

Setting the first-order derivative of the energy function with respect to $\boldsymbol{s}^*$ to be 0, we can obtain

$$\frac{\partial \mathbf{E}}{\partial \boldsymbol{s}^*} = 2\big(\boldsymbol{s}^* - \boldsymbol{s}\big) + 2(\mathbf{D}_c - \mathbf{W}_c) \cdot \boldsymbol{s}^* + 2(\mathbf{D}_d - \mathbf{W}_d) \cdot \boldsymbol{s}^* = 0. \tag{14}$$

Combining the like terms, the solution is given by

$$\boldsymbol{s}^* = [\mathbf{I} + (\mathbf{D}_c - \mathbf{W}_c) + (\mathbf{D}_d - \mathbf{W}_d)]^{-1} \cdot \boldsymbol{s} \tag{15}$$

where $\mathbf{I}$ is an identity matrix with size $N \times N$.

IV. Experiments

In this section, we evaluate the proposed DTM on the NJUD dataset, STEREO dataset, and NLPR dataset. The qualitative and quantitative comparisons and ablation studies are discussed.

A. Experimental Settings

In the experiments, three public RGBD saliency detection datasets are used to evaluate the effectiveness of the proposed transformation model. The STEREO dataset [51] is also an image pair dataset that is distributed in indoor and outdoor scenes, which contains 797 pairs of binocular images. The corresponding estimated depth map and pixel-level ground truth are provided. The NJUD dataset [41] contains 2000 stereo image pairs, which were collected from the Internet, 3-D movies, and photographs taken by stereo cameras. The provided depth map in the NJUD dataset is estimated by the optical-flow method, and the corresponding pixel-level ground truth is given. The NLPR dataset [44] includes 1000 RGBD images with pixel-level ground truth, where the depth maps are captured by Microsoft Kinect. In this article, the number of superpixels for each image is set to 200, and the number of initial salient seeds in the RDSP component is set to 30. The project is available on our website.[1]

For quantitative evaluation, four criteria, including the Precision–Recall (PR) curve, F-measure, area under ROC curve (AUC), and S-measure are used. Comparing the binary saliency map with the ground truth, the precision and recall scores can be calculated, where the precision represents the percentage of salient pixels correctly allocated, and the recall denotes the ratio of detected salient pixels with respect to the salient pixels in the ground truth. Thus, the PR curve represents the tradeoff relationship between the precision and recall scores. As a comprehensive performance measurement,

TABLE I
QUANTITATIVE COMPARISONS ON THE NJUD DATASET. △PG IS THE PERCENTAGE GAIN BETWEEN THE RGB SALIENCY AND PROPOSED DTM

	DSR [7]	DTM	△PG	RBD [8]	DTM	△PG
F_β	0.6457	0.7566	17.2%	0.6433	0.7490	16.4%
AUC	0.8634	0.9174	6.3%	0.8498	0.9138	7.5%
S_m	0.6321	0.7063	11.7%	0.6542	0.7059	7.9%
	DCLC [9]	DTM	△PG	HDCT [10]	DTM	△PG
F_β	0.6527	0.7514	15.1%	0.6581	0.7602	15.5%
AUC	0.8526	0.9081	6.5%	0.8655	0.9220	6.5%
S_m	0.6188	0.7010	13.3%	0.6391	0.7099	11.1%
	SMD [11]	DTM	△PG	RCRR [12]	DTM	△PG
F_β	0.6900	0.7633	10.6%	0.6508	0.7548	16.0%
AUC	0.8535	0.9204	6.6%	0.8500	0.9153	7.7%
S_m	0.6782	0.7170	5.7%	0.6412	0.7079	10.4%

TABLE II
QUANTITATIVE COMPARISONS ON THE STEREO DATASET. △PG IS THE PERCENTAGE GAIN BETWEEN THE RGB SALIENCY AND PROPOSED DTM

	DSR [7]	DTM	△PG	RBD [8]	DTM	△PG
F_β	0.6974	0.7973	14.3%	0.7157	0.7997	11.7%
AUC	0.9126	0.9470	3.8%	0.9149	0.9497	3.8%
S_m	0.6674	0.7335	9.9%	0.7136	0.7449	4.4%
	DCLC [9]	DTM	△PG	HDCT [10]	DTM	△PG
F_β	0.7150	0.7925	10.8%	0.7005	0.7954	13.5%
AUC	0.9052	0.9416	4.0%	0.9087	0.9495	4.5%
S_m	0.6614	0.7314	10.6%	0.6775	0.7372	8.8%
	SMD [11]	DTM	△PG	RCRR [12]	DTM	△PG
F_β	0.7620	0.8077	6.0%	0.7360	0.7996	8.6%
AUC	0.9260	0.9532	2.9%	0.9066	0.9490	4.7%
S_m	0.7398	0.7528	1.8%	0.7039	0.7430	5.6%

