Review of Visual Saliency Detection with Comprehensive Information

Runmin Cong, Jianjun Lei, *Senior Member, IEEE,* Huazhu Fu, *Senior Member, IEEE,*
Ming-Ming Cheng, Weisi Lin, *Fellow, IEEE,* and Qingming Huang, *Fellow, IEEE*

Abstract—Visual saliency detection model simulates the human visual system to perceive the scene, and has been widely used in many vision tasks. With the development of acquisition technology, more comprehensive information, such as depth cue, inter-image correspondence, or temporal relationship, is available to extend image saliency detection to RGBD saliency detection, co-saliency detection, or video saliency detection. RGBD saliency detection model focuses on extracting the salient regions from RGBD images by combining the depth information. Co-saliency detection model introduces the inter-image correspondence constraint to discover the common salient object in an image group. The goal of video saliency detection model is to locate the motion-related salient object in video sequences, which considers the motion cue and spatiotemporal constraint jointly. In this paper, we review different types of saliency detection algorithms, summarize the important issues of the existing methods, and discuss the existent problems and future works. Moreover, the evaluation datasets and quantitative measurements are briefly introduced, and the experimental analysis and discission are conducted to provide a holistic overview of different saliency detection methods.

Index Terms—Salient object, RGBD saliency detection, depth attribute, co-saliency detection, inter-image correspondence, video saliency detection, spatiotemporal constraint.

I. INTRODUCTION

HUMAN visual system works as a filter to allocate more attention to the attractive and interesting regions or objects for further processing. Humans can exhibit visual fixation, which is maintaining of the visual gaze on a single location. Inspired by this visual perception phenomena, some visual saliency models focus on predicting human fixations [1]. In addition, driven by computer vision applications,

some visual saliency models aim at identifying the salient regions from the image or video [2]. In this survey, we mainly review the latest progress of salient object detection, which has been applied in image/video segmentation [3], [4], image/video retrieval [5], [6], image retargeting [7], [8], image compression [9], image enhancement [10]–[12], video coding [13], foreground annotation [14], quality assessment [15], [16], thumbnail creation [17], action recognition [18], and video summarization [19].

The last decade has witnessed the remarkable progress of image saliency detection, and a plenty of methods have been proposed and achieved the superior performances, especially the deep learning based methods have yielded a qualitative leap in performances. Following [2], image saliency detection methods can be classified into bottom-up model [20]–[33] and top-down model [34]–[49]. Bottom-up model is stimulus-driven, which focuses on exploring low-level vision features. Some visual priors are utilized to describe the properties of salient object based on the visual inspirations from the human visual system, such as contrast prior [20], background prior [24], [26], [31], and compactness prior [28]. In addition, some traditional techniques are also introduced to achieve image saliency detection, such as frequency domain analysis [50], sparse representation [22], cellular automata [25], random walks [26], low-rank recovery [29], and Bayesian theory [30]. Top-down model is task-driven, which utilizes supervised learning with labels and achieves high performance. Especially, deep learning technique has been demonstrated the powerful ability in saliency detection. Some hierarchical deep networks for saliency detection are proposed, such as SuperCNN [38], and DHSNet [41]. In addition, the multi-scale or multi-context deep saliency network is proposed to learn more comprehensive features, such as deep contrast network [40], multi-context deep learning framework [46], multi-scale deep network [47], and network with short connections [43]. The symmetrical network is also introduced in saliency detection, such as the encoder-decoder fully cnvolutional networks [44]. Moreover, some deep weakly supervised methods for salient object detection are proposed by using the image-level supervision [48] or noisy annotation [49].

In fact, the human visual system can not only perceive the appearance of the object, but also be affected by the depth information from the scene. With the development of imaging devices, the depth map can be acquired conveniently, which lays the data foundation for RGBD saliency detection [51]. Generally, there are three options for 3D depth imaging, i.e., structured light [52], TOF (Time-of-Flight) [53], and
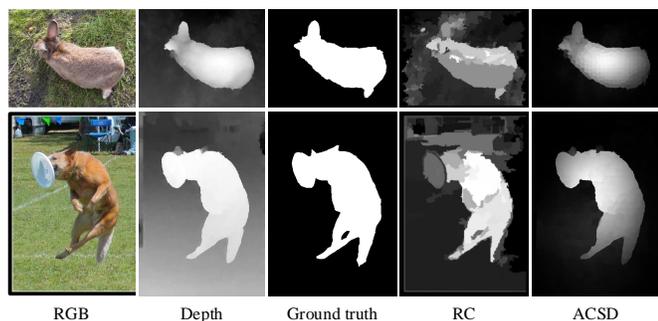
| RGB | Depth | Ground truth | RC | ACSD |

Fig. 1. Some illustrations of saliency detection with and without depth cue. The first three columns shows the RGB image, depth map, and ground truth. The fourth column shows the image saliency detection result using the RC method [20]. The fifth column represents the RGBD saliency detection result using the ACSD method [64].

binocular imaging [54]. The structured light pattern (e.g., Kinect) captures the depth information via the change of light signal projected by the camera, which can obtain high-resolution depth map. The TOF system (e.g., Camcube) estimates the depth through the round-trip time of the light pulses, which has good anti-jamming performance and wider viewing angle. The stereo imaging system takes photo pair via stereo camera and calculates the object's disparity based on two-view geometry. Depth map can provide many useful attributes for foreground extraction from the complex background, such as shape, contour, and surface normal. Some examples of saliency detection with and without depth cue are shown in Fig. 1. As can be seen, utilizing the depth cue, RGBD saliency model achieves superior performance with consistent foreground enhancement. However, how to effectively exploit the depth information to enhance the identification of salient object has not yet reached a consensus, and still needs to be further investigated. Considering the ways of using depth information, we divide the RGBD saliency detection model into depth feature based method [55]–[62] and depth measure based method [63]–[70]. Depth feature based method focuses on taking the depth information as a supplement to color feature, and depth measure based method aims at capturing comprehensive attributes from the depth map (e.g., shape) through the designed depth measurements.

In recent years, with the explosive growth of data volume, human need to process multiple relevant images collaboratively. As an emerging and challenging issue, co-saliency detection gains more and more attention from researchers, which aims at detecting the common and salient regions from an image group containing multiple related images, while the categories, intrinsic attributes, and locations are entirely unknown [71]. In general, three properties should be satisfied by the co-salient object, i.e., (1) the object should be salient in each individual image, (2) the object should be repeated in most of the images, and (3) the object should be similar in appearance among multiple images. Some visual examples of co-saliency detection are provided in Fig. 2. In the individual image, all the cows should be detected as the salient objects. However, only the brown cow is the common object from the image group. Therefore, the inter-image correspondence
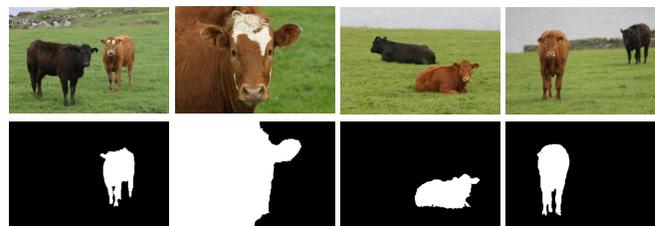

Fig. 2. Examples of the co-saliency detection model on the iCoseg dataset. The first row presents the input images, and the second row shows the ground truth for co-saliency detection.
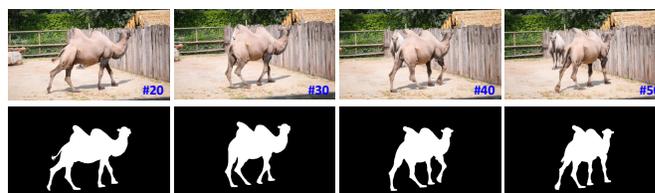

Fig. 3. Examples of the video saliency detection model on the DAVIS dataset. The first row is the input video frames, and the second row shows the ground truth for video saliency detection.

among multiple images plays a useful role in representing the common attribute. On the whole, co-saliency detection methods are roughly grouped into two categories according to whether the depth cue is introduced, i.e., RGB co-saliency detection [72]–[88] and RGBD co-saliency detection [89]–[92]. Then, the RGB co-saliency detection methods are further divided into some sub-classes based on different correspondence capturing strategies, i.e., matching based method [72]–[79], clustering based method [80], rank analysis based method [81], [82], propagation based method [83], [84], and learning based method [85]–[88].

Different from image data, video sequences contain more abundant appearance information and continuous motion cue, which can better represent the characteristics of the target in a dynamic way. However, the clustered backgrounds, complex motion patterns, and changed views also bring new challenges to interpret video content effectively. Video saliency detection aims at continuously locating the motion-related salient object from the given video sequences by considering the spatial and temporal information jointly. The spatial information represents the intra-frame saliency in the individual frame, while the temporal information provides the inter-frame constraints and motion cues. Fig. 3 illustrates some examples of video saliency detection. In this camel video, both two camels appeared from 40th frame should be detected as the salient objects through a single image saliency model. However, only the front one is continuously moving and repeating, which is the salient object in this video. The differences between co-saliency detection and video saliency detection lie in two aspects, i.e., (1) The inter-frame correspondence has the temporal property in video saliency detection rather than in co-saliency detection. For co-saliency detection in an image group, the common salient objects have the consistent semantic category, but are not necessarily the same object. By contrast, the salient objects in video are continuous in the time axis and consistent among
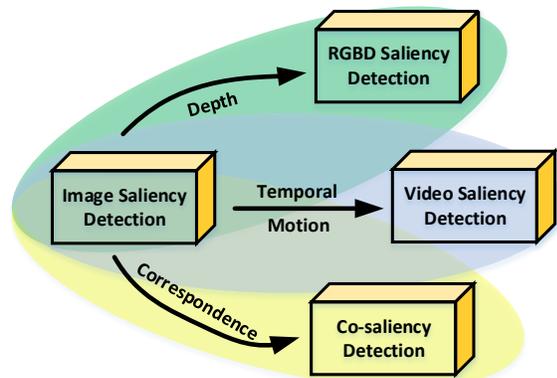
Fig. 4.   Relationships between different visual saliency detection models.

different frames; (2) In video saliency detection model, motion cue is essential to distinguish the salient object from the complex scene. However, this cue is not included in co-saliency detection model. Similar to the classification strategy of image saliency detection, we divide the video saliency detection methods into two categories, i.e., low-level cue based method [93]–[105] and learning based method [106]–[110]. For clarity, the low-level cue based method is further grouped into fusion model and direct-pipeline model according to feature extraction method, and the learning based method is further divided into supervised method and unsupervised method.

As stated above, the major relationships among four different visual saliency detection models are summarized in Fig. 4, where the image saliency detection model is the basis for other three models. With the depth cue, RGBD saliency map can be obtained from an image saliency detection model. Introducing the inter-image correspondence, image saliency detection model can be transformed into a co-saliency detection method. Video saliency detection can be derived from an image saliency detection model by combining the temporal correspondence and motion cue, or from a co-saliency detection method by integrating the motion cue. In practice, in order to obtain superior performance, it is necessary to design a specialized algorithm to achieve co-saliency detection or video saliency detection, rather than directly transplanting the image saliency detection algorithms.

In this paper, we review different saliency detection models with comprehensive information, and the rest of this paper is organized as follows. Section II surveys the existing RGBD saliency detection models. Section III introduces some co-saliency detection methods. Section IV summarizes the related works of video saliency detection. The experimental comparisons and discussions are presented in Section V. Finally, the conclusion and future work are summarized in Section VI.

