RGBT Salient Object Detection: Benchmark and A Novel Cooperative Ranking Approach

Jin Tang, Dongzhe Fan, Xiaoxiao Wang, Zhengzheng Tu, and Chenglong Li

***Abstract*— Despite significant progress, image saliency detection still remains a challenging task in complex scenes and environments. Integrating multiple different but complementary cues, like RGB and Thermal infrared (RGBT), may be an effective way for boosting saliency detection performance. This work contributes a RGBT image dataset, which includes 821 spatially aligned RGBT image pairs and their ground truth annotations for saliency detection purpose. Moreover, 11 challenges are annotated on these image pairs for performing the challenge-sensitive analysis and 3 kinds of baseline methods are implemented to provide a comprehensive comparison platform. With this benchmark, we propose a novel approach based on a cooperative ranking algorithm for RGBT saliency detection. In particular, we introduce a weight for each modality to describe the reliability and a ℓ_1-based cross-modal consistency in a unified ranking model, and design an efficient solver to iteratively optimize several subproblems with closed-form solutions. Extensive experiments against baseline methods demonstrate the effectiveness of the proposed approach on both our introduced dataset and a public dataset.**

***Index Terms*— RGBT image saliency detection, cooperative ranking, cross-modal consistency, reliability weight, joint optimization.**

I. Introduction

IMAGE Saliency Detection aims at highlighting salient foreground objects automatically from background, and has received increasing attentions due to its wide range of applications in computer vision and graphics, such as object recognition, content-aware retargeting, video compression, and image classification. Despite significant progress, image saliency detection still remains a challenging task in complex scenes and environments.

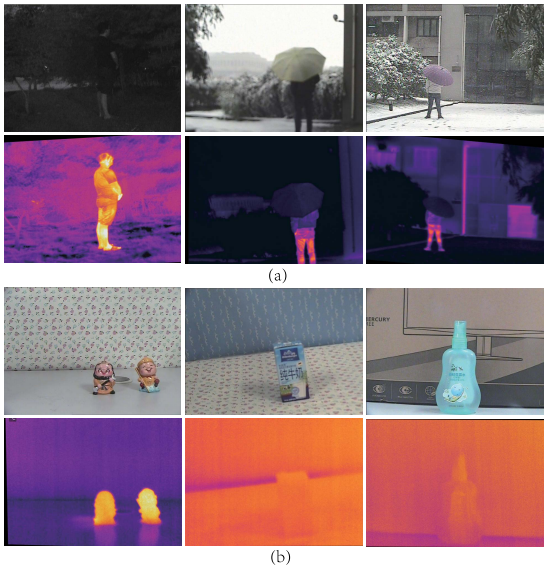


Fig. 1. (a) Benefits of thermal sources over RGB ones, where visible spectrum is disturbed by high illumination, snow weather and low illumination. (b) Benefits of RGB sources over thermal ones, where thermal spectrum is disturbed by thermal crossover and reflection.

One of major reasons is the limitations of visible imaging in some complex scenarios, such as low illumination, background clutters, as well as bad weathers, and thus numerous approaches utilize other modalities to fuse multiple cues for high accurate detection results. For example, depth data can provide dependable depth information to improve the saliency detection results [1], [2]. Although the depth sensors can provide valuable additional depth information to improve saliency results by robust distance awareness, these sensors suffer from the limited range (e.g., 4-5 meters at most) and imaging noises. More importantly, we think that the depth sensors capture the distance information of a scene and thus are hard to provide reliable information that identifies salient objects.

Thermal infrared cameras capture infrared radiation (0.75-$13\mu m$) emitted by objects with a temperature above absolute zero and their imaging procedure is insensitivity to lighting and weather conditions and could also suppress background clutter to some extent. However, resolutions of thermal images are low, and edge details are fuzzy. Also, the imaging quality is easily affected by thermal crossover and reflection. Fig. 1 shows some examples that reflect the

Manuscript received June 28, 2019; revised September 12, 2019; accepted October 23, 2019. Date of publication November 5, 2019; date of current version December 4, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61702002, Grant 61976003, Grant 61872005, and Grant 61602006, in part by the Natural Science Foundation of Anhui Province under Grant 1808085QF187, in part by the Natural Science Foundation of Anhui Higher Education Institutions of China under Grant KJ2018A0023, and in part by the Open fund for Discipline Construction, Institute of Physical Science and Information Technology, Anhui University. This article was recommended by Associate Editor Y. Peng. *(Jin Tang, Dongzhe Fan, and Xiaoxiao Wang contributed equally to this work.) (Corresponding author: Jin Tang.)*

J. Tang, D. Fan, X. Wang, and Z. Tu are with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: tangjin@ahu.edu.cn; fdz.ahu@foxmail.com; xiaoxiao9212@foxmail.com; zhengzhengahu@163.com).

C. Li is with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China, and also with the Institute of Physical Science and Information Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com.)

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

Digital Object Identifier 10.1109/TCSVT.2019.2951621

complementary benefits of RGB and thermal images. In this work, we make the first attempt to introduce the thermal information in the task of image saliency detection, called "RGBT saliency object detection" (RGBT SOD) in this paper.

Focusing on a comprehensive evaluation platform and a cooperative model of RGBT SOD, we address following issues in this paper through existing works.

There are several RGBT datasets for different vision tasks, such as moving object detection [3], [4] and visual tracking [5], [6]. There is no dataset for the task of RGBT SOD, and existing RGBT datasets are designed for a certain vision task and unsuitable to extend them to the field of SOD. Therefore, we contribute a RGBT image dataset with reasonable size, high diversity and low bias for RGBT SOD. In a specific, we collect 821 aligned RGB-T image pairs with ground truths in different scenes and environmental conditions. The category, size, number and spatial information of salient objects are also taken into account for enhancing the diversity and challenge. To analyze the challenge-sensitive performance of different algorithms, we also annotate 11 different challenges. In addition to this, we implement some baseline methods to provide a comparison platform. On one hand, we regard RGB or thermal images as inputs in recent popular methods to achieve single-modality saliency detection. These baselines can be utilized to identify the importance and complementarity of RGB and thermal information with comparison with RGB-T saliency detection methods. On the other hand, we extend some RGB methods to RGBT ones by fusing the results from RGB and thermal data.