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

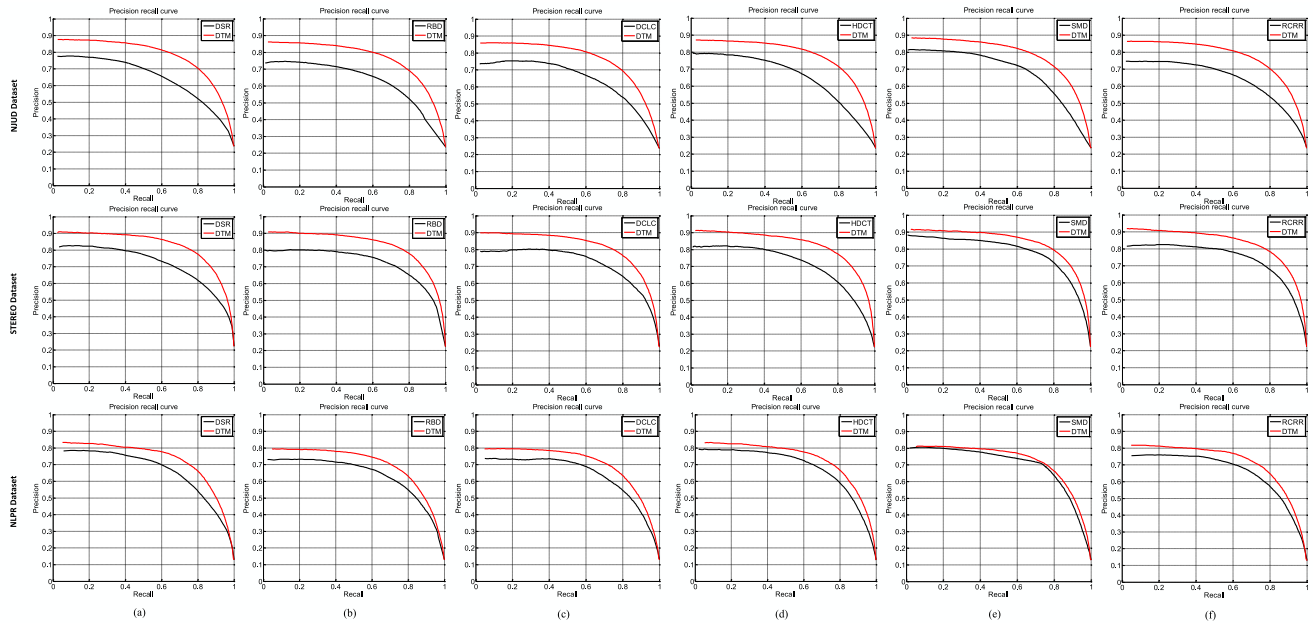

Fig. 4. PR curves of different RGB saliency methods and corresponding DTM results on three RGBD saliency detection datasets, where the first row is the results on the NJUD dataset, the second row shows the results on the STEREO dataset, and the third row shows the results on the NLPR dataset. In each subfigure, the black line denotes the PR curve of the RGB saliency result, and the red line corresponds to the PR curve of the DTM result. (a) DSR and corresponding DTM results. (b) RBD and corresponding DTM results. (c) DCLC and corresponding DTM results. (d) HDCT and corresponding DTM results. (e) SMD and corresponding DTM results. (f) RCRR and corresponding DTM results.

the F-measure is defined as the weighted harmonic mean of precision and recall [52]

$$F_\beta = \frac{\left(1+\beta^2\right)\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{16}$$

where $\beta^2 = 0.3$ for emphasizing the precision. The ROC curve describes the relationship between the false positive rate and true positive rate, and the area under the ROC curve is defined as the AUC score, and the larger, the better. In addition, S-measure [19] evaluates the structural similarity between the saliency map and ground truth as

$$S_m = \alpha \times S_o + (1-\alpha) \times S_r \tag{17}$$

where α is set to 0.5 for assigning the equal contribution to both region (S_r) and object (S_o) similarity.