II. RGBD SALIENCY DETECTION

Different from image saliency detection, RGBD saliency detection model considers the color information and depth cue

together to identify the salient object. As a useful cue for saliency detection, depth information is usually utilized in two ways, i.e., directly incorporating as the feature and designing as the depth measure. Depth feature based method [55]–[62] focuses on using the depth information as a supplement to color feature. Depth measure based method [63]–[70] aims at capturing comprehensive attributes from the depth map (e.g. shape and structure) through the designed depth measures.

A. Depth Feature Based RGBD Saliency Detection

To achieve RGBD saliency detection, the depth feature is directly embedded into the feature pool as the supplement of color information. In [57], color, luminance, texture, and depth features were extracted from the RGBD images to calculate the feature contrast maps. Then, the fusion and enhancement schemes were utilized to produce the final 3D saliency map. In [59], multi-level features were used to generate the various saliency measures at different scales, then the discriminative saliency fusion strategy was designed to fuse the multiple saliency maps and obtain the final saliency result. Moreover, a bootstrap learning based salient object segmentation method was proposed. In addition, inspired by the observation that the salient regions are distinctly different from the local or global backgrounds in the depth map, "depth contrast" was calculated as a common depth property. In [55], global depth contrast and domain knowledge were calculated to measure the stereo saliency. Peng et al. [58] calculated the depth saliency through a multi-contextual contrast model, which considers the contrast prior, global distinctiveness, and background cue of depth map. Moreover, a multi-stage RGBD saliency model combining the low-level feature contrast, mid-level region grouping, and high-level prior enhancement was proposed.

Recently, deep learning is also successfully applied to RGB-D saliency detection [60]–[62]. Qu et al. [60] designed a CNN to automatically learn the interaction between low-level cues and saliency result for RGBD saliency detection. The local contrast, global contrast, background prior, and spatial prior were combined to generate the raw saliency feature vectors, which are embedded into a CNN to produce the initial saliency map. Finally, Laplacian propagation was introduced to further refine the initial saliency map and obtain the final saliency result. In addition to the multi-modal fusion problem that previous RGBD salient object detection focus on, Han et al. [61] firstly exposed the cross-modal discrepancy in the RGBD data and proposed two cross-modal transfer learning strategies to better explore modal-specific representations in the depth modality. This work is the pioneering one that involves the cross-modal transfer learning problem in RGBD salient object detection. In [62], Chen et al. innovatively modelled the cross-modal complementary part including the RGB and depth data as a residual function for RGBD saliency detection. Such a re-formulation elegantly posed the problem of exploiting cross-modal complementarity as approximating the residual, making the multi-modal fusion network to be really complementarity-aware. In this work, the high-level contexts and low-level spatial cues were well-integrated, and the saliency maps were enhanced progressively.

Fig. 5. Different quality depth maps. (a) Good depth map, $\lambda_d = 0.8014$. (b) Common depth map, $\lambda_d = 0.3890$. (c) Poor depth map, $\lambda_d = 0.0422$.

B. Depth Measure Based RGBD Saliency Detection

In order to capture the comprehensive and implicit attributes from the depth map and enhance the identification of salient object, some depth measures, such as anisotropic center-surround difference measure [64], local background enclosure measure [66], and depth contrast increased measure [68], are designed in different methods. Ju *et al.* [64] proposed an Anisotropic Center-Surround Difference (ACSD) measure with 3D spatial prior refinement to calculate the depth-aware saliency map. Combining the ACSD measure with color saliency map, Guo *et al.* [65] proposed an iterative propagation method to optimize the initial saliency map and generate the final result. Since the backgrounds always contain the regions that are highly variable in depth map, some high contrast background regions may induce false positives. To overcome this problem, Feng *et al.* [66] proposed a Local Background Enclosure (LBE) measure to directly capture salient structure from depth map, which quantifies the proportion of object boundary located in front of the background. The salient objects are always placed at different depth levels and occupy small areas according to the domain knowledge in photography. Based on this observation, Sheng *et al.* [68] proposed a depth contrast increased measure to pop-out the salient object through increasing the depth contrast between the salient object and distractors. Wang *et al.* [69] proposed a multistage salient object detection framework for RGBD images via Minimum Barrier Distance (MBD) transform and multilayer cellular automata based saliency fusion. The depth-induced saliency map was generated through the FastMBD method, and the depth bias and 3D spatial prior were used to fuse different saliency maps at multiple stages.

C. Discussion

Depth feature based method is an intuitive and explicit way to achieve RGBD saliency detection, which uses the depth information as an additional feature to supplement color feature, but ignores the potential attributes (e.g., shape and contour) in the depth map. By contrast, depth measure based method aims at exploiting these implicit information to refine the saliency result. However, how to effectively exploit the depth information to enhance the identification of salient object is a relatively difficult work. In addition, limited by the depth imaging techniques, sometimes the quality of the depth map is not satisfactory, as shown in Fig. 5(c). A good depth map benefits for the saliency detection, whereas a poor depth map may degenerate the saliency measurement. Therefore, Cong *et al.* [70] proposed a depth confidence measure to evaluate the quality of depth map, which works as a controller to constrain the introduction of depth information in the saliency model. The depth confidence measure λ_d is defined as follows:

$$\lambda_d = \exp((1 - m_d) \cdot CV \cdot H) - 1 \tag{1}$$

where m_d is the mean value of depth map, CV denotes the coefficient of variation, and H represents the depth frequency entropy, which describes the randomness of depth distribution. With the depth confidence measure, the RGBD saliency map was generated by combining the depth-aware compactness saliency and depth-guided foreground saliency. Fig. 5 illustrates some different quality depth maps. As visible, the depth confidence measure λ_d effectively distinguishes different quality depth maps according to the statistical characteristics of depth map.

III. CO-SALIENCY DETECTION

Co-saliency detection aims at detecting the common and salient regions from an image group containing multiple related images, which has been applied in foreground co-segmentation [73], object co-detection [111], and image matching [112]. In co-saliency detection, the inter-image correspondence is introduced as the common attribute constraint to discriminate the common objects from all the salient objects. To achieve co-saliency detection, some low-level or high-level features are firstly extracted to represent each image unit (e.g., superpixels), where the low-level feature describes the heuristic characteristics (e.g., color, texture, luminance), and the high-level feature captures the semantic attributes through some deep networks. Then, using these features, intra and inter saliency models are designed to explore the saliency representation from the perspectives of the individual image and inter image, respectively. For inter-image constraints capturing, different techniques are introduced, such as clustering, similarity matching, low rank analysis, and propagation. Finally, fusion and optimization schemes are utilized to generate the final co-saliency map.

We discuss two categories of co-saliency detection methods according to the different data, i.e., RGB co-saliency detection and RGBD co-saliency detection. Obviously, different from the RGB co-saliency detection, RGBD co-saliency detection model needs to combine the depth constraint with inter-image correspondence jointly. In addition, similar to the RGBD saliency detection, the depth cue can be used as an additional feature or a measure in RGBD co-saliency detection methods.

A. RGB Co-saliency Detection

As mentioned earlier, the inter-image correspondence plays an important role in co-saliency detection. In this subsection, we review some RGB co-saliency detection models based on different correspondence capturing strategies, i.e, matching based method [72]–[79], clustering based method [80], rank analysis based method [81], [82], propagation based method [83], [84], and learning based method [85]–[88]. A brief summary is presented in Table I.

Similarity Matching. In most of the existing methods, the inter-image correspondence is simulated as a similarity matching process among basic units. As a pioneering work, Li

TABLE I
BRIEF INTRODUCTION OF RGB CO-SALIENCY DETECTION

Model	Year	Inter-image capturing	Main technique
CSP [72]	2011	matching	normalized single-pair SimRank
UEM [73]	2011	matching	repeatedness representation
SA [74]	2013	matching	superpixel-level graph matching
CSM [75]	2013	matching	similarity ranking and matching
RFPR [76]	2014	matching	inter-region dissimilarity
HSCS [77]	2014	matching	global similarity
SCS [78]	2015	matching	ranking scheme
HCM [79]	2018	matching	hierarchical consistency measure
CCS [80]	2013	clustering	clustering with multiple cues
SAW [81]	2014	rank analysis	rank one constraint
LRMF [82]	2015	rank analysis	multiscale low-rank fusion
CSP [83]	2016	propagation	two-stage propagation
CFR [84]	2017	propagation	color feature reinforcement
LDW [85]	2015	learning	deep learning, Bayesian
GCS [86]	2017	learning	FCN framework, end-to-end
SPMI [87]	2015	learning	self-paced multi-instance learning
UML [88]	2017	learning	metric learning

and Ngan [72] proposed a co-saliency detection model for an image pair, where the inter-image correspondence is formulated as the similarity between two nodes through the normalized single-pair SimRank on a co-multilayer graph. However, this method is only applicable to image pairs. Tan *et al.* [74] proposed a self-contained co-saliency detection model based on affinity matrix, which evaluates the co-saliency according to the bipartite superpixel-level graph matching across image pairs. Li *et al.* [75] combined the intra and inter saliency maps to achieve co-saliency detection, where the inter-image corresponding relationship is measured by pairwise similarity ranking with pyramid features and minimum spanning tree image matching. Liu *et al.* [77] proposed a hierarchical segmentation based co-saliency detection model, where the inter-image correspondence is formulated as the global similarity of each region. Li *et al.* [78] proposed a saliency-guided co-saliency detection method, where the first stage recovers the co-salient parts missing in the single saliency map through the efficient manifold ranking, and the second stage captures the corresponding relationship via a ranking scheme with different queries. This model can make the existing saliency models work well in co-saliency scenarios.

Clustering is an effective way to build the inter-image correspondence, where the co-salient regions should be assigned to the same category. A cluster-based co-saliency detection algorithm without heavy learning for multiple images was proposed in [80]. Taking the cluster as the basic unit, an inter-image clustering model was designed to represent the multi-image relationship by integrating the contrast, spatial, and corresponding cues. The proposed method achieved a substantial improvement in efficiency.

Rank Analysis. Ideally, feature representations of co-salient objects should be similar and consistent, thus, the rank of feature matrix should appear low. Cao *et al.* [81] proposed a fusion framework for co-saliency detection based on rank

constraint, which is valid for multiple images and also works well on single image saliency detection. The self-adaptive weights for fusion process were determined by the low-rank energy. Moreover, this method can be used as a universal fusion framework for multiple saliency maps. Huang *et al.* [82] proposed a multiscale low-rank saliency fusion method for single image saliency detection, and the Gaussian Mixture Model (GMM) was used to generate the co-saliency map via a co-saliency prior.

Propagation scheme among multiple images is presented to capture the inter-image relationship. Ge *et al.* [83] proposed a co-saliency detection method based on two-stage propagation, where the inter-saliency propagation stage is utilized to discover common properties and generate the pairwise common foreground cue maps, and the intra-saliency propagation stage aims at further suppressing the backgrounds and refining the inter-saliency propagation maps. Based on the observation that co-salient objects appear similar color distributions in an abundant color feature space, Huang *et al.* [84] proposed a co-saliency detection method without single saliency residuals by using color feature reinforcement. In this method, eight color features with four exponents were formed into an abundant color feature space, and co-saliency indication maps were obtained with the help of feature coding coefficients and salient foreground dictionary.

Learning Model. Recently, learning based methods for RGB co-saliency detection attract more and more attention and achieve competitive performance.