For the past years, numerous outstanding saliency detection methods have emerged, where the graph-based ranking approaches have received much attention. For example, Yang *et al.* [7] utilize a manifold ranking algorithm for salient object detection with both background and foreground queries. Ji *et al.* [8] propose a graph model-based bottom-up salient object detection framework by fusing multiple saliency maps using low-level features and objectness features. These graph-based methods achieve good balance of accuracy and efficiency trade-off, but how to extend it for the usage of RGBT SOD is still investigated.

In this paper, we proposed a novel graph-based cooperative ranking approach for RGBT SOD. RGB and thermal sources reflect different properties of a scene. Existing methods [5], [9], [10] often model their collaboration but ignore the heterogeneity, which is crucial for effective fusion of different modalities. To handle this problem, we take both collaboration and heterogeneity into account into the graph-based ranking model, and collaboratively computes the ranking functions of multiple modalities while allowing sparse discordance to model their heterogeneous. To this end, we formulate it as a ℓ_1-based sparse learning problem. Moreover, motivated by recent works [5], [10], we introduce the weight variables in our cooperative ranking model to adaptively integrate different modalities based on their reliabilities.

For the optimization, we design an efficient algorithm based on the alternating direction method of multipliers (ADMM) [11] to optimize the proposed model. Each subproblem has a closed-form solution and the convergence of the proposed algorithm is thus guaranteed. Extensive experiments against baseline methods are conducted to demonstrate the effectiveness of the proposed approach on both our introduced dataset and a public dataset.

The major contributions of this work are summarized as follows.

- We are the first to propose a RGBT image dataset for SOD, and provide a new direction to accelerate the research progress of SOD. The dataset consists of 821 aligned RGB-T image pairs with ground truths in different scenes and environmental conditions, and has been released to public for free academic usage.[1]
- We propose a novel cooperative ranking approach for RGBT SOD. In particular, the proposed ranking model could perform effective fusion of different modalities by taking both collaboration and heterogeneity into account.
- We design an efficient ADMM-based solver to optimize the proposed model. The closed-form solution of each subproblem guarantee the convergence of the designed optimization algorithm.
- Extensive experiments on benchmark datasets suggest that the thermal information plays a critical role in boosting SOD performance and our approach performs favorably against the state-of-the-art methods. We also provide some insights and potential research directions for RGBT SOD.

This paper provides a more complete understanding of the early results [12], with more background, insights, analysis, and evaluation. In particular, our work advances the previous one in several aspects. First, we make our evaluation platform more comprehensive by adding more recent state-of-the-art methods. Second, we employ an effective consistent constraint to fuse different modalities effectively with both collaboration and heterogeneity considerations. Third, we integrate this consistent constraint into our cooperative ranking model and design an efficient algorithm to optimize it. Fourth, we carry out extensive experiments on an extra dataset to demonstrate the effectiveness of the proposed algorithm. Finally, we provide some insights and potential research directions for RGBT SOD through an in-depth analysis of experimental results.

II. Related Work

In past decades, salient object detection has been a tremendous growth, numerous saliency models and algorithms have been proposed on the basis of different principles. It can be categorized to two classes: bottom-up approaches and top-down approaches. The former is data-driven and pre-attentive, while the latter is task-driven and supervised learning problem. Since our work is a bottom-up model, here we mainly review bottom-up methods of salient object detection as follows.

Bottom-up data-driven models take the underlying image features and some priors [13]–[15] into consideration, such as color, orientation, texture, boundary, and contrast. Among those bottom-up approaches, graph-based approaches have

[1]RGB-T Image Saliency Detection Dataset: http://chenglongli.cn/people/lcl/dataset-code.html

obtained significant popularity due to the simplicity and efficiency of graph algorithms. Follow this issue, Jiang *et al.* [16] calculate saliency value via an absorbing Markov chain, and the time taken by each random walk to the image boundary as the saliency value. As a graph based method likewise, Yang *et al.* [7] utilize a manifold ranking algorithms to surveillance salient object, which consider both background and foreground queries to obtain the saliency maps. Zhu *et al.* [17] propose a new salient object detection algorithm via ranking structured trees, it takes the structural representation of rigid grids as receptive field which considers neighboring relationships in both spatial and feature spaces. Closely related to graph-based manifold ranking mentioned above, Wang *et al.* [18] propose an unsupervised bottom-up saliency detection approach by applying graph techniques, they optimize the saliency map by using a robust background measure to obtain more accurate saliency estimations. Ji *et al.* [8] propose a graph model-based bottom-up salient object detection framework by fusing multiple saliency maps using low-level features and objectness features under a manifold ranking framework. Li *et al.* [19] utilize a graph-based fusion method with superpixel-level belief propagation for 3D fixation prediction on stereoscopic video.

In recent years, a good deal of deep learning based methods have emerged in the field of significance detection, by virtue of its deep nonlinear network structure and puissant learning capacity. Li *et al.* [20] propose a FCN network with global input and global output in salient object detection, which takes a data-driven strategy for encoding the underlying saliency prior information. Then, Wang *et al.* [21] develop a new salient object detection method by exploiting recurrent fully convolutional networks which can incorporate saliency prior knowledge and receives better performance. As FCN-based methods, in [22], Zhang *et al.* propose an attention-guided network which can selectively integrate different layers of content information in a step-by-step manner. Furthermore, Hou *et al.* [23] introduced short connections to the skip-layer structure by transforming high-level features to shallower side output layers and thus obtain ideal results.

As a new proposed RGBT method, Tu *et al.* [24] propose an effective approach SDGL for RGBT image saliency detection, which takes superpixels as graph nodes, and collaboratively uses hierarchical deep features to jointly learn graph affinity and node saliency in a unified optimization framework. It should be noted that our model is different from this method in the following aspects. i) SDGL employs multiple graphs constructed by different features to learn a collaborative graph, and then use it to compute saliency results. While we construct a graph for each modality, and then compute saliency results for different modalities. In particular, we integrate the cross-modality consistent constraint matrix into our model both collaboration and heterogeneity of different modalities. ii) The modality weights in SDGL are dependent in the optimization since the sum of all weights is equal to 1. While our model makes them to be optimized independently. iii) Our model has much little computational complexity and is thus more efficient than SDGL.