TABLE III
QUANTITATIVE COMPARISONS ON THE NLPR DATASET. △PG IS THE PERCENTAGE GAIN BETWEEN THE RGB SALIENCY AND PROPOSED DTM

	DSR [7]	DTM	△PG	RBD [8]	DTM	△PG
F_β	0.6743	0.7326	8.6%	0.6542	0.7090	8.4%
AUC	0.9252	0.9340	1.0%	0.9201	0.9295	1.0%
S_m	0.7037	0.7320	4.0%	0.7124	0.7260	1.9%
	DCLC [9]	DTM	△PG	HDCT [10]	DTM	△PG
F_β	0.6662	0.7134	7.1%	0.6914	0.7326	6.0%
AUC	0.8992	0.9246	2.8%	0.9385	0.9397	0.1%
S_m	0.6829	0.7245	6.1%	0.7108	0.7338	3.2%
	SMD [11]	DTM	△PG	RCRR [12]	DTM	△PG
F_β	0.7138	0.7290	2.1%	0.6783	0.7252	6.9%
AUC	0.9229	0.9334	1.1%	0.9017	0.9270	2.8%
S_m	0.7303	0.7355	0.7%	0.6919	0.7273	5.1%

B. Performance Evaluation of the Proposed Model

In our model, the existing RGB saliency detection model is used to generate the RGB saliency baseline. We first use six recent and representative RGB saliency detection methods to generate the baselines, including DSR [7], RBD [8], DCLC [9], HDCT [10], SMD [11], and RCRR [12]. Then, the RGB image, depth map, and RGB saliency map are, respectively, embedded into the proposed model to produce the corresponding improved RGBD saliency map. In this way, we can obtain six groups of results on one dataset. Some visual examples of the RGB saliency map and corresponding DTM saliency map are shown in Fig. 3. The quantitative evaluations, including PR curves, F-measure, AUC scores, and S-measure are reported in Fig. 4 and Tables I–IV.

In Fig. 3, four visual groups, including different RGB saliency maps and the corresponding DTM results are presented. In the first horse image, some backgrounds, such as the distant trees, are misdetected by the RGB saliency methods (i.e., RBD, SMD, and RCRR), whereas these regions are effectively suppressed with the help of depth information. Through the proposed DTM, backgrounds are removed successfully. In the second image, in addition to the misdetected backgrounds, the horseman cannot be completely highlighted by all RGB saliency methods due to the low contrast against the background in the color space. In contrast, these problems are effectively addressed through the DTM. In the third image, the legs of the rider are not highlighted as the body through the RGB saliency detection methods, and the distant background regions are wrongly retained. Going through the proposed model, the rider is highlighted completely and consistently with clear backgrounds. In the last image, the haystack is

TABLE IV
QUANTITATIVE COMPARISONS OF DIFFERENT RGBD SALIENCY DETECTION METHODS ON THREE DATASETS

	NJUD Dataset			STEREO Dataset			NLPR Dataset		
	F_β	AUC	S_m	F_β	AUC	S_m	F_β	AUC	S_m
SS [34]	0.6128	0.8103	0.5755	0.5478	0.7943	0.5412	0.4712	0.8007	0.5737
ACSD [36]	0.7459	0.9259	0.6987	0.7467	0.9333	0.7082	0.6695	0.9229	0.6825
WSC [37]	0.6418	0.7579	0.6325	0.6987	0.8034	0.6727	0.6586	0.8494	0.6955
CDCP [42]	0.6673	0.8699	0.6689	0.7168	0.9065	0.7181	0.6863	0.9175	0.7266
MBP [43]	0.6025	0.7231	0.5272	0.6627	0.7701	0.5574	0.6015	0.7852	0.6050
LMH [44]	0.7029	0.8489	0.5137	0.5862	0.7360	0.4773	0.7057	0.8947	0.6141
DF [45]	0.6384	0.8338	0.5881	0.6961	0.8804	0.6279	0.6407	0.8801	0.6610
ours	0.7633	0.9204	0.7170	0.8077	0.9532	0.7528	0.7290	0.9334	0.7355

TABLE V
COMPARISONS OF THE AVERAGE RUNNING TIME (SECONDS PER IMAGE)

Method	SS	ACSD	WSC	CDCP	MBP	LMH	DF	ours
Platform	C++	exe	Matlab	Matlab	Matlab	Matlab	Matlab	Matlab
Time	4.39	0.46	10.30	16.75	45.50	2.97	10.77	2.22

detected accurately and backgrounds are suppressed effectively through the DTM method. In particular, when the RCRR method completely fails in detecting the salient object, benefiting from the introduction of multilevel saliency initialization designed in our model, we still obtain superior detection results. This also illustrates the robustness of our algorithm.