Deep learning has been demonstrated to be powerful in learning the high-level semantic representation, and some heuristic studies of co-saliency detection based on deep learning have been proposed. Zhang *et al.* [85] proposed a co-saliency detection model from deep and wide perspectives under the Bayesian framework. From the deep perspective, some higher-level features extracted by the convolutional neural network with additional adaptive layers were used to explore better representations. From the wide perspective, some visually similar neighbors were introduced to effectively suppress the common background regions. This method is a pioneering work to achieve co-saliency detection by using deep learning, which mainly uses the convolutional network to extract better feature representations of the target. With the FCN framework, Wei *et al.* [86] proposed an end-to-end group-wise deep co-saliency detection model. First, the semantic block with 13 convolutional layers was utilized to obtain the basic feature representation. Then, the group-wise feature representation and single feature representation were captured to represent the group-wise interaction information and individual image information, respectively. Finally, the collaborative learning structure with the convolution-deconvolution model was used to output the co-saliency map. The overall performance of this method is satisfactory, but the boundary of the target needs to be sharper.

The Multi-Instance Learning (MIL) model aims to learn a predictor for each instance through maximizing inter-class distances and minimizing intra-class distances. The Self-Paced Learning (SPL) theory is to gradually learn from the easy/faithful samples to more complex/confusable ones.

(a) Dog   (b) Sculpture

Fig. 6. Illustration of DSP descriptor. From the left to right in a group are the RGB image, depth map, and DSP map.

Integrating the MIL regime into SPL paradigm, Zhang *et al.* [87] proposed a novel framework for co-saliency detection.

Metric learning works on learning a distance metric to make the same-class samples closer and different-class samples as far as possible. Han *et al.* [88] introduced metric learning into co-saliency detection, which jointly learns discriminative feature representation and co-salient object detector via a new objective function. This method has the capacity to handle the wide variation in image scene and achieves superior performance.

B. RGBD Co-saliency Detection

The superiority of depth cue has been proved in RGBD saliency detection. Combining the depth cue with inter-image correspondence, RGBD co-saliency detection can be achieved. For this task, there are two commonly used datasets, i.e., RGBD Coseg183 dataset [89] and RGBD Cosal150 dataset [91]. Limited by the data sources, only a few of methods are proposed to achieve RGBD co-saliency detection.

Clustering with Depth Feature. In [90], Song *et al.* proposed a bagging-based clustering method for RGBD co-saliency detection. The inter-image correspondence was explored via feature bagging and regional clustering. Moreover, three depth cues, including average depth value, depth range, and the Histogram of Oriented Gradient (HOG) on the depth map, were extracted to represent the depth attributes of each region.

Similarity Matching with Depth Feature. Fu *et al.* [89] introduced the RGBD co-saliency map into an object-based RGBD co-segmentation model with mutex constraint, where the depth cue is utilized to enhance identification of common foreground objects and provide local features for region comparison. Introducing the depth cue as an additional feature, Cong *et al.* [91] proposed a co-saliency detection method for RGBD images by using the multi-constraint feature matching and cross label propagation. The inter-image relationship was modeled at two scales, i.e., multi-constraint based superpixel-level similarity matching and hybrid feature based image-level similarity matching. Finally, cross label propagation scheme was designed to refine the intra and inter saliency maps in a cross way and generate the final co-saliency map.

Similarity Matching with Depth Measure. In [92], an iterative co-saliency detection framework for RGBD images was proposed, which integrates the addition scheme, deletion scheme, and iterative scheme. The addition scheme aimed at introducing the depth information and improving the performance of single saliency map. The deletion scheme focused on capturing the inter-image correspondence via the designed common probability function. The iterative scheme was served

as an optimization process through a refinement-cycle to further improve the performance. Notably, a novel depth descriptor, named Depth Shape Prior (DSP), was designed to exploit the depth shape attribute and convert the RGB saliency map into RGBD scenarios. Fig. 6 provides an illustration of DSP descriptor, which effectively describes the shape of salient object from the depth map. In other words, any RGB saliency map can be converted to an RGBD saliency map by using the DSP descriptor.

C. Discussions

Compared with image saliency detection, co-saliency detection is still an emerging topic, where the inter-image correspondence is crucial to represent the common attribute. The accurate inter-image constraint can effectively eliminate non-common saliency interference and improve the accuracy. On the contrary, the inaccurate inter-image correspondence, like noise, will degenerate the performance. The matching and propagation based methods usually capture relatively accurate inter-image relationship, but they are very time-consuming. In addition, the inter-image modeling among multiple images is a problem worth pondering in learning based method, and the stack strategy may be not a good choice for performance improvement. Of course, how to exploit the depth attribute to enhance the identification of co-salient object also needs to be further investigated.

IV. VIDEO SALIENCY DETECTION

Video sequences provide the sequential and motion information in addition to the color appearance, which benefit for the perception and identification of scene. The salient object in video is defined as the repeated, motion-related, and distinctive target. The repeated attribute constrains the salient object that should appear in most of the video frames. The motion-related characteristic is consistent with the human visual mechanism that the moving object attracts more attention than the static one. The distinctive property indicates the object should be prominent with respect to the background in each frame.

Most of the video saliency detection methods are dedicated to exploiting the low-level cues (e.g., color appearance, motion cue, and prior constraint) [93]–[105]. Only a few works focus on learning the high-level features and extracting the salient object in video through a learning network [106]–[110]. In the following, we will detail these two types of methods.

A. Low-level Cue Based Video Saliency Detection

According to the way of spatiotemporal extraction, low-level cue based video saliency detection method is classified into fusion model and direct-pipeline model, as shown in Fig. 7. For the fusion model, the spatial and temporal features are extracted to generate the spatial saliency and temporal saliency respectively, then they are combined to produce the final spatiotemporal saliency. By contrast, the direct-pipeline model directly extracts the spatiotemporal feature to generate the final spatiotemporal saliency in a straightforward and progressive way without any branches. A brief summary of the state-of-the-art methods is presented in Table II.
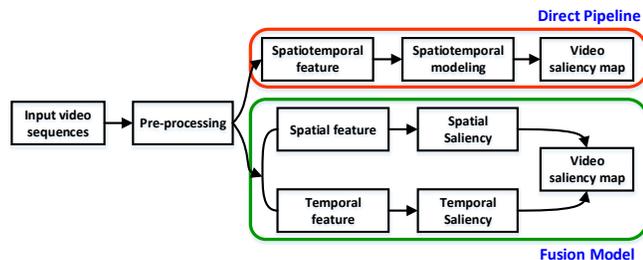
Fig. 7. Framework of low-level cue based video saliency detection.

TABLE II
BRIEF SUMMARY OF LOW-LEVEL CUE BASED VIDEO SALIENCY
DETECTION METHODS

Model	Year	Low-level cue	Type
SSR [93]	2012	sparse reconstruction, motion trajectory	fusion model
VSCD [95]	2014	DCT coefficients, luminance, color, texture, motion features	
SCUW [96]	2014	luminance, color, texture, optical flow, spatiotemporal uncertainty	
SS [97]	2014	global contrast, spatial sparsity, motion histogram	
STBP [102]	2017	spatiotemporal background prior, SIFT flow	
SFLR [103]	2017	spatiotemporal gradient contrast	
LRSD [94]	2012	low-rank, sparse decomposition	direct-pipeline model
CVS [98]	2015	gradient flow field, local and global contrasts	
RWRV [99]	2015	random walk with restart, motion distinctiveness, temporal consistency, abrupt change	
SG [101]	2018	spatial edge, motion boundary, geodesic distance	
SGSP [104]	2017	global motion histogram, shortest path on graph	
VSOP [105]	2017	object proposals, background and contrast priors, optical flow	

Fusion model fuses the spatial saliency and temporal saliency to achieve video saliency, where the spatial cue represents the intra-frame information in each frame and the temporal cue describes the inter-frame relationship among multiple frames.

For spatial saliency detection, some techniques and priors in image saliency detection can be used, such as sparse reconstruction, low-rank analysis, center-surround contrast prior, and background prior. In [93], the sparse reconstruction was utilized to discover the regions with high center-surround contrast. Fang *et al.* [95] generated the static saliency map via feature contrast in compressed domain by using the luminance, color, and texture features. In [97], the global contrast and spatial sparsity were used to measure the spatial saliency of each superpixel. Xi *et al.* [102] utilized the background prior to calculate the spatial saliency. In [103], color contrast was used to define the color saliency.

For temporal saliency, the motion cue is exploited to represent the moving objects in the video. In [93], the target patch was reconstructed by overlapping patches in neighboring frames. In [95], the motion vectors extracted from the video bitstream were used to calculate the feature differences between DCT blocks. In [97], the superpixel-level temporal saliency was evaluated by motion distinctiveness of motion histograms. In [102], the SIFT flow and bidirectional consistency propagation were used to define the temporal background prior. In [103], the motion gradient guided contrast computation was used to define the temporal saliency.

In most of the fusion based models, fusion strategy is not a key issue. In [95], a fusion scheme considering the saliency characteristic was designed. In [97], an adaptive fusion method at the pixel level was utilized to generate the pixel-level spatiotemporal saliency map. In [102], the spatial and temporal saliency maps were fused via a simple addition strategy. In [103], the modeling-based saliency adjustment and low-level saliency fusion were conducted to produce the fusion result. Furthermore, the low-rank coherency guided spatial-temporal saliency diffusion and saliency boosting strategies were adopted to improve the temporal smoothness and saliency accuracy.

Direct-pipeline model directly extracts the spatiotemporal feature to generate the final spatiotemporal saliency in a straightforward and progressive way.

In [94], the stacked temporal slices along X-T and Y-T planes were used to represent the spatiotemporal feature, and the motion saliency was calculated by low-rank and sparse decomposition, where the low-rank component corresponds to the background, and the sparse proportion represents the moving foreground object.

Optical flow and its deformations are utilized to define the spatiotemporal feature. Wang *et al.* [98] presented a spatiotemporal saliency model based on gradient flow field and energy optimization, which is robust to complex scenes, various motion patterns, and diverse appearances. The gradient flow field represented the salient regions by incorporating the intra-frame and inter-frame information. Liu *et al.* [104] presented a progressive pipeline for video saliency detection, including the superpixel-level graph based motion saliency, temporal propagation, and spatial propagation. The motion saliency was measured by the shortest path on the superpixel-level graph with global motion histogram feature. Guo *et al.* [105] introduced a salient object detection method for video from the perspective of object proposal via a more intuitive visual saliency analysis. The salient proposals were firstly determined by spatial saliency stimuli and contrast-based motion saliency cue. Then, proposal ranking and voting schemes were conducted to screen out non-salient regions and estimate the initial saliency. Finally, temporal consistency and appearance diversity were considered to refine the initial saliency map. It is worth learning that object proposal provides a more comprehensive and high-level representation to detect the salient object.

In addition, motion knowledge is used to capture the spatiotemporal feature. In [99], the random walk with restart was

exploited to detect the salient object in video, where the temporal saliency calculated by motion distinctiveness, temporal consistency, and abrupt change is employed as the restarting distribution of random walker. In [100], [101], the spatial edge and motion boundary were incorporated as the spatiotemporal edge probability cue to estimate the initial object on the intra-frame graph, and the spatiotemporal saliency was calculated by the geodesic distance on the inter-frame graph.

In summary, the fusion model is a more intuitive method compared with the direct-pipeline model. Moreover, the existing image saliency methods can be directly used to compute the spatial saliency, which lays the foundation for spatiotemporal saliency calculation. Therefore, most of the methods pay more attention to this type.

B. Learning Based Video Saliency Detection

Recently, learning based video saliency detection has achieved more competitive performance, which can be roughly divided into supervised learning method [107]–[109] and unsupervised learning method [110].