TABLE I
LIST OF THE ANNOTATED CHALLENGES OF OUR RGB-T DATASET

Challenge	Description
BSO	Big Salient Object - the ratio of ground truth salient objects over image is more than 0.26.
SSO	Small Salient Object - the ratio of ground truth salient objects over image is less than 0.05.
LI	Low Illumination - the environmental illumination is low.
BW	Bad Weather - the image pairs are recorded in bad weathers, such as snowy, rainy, hazy and cloudy.
MSO	Multiple Salient Objects - the number of the salient objects in the image is more than 1.
CB	Center Bias - the centers of salient objects are far away from the image center.
CIB	Cross Image Boundary - the salient objects cross the image boundaries.
SA	Similar Appearance - the salient objets have similar color or shape to the background.
TC	Thermal Crossover - the salient objects have similar temperature to the background.
IC	Image Clutter - the image is cluttered.
OF	Out of Focus - the image is out-of-focus.

III. RGBT IMAGE SALIENCY BENCHMARK

In this section, we introduce our newly created RGBT saliency benchmark, which includes a dataset with statistic analysis, kinds of baseline methods with different inputs and four evaluation metrics.

A. Dataset

Our RGBT image dataset is collected under different scenarios and environmental conditions. The indoor scenes include offices, apartments, library, *etc*, while outdoor locations contain roads, lawn, corridor, streets, buildings, *etc*.

1) Hardware Setup: Our imaging hardware consists of an online thermal imager (MAG32) and a CCD camera (SONY TD-2073). We collect 821 RGBT image pairs with the resolution of 480×640 by our recording system, and thus call it VT821 in this paper.

2) Alignment: Different from industry registration in RGBD sensors, we manually construct the recording system, and develop an annotation tool to align RGBT images in following way. We uniformly select a number of point correspondences in each image pairs, and compute the homography matrix by the least-square algorithm. Herein, we manually choose these points for accurate matching since these two modalities are heterogeneous and the points extracted from existing keypoint detection methods are thus hard to match for different modalities. Then, the image pairs can be aligned by applying the computed homography matrix. It's worth noting that this registration method can accurately align image pairs due to the following three reasons: i) we carefully choose the planar and non-planar scenes to make the homography assumption effective. ii) since two camera views are almost coincident as we made, the transformation between two views is simple. As each image pair is aligned, we annotate the pixel-level ground truth using more reliable modality. Fig. 2 shows some sample image pairs and their ground truths.

3) Annotation: We capture more than 1500 natural RGBT image pairs, and manually choose 1000 image pairs. We struck

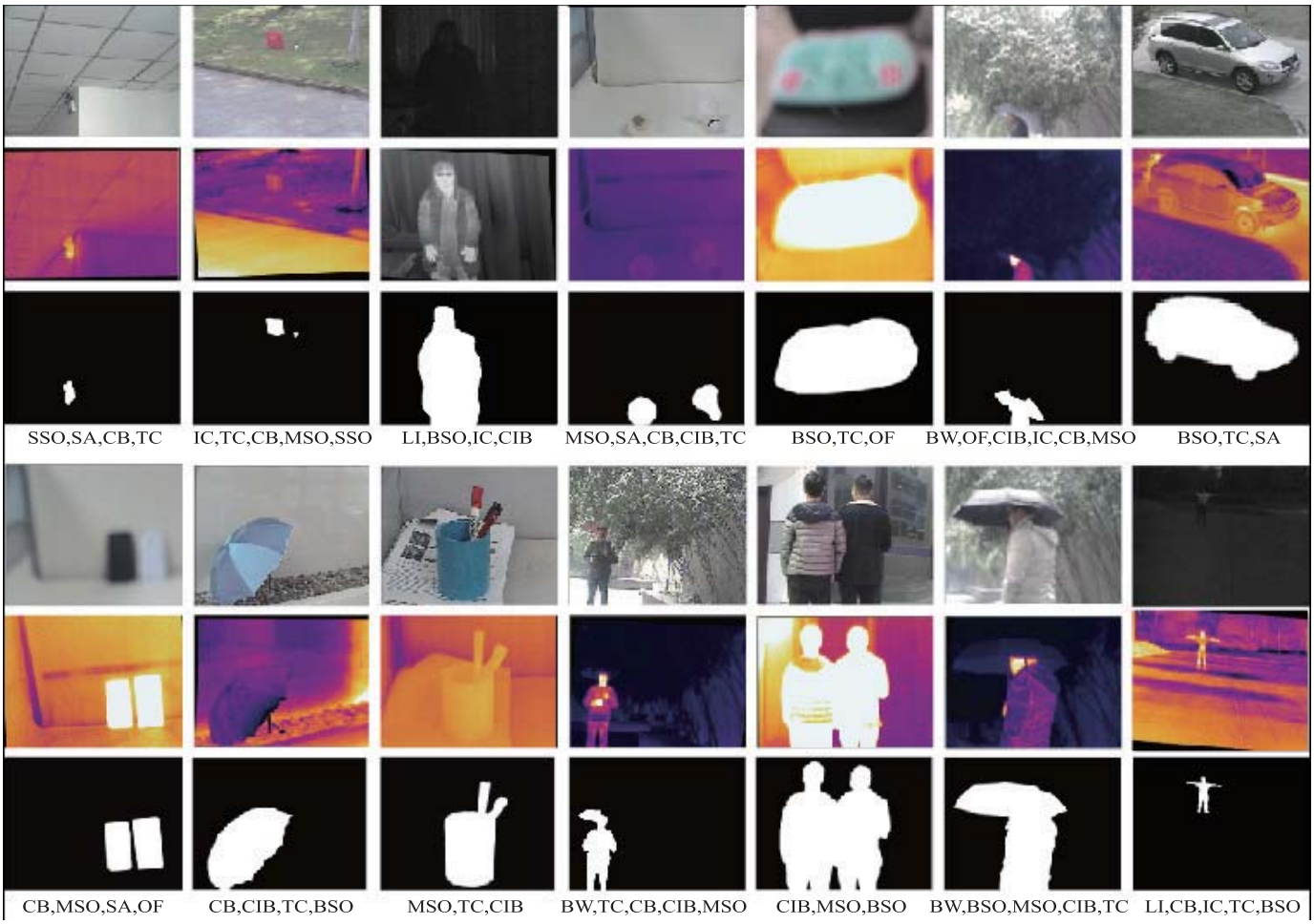


Fig. 2. Sample image pairs with annotated ground truths and challenges from our RGB-T dataset.

down those images with low labeling consistency and choose top 821 image pairs. Eventually, three volunteers utilize the software of Adobe Photoshop to cut out the RGB images which are totally overlapped with thermal images, and then segment the salient object manually from each image to obtain the pixel-level ground truth masks.