The PR curves of different RGB saliency methods and corresponding DTM results on three RGBD saliency detection datasets are shown in Fig. 4. Compared with the RGB saliency result (marked in black line) with the updated DTM result (marked in red line), we can see that the PR curve of DTM is much higher than the curve of the initial RGB saliency method, which demonstrates superior performance improvement of the proposed DTM. The numerical quantitative measurements, including *F*-measure, AUC score, and *S*-measure are listed in Tables I–III. On the NJUD dataset, the *F*-measure of DTM reaches 0.7633, originating from 0.69; the AUC score achieves 0.9220 from 0.8655, and *S*-measure reaches 0.7170 from 0.6782. The maximum percentage gain reaches 17.2% in terms of the *F*-measure compared with the initial RGB saliency map, and average percentage gain also reaches 15.1%. From the tables, we can see that the average percentage gain of AUC reaches 6.85%, and the average percentage gain of *S*-measure achieves 10.01%. On the STEREO dataset, the *F*-measure can be improved from 0.6974 to 0.7973 with the percentage gain of 14.3%, the AUC score is updated from 0.9066 to 0.9490 with the percentage gain of 4.7%, and *S*-measure is increased from 0.6674 to 0.7335 with the percentage gain of 9.9%. On this dataset, the average percentage gains reach 10.8% in terms of the *F*-measure, 4.0% in terms of the AUC score, and 6.9% in terms of the *S*-measure. On the NLPR dataset, compared with the original saliency baseline, the maximum percentage gain reaches 8.6% in terms of the *F*-measure, 2.8% in terms of the AUC score, and 6.1% in terms of the *S*-measure. In general, the more accurate the RGB saliency baseline is, the better it is for the RGBD saliency computation using our model. Therefore, in order to achieve better performance in practical applications, we can select the state-of-the-art method to generate a superior baseline.

In addition, the quantitative comparisons on three datasets with seven RGBD saliency detection methods, including SS [34], ACSD [36], WSC [37], CDCP [42], MBP [43], LMH [44], and DF [45] are reported in Table IV, where the proposed method is derived from the SMD method [11]. On these three datasets, the proposed method achieves the highest *F*-measure and *S*-measure. Moreover, the proposed method is superior to other methods in terms of the AUC score on the STEREO and NLPR datasets. On the NJUD dataset, the maximum percentage gain of *F*-measure achieves 26.7% compared with other methods, and the minimum percentage gain also reaches 2.3%. On the NLPR dataset, the maximum and minimum percentage gains of the *F*-measure reach 54.7% and 3.3%, respectively. On the STEREO dataset, the percentage gain is more significant compared with the second best method, that is, the gains of *F*-measure, AUC score, and *S*-measure reach 8.2%, 5.3%, and 4.8%, respectively. We also test the running time of all methods on a Quad Core 3.4-GHz PC with 16-GB RAM. The average running time is reported in Table V. As can be seen, the proposed method costs 2.22 s, on average, to process one image, where the preprocessing (such as data loading, SLIC, graph construction, etc.) costs 74.72% running time, while multilevel RGBD saliency initialization uses 6.61% running time, depth-guided saliency refinement costs 3.73% running time, and saliency optimization with depth constraints consumes 14.94% running time. Compared with other RGBD saliency detection methods, our method ranks second, that is, it is only slower than the ACSD method implemented by exe integration. All of these visual examples and quantitative measures demonstrate the effectiveness and computational efficiency of the proposed DTM from RGB to RGBD saliency.

C. Module Analysis

The proposed DTM going from RGB to RGBD saliency is composed of three modules. The multilevel RGBD saliency initialization is proposed to integrate the global compactness

TABLE VI
F-MEASURE OF DIFFERENT MODULES ON THE STEREO DATASET

Modules	*F*-measure
RGB Saliency (DSR)	0.6974
Multi-level RGBD Saliency Initialization	0.7050
Depth-guided Saliency Refinement	0.7697
Saliency Optimization with Depth Constraints	0.7973

and the local geodesic saliency cues, where the depth feature is used as a supplement to the color information. The depth-guided saliency refinement is used to further highlight the salient objects and suppress the background regions by introducing the depth domain knowledge prior and RDSP. The saliency optimization with depth constraints is designed to improve the accuracy and consistency through an energy function considering the depth consistency. We comprehensively evaluate each module on the STEREO dataset, and the *F*-measure is presented in Table VI.