The supervised learning method aims at learning the spatiotemporal features for video saliency detection by means of a large number of labelled video sequences. Le *et al.* [107] proposed a deep model to capture the SpatioTemporal deep Feature (STF), which consists of the local feature produced by a region-based CNN and the global feature computed from a block-based CNN with temporal-segments embedding. Using the STF feature, random forest and spatiotemporal conditional random field models were introduced to obtain the final saliency map. Wang *et al.* [108] designed a deep saliency detection model for video, which captures the spatial and temporal saliency information simultaneously. The static network generated the static saliency map for each individual frame via the FCNs, and the dynamic network employed frame pairs and static saliency map as input to obtain the dynamic saliency result. It is worth mentioning that, a video augmentation technique was proposed to generate the labeled video training data from the existing annotated image datasets, which effectively alleviates the problem of insufficient training samples.

Most of the deep learning based video saliency detection methods focus on designing a separated network rather than an end-to-end network. In [109], Le *et al.* firstly proposed an **end-to-end** 3D fully convolutional network for salient object detection in video. The Deeply Supervised 3D Recurrent Fully Convolutional Network (DSRFCN3D) contained an encoder network and a decoder network. The encoder network was used to extract the 3D deep feature from the input video block, and the decoder network aimed at computing the accurate saliency voxel. Moreover, a refinement mechanism with skip-connection and 3D Recurrent Convolution Layer (RCL3D) was designed to learn the contextual information. The loss function combined the saliency prediction loss and 3D deconvolution loss jointly, which is represented as:

$$L(\theta, \omega) = \zeta_{pred}(\theta, \omega_{pred}) + \sum_{m=1}^{M} \zeta_{dec3D}(\theta, \omega_{dec3D}^{m}) \quad (2)$$

where θ is the overall network parameters, ω_{pred} denotes the weights of saliency prediction network, ω_{dec3D} represents the weights of 3D deconvolution network, M is the number of 3D deconvolution layers, and ζ denotes the binary cross-entropy function.

Compared with supervised learning methods, only a few works focus on unsupervised learning model. As a pioneering work, Li *et al.* [110] proposed an **unsupervised** approach for video salient object detection by using the saliency-guided stacked autoencoders. First, saliency cues extracted from the spatiotemporal neighbors at three levels (i.e., pixel, superpixel, and object levels) were combined as a high-dimensional feature vector. Then, the stacked autoencoders were learned in an unsupervised manner to obtain the initial saliency map. Finally, some post-processing operations were used to further highlight the salient objects and suppress the distractors. In this method, manual intervention will be further reduced if the hand-crafted saliency cues are automatically learned from the network.

C. Discussions

For video saliency detection, motion cue is crucial to suppress the backgrounds and static salient objects, especially in the case of multiple objects. In general, optical flow is a common technique to represent the motion attribute. However, it is time-consuming and sometimes inaccurate, which will degenerate the efficiency and accuracy. Therefore, some deep learning based methods directly embed the continuous multiple frames into the network to learn the motion information and avoid the optical flow calculation. Of course, the video frame and optical flow can be simultaneously embedded into the network to learn the spatiotemporal feature. However, the first option may be better in terms of efficiency. In addition, the salient objects should be consistent in appearance among different frames. Therefore, some techniques, such as energy function optimization, are adopted to improve the consistency of the salient object.

In Table III, we further summarize the characteristics of different types of saliency models, including RGBD saliency detection, co-saliency detection, and video saliency detection.

V. EVALUATION AND DISCUSSION

A. Evaluation Metrics

In addition to directly comparing the saliency map with ground truth, some evaluation metrics are developed to quantitatively evaluate the performance of saliency detection methods, such as Precision-Recall (PR) curve, F-measure, Receive Operator Characteristic (ROC) curve, Area Under the Curve (AUC) score, and Mean Absolute Error (MAE).

Precision-Recall (PR) curve and F-measure. By thresholding the saliency map with a series of fixed integers from 0 to 255, the binary saliency masks are achieved. Therefore, the precision and recall scores are calculated by comparing the binary mask with the ground truth. The PR curve is drawn under different precision and recall scores, where the vertical axis denotes the precision score, and the horizontal axis corresponds to the recall score. The closer the PR curve

TABLE III
A SUMMARY OF SALIENCY DETECTION MODELS WITH COMPREHENSIVE INFORMATION

Model	Category	Key points	Descriptions
RGBD saliency	depth feature	use as an additional feature, a heuristic and intuitive strategy	directly embed into the feature pool as the supplement of color information or calculate the "depth contrast" as a common depth property [55]–[62]
RGBD saliency	depth measure	design a depth measure to capture the shape and structure attribute	design some depth measures to fully exploit the effective information from depth map (e.g., shape and structure) [63]–[70]
Co-saliency	RGB co-saliency	inter-image constraint, low-level appearance and high-level semantic features	inter-image correspondence can be modelled as a similarity matching, clustering, rank analysis, propagation, or learning process [72]–[88]
Co-saliency	RGBD co-saliency	a new topic, inter-image constraint, depth attribute (feature or measure)	depth cue as a feature or a measure to enhance identification performance, combine the depth cue with inter-image correspondence [89]–[92]
Video saliency	low-level cue based	direct pipeline: spatiotemporal feature fusion model: spatial, temporal saliencies	explore the inter-frame constraint and motion information (e.g., optical flow), fusion model is more intuitive and popular [93]–[105]
Video saliency	learning based	unsupervised: stacked autoencoder supervised: symmetrical deep structure	learn to the spatiotemporal features and achieve competitive performance, multiple frames or optical flow is embedded to represent motion cue [106]–[110]

is to the upper left, the better performance achieves. In order to comprehensively evaluate the saliency map, a weighted harmonic mean of precision and recall is defined as F-measure [2], which is expressed as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Reall}{\beta^2 \times Precision + Recall} \qquad (3)$$

where β^2 is generally set to 0.3 for emphasizing the precision as suggested in [50].

Receive Operator Characteristic (ROC) curve and AUC score. The ROC curve describes the relationship between the false positive rate (FPR) and true positive rate (TPR), which is represented as:

$$TPR = \frac{|S_F \bigcap G_F|}{|G_F|}, FPR = \frac{|S_F \bigcap G_B|}{|G_B|} \qquad (4)$$

where S_F, G_F, and S_B denote the set of detected foreground pixels in the binary saliency mask, the set of foreground pixels in the ground truth, and the set of background pixels in the ground truth, respectively. The closer the ROC curve is to the upper right, the better performance achieves. AUC score is the area under the ROC curve, and the larger, the better.

Mean Absolute Error (MAE) score. MAE score directly evaluates the difference between the continuous saliency map S and ground truth G directly:

$$MAE = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} |S(x,y) - G(x,y)| \qquad (5)$$

where w and h represent the width and height of the image, respectively. The smaller the MAE score is, the more similar to the ground truth, and the better performance achieves.

B. Datasets

In this section, we introduce the datasets for (RGBD) image saliency detection, co-saliency detection, and video saliency detection, respectively.

For image saliency detection, a number of datasets have been constructed over the past decade, including some large datasets with pixel-level annotations, such as DUT-OMRON [32], MSRA10K [113], HKU-IS [47], and XPIE [114], as listed in Table IV. Benefiting from the growth of data volume, deep learning based RGB saliency detection methods have achieved superior performance.

In contrast, the datasets with pixel-wise ground truth annotations for RGBD saliency detection are relatively inadequate, which only consist of NLPR dataset [58] and NJUD dataset [64], as listed in the last two rows of Table IV. The NLPR dataset includes 1000 RGBD images with the resolution of 640×640, where the depth maps are captured by Microsoft Kinect. The NJUD dataset is released on 2015, which includes 2000 RGBD images with the resolution of 600×600. The depth map in the NJUD dataset is estimated by the stereo images.

For co-saliency detection, five RGB datasets and two RGBD datasets are commonly used for evaluation, as listed in Table V. MSRC [115] is a challenging dataset with complex background, which contains 7 image groups of totally 240 images with manually pixel-wise ground truth. The iCoseg [116] dataset consists of 38 image groups of totally 643 images, and the manually labeled pixel-wise ground-truth masks is also provided. Image Pair [72] dataset only contains image pairs, whereas other datasets usually include more than two images in each group. A larger co-saliency detection dataset named RGBD Cosal2015 is constructed in [117], which consists of 2015 RGB images distributed in 50 image groups with pixel-wise ground truth. INCT2016 [118] is a more challenging dataset with larger appearance variation, indefinite number of targets, and complicated backgrounds, which contains 291 images distributed in 12 categories with pixel-level ground truth. There are two commonly used datasets with pixel-level hand annotations for RGBD co-saliency detection. One is the RGBD Coseg183 dataset [89], which contains 183 RGBD images in total that distributed in 16 image groups. The other

TABLE IV
BRIEF INTRODUCTION OF SALIENCY DETECTION FOR RGB IMAGE AND RGBD IMAGES

Dataset	Image number	Max Resolution	Depth attribute	Object property	Background property	Publish year	Best performance
ACSD [50]	1000	400×400	–	single, moderate	clean, simple	2009	F_β: 0.94; MAE: 0.03
ECSSD [21]	1000	400×400	–	single, large	clean, simple	2012	F_β: 0.88; MAE: 0.08
DUT-OMRON [32]	5168	400×400	–	single, small	complex	2013	F_β: 0.77; MAE: 0.07
MSRA10K [113]	10000	400×400	–	single, largre	clean, simple	2014	F_β: 0.93; MAE: 0.04
PASCAL-S [33]	1000	500×500	–	multiple, moderate	simple	2014	F_β: 0.81; MAE: 0.11
HKU-IS [47]	850	400×400	–	multiple, moderate	clean	2015	F_β: 0.86; MAE: 0.06
XPIE [114]	4447	300×300	–	single, moderate	complex	2017	F_β: 0.72; MAE: 0.12
NLPR [58]	1000	640×640	Kinect capturing	single, moderate	diverse	2014	F_β: 0.82; AUC: 0.98
NJUD [64]	2000	600×600	depth estimation	single, moderate	diverse	2015	F_β: 0.81; AUC: 0.98

TABLE V
BRIEF INTRODUCTION OF CO-SALIENCY DETECTION DATASETS

Dataset	Image number	Group number	Group size	Depth attribute	Resolution	Object property	Background property	Publish year	Best performance
MSRC [115]	240	7	30-53	–	320×210	complex	clean, simple	2005	F_β: 0.84; AUC: 0.70
iCoseg [116]	643	38	4-42	–	500×300	multiple	diverse	2010	F_β: 0.85; AUC: 0.85
Image Pair [72]	210	115	2	–	128×100	single	clustered	2011	F_β: 0.93; AUC: 0.97
Cosal2015 [117]	2015	50	26-52	–	500×333	multiple	clustered	2016	F_β: 0.71; AUC: 0.90
INCT2016 [118]	291	12	15-31	–	500×375	multiple	complex	2016	–
RGBD Coseg183 [89]	183	16	12-36	Kinect capturing	640×480	multiple	clustered, complex	2015	F_β: 0.71; MAE: 0.06
RGBD Cosal150 [91]	150	21	2-20	depth estimation	600×600	single	diverse	2018	F_β: 0.84; MAE: 0.14

TABLE VI
BRIEF INTRODUCTION OF VIDEO SALIENCY DETECTION DATASETS

Dataset	Frame number	Video number	Video size	Resolution	Object property	Background property	Publish year	Best performance
SegTrackV1 [119]	244	6	21-71	414×352	single	diverse	2010	F_β: 0.88; MAE: 0.10
SegTrackV2 [120]	1065	14	21-279	640×360	single	diverse	2013	F_β: 0.92; MAE: 0.02
ViSal [98]	963	17	30-100	512×228	single	diverse	2015	F_β: 0.85; MAE: 0.03
MCL [99]	3689	9	131-789	480×270	single, small	complex	2015	–
DAVIS [121]	3455	50	25-104	1920×1080	multiple	complex	2016	F_β: 0.82; MAE: 0.03
UVSD [104]	6524	18	71-307	352×288	single, small	clustered, complex	2017	F_β: 0.51; MAE: 0.10
VOS [110]	116103	200	~ 500	800×800	single	complex	2018	F_β: 0.78; MAE: 0.05

one is the RGBD Cosal150 dataset [91], which collects 21 image groups containing a total of 150 RGBD images.