4) Statistics: The image pairs in our dataset are recorded in approximately 60 scenes with different environmental conditions, and the category, size, number and spatial information of salient objects are also taken into account for enhancing the diversity and challenge. Specifically, the following main aspects are considered in creating the RGB-T image dataset.

- *Illumination condition*. The image pairs are captured under different light conditions, such as sunny, snowy, and nighttime. The low illumination and illumination variation caused by different light conditions usually bring big challenges in RGB images.
- *Background factor*. Two background factors are taken into account for our dataset. First, similar background to the salient objects in appearance or temperature will introduce ambiguous information. Second, it is difficult to separate objects accurately from cluttered background.
- *Thermal crossover*. When the targets have similar temperature with other objects or background, thermal crossover will occur in thermal images. In such circumstance, thermal information will be ambiguous in salient objects.
- *Salient object attribute*. We take different attributes of salient objects, including category (more than 60 categories), size and number, into account in constructing our dataset for high diversity.
- *Object location*. Most of methods employ the spatial information (center and boundaries of an image) of the salient objects as priors, which is verified to be effective. However, some salient objects are not at center or cross image boundaries, and these situations isolate the spatial priors. We incorporate these factors into our dataset construction to bring its challenge.

Considering the above-mentioned factors, we annotate 11 challenges for our dataset to facilitate the challenge-sensitive performance of different algorithms. They are: big salient object (BSO), small salient object (SSO), multiple salient object (MSO), low illumination (LI), bad weather (BW), center bias (CB), cross image boundary (CIB), similar appearance (SA), thermal crossover (TC), image clutter (IC), and out of focus (OF). In particular, Table I shows the details.

B. Baseline Methods

For comprehensively demonstrating the effectiveness of our methods, we compare the results of our methods with

15 popular methods, including MR [7], CA [25], RRWR [26], RBD [27], Markov [16], MST [28], BFS [29], BL [30], DSS [23], MDF [31], MSS [32], MILPS [33], FCNN [20], SDGL [24] and MTMR [12], where MDF, FCNN, DSS and SDGL are deep learning approaches. On one hand, we regard RGB or thermal images as inputs in recent popular methods to achieve single-modality saliency detection. These baselines can be utilized to identify the importance and complementarity of RGB and thermal information with the comparison with RGB-T saliency detection methods. On the other hand, we extend some RGB methods to RGBT ones by fusing the results from RGB and thermal data. These baseline methods are applied to salient object detection with RGB or thermal input, which can indicate the benefits of fusion of RGB and thermal modalities. Note that SDGL and MTMR are RGBT-based SOD methods, where MTMR is the previous version of the proposed approach.

C. Evaluation Metrics

There are several metrics to evaluate the results of salient object detection. In this work, we utilize PR curves and $F_{0.3}$ metric and Mean Absolute Error (MAE) to evaluate all the algorithms. Given the binarized saliency map via the threshold value within [0, 255], precision means the ratio of the correctly assigned salient pixel number in relation to all the detected salient pixel number, and recall means the ratio of the correct salient pixel number in relation to the ground truth number. The F-measure (F) is defined as $F_{\beta^2} = \frac{(1+\beta^2)\times P\times R}{\beta^2\times P+R}$, where β^2 is set to 0.3 to emphasize the precision [34]. PR curves and $F_{0.3}$ metric are aimed at quantitative comparison, while MAE are better than them for taking visual comparison into consideration to estimate dissimilarity between a saliency map S and the ground truth G, which is defined as $MAE = \frac{1}{w\times h}\sum_{i=1}^{w}\sum_{j=1}^{h}|S(i,j)-G(i,j)|$, where w and h denote the width and height of an image, respectively.

IV. Cross-Modal Cooperative Ranking Algorithm

The graph-based ranking problem is described as follows: Given a graph and a node in this graph as the query, the remaining nodes are ranked based on their affinities to the given query. The goal is to learn a ranking function that defines the relevance between unlabeled nodes and queries. This section will introduce the proposed cooperative ranking model and the associated optimization algorithm. The optimized modality weights and ranking scores will be utilized for RGBT SOD in next section.

A. Graph Construction

Given a pair of RGB-T images, we regard the thermal image as one of image channels, and then employ SLIC algorithm [35] to generate n non-overlapping superpixels. We take these superpixels as nodes to construct a graph $G = (V, E)$, where V is a node set and E is a set of undirected edges. In this work, any two nodes in V are connected if they follow one of the conditions in [7] for capturing local smoothness cues and reducing the geodesic distance of similar superpixels. It is worth noting that we can explore more cues in RGB and thermal data to construct an adaptive graph that makes best use of intrinsic relationship among superpixels. In a specific, if nodes V_i and V_j is connected, we assign it with an edge weight as: $\mathbf{W}_{ij}^k = e^{-\gamma^k||\mathbf{c}_i^k-\mathbf{c}_j^k||}$, where $\mathbf{c}_i^k$ denotes the mean of the i-th superpixel in the k-th modality, and γ^k is the scaling parameter for k-th modality which controls the edge strength between two superpixel nodes.