The RGB saliency method (i.e., DSR) is used to produce the initial saliency result with the *F*-measure of 0.6974. In order to exploit the explicit depth feature and generate stable saliency initialization, the multilevel RGBD saliency model is designed by combining the global compactness saliency and local geodesic saliency, and achieves the *F*-measure of 0.7050. In addition to the explicit depth feature, the implicit depth information, including the depth domain knowledge prior and RDSP, are introduced in the depth-guided saliency refinement to further highlight the salient objects and suppress the background regions. Benefiting from the superior ability of depth representation, the *F*-measure of depth-guided saliency refinement model reaches 0.7697 with the percentage gain of 9.2% in comparison with the saliency initialization. Inspired by the fact that the salient object in the depth map has high consistency, the saliency optimization model with depth constraints is designed to attain more consistent and accurate saliency results. After the optimization model in module 3, the *F*-measure is further improved to 0.7973, with the percentage gain of 3.6% when compared with the saliency refinement model. In summary, the performance improvement of our model primarily comes from these three effective depth-guided modules.

D. Evaluation of Refined Depth Shape Prior

For the depth map, the implicit attribute will be beneficial for the identification of the salient object, such as the shape prior. In order to improve the independence of the salient object and enhance the exploitation of depth shape, an RDSP is proposed in this article. Different from the depth shape prior (DSP), RDSP adds the color constraint to refine the smoothness and consistency decisions, and refreshes the propagation seeds selection to enhance the robustness. We conduct some experiments on the STEREO dataset to evaluate the performance of RDSP. The PR curves and *F*-measure are shown in Fig. 5, where the black line is the PR curve of DSP, and the blue line denotes the PR curve of RDSP. As can be seen, RDSP reaches a higher position than DSP in

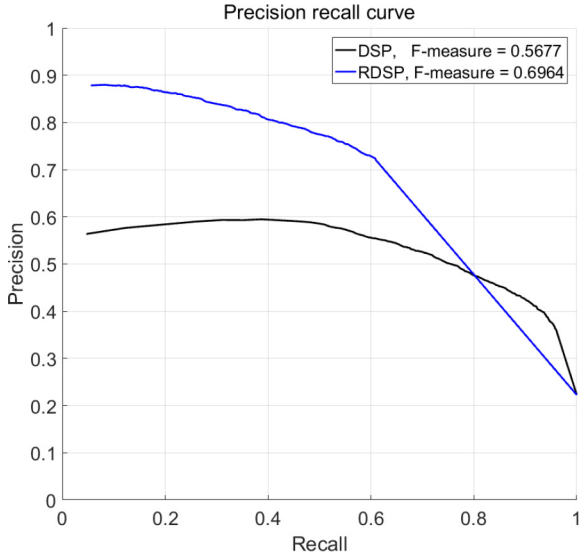

Fig. 5. Quantitative comparisons between DSP and RDSP on the STEREO dataset.

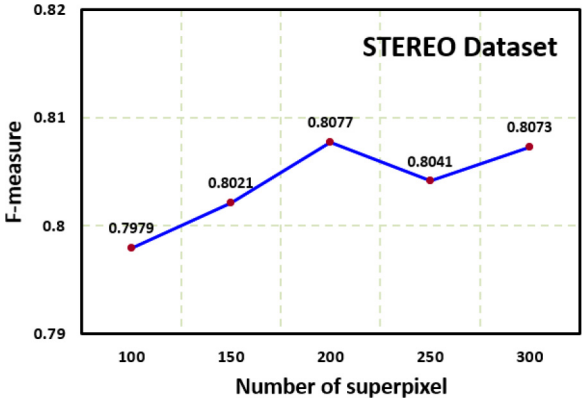

Fig. 6. *F*-measure of our method with different numbers of superpixels on the STEREO dataset.

the PR curves, which demonstrates the performance superiority of RDSP. In terms of the *F*-measure, the DSP descriptor reaches 0.5677, and the RDSP descriptor reaches 0.6964 with a percentage gain of 22.7% compared to DSP. In addition, some visual comparisons are shown in Fig. 2. Compared with the DSP result shown in the third column, the shape of the salient object is effectively highlighted with a sharp boundary, complete shape, and little interference. In the second image, although the salient object has prominent properties in the depth map, the DSP algorithm still fails to capture its shape attribute accurately. In contrast, the RDSP method suppresses the background and highlights the shape of the salient region from the depth map. In addition, especially, for the low-depth contrast image shown in the last row of Fig. 2, the DSP descriptor cannot accurately describe the object shape; however, the RDSP algorithm effectively depicts the shape with clean background noise.