For video saliency detection, many datasets have been released, such as ViSal [98], MCL [99], UVSD [104], VOS [110], SegTrackV1 [119], SegTrackV2 [120], and DAVIS [121], as listed in Table VI. The DAVIS dataset is a commonly used and challenging dataset, which contains 50 video sequences with the fully-annotated pixel-level ground truth for each frame. The UVSD dataset is a specially designed and newly established dataset for video saliency detection, which

consists of 18 unconstrained videos with complicated motion patterns and cluttered scenes, and the pixel-wise ground truth for each frame is available. A very large video saliency detection dataset named VOS is constructed, which consists of 116103 frames in total that distributed in 200 video sequences. In this dataset, 7467 frames are annotated into binary ground truth, which is suitable for training and learning a deep model to extract the salient object in video.
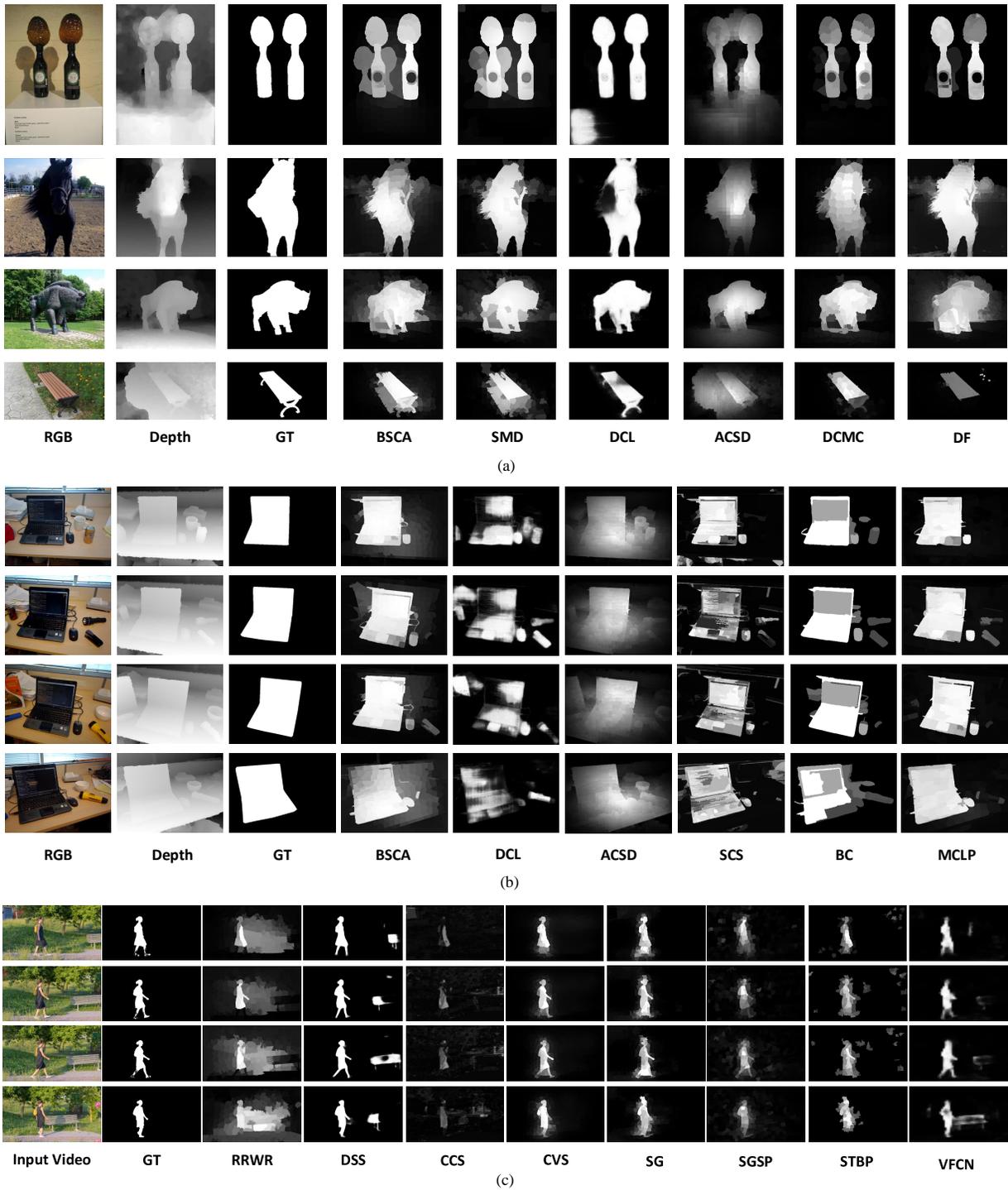
Fig. 8. Visual examples of different methods on different datasets. (a) NJUD dataset. (b) RGBD Coseg183 dataset. (c) DAVIS dataset.

C. Comparison and Analysis

We report some visual examples and quantitative comparisons in this section, and the related results on different datasets are shown in Figs. 8-9 and Tables VII-IX. All the results are directly provided by the authors or implemented by the source codes.

Image Saliency Detection vs RGBD Saliency Detection. We evaluate the RGB saliency detection methods (HS [21],

BSCA [25], RRWR [26], DCLC [28], SMD [29], DCL [40], and DSS [43]) and RGBD saliency detection methods (SS [55], ACSD [64], DCMC [70], and DF [60]) on the NJUD and NLPR datasets, where the DCL, DSS, and DF are the deep learning based methods.

Fig. 8(a) presents some visual examples on the NJUD dataset, where the first three columns correspond to the input RGBD images and ground truth. The foreground object is effectively popped out, and the background is suppressed in
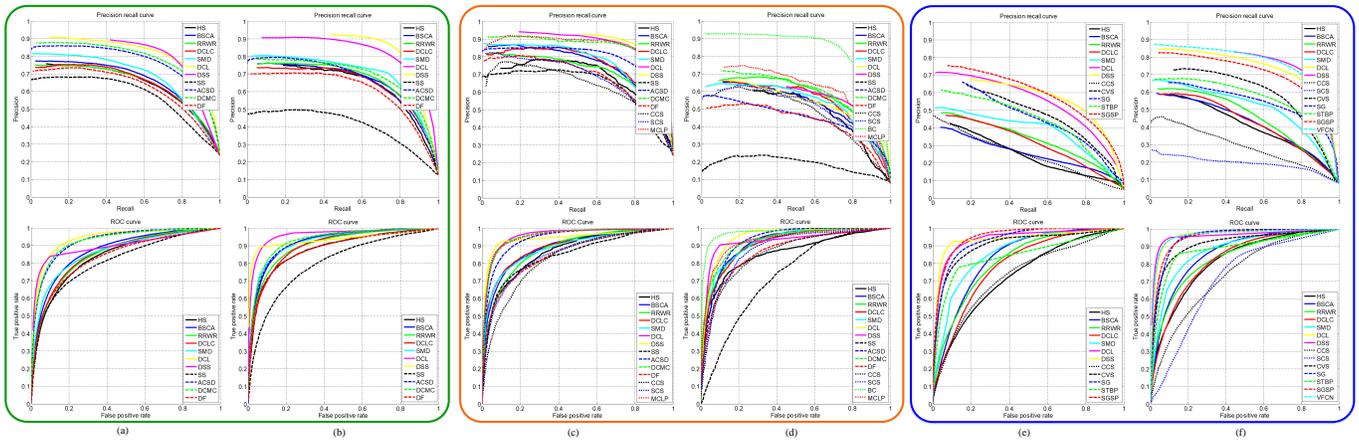
Fig. 9. PR and ROC curves of different methods on different datasets. (a) NJUD dataset. (b) NLPR dataset. (c) RGBD Cosal150 dataset. (d) RGBD Coseg183 dataset. (e) UVSD dataset. (f) DAVIS dataset.

TABLE VII
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON NJUD AND NLPR DATASETS, WHERE "*" DENOTES THE DEEP LEARNING BASED METHODS

	NJUD Dataset			NLPR Dataset		
	F_β	AUC	MAE	F_β	AUC	MAE
HS [21]	0.6494	0.8390	0.2516	0.6659	0.8785	0.1918
BSCA [25]	0.6672	0.8709	0.2148	0.6702	0.9207	0.1768
RRWR [26]	0.6520	0.8510	0.2161	0.6804	0.9038	0.1575
DCLC [28]	0.6527	0.8526	0.2007	0.6662	0.8992	0.1381
SMD [29]	0.6900	0.8635	0.1950	0.7138	0.9229	0.1300
DCL* [40]	0.7863	0.9393	0.1236	0.7995	0.9617	0.0727
DSS* [43]	0.7971	0.8940	0.1147	0.8384	0.9360	0.0583
SS [55]	0.6128	0.8103	0.2227	0.4712	0.8007	0.1752
ACSD [64]	0.7459	0.9259	0.1939	0.6695	0.9229	0.1635
DCMC [70]	0.7591	0.9258	0.1716	0.6975	0.9289	0.1168
DF* [60]	0.6383	0.8338	0.2022	0.6407	0.8801	0.1156

TABLE VIII
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON RGBD COSAL150 AND RGBD COSEG183 DATASETS, WHERE "*" DENOTES THE DEEP LEARNING BASED METHODS

	RGBD Cosal150 Dataset			RGBD Coseg183 Dataset		
	F_β	AUC	MAE	F_β	AUC	MAE
HS [21]	0.7101	0.8644	0.2375	0.5645	0.8540	0.2018
BSCA [25]	0.7318	0.8914	0.1925	0.5678	0.9164	0.1877
RRWR [26]	0.7106	0.8797	0.1967	0.6089	0.9163	0.1504
DCLC [28]	0.7385	0.8913	0.1728	0.5994	0.9073	0.1097
SMD [29]	0.7494	0.8863	0.1774	0.5760	0.9161	0.1229
DCL* [40]	0.8345	0.9580	0.1056	0.5531	0.9448	0.0967
DSS* [43]	0.8540	0.9404	0.0869	0.5972	0.9200	0.0783
SS [55]	0.6744	0.8453	0.2052	0.2567	0.7295	0.1716
ACSD [64]	0.7788	0.9410	0.1806	0.4787	0.9226	0.1940
DCMC [70]	0.8348	0.9551	0.1498	0.6169	0.9253	0.1009
DF* [60]	0.6844	0.8510	0.1945	0.4840	0.8654	0.1077
CCS [80]	0.6311	0.8242	0.2138	0.5383	0.8563	0.1210
SCS [78]	0.6724	0.8515	0.1966	0.5553	0.8797	0.1616
BC [90]	—	—	—	0.8262	0.9746	0.0541
MCLP [91]	0.8403	0.9550	0.1370	0.6365	0.9294	0.0979

the first three depth maps. In other words, the depth map can provide useful information to enhance the identification of the salient object. For the unsupervised RGB saliency detection methods (BSCA [25] and SMD [29]), some backgrounds are wrongly detected, such as the shadow in the first image and the trees in the third image. Introducing the depth cue, the consistency of salient object and the false positive in background regions are obviously improved. For example, the trees in the second and third images are effectively suppressed by the DCMC method [70]. For the supervised learning method, the DCL method [40] shows the competitive performance benefitting from the deep learning technique with a large number of labelled training data. However, limited by the annotated RGBD saliency data, the DF method [60] cannot completely suppress the backgrounds (e.g., the trees in the third image) and lose some foreground details (e.g., the chair legs in the fourth image).