B. Model Formulation

We first review the algorithm of graph-based manifold ranking that exploits the intrinsic manifold structure of data for graph labeling [36]. Given a superpixel feature set $X = \{\mathbf{x}_1^k, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{d\times n}$, some superpixels are labeled as queries and the rest need to be ranked according to their affinities to the queries. Let $\mathbf{s} : X \to \mathbb{R}^n$ denote a ranking function that assigns a ranking value $\mathbf{s}_i$ to each superpixel $\mathbf{x}_i$, and $\mathbf{s}$ can be viewed as a vector $\mathbf{s} = [s_1, \ldots, s_n]^T$. In this work, we regard the query labels as initial superpixel weights, and $\mathbf{s}$ is thus a superpixel weight vector. Let $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$ denotes an indication vector, in which $\mathbf{y}_i = 1$ if $\mathbf{x}_i$ is a query, and $\mathbf{y}_i = 0$ otherwise. Given G, the optimal ranking of queries are computed by solving the following problem:

$$\min_{\mathbf{s}} \frac{1}{2}\Big(\sum_{i,j=1}^{n}\mathbf{W}_{ij}\Big|\Big|\frac{s_i}{\sqrt{\mathbf{D}_{ii}}}-\frac{s_j}{\sqrt{\mathbf{D}_{jj}}}\Big|\Big|^2+\mu||\mathbf{s}-\mathbf{y}||^2\Big), \quad (1)$$

where $\mathbf{D} = diag\{\mathbf{D}_{11}, \ldots, \mathbf{D}_{nn}\}$ is the degree matrix, and $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. $diag$ indicates the diagonal operation. μ is a parameter to balance the smoothness term and the fitting term. Then, we apply the manifold ranking on multiple modalities. Each modality can be viewed as a individual task, and multi-task ranking model is thus as follows:

$$\min_{\mathbf{s}^k} \frac{1}{2}\Big(\sum_{i,j=1}^{n}\mathbf{W}_{ij}^k\Big|\Big|\frac{s_i^k}{\sqrt{\mathbf{D}_{ii}^k}}-\frac{s_j^k}{\sqrt{\mathbf{D}_{jj}^k}}\Big|\Big|^2+\mu||\mathbf{s}^k-\mathbf{y}||^2\Big), \quad k = 1, 2, \ldots, K. \quad (2)$$

From Eq. (2), we can see that it inherently indicates that available modalities are independent and contribute equally. This may significantly limit the performance in dealing with occasional perturbation or malfunction of individual sources. Li *et al.* [5] pursue a collaborative sparse representation for adaptive object tracking by introducing the weight variables to represent modality reliabilities. Similar to their approach, we also introduce modality weights into our model to achieve adaptive fusion of different modalities, and the effectiveness of modality weights are shown in Fig. 3. Therefore, the formulation of the cooperative ranking model is proposed as follows:

$$\min_{\mathbf{s}^k,\mathbf{r}^k} \frac{1}{2}\sum_{k=1}^{K}((r^k)^2\sum_{i,j=1}^{n}\mathbf{W}_{ij}^k\Big|\Big|\frac{s_i^k}{\sqrt{\mathbf{D}_{ii}^k}}-\frac{s_j^k}{\sqrt{\mathbf{D}_{jj}^k}}\Big|\Big|^2) + \mu\sum_{k=1}^{K}||\mathbf{s}^k-\mathbf{y}||^2+||\Gamma\circ(\mathbf{1}-\mathbf{r})||^2, \quad (3)$$

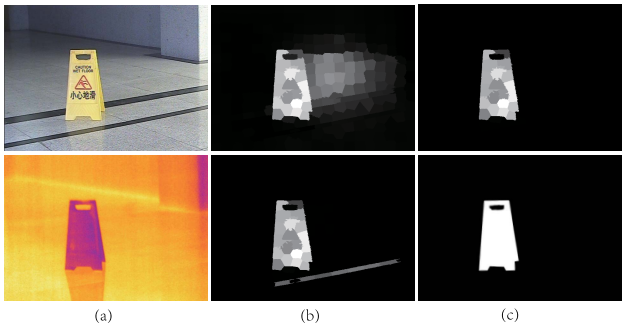


Fig. 3. Illustration of the effectiveness of the introduced modality weights and the cross-modal consistency. (a) Input RGB and thermal images. (b) Results of our method without modality weights and cross-modal consistency are shown in the first and second rows, respectively. (c) Our results and the corresponding ground truth.

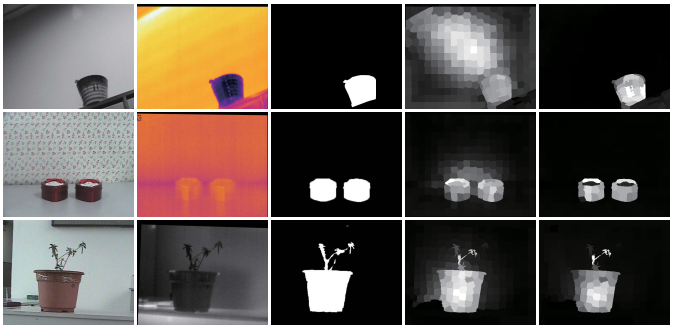


Fig. 4. Advantages of our ranking model over the previous version MTMR [12] without the heterogeneity considerations. (a)-(c) Input RGB and thermal images and their ground truth. (d) Results of MTMR. (e) The results of our approach.

where $\Gamma = [\Gamma^1, \ldots, \Gamma^K]^T$ is an adaptive parameter vector, which is initialized after the first iteration (see Algorithm 1), and $\mathbf{r} = [r^1, \ldots, r^K]^T$ is the modality weight vector. $\circ$ denotes the element-wise product, and λ is a balance parameter. The third term is to avoid overfitting of $\mathbf{r}$, and the last term is the cross-modality consistent constraint.