E. Evaluation of Different Numbers of Superpixels

In this section, we discuss the influence of different numbers of superpixels on the STEREO dataset, and the *F*-measures are

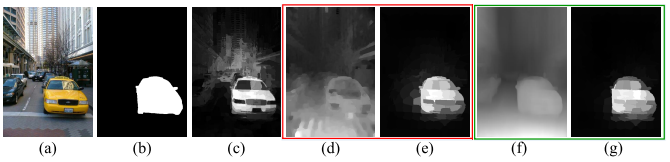

Fig. 7. Visual comparisons with different qualities of the depth maps. (a) RGB image. (b) Ground truth. (c) RGB saliency map by using the HDCT method. (d) and (e) Original depth map from the dataset and the corresponding DTM result. (f) and (g) Generated depth map by using the MegaDepth method [53] and the corresponding DTM result.

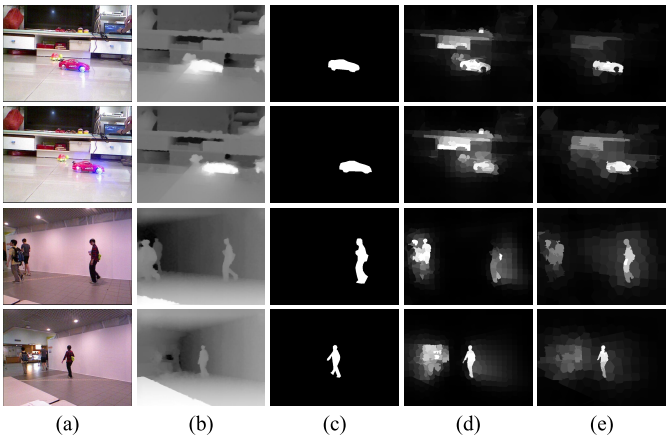

Fig. 8. Visual examples of the proposed method applied to RGBD videos. (a) Input RGB video sequences. (b) Input depth video sequences. (c) Ground truth. (d) RGB video saliency results produced by the CVS method [25]. (e) RGBD video saliency results by using our transformation model.

shown in Fig. 6. As can be seen, when the number of superpixel is set to 200, our model achieves the best performance. In fact, the consistent performance of our model with different numbers of superpixels indicates that the proposed algorithm is insensitive to the number of superpixels. Considering the effectiveness and efficiency, the number of superpixels is set to 200 in the experiments.

F. Discussion

We show some visual comparisons with different qualities of the depth maps in Fig. 7. In this example, the quality of the original depth map from the dataset is relatively poor, while the generated depth map using the MegaDepth method [53] is more homogeneous and accurate. With these two depth maps as an input, the proposed model can still optimize the RGB saliency result and suppress the backgrounds. Moreover, our model with a high-quality depth map as input yields better results [see Fig. 7(g)], as the background regions are suppressed more effectively. It is worth mentioning at this point that we introduce the depth confidence measure to control the volume of the depth information, which reduces the negative influence of the poor depth map to some extent. All of these examples demonstrate the robustness of the proposed model.

In addition, we evaluate the performance of our model in dynamic scenarios. Some visual examples of the proposed method that applied to RGBD videos [54] are shown in Fig. 8, where the RGB video saliency detection method denoted as CVS [25] is used to generate the input baseline. In the toy car sequences, some backgrounds (e.g., TV) are misdetected as the salient regions by the CVS method. Through our model, these regions are suppressed to a certain extent, and the toy car is also further highlighted. Similarly, the background regions on the left in the walking sequences are effectively suppressed by our method. In this experiment, we just made a simple attempt in the video by using our model. In fact, the motion constraint and spatiotemporal information should be considered in the transformation model for the RGBD video to further improve saliency performance. This is a valuable and promising research direction in the future.