The PR and ROC curves are shown in Fig. 9(a)-(b). As

can be seen, on both two datasets, the deep learning based RGB saliency detection methods (DSS [43] and DCL [40]) achieve the top two performances on the PR and ROC curves, and the unsupervised RGBD saliency model (DCMC [70]) reaches the third precision on the ROC curve. Table VII reports the quantitative measures of different saliency methods on these two datasets, including the F-measure, AUC score, and MAE score. The overall trend of quantitative comparisons is consistent with the visualization results, that is, the performances of the unsupervised RGBD saliency models (e.g., ACSD [64] and DCMC [70]) are significantly superior to the unsupervised RGB saliency detection methods, with the maximum percentage gain of 18% on the NJUD dataset in terms of the F-measure. Benefitting from the supervised learning with a large number of labelled data, deep learning based image saliency detection methods (DCL [40] and

TABLE IX
QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON UVSD AND DAVIS DATASETS, WHERE "*" DENOTES THE DEEP LEARNING BASED METHODS

	UVSD Dataset			DAVIS Dataset		
	F_β	AUC	MAE	F_β	AUC	MAE
HS [21]	0.3258	0.7288	0.2727	0.4552	0.8170	0.2495
BSCA [25]	0.3059	0.8320	0.2227	0.4731	0.8602	0.1945
RRWR [26]	0.3931	0.8152	0.1842	0.5138	0.8328	0.1678
DCLC [28]	0.3878	0.7899	0.1249	0.4812	0.8250	0.1337
SMD [29]	0.4521	0.8599	0.1347	0.5434	0.8774	0.1506
DCL* [40]	0.5759	0.9376	0.0593	0.7200	0.9647	0.0630
DSS* [43]	0.5967	0.9362	0.0480	0.7564	0.9576	0.0500
CCS [80]	0.3124	0.7303	0.1107	0.3485	0.7418	0.1506
SCS [78]	–	–	–	0.2305	0.7166	0.2569
CVS [98]	0.5122	0.9052	0.1031	0.6251	0.9162	0.0995
SG [100]	0.4851	0.9310	0.1050	0.5600	0.9485	0.1027
STBP [102]	0.4914	0.8443	0.0840	0.5859	0.8842	0.1016
SGSP [104]	0.6016	0.9505	0.1585	0.6944	0.9504	0.1375
VFCN* [108]	–	–	–	0.7488	0.9637	0.0588

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

Method	DCLC [28]	SMD [29]	DF [60]	CCS [80]	MCLP [91]
Time	1.96	7.49	12.95	2.65	41.03

SCS [78]). Moreover, the performance of RGBD co-saliency detection method with the depth constraint is superior to the RGB co-saliency detection method. For example, the MCLP method [91] achieves the best visual performance compared with other methods in Fig. 8(b). The computer in each image is highlighted more consistent and homogeneous, while the backgrounds (e.g., the desk) and non-common objects (e.g., the yellow flashlight, red hat, and orange can) are effectively eliminated. However, the flashlights with the same color as the computer are mistakenly reserved. The main reason is that, the low-level feature based method primarily focuses on capturing the color appearances from the image, while ignoring the high-level semantic attributes. The consistent conclusion can be drawn from the quantitative results in Fig. 9(c)-(d) and Table VIII. On the RGBD Coseg183 dataset, compared with the deep learning based DCL method [40], RGB co-saliency detection SCS method [78] achieves better performance in terms of F-measure. Moreover, benefitting from the depth cue and inter-image constraint, RGBD co-saliency models achieve more competitive performances. For example, the percentage gain of the BC method [90] reaches at least 33.9% in terms of F-measure compared with others on the RGBD Coseg183 dataset, which indirectly proves the importance role of these information in co-saliency detection.

For evaluating the running time, we tested the typical saliency detection methods, including single image saliency (DCLC [28] and SMD [29]), RGBD saliency (DF [60]), co-saliency (CCS [80]), and RGBD co-saliency (MCLP [91]), on a Quad Core 3.7GHz workstation with 16GB RAM. The codes are provided by the authors, which are implemented by using MATLAB 2014a. The comparisons of the average running time on the RGBD Cosal150 dataset are listed in Table X. The single image saliency detection method only considers the visual information from the individual image and costs less running time. For example, DCLC method [28] only costs 1.96 seconds to process one image. As a deep learning based RGBD saliency detection method, DF method [60] takes 12.95 seconds for testing one image, which is relatively slow. Co-saliency detection algorithm needs to build the global corresponding constraint from the multiple images, thus it generally requires more computation time, especially for the matching based methods (such as MCLP [91]). Although the matching based co-saliency detection algorithm may be slower, it tends to achieve better performance. Moreover, the computations can be further accelerated on GPUs using C++.

Image Saliency Detection / Co-saliency Detection vs Video Saliency Detection. We compare the video saliency detection methods with image saliency detection and co-saliency detection models on the UVSD and DAVIS datasets, and the

DSS [43]) yield the decent performance, even superior to the RGBD saliency methods. For different RGBD saliency detection methods, SS method [55] only focuses on some straightforward domain knowledge from the depth map, thus the performance is unsatisfactory. ACSD method [64] designs a novel depth measure to fully capture the depth attributes, and achieves appreciable performance. However, it does not have the ability to distinguish different quality of depth map. Introducing the depth confidence measure, DCMC method [70] is more robust to the poor depth map, and achieves more stable performance. For the deep learning based RGBD saliency detection method (i.e., DF [60]), due to the lack of labelled RGBD images, the performance is not satisfactory.

Image Saliency Detection vs Co-saliency Detection. We evaluate four types of saliency detection methods on the RGB-D Cosal150 and RGBD Coseg183 datasets, including image saliency detection methods (HS [21], BSCA [25], RRWR [26], DCLC [28], SMD [29], DCL [40], and DSS [43]), RGBD saliency detection methods (SS [55], ACSD [64], DCMC [70], and DF [60]), co-saliency detection methods (CCS [80] and SCS [78]), and RGBD co-saliency detection methods (BC [90] and MCLP [91]).

In Fig. 8(b), we present an image group with one common salient object (i.e., black computer) and cluttered backgrounds. From the figure, we can see that the image saliency detection methods (i.e., BSCA [25] and DCL [40]) cannot achieve better visual result with consistently highlighted salient objects and effectively suppressed background regions. For example, the desk with high luminance is wrongly detected by the unsupervised BSCA method [25], and the salient objects are not effectively highlighted by the deep learning based DCL method [40] due to the complex and cluttered backgrounds. By contrast, considering the inter-image corresponding relationship, some backgrounds (e.g., the desk) are effectively suppressed by the RGB co-saliency detection method (e.g.,

results are presented in Figs. 8-9 and Table IX. Video saliency detection is a more challenging task due to the complex motion patterns, cluttered backgrounds, and diversity spatiotemporal features. Five video saliency detection methods, including four unsupervised methods (CVS [98], SG [100], STBP [102], SGSP [104]), and a deep learning based method (VFCN [108]) are used for comparison.

From the qualitative examples show in Fig. 8(c), the unsupervised image saliency detection methods cannot achieve superior performances due to the lack of temporal and motion constraints. For example, the background regions cannot be effectively suppressed by the RRWR method [26], as shown in the third column of Fig. 8(c). Although the co-saliency detection model considers the inter-image relationship, it is still insufficient to fully represent the continuous inter-frame correspondence. Coupled with the lack of motion description, the salient object in video is not vigorously highlighted by the co-saliency model, such as the fifth column of Fig. 8(c). By contrast, video saliency detection methods achieve more satisfying performances both qualitatively and quantitatively. For example, the moving woman is highlighted homogeneously with clean background interference through the CVS method [98].

From the quantitative results reported in Fig. 9(e)-(f) and Table IX, all the measurements of unsupervised video saliency models are superior to the unsupervised image saliency and co-saliency detection methods on these two datasets. For example, on the DAVIS dataset, the maximum percentage gain reaches 52.5% in terms of F-measure. On the UVSD dataset, the F-measure can be improved from 0.3059 to 0.6016, with the percentage gain of 96.7%. Notably, the deep learning based methods demonstrate excellent performance improvement, especially including the image saliency detection method (DCL [40] and DSS [43]). On the DAVIS dataset, the VFCN and DSS methods are comparable in performance and superior to the DCL method. Compared the visual examples of the DSS and VFCN methods in Fig. 8(c), the DSS method obtains more consistent salient regions, while the static object (i.e., the bench) is mistakenly reserved. By contrast, the bench can be partially suppressed by the VFCN method due to the introduction of motion cue and inter-frame continuity.

Summary. Taking the depth information as a supplementary feature of color information is a heuristic and intuitive strategy to achieve RGBD saliency detection. This type of method is easy to implement, and the performance can be improved by a reasonable saliency framework with depth cue. By contrast, the depth measure based method often yields better performance, because of it can further capture effective depth attributes from the original depth map rather than only focusing on some low-level statistical features. However, it is a tricky question, which requires researchers to gain insight into the characteristics of depth data, and to comprehensively explore the depth attributes of salient objects. In addition, there is an interesting phenomenon, the performance of deep learning based RGBD saliency detection method did not exceed the unsupervised methods due to the lack of annotated RGBD saliency training data. Thus, the data augmentation and network designing need to be further investigated for the deep learning based RGBD saliency detection method.

Compared with image saliency detection, co-saliency detection is a more challenging task because the multiple images need to be processed jointly. Therefore, the inter-image correspondence is crucial to determine the common attribute of salient objects and suppress the background regions. However, the inaccurate inter-image correspondence, like noise, may degenerate the detection performance, even not as good as some image saliency detection methods. Combining the inter-image constraint and depth cue, RGBD co-saliency detection is achieved, where the depth cue is utilized as an additional feature rather than a depth measure in most of the existing methods. Benefitting from the introduction of depth cue and inter-image constraint, the performance of RGBD co-saliency detection model is obviously improved, and the percentage gain reaches more than 30% as shown in Table VIII. At present, the research on RGBD co-saliency detection is relatively preliminary, and mainly focuses on unsupervised methods. Therefore, how to extract the depth and color features, capture the inter-image constraint relationship, and guarantee the consistency of salient regions are the research priorities of co-saliency detection in the future.

Taking the video sequences as some independent images, the image saliency detection models cannot obtain satisfying performance due to ignoring the inter-frame constraint and motion cue, especially the unsupervised methods. Considering the correspondence relationship between frames, the performance of co-saliency models are also disillusionary. There are two main reasons, i.e., (1) The salient objects in video are continuous in temporal axis and consistent among different frames. Thus, the inter-image correspondence in image group is not equal to the inter-frame constraint in video. (2) Motion information is essential to distinguish the salient object from the complex scene. Therefore, the temporal and motion cues should be fully utilized to highlight the salient object and suppress the backgrounds. Furthermore, the deep learning based methods have demonstrated the great superiority in performance. The continuous multiple frames or optical flow are embedded in the symmetrical network (e.g., convolution-deconvolution, encoder-decoder) to learn the spatiotemporal information and completely recover the salient regions. Although only a few of the deep learning based video saliency detection methods are available, it also points out an effort direction for future research.

D. Applications

We briefly introduce two intuitive and novel applications of saliency detection, i.e., Region-of-Interest (ROI) extraction in remote sensing image and primary object detection in video. The first task aims at extracting the ROI regions in remote sensing image, which is similar to the image saliency detection. The second one focuses on consecutively discovering the primary object in video, which is analogous to the video saliency detection.