To take advantage of the complementary benefits of RGB and thermal data, we need to impose the modality consistency on the ranking process. Note that RGB and thermal sources reflect different properties of a scene. Existing methods [5], [9] often model their collaboration but ignore the heterogeneity, which is crucial for effective fusion of different modalities. To handle this problem, we take both collaboration and heterogeneity into account into the graph-based ranking model, and collaboratively computes the ranking functions of multiple modalities while allowing sparse discordance to model their heterogeneous. Fig. 4 shows the effectiveness of the introduced consistent constraint. To this end, we formulate it as a ℓ_1-based sparse learning problem as follows:

$$\min_{\mathbf{s}^k} \lambda \sum_{k=1}^{K} ||\mathbf{s}^k - \mathbf{s}^{k-1}||_1 = \min_{\mathbf{s}^k} \lambda ||\mathbf{CS}||_1, \tag{4}$$

To integrate Eq. (4) into Eq. (3), the final cooperating ranking model can be formulated as:

$$\min_{\mathbf{s}^k, \mathbf{r}^k} \frac{1}{2} \sum_{k=1}^{K} ((r^k)^2 \sum_{i,j=1}^{n} \mathbf{W}_{ij}^k || \frac{s_i^k}{\sqrt{\mathbf{D}_{ii}^k}} - \frac{s_j^k}{\sqrt{\mathbf{D}_{jj}^k}} ||^2) + \mu ||\mathbf{S} - \mathbf{Y}||^2 + ||\Gamma \circ (\mathbf{1} - \mathbf{r})||^2 + \lambda ||\mathbf{CS}||_1, \tag{5}$$

where $\mathbf{S} = [\mathbf{s}^1; \mathbf{s}^2; \ldots; \mathbf{s}^K] \in \mathbb{R}^{nK \times 1}$, $\mathbf{Y} = [\mathbf{y}^1; \mathbf{y}^2; \ldots; \mathbf{y}^K] \in \mathbb{R}^{nK \times 1}$, and $\mathbf{C} \in \mathbb{R}^{n(K-1) \times nK}$ denotes the cross-modality consistent constraint matrix, which is defined as:

$$\begin{bmatrix} \mathbf{I}^{2,1} & -\mathbf{I}^2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^{3,2} & -\mathbf{I}^3 & \cdots & \mathbf{0} & \mathbf{0} \\ & \cdots & & & \cdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}^{K,K-1} & -\mathbf{I}^K \end{bmatrix}, \tag{6}$$

where $\mathbf{I}^k$ and $\mathbf{I}^{k,k-1}$ are the identity matrices with the size of $n \times n$.

C. Optimization Algorithm

Although the variables of (5) are not joint convex, the subproblem of each variable with others fixed is convex and has a closed form solution. The ADMM (alternating direction method of multipliers) algorithm [11] is efficient and effective solver of the problems like (5). To apply ADMM to our problem, we introduce two auxiliary variables $\mathbf{P} = \mathbf{CS}$ to make (5) separable. With some algebra, we have

$$\min_{\mathbf{s}^k, \mathbf{r}^k} (r^k)^2 (\mathbf{s}^k)^T \mathbf{A}^k \mathbf{s}^k + \mu ||\mathbf{S} - \mathbf{Y}||^2 + ||\Gamma \circ (\mathbf{1} - \mathbf{r})||^2 + \lambda ||\mathbf{P}||_1, s.t. \quad \mathbf{P} = \mathbf{CS}. \tag{7}$$

where $\mathbf{A}^k$ is a block-diagonal matrix defined as $\mathbf{A} = diag\{\mathbf{A}^1, \mathbf{A}^2, \ldots, \mathbf{A}^K\} \in \mathbb{R}^{nK \times nK}$, and $\mathbf{A}^k = \mathbf{I} - (\mathbf{D}^k)^{-\frac{1}{2}} \mathbf{W}^k (\mathbf{D}^k)^{-\frac{1}{2}}$. is the normalized Laplacian matrix of m-th modality. $\mathbf{R} = [r^1; \ldots; r^1; r^2; \ldots; r^2; \ldots; r^K; \ldots; r^K] \in \mathbb{R}^{nK \times 1}$, where $\mathbf{I}$ is an identity matrix with the size of $nK \times nK$. The augmented Lagrange function of (7) is:

$$\begin{aligned} &J(\mathbf{s}^k, r^k, \mathbf{P}, \mathbf{Y}_1, \mu_1) \\ &\quad \min_{\mathbf{s}^k, \mathbf{P}_k, \mathbf{r}^k} (r^k)^2 (\mathbf{s}^k)^T \mathbf{A}^k \mathbf{s}^k + \mu ||\mathbf{S} - \mathbf{Y}||^2 + ||\Gamma \circ (\mathbf{1} - \mathbf{r})||^2 \\ &\quad + \lambda ||\mathbf{P}||_1 + < \mathbf{Y}_1, \mathbf{P} - \mathbf{CS} > + \frac{\mu_1}{2} \|\mathbf{P} - \mathbf{CS}\|_F^2 \\ &= \min_{\mathbf{s}^k, \mathbf{r}^k} (r^k)^2 (\mathbf{s}^k)^T \mathbf{A}^k \mathbf{s}^k + \mu ||\mathbf{S} - \mathbf{Y}||^2 + ||\Gamma \circ (\mathbf{1} - \mathbf{r})||^2 \\ &\quad + \lambda ||\mathbf{P}||_1 + \frac{\mu_1}{2} \|\mathbf{P} - \mathbf{CS} + \frac{\mathbf{Y}_1}{\mu_1}\|_F^2 - \frac{1}{2\mu_1} \|\mathbf{Y}_1\|_F^2. \end{aligned} \tag{8}$$

We then alternatively update one variable by minimizing (8) with fixing other variables. Besides the Lagrangian multipliers, there are two variables, including $\mathbf{S}$, $\mathbf{P}$, to solve. The solutions of these subproblems are as follows.

Solving $\mathbf{P}$: $\mathbf{P}$ can be solved by the soft-thresholding method with closed-form solution:

$$\min_{\mathbf{P}} \lambda \|\mathbf{P}\|_1 + \frac{\mu_1}{2} \|\mathbf{P} - \mathbf{CS} + \frac{\mathbf{Y}_1}{\mu_1}\|_F^2 \Rightarrow \mathbf{P} = soft\_thr(\mathbf{CS} - \frac{\mathbf{Y}_1}{\mu_1}, \frac{\lambda}{\mu_1}) \tag{9}$$

Algorithm 1 Optimization Procedure to Eq. (8)

Input: The matrix $\mathbf{A}^k = \mathbf{I} - (\mathbf{D}^k)^{-\frac{1}{2}}\mathbf{W}^k(\mathbf{D}^k)^{-\frac{1}{2}}$, the indication vector $\mathbf{Y}$, and the parameters μ and λ;
Set $r^k = \frac{1}{K}$; $\varepsilon = 10^{-4}$, $maxIter = 50$.
Output: $\mathbf{S}$, $\mathbf{r}$, $\mathbf{P}$.

```
for t = 1 : maxIter do
  Update P_k by Eq. (9);
  Update S_k by Eq. (12);
  Update r_k by Eq. (14);
  if Convergence condition is true then
    Terminate the loop.
  end if
end for
```