V. Conclusion

Different from the existing RGBD saliency methods focusing on designing a straightforward and comprehensive model, in this article, we proposed a DTM from RGB to RGBD saliency, which pays more attention to capture the explicit and implicit information from the depth map. First, the explicit depth feature is used to generate multilevel RGBD saliency initialization, which combines the global compactness saliency and local geodesic saliency cues. Then, the implicit attributes of the depth map, including depth domain knowledge prior and RDSP are captured to refine the saliency result. Finally, inspired by the depth consistency in the interior of the object, a saliency optimization strategy with depth constraint is designed to further improve the consistency and accuracy, which introduces the depth smoothness relationship as an additional term in the energy optimization function. The proposed model can effectively exploit any existing RGB saliency model to work well in RGBD saliency scenarios. The comprehensive comparisons and ablation studies on three RGBD saliency detection datasets have demonstrated the effectiveness of the proposed method both qualitatively and quantitatively.

In this article, we focus on designing an unsupervised framework that transforms the RGB saliency to RGBD saliency with the help of depth constraints. The past decade has witnessed the vigorous development and qualitative leap in learning-based saliency detection methods. Thus, depicting the handcrafted features designed in this article using the learning methods (e.g., deep learning [17], [18], [48], [49]; extreme learning [55]–[59]; and zero-shot learning [60]) is a very interesting and promising research topic in the future.

References

[1] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," *arXiv 1904.09146*, Apr. 2019. [Online]. Available: http://arxiv.org/abs/1904.09146

[2] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.

[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] C.-Y. Li, J.-C. Guo, R.-M. Cong, Y.-W. 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.

[5] Q. Jiang, F. Shao, W. Gao, Z. Chen, G. Jiang, and Y.-S. Ho, "Unified no-reference quality assessment of singly and multiply distorted stereoscopic images," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1866–1881, Apr. 2019.

[6] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, Florence, Italy, 2012, pp. 29–42.

[7] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. ICCV*, Sydney, NSW, Australia, 2013, pp. 2976–2983.

[8] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 2814–2821.

[9] 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.

[10] 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. 2015.

[11] H. Peng, B. Li, H. Ling, W. Hua, W. Xiong, and S. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[12] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.

[13] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[14] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 478–487.

[15] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. ICCV*, Venice, Italy, 2017, pp. 212–221.

[16] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 5300–5309.

[17] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. AAAI*, New Orleans, LA, USA, 2018, pp. 6943–6950.

[18] Z. Deng *et al.*, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. IJCAI*, Stockholm, Sweden, 2018, pp. 684–690.

[19] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. ICCV*, Venice, Italy, 2017, pp. 4548–4557.

[20] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published.

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

[22] Y. Zhang, L. Li, R. Cong, X. Guo, H. Xu, and J. Zhang, "Co-saliency detection via hierarchical consistency measure," in *Proc. ICME*, San Diego, CA, USA, 2018, pp. 1–6.

[23] 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.

[24] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "HSCS: Hierarchical sparsity based co-saliency detection for RGBD images," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1660–1671, Jul. 2019.

[25] 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.

[26] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4819–4831, Oct. 2019.

[27] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 8554–8564.

[28] Y. Yang, B. Li, P. Li, and Q. Liu, "A two-stage clustering based 3D visual saliency model for dynamic scenarios," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 809–820, Apr. 2019.

[29] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[30] Y. Yang, Q. Liu, X. He, and Z. Liu, "Cross-view multi-lateral filter for compressed multi-view depth video," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 302–315, Jan. 2019.

[31] H. Fu, D. Xu, S. Lin, and J. Liu, "Object-based RGBD image co-segmentation with mutex constraint," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 4428–4436.

[32] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, and X. Fan, "Color-guided depth map super resolution using convolutional neural network," *IEEE Access*, vol. 2, pp. 26666–26672, 2017.

[33] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, May 2019.

[34] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 454–461.

[35] Y. Fang, J. Wang, M. Narwaria, P. L. Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2625–2636, Jun. 2014.

[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] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 5216–5223.

[38] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 2343–2350.

[39] 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.

[40] H. Song, Z. Liu, H. Du, G. Sun, O. L. Meur, and T. Ren, "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.

[41] 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.

[42] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. ICCVW*, Venice, Italy, 2017, pp. 1509–1515.

[43] C. Zhu and G. Li, "A multilayer backpropagation saliency detection algorithm and its applications," *Multimedia Tools Appl.*, vol. 77, no. 19, pp. 25181–25197, 2018.

[44] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, Zürich, Switzerland, 2014, pp. 92–109.

[45] 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.

[46] 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.

[47] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018, pp. 3051–3060.