ROI Extraction in Remote Sensing Image. With the development of imaging devices and sensors, the acquisition of high-resolution remote sensing image becomes more and
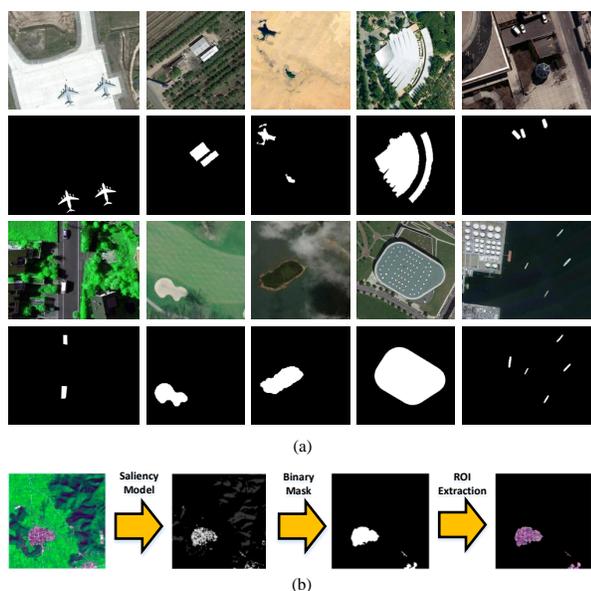
15

Fig. 10. Examples. (a) Some remote sensing images and the corresponding saliency region masks. (b) Flowchart of ROI extraction in remote sensing image.
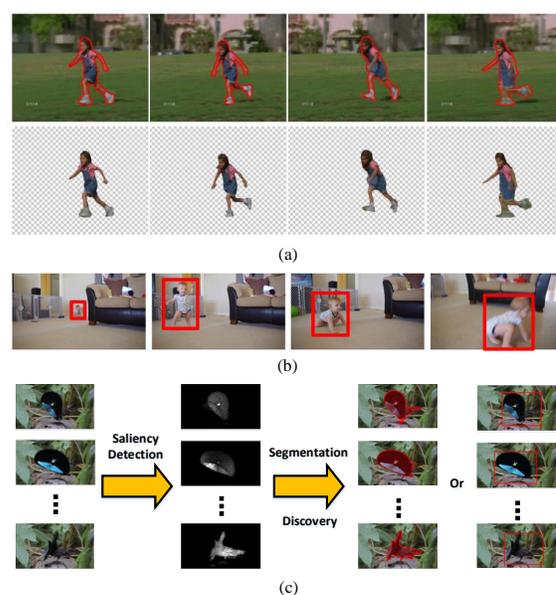

Fig. 11. Examples of primary object detection in video. (a) Primary object segmentation in video. (b) Primary object discovery in video. (c) Flowchart of primary object detection in video.

more convenient and accurate. Fig. 10(a) shows some remote sensing images and the corresponding saliency masks. As visible, the remote sensing image is very similar to the conventional color image, except that the remote sensing image is mainly photographed from a high angle shot. Therefore, there are many small targets in remote sensing images. ROI extraction technique in remote sensing image has been applied in a variety of perception tasks, such as object detection, land cover classification, and object recognition. Generally, the saliency attribute of the object is used in most of the existing methods to constrain the ROI extraction in remote sensing image [122]–[128], and the general flowchart is shown in Fig. 10(b). However, due to complex imaging mechanisms and different image characteristics, it is difficult to achieve satisfactory performance by directly transplanting the traditional RGB image saliency detection method to remote sensing image. There are two urgent issues that need to be addressed, i.e., (1) Extract the unique expression of salient object in remote sensing image. (2) Handle the small targets in remote sensing image. At present, this task has a very broad space for development.

Primary Object Detection in Video. Video sequence can be divided into two components, i.e., primary objects and background regions. Primary object segmentation in video aims at obtaining the pixel-level result of the primary object, while primary object discovery locates the primary object through a bounding box, as shown in Fig. 11(a-b). Fig. 11(c) provides the flowchart of primary object detection in video, which mainly contains saliency model construction and primary object detection. In most of the existing methods, the primary object is directly defined as the salient object in video [129]–[134]. In fact, there are several differences and connections between them. First, the primary object may not be the most salient one in all frames. Secondly, inter-frame corre-

spondence works as an important temporal cue for highlighting the primary object and suppressing distractors. Thirdly, some commonly used visual priors in saliency detection may no longer be valid in video due to camera and object motion, such as background prior. Last but not least, primary object should be consistently detected from a varying scene, whereas salient object detection only considers the individual and fixed scene. Therefore, it is insufficient to only consider the image saliency attribute. The motion cue and global appearance should be introduced jointly to constrain the initial result generation. In the future, some machine learning techniques, such as deep learning and reinforcement learning, can be incorporated into the model to achieve superior performance.

E. Challenges and Problems

In the last decades, a plenty of saliency detection methods have been proposed to obtain the remarkable progresses and performance improvements. However, there still exist many issues that are not well resolved and needed to be further investigated in the future.

For RGBD saliency detection, how to capture the accurate and effective depth representation to assist in saliency detection is a challenge. Taking the depth information as an additional feature to supplement color feature is an intuitive and explicit way, but it ignores the potential attributes in the depth map, such as shape and contour. By contrast, depth measure based method aims at exploiting these implicit information to refine the saliency result. For example, the depth shape can be used to highlight the salient object and suppress the background, and the depth boundary can be utilized to refine the object boundary and obtain sharper saliency result. In addition, the whole object usually has high consistency in the depth map. Therefore, the depth information can be used

to improve the consistency and smoothness of the acquired saliency map. Generally, depth measure based methods can achieve a better performance. However, how to effectively exploit the depth information to enhance the identification of salient object has not yet reached a consensus. On the whole, combining the explicit and implicit depth information to obtain a more comprehensive depth representation is a meaningful attempt for RGBD saliency detection.

For co-saliency detection, how to explore inter-image correspondence among multiple images to constrain the common properties of salient object is a challenge. Inter-image corresponding relationship plays an essential role in determining the common object from all the salient objects, which can be formulated as a clustering process, a matching process, a propagation process, or a learning process. However, these methods may either be noise-sensitive or time-consuming. The accuracy of corresponding relationship is directly related to the performance of the algorithm. Thus, capturing the accurate inter-image correspondence is an urgent problem to be addressed. At present, there have been some attempts to detect co-salient object using deep learning network. However, these methods often simply cascade the features produced from the single image and re-learn, rather than designing a specific inter-image network to learn the effective inter-image correspondence.

For video saliency detection, how to combine more information and constraints, such as motion cue, inter-frame correspondence, and spatiotemporal consistency, is a challenge. Motion cue plays more important role in discovering the salient object from the clustered and complex scene. The inter-frame correspondence represents the relationship among different frames, which is used to capture the common attribute of salient objects from the whole video. The spatiotemporal consistency constrains the smoothness and homogeneity of salient objects from the spatiotemporal domain. The main contributions of the existing methods are often concentrated in these three aspects. In addition, the video saliency detection algorithm based on deep learning is still immature, and only a few methods have been proposed, which is a relatively underexplored area. However, it is a challenging task to learn the comprehensive features including intra-frame, inter-frame, and motion through a deep network under the limited training samples.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have conducted a survey on visual saliency detection with comprehensive information, including depth cue for RGBD saliency detection, inter-image correspondence for co-saliency detection, and temporal constraint for video saliency detection. We have reviewed the recent progress of saliency detection, analyzed the different types of saliency detection algorithms, and conducted experimental comparisons and analysis. It has been demonstrated that the performance has been improved by introducing the comprehensive information in an appropriate way. For example, the depth measure based method often yields better performance, because of it can further capture effective depth attributes from the depth map. Combining the inter-image correspondence and depth cue, RGBD co-saliency detection models achieve better performance. Through a symmetrical structure (e.g., convolution-deconvolution, encoder-decoder) with continuous multiple frames input, the deep learning based video saliency detection methods learn the high-level spatiotemporal feature and improve the efficiency.

In the future, some research directions and emphases of saliency detection can be focused on, i.e., (1) New attempts in learning based saliency detection methods, such as small samples training, weakly supervised learning, and cross-domain learning. Limited by the labelled training data, more work, such as designing a special network, can be explored in the future to achieve high-precision detection with small training samples. In addition, weakly supervised salient object detection method is a good choice to address the insufficient pixel-level saliency annotations. Furthermore, the cross-domain learning is another direction that needs to be addressed for learning based RGBD saliency detection method. (2) Extending the saliency detection task in different data sources, such as light filed image, RGBD video, and remote sensing image. In the light filed image, the focusness prior, multi-view information, and depth cue should be considered jointly. For the RGBD video data, the depth constraint should be introduced to assist in the spatiotemporal saliency. In the remote sensing image, due to the high angle shot photographed, some small targets and shadows are included. Thus, how to suppress the interference effectively and highlight the salient object accurately should be further investigated in the future.

REFERENCES

[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.
[2] 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, 2015.
[3] 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, 2017.
[4] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Image cosegmentation via saliency-guided constraint clustering with cosine similarity," in *AAAI*, 2017, pp. 4285–4291.
[5] J. Sun, X. Liu, W. Wan, J. Li, D. Zhao, and H. Zhang, "Database saliency for fast image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359–369, 2015.
[6] Y. Gao, M. Shi, D. Tao, and C. Xu, "Video hashing based on appearance and attention features fusion via DBN," *Neurocomputing*, vol. 213, pp. 84–94, 2016.
[7] J. Lei, M. Wu, C. Zhang, F. Wu, N. Ling, and C. Hou, "Depth-preserving stereo image retargeting based on pixel fusion," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1442–1453, 2017.
[8] Y. Fang, K. Zeng, Z. Wang, W. Lin, Z. Fang, and C.-W. Lin, "Objective quality assessment for image retargeting based on structural similarity," *IEEE J. Emerg. Sel. Topic Circuits Syst.*, vol. 4, no. 1, pp. 95–105, 2014.
[9] S. Han and N. Vasconcelos, "Image compression using object-based regions of interest," in *ICIP*, 2006, pp. 3097–3100.
[10] J. Lei, C. Zhang, Y. Fang, Z. Gu, N. Ling, and C. Hou, "Depth sensation enhancement for multiple virtual view rendering," *IEEE Trans. Multimedia*, vol. 17, no. 4, pp. 457–469, 2015.
[11] 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. on Image Process.*, vol. 25, no. 12, pp. 5664–5677, 2016.
[12] J. Lei, L. Li, H. Yue, F. Wu, N. Ling, and C. Hou, "Depth map super-resolution considering view synthesis quality," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1732–1745, 2017.

[13] J. Lei, J. Duan, F. Wu, N. Ling, and C. Hou, "Fast mode decision based on grayscale similarity and inter-view correlation for depth map coding in 3D-HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. 99, pp. 1–13, 2017.

[14] 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, 2016.

[15] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[16] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, and B. Chen, "No-reference quality assessment of deblocked images," *Neurocomputing*, vol. 177, pp. 572–584, 2016.

[17] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Vis. Comput. Graph*, vol. 23, no. 8, pp. 2014–2027, 2017.

[18] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 510–514, 2017.

[19] H. Jacob, F. Padua, A. Lacerda, and A. Pereira, "Video summarization approach based on the emulation of bottom-up mechanisms of visual attention," *J. Intell. Information Syst.*, vol. 49, no. 2, pp. 193–211, 2017.

[20] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.

[21] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 717–729, 2016.

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

[23] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *ICCV*, 2013, pp. 1976–1983.

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

[25] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015, pp. 110–119.

[26] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Feng, "Robust saliency detection via regularized random walks ranking," in *CVPR*, 2015, pp. 2710–2717.