Solving $\mathbf{S}$: Given $\mathbf{r}$, Eq. (7) can be written as:

$$J(\mathbf{S}) = \frac{1}{2}\sum_{k=1}^{K}((r^k)^2 \sum_{i,j=1}^{n} \mathbf{W}_{ij}^k || \frac{s_i^k}{\sqrt{\mathbf{D}_{ii}^k}} - \frac{s_j^k}{\sqrt{\mathbf{D}_{jj}^k}} ||^2) + \mu ||\mathbf{S} - \mathbf{Y}||^2, \quad (10)$$

and we reformulate it as follows:

$$J(\mathbf{S}) = (\mathbf{R} \circ \mathbf{S})^T \mathbf{A}(\mathbf{R} \circ \mathbf{S}) + \mu ||\mathbf{S} - \mathbf{Y}||^2, \quad (11)$$

Taking the derivative of $J(\mathbf{S})$ with respect to $\mathbf{S}$, we have

$$\mathbf{S} = \left(\frac{\mathbf{R}\mathbf{R}^T \circ \mathbf{A}}{\mu} + \mathbf{I}\right)^{-1} \mathbf{Y} \quad (12)$$

Solving $\mathbf{r}$: Given $\mathbf{S}$, Eq. (7) can be written as:

$$J(r^k) = (r^k)^2 (\mathbf{s}^k)^T \mathbf{A}^k \mathbf{s}^k + ||\Gamma^k \circ (1 - r^k)||_F^2, \quad (13)$$

and we take the derivative of $J(r^k)$ with respect to r^k, and obtain

$$r^k = \frac{1}{1 + \frac{(\mathbf{s}^k)^T \mathbf{A}^k \mathbf{s}^k}{(\Gamma^k)^2}}, \quad k = 1, 2, \ldots, K. \quad (14)$$

Algorithm 1 summarizes the optimization procedures. Although the whole problem is non-convex, each subproblem is convex. Therefore, the solution by the proposed algorithm satisfies the Nash equilibrium conditions and the convergence of the proposed algorithm is thus guaranteed [37].

V. Two-Stage RGB-T Saliency Detection

In this section, we present the two-stage ranking scheme for RGBT SOD using the proposed algorithm with boundary priors and foreground queries.

Given an pair of input RGBT images represented as a graph and some salient query nodes, the saliency of each node is defined as its ranking score computed by Eq. (5). In the conventional ranking problems, the queries are manually labeled with the ground truth annotations. In this work, to pop out the salient regions of the RGBT frame, we first employ the boundary prior widely used in other works [7], [38] and screen out its highly confident superpixels (low ranking scores in all modalities) belonging to the conspicuous objects as the foreground queries. Then, we perform the proposed algorithm to produce the final saliency map with the corresponding modality weights.

A. The First Ranking Stage

Based on the attention theories for visual saliency [39], we first combine four borders' superpixels together to generate a region and regard all the nodes of the region as background seeds. Once the queries are obtained, we can label the corresponding superpixels to be background priors. Thus, we can exploit the priors and proposed model to rank the relevances of all other superpixels in the first stage.

Taking the right image boundary as an example, we utilize the superpixels on this side as labelled queries and the rest as the unlabeled data, and initilize the indicator $\mathbf{y}$ in (3). Given $\mathbf{y}$, the ranking values $\{\mathbf{s}_r^k\}$ are computed by employing the proposed ranking algorithm, and we normalize $\{\mathbf{s}_r^k\}$ as $\{\hat{\mathbf{s}}_r^k\}$ with the range between 0 and 1. Similarly, given the top, left and bottom image boundaries, we can obtain the respective ranking values $\{\hat{\mathbf{s}}_t^k\}$, $\{\hat{\mathbf{s}}_l^k\}$, $\{\hat{\mathbf{s}}_b^k\}$. We integrate them to compute the initial saliency map for each modality in the first stage:

$$\mathbf{s}_f^k = (\mathbf{1} - \hat{\mathbf{s}}_b^k) \circ (\mathbf{1} - \hat{\mathbf{s}}_t^k) \circ (\mathbf{1} - \hat{\mathbf{s}}_l^k) \circ (\mathbf{1} - \hat{\mathbf{s}}_r^k), \quad k = 1, 2, \ldots, K. \quad (15)$$

Thus, the final saliency map in the first stage is computed as $\bar{\mathbf{s}}_1 = \sum_{k=1}^{K}(r^k \mathbf{s}_f^k)$.

B. The Second Ranking Stage

Given $\mathbf{s}_f^1$ for RGB modality and $\bar{\mathbf{s}}_1$, we dug diverse modality properties and subtly set an adaptive threshold $T_1 = max(\mathbf{s}_f^1) - \beta_1$, $T_2 = max(\bar{\mathbf{s}}_1) - \beta_2$ to generate the foreground queries, where $max(\cdot)$ indicates the maximum operation, and β_k is a constant of the k-th modality. In a specific, we select the i-th superpixel as the foreground query of the k-th modality if $\mathbf{s}_{f,i}^k > T_k$. Therefore, we compute the ranking values $\mathbf{s}_s^k$ and the modality weights $\mathbf{r}$ in the second stage by employing our ranking algorithm. Similar to the first stage, we normalize the ranking value $\mathbf{s}_{ss}^k$ of the k-th modality as $\hat{\mathbf{s}}_{ss}^k$ with the range between 0 and 1. Eventually, the final saliency map $\bar{\mathbf{s}}_2$ can be obtained by combining the ranking values with the modality weights $\bar{\mathbf{s}}_2 = \sum_{k=1}^{K}(r^k \hat{\mathbf{s}}_{ss}^k)$.

VI. Experiments

In this section, we evaluate our method on two datasets, VT821 (created by us) and VT1000 [24]. The experiments are carried out on a PC with an Intel i7 4.0GHz CPU and 32GB RAM, and implemented in the mixture platform of C++ and MATLAB. The proposed algorithm costs about 1.41 second per image pair.