[48] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multiscale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, Feb. 2019.

[49] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[50] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "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.

[51] J. Guo, T. Ren, and J. Bei, "Salient object detection in RGB-D image via saliency evolution," in *Proc. ICME*, Seattle, WA, USA, 2016, pp. 1–6.

[52] 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.

[53] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from Internet photos," in *Proc. CVPR*, Salt Lake City, UT, USA, 2018, pp. 2041–2050.

[54] H. Fu, D. Xu, and S. Lin, "Object-based multiple foreground segmentation in RGBD video," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1418–1427, Mar. 2017.

[55] X.-Z. Wang, T. Zhang, and R. Wang, "Noniterative deep learning: Incorporating restricted Boltzmann machine into multilayer random weight neural networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7, pp. 1299–1308, Jul. 2019.
[56] R. Wang *et al.*, "Taxirec: Recommending road clusters to taxi drivers using ranking-based extreme learning machines," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 3, pp. 585–598, Mar. 2018.
[57] X.-Z. Wang, R. Wang, and C. Xu, "Discovering the relationship between generalization and uncertainty by incorporating complexity of classification," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 703–715, Feb. 2018.
[58] R. Wang, X.-Z. Wang, S. Kwong, and C. Xu, "Incorporating diversity and informativeness in multiple-instance active learning," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1460–1475, Dec. 2017.
[59] R. Wang, C.-Y. Chow, and S. Kwong, "Ambiguity based multiclass active learning," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 1, pp. 242–248, Feb. 2016.
[60] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5652–5667, Nov. 2018.

Runmin Cong (M'19) received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2019.

He is currently an Associate Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. He was a visiting student with Nanyang Technological University, Singapore, from 2016 to 2017. In 2018, he has spent one year as a Research Associate with the Department of Computer Science, City University of Hong Kong, Hong Kong. His current research interests include computer vision, image processing, saliency detection, and 3-D imaging.

Dr. Cong was a recipient of the Best Student Paper Runner-Up at IEEE ICME in 2018. He is a Reviewer of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

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.

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

Prof. Lei is on the editorial boards 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. And 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, UAE. His current research interests include computer vision, image processing, and medical image analysis.

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

Junhui Hou (S'13–M'16) received the B.Eng. degree in information engineering (Talented Students Program) from the South China University of Technology, Guangzhou, China, in 2009, the M.Eng. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2012, and the Ph.D. degree in electrical and electronic engineering from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2016.

He has been an Assistant Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong, since 2017. His current research interests include visual signal processing, such as adaptive image/video representation and analysis (RGB/depth/light field/hyperspectral), static/dynamic 3-D geometry representation and processing (mesh/point cloud/MoCap), and semisupervised modeling for clustering/classification.

Dr. Hou was a recipient of several prestigious awards, including the Chinese Government for Outstanding Self-Financed Students Abroad (China Scholarship Council in 2015) and the Early Career Award from the Hong Kong Research Grants Council in 2018. He serves/served as an Associate Editor for *The Visual Computer*, an Area Editor for *Signal Processing: Image Communication*, and the Guest Editor for the *Journal of Visual Communication and Image Representation*. He is an Area Chair of the ACM International Conference on Multimedia 2019. He was/is also involved in the organization of some international conferences, such as the Local Arrangement Chair of the 26th Pacific Conference on Computer Graphics and Applications in 2018 and the Publication Co-Chair of the IEEE International Conference on Visual Communication and Image Processing in 2020.

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 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. He has published over 400 academic papers in prestigious international journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and top-level conferences, such as ACM Multimedia, ICCV, CVPR, IJCAI, and VLDB. His current research interests include multimedia video analysis, image processing, computer vision, and pattern recognition.

Prof. Huang is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Acta Automatica Sinica*, and a Reviewer of various international journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as the General Chair, the Program Chair, the Track Chair, and a TPC Member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, and PSIVT.

Sam Kwong (M'93–SM'04–F'13) received the B.S. degree in electrical engineering from the State University of New York at Buffalo, Buffalo, NY, USA, in 1983, the M.S. degree from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree in electrical engineering from the University of Hagen, Hagen, Germany, in 1996.

From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada, Ottawa, ON, Canada. He joined Bell Northern Research Canada, Ottawa, as a member of the Scientific Staff. In 1990, he became a Lecturer with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he is currently a Professor with the Department of Computer Science. His current research interests include video and image coding and evolutionary algorithms.