[27] 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, 2015.

[28] 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, 2015.

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

[30] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, 2016.

[31] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 750–762, 2017.

[32] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[33] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.

[34] T. L. abd Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011.

[35] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013, pp. 2083–2090.

[36] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, 2017.

[37] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, 2015.

[38] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 330–344, 2015.

[39] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016, pp. 660–668.

[40] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.

[41] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.

[42] J. Zhang, Y. Dai, and F. Porikli, "Deep salient object detection by integrating multi-level cues," in *WACV*, 2017, pp. 1–10.

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

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

[45] J. Zhang, B. Li, Y. Dai, F. Porikli, and M. He, "Integrated deep and shallow networks for salient object detection," in *ICIP*, 2017, pp. 271–276.

[46] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[47] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015, pp. 5455–5463.

[48] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017, pp. 3796–3806.

[49] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *CVPR*, 2018, pp. 9029–9038.

[50] R. Achanta, S. Hemami, F. Estrada, and S. Ssstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[51] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*, 2012, pp. 101–115.

[52] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. on Cybern.*, vol. 43, no. 5, pp. 1318–1334, 2013.

[53] S. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor system description, issues and solutions," in *CVPRW*, 2004, pp. 35–45.

[54] "Stereo camera," http://en.wikipedia.org/wiki/Stereocamera.

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

[56] J. Lei, H. Zhang, L. You, C. Hou, and L. Wang, "Evaluation and modeling of depth feature incorporated visual attention for salient object segmentation," *Neurocomputing*, vol. 120, pp. 24–33, 2013.

[57] 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, 2014.

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

[59] 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, 2017.

[60] 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, 2017.

[61] 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. on Cybern.*, vol. PP, no. 99, pp. 1–13, 2017.

[62] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *CVPR*, 2018, pp. 3051–3060.

[63] J. Guo, T. Ren, J. Bei, and Y. Zhu, "Salient object detection in RGBD image based on saliency fusion and propagation," in *ICIMCS*, 2015, pp. 1–5.

[64] 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, 2015.

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

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

[67] H. Song, Z. Liu, H. Du, and G. Sun, "Depth-aware saliency detection using discriminative saliency fusion," in *ICASSP*, 2016, pp. 1626–1630.

[68] H. Sheng, X. Liu, and S. Zhang, "Saliency analysis based on depth contrast increased," in *ICASSP*, 2016, pp. 1347–1351.

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

[70] 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. Letters.*, vol. 23, no. 6, pp. 819–823, 2016.

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

[72] H. Li and K. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, 2011.

[73] K. Chang, T. Liu, and S. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *CVPR*, 2011, pp. 2129–2136.

[74] Z. Tan, L. Wan, W. Feng, and C.-M. Pun, "Image co-saliency detection by propagating superpixel affinities," in *ICASSP*, 2013, pp. 2114–2118.

[75] H. Li, F. Meng, and K. Ngan, "Co-salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1869–1909, 2013.

[76] L. Li, Z. Liu, W. Zou, X. Zhang, and O. L. Meur, "Co-saliency detection based on region-level fusion and pixel-level refinement," in *ICME*, 2014, pp. 1–6.

[77] 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, 2014.

[78] 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, 2015.

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

[80] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, 2013.

[81] 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, 2014.

[82] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *ICME*, 2015, pp. 1–6.

[83] 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, 2016.

[84] 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, 2017.

[85] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *CVPR*, 2015, pp. 2994–3002.

[86] L. Wei, S. Zhao, O. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *IJCAI*, 2017, pp. 3041–3047.

[87] D. Zhang, D. Meng, C. Lia, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *ICCV*, 2015, pp. 594–602.

[88] 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. PP, no. 99, pp. 1–11, 2017.

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

[90] 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, 2016.

[91] 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, 2018.

[92] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for RGBD images," *IEEE Trans. on Cybern.*, vol. PP, no. 99, pp. 1–14, 2017.

[93] Z. Ren, S. Gao, D. Rajan, L.-T. Chia, and Y. Huang, "Spatiotemporal saliency detection via sparse representation," in *ICME*, 2012, pp. 158–163.

[94] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in *ICASSP*, 2012, pp. 1485–1488.

[95] Y. Fang, W. Lin, Z. Chen, C. Tsai, and C. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, 2014.

[96] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3910–3921, 2014.

[97] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, 2014.

[98] 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, 2015.

[99] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, 2015.

[100] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *CVPR*, 2015, pp. 3395–3402.

[101] W. Wang, J. Shen, R. Yang, and F. Porikli, "A unified spatiotemporal prior based on geodesic distance for video object segmentation," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, 2018.

[102] T. Xi, W. Zhao, H. Wang, and W. Lin, "Salient object detection with spatiotemporal background priors for video," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3425–3436, 2017.

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

[104] 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. PP, no. 99, pp. 1–17, 2017.

[105] F. Guo, W. Wang, J. Shen, L. Shao, J. Yang, D. Tao, and Y. Y. Tang, "Video saliency detection using object proposals," *IEEE Trans. Cybern.*, vol. PP, no. 99, pp. 1–12, 2017.

[106] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011.

[107] T.-N. Le and A. Sugimoto, "Spatiotemporal utilization of deep features for video saliency detection," in *ICMEW*, 2017, pp. 465–470.

[108] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, 2018.

[109] T.-N. Le and A. Sugimoto, "Deeply supervised 3D recurrent FCN for salient object detection in videos," in *BMVC*, 2017, pp. 1–13.

[110] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349–364, 2018.

[111] H. Fu, D. Xu, S. Lin, and J. Liu, "Robust object co-detection," in *CVPR*, 2013, pp. 3206–3213.

[112] A. Toshev, J. Shi, and K. Daniilidis, "Image matching via saliency region correspondences," in *CVPR*, 2007, pp. 1–8.

[113] "MSRA10K," http://mmcheng.net/gsal/.

[114] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors," in *CVPR*, 2017, pp. 4142–4150.

[115] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *ICCV*, 2005, pp. 1800–1807.

[116] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *CVPR*, 2010, pp. 3169–3176.

[117] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vision*, vol. 120, no. 2, pp. 215–232, 2016.

[118] K. Li, J. Zhang, and W. Tao, "Unsupervised co-segmentation for indefinite number of common foreground objects," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1898–1909, 2016.

[119] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *BMVC*, 2010, pp. 1–11.

[120] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *ICCV*, 2013, pp. 2192–2199.

[121] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016, pp. 724–732.

[122] L. Zhang and K. Yang, "Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 916–920, 2014.

[123] Y. Zhang, L. Zhang, and X. Yu, "Region of interest extraction based on multiscale visual saliency analysis for remote sensing images," *J. Applied Remote Sens.*, vol. 9, pp. 1–15, 2015.

[124] L. Zhang, X. Lv, J. Chen, and L. Zhang, "Region of interest detection based on salient feature clustering for remote sensing images," in *IGARSS*, 2016, pp. 88–91.

[125] L. Ma, B. Du, H. Chen, and N. Q. Soomro, "Region-of-interest detection via superpixel-to-pixel saliency analysis for remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1752–1756, 2016.

[126] L. Zhang, Y. Wang, X. Li, and S. Wang, "Region-of-interest extraction based on spectrum saliency analysis and coherence-enhancing diffusion model in remote sensing images," *Neurocomputing*, vol. 207, pp. 630–640, 2016.

[127] T. Li, J. Zhang, X. Lu, and Y. Zhang, "SDBD: A hierarchical region-of-interest detection approach in large-scale remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 699–703, 2017.

[128] L. Zhang and S. Wang, "Region-of-interest extraction based on local-global contrast analysis and intra-spectrum information distribution estimation for remote sensing images," *Remote Sens.*, vol. 9, no. 6:597, pp. 1–22, 2017.

[129] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *CVPR*, 2013, pp. 628–635.

[130] W.-D. Jang, C. Lee, and C.-S. Kim, "Primary object segmentation in videos via alternate convex optimization of foreground and background distributions," in *CVPR*, 2016, pp. 696–704.

[131] J. Yang, B. Price, X. Shen, Z. Lin, and J. Yuan, "Fast appearance modeling for automatic primary video object segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 503–515, 2016.

[132] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, and J. Brandt, "Discovering primary objects in videos by saliency fusion and iterative appearance estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1070–1083, 2016.

[133] Y. J. Koh, W.-D. Jang, and C.-S. Kim, "POD: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models," in *CVPR*, 2016, pp. 1068–1076.

[134] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From key frames to key objects: Video summarization by representative object proposal selection," in *CVPR*, 2016, pp. 1039–1048.

Runmin Cong received the M.S. degree from the Civil Aviation University of China, Tianjin, China, in 2014. He is currently pursuing his Ph.D. degree in information and communication engineering with Tianjin University, Tianjin, China.

He was a visiting student at Nanyang Technological University (NTU), Singapore, from Dec. 2016 to Feb. 2017. Since May 2018, he has been working as a Research Associate at the Department of Computer Science, City University of Hong Kong (CityU), Hong Kong. He is a Reviewer for the IEEE TIP, TMM, and TCSVT, etc. His research interests include computer vision, image processing, saliency detection, and 3-D imaging.

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

He was a visiting researcher at the Department of Electrical Engineering, University of Washington, Seattle, WA, from August 2012 to August 2013. He is currently a Professor at Tianjin University, Tianjin, China. He is on the editorial boards of Neurocomputing and China Communications. His research interests include 3D video processing, virtual reality, and artificial intelligence.

Huazhu Fu (SM'18) received the Ph.D. degree in computer science from Tianjin University, China, in 2013. From 2013 to 2015, he worked as the research fellow at Nanyang Technological University, Singapore. And from 2015 to 2018, he worked as a Research Scientist at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. He is currently the 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. He is the Associate Editor of IEEE Access and BMC Medical Imaging.

Ming-Ming Cheng received his Ph.D. degree from Tsinghua University in 2012. Then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests includes computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program, etc.
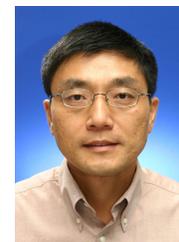
Weisi Lin (M'92-SM'98-F'16) received his Ph.D. from King's College, London University, U.K. He is a Professor in the School of Computer Science and Engineering, Nanyang Technological University. His areas of expertise include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has published 180+ journal papers, 230+ conference papers, filed 7 patents, and authored 2 books. He has been an AE for IEEE Trans. on Image Processing, IEEE Trans. on Circuits and Systems for Video Tech., IEEE Trans. on Multimedia, and IEEE Signal Processing Letters. He has been a Technical Program Chair for IEEE ICME 2013, PCM 2012, QoMEX 2014 and IEEE VCIP 2017. He has been an invited/panelist/keynote/tutorial speaker in 20+ international conferences, as well as a Distinguished Lecturer of IEEE Circuits and Systems Society 2016-2017, and Asia-Pacific Signal and Information Processing Association (APSIPA), 2012-2013. He is a Fellow of IEEE and IET, and an Honorary Fellow of Singapore Institute of Engineering Technologists.

Qingming Huang (SM'08-F'18) is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China.

His research areas include multimedia video analysis, image processing, computer vision and pattern recognition He has published more than 300 academic papers in prestigious international journals including IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on Circuits and Systems for Video Tech., etc, and top-level conferences such as ACM Multimedia, ICCV, CVPR, IJCAI, VLDB, etc. He is the associate editor of IEEE Trans. on Circuits and Systems for Video Tech., and Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. on Multimedia, IEEE Trans. on Circuits and Systems for Video Tech., IEEE Trans. on Image Processing, etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PCM, PSIVT, etc.