A. Parameter Settings

For fair comparisons, we fix all parameters and other settings of our approach in the experiments, and use the default parameters released in their public codes for other baseline methods. In graph construction, we empirically generate $n = 300$ superpixels and set the affinity parameters $\gamma^1 = 29$ and $\gamma^2 = 23$. In Algorithm 1 and two-stage scheme, we empirically set $\lambda = 0.02$, $\mu = 0.02$ and $\beta_1 = 0.04$(the first stage), $\mu_1 = 0.02$ and $\beta_2 = 0.189$ (the second stage). Herein, we use bigger balance weight in the second stage than in the

TABLE II

AVERAGE PRECISION, RECALL, AND F-MEASURE OF OUR METHOD AGAINST DIFFERENT KINDS OF BASELINE METHODS ON THE VT821 DATASET. WHERE THE BEST,THE SECOND AND THE THIRD BEST RESULTS ARE IN RED, GREEN AND BLUE COLORS, RESPECTIVELY

Algorithm	RGB			Thermal			RGB-T		
	P	R	F	P	R	F	P	R	F
MR [7]	0.643	0.602	0.586	0.699	0.573	0.602	0.732	0.653	0.664
CA [25]	0.592	0.668	0.568	0.625	0.612	0.576	0.644	0.666	0.608
RRWR [26]	0.640	0.608	0.587	0.688	0.580	0.595	0.695	0.616	0.627
RBD [27]	0.610	0.738	0.601	0.549	0.782	0.554	0.611	0.804	0.621
BL [30]	0.623	0.699	0.599	0.664	0.587	0.583	0.596	0.623	0.556
DSS [23]	0.740	0.727	0.693	0.462	0.24	0.307	0.710	0.673	0.640
MDF [31]	0.692	0.699	0.654	0.631	0.585	0.549	-	-	-
MSS [32]	0.587	0.742	0.575	0.559	0.676	0.548	0.589	0.722	0.579
Markov [16]	0.598	0.640	0.569	0.624	0.553	0.547	-	-	-
BFS [29]	0.560	0.648	0.541	0.542	0.614	0.519	-	-	-
MST [28]	0.623	0.737	0.607	0.662	0.654	0.595	-	-	-
MILPS [33]	0.637	0.691	0.612	0.643	0.680	0.612	0.664	0.753	0.644
FCNN [20]	0.636	0.806	0.642	0.627	0.711	0.615	0.647	0.820	0.653
SDGL [24]	-	-	-	-	-	-	0.794	0.724	0.744
MTMR [12]	-	-	-	-	-	-	0.716	0.713	0.680
Ours	-	-	-	-	-	-	0.819	0.661	0.739

TABLE III

AVERAGE PRECISION, RECALL, AND F-MEASURE OF OUR METHOD AGAINST DIFFERENT KINDS OF BASELINE METHODS ON THE VT1000 DATASET. WHERE THE BEST,THE SECOND AND THE THIRD BEST RESULTS ARE IN RED, GREEN AND BLUE COLORS, RESPECTIVELY

Algorithm	RGB			Thermal			RGB-T		
	P	R	F	P	R	F	P	R	F
MR [7]	0.766	0.588	0.635	0.706	0.555	0.586	0.783	0.630	0.667
CA [25]	0.718	0.644	0.621	0.676	0.598	0.581	0.701	0.637	0.610
RRWR [26]	0.766	0.594	0.637	0.703	0.557	0.596	0.730	0.592	0.616
RBD [27]	0.717	0.680	0.628	0.649	0.677	0.576	0.718	0.745	0.650
BL [30]	0.732	0.643	0.620	0.717	0.554	0.564	0.714	0.607	0.596
DSS [23]	0.877	0.676	0.721	0.660	0.357	0.416	0.808	0.594	0.634
MDF [31]	0.804	0.633	0.656	0.675	0.612	0.655	-	-	-
MSS [32]	0.695	0.666	0.601	0.673	0.659	0.588	0.703	0.685	0.615
Markov [16]	0.756	0.616	0.639	0.681	0.572	0.579	-	-	-
BFS [29]	0.686	0.590	0.577	0.654	0.543	0.545	-	-	-
MST [28]	0.733	0.667	0.640	0.671	0.601	0.535	-	-	-
MILPS [33]	0.769	0.664	0.663	0.714	0.608	0.610	0.762	0.686	0.661
FCNN [20]	0.771	0.746	0.689	0.688	0.635	0.590	0.750	0.739	0.671
SDGL [24]	-	-	-	-	-	-	0.853	0.649	0.727
MTMR [12]	-	-	-	-	-	-	0.792	0.628	0.673
Ours	-	-	-	-	-	-	0.859	0.576	0.693

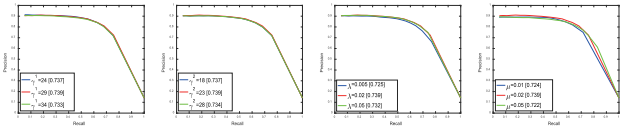


Fig. 5. Precision-recall curves on the VT821 dataset by the proposed algorithm with different parameter values. The representative scores of F-measure are presented in the legend.

first stage as the refined foreground queries are more confident than the background queries according to the boundary prior. In addition, we demonstrate the parameter insensitivity in Table 5. From the results we can see that our algorithm is insensitive to the parameter variations, demonstrating the robustness of the proposed algorithm.

B. Evaluation on VT821

1) Overall Performance: In the first place, we compare our proposed approach against other baseline methods mentioned above on the entire dataset, as shown in Fig. 6 and Table II. Obviously, we can observe that our approach obtains remarkable results in three evaluation metrics. Furthermore, it is worth noting that our approach achieves a significant improvement over our previous work, MTMR [12], about 6% in F-measure over it. Meanwhile, comparing with other traditional methods, the preferable evaluation results further validate the effectiveness of the proposed approach, the importance of thermal information and the complementary benefits of RGBT data. Notice that the latest RGBT method SDGL [24] gains 0.5% higher than our approach, because they extract deep learning features from different layers and propose a more complex model for RGBT SOD. We only utilize the color features and propose a simple model. In spite of the simplicity, our approach is not that far behind of SDGL and much faster than it.

Fig. 6 and Table II also show that F-measure of ours is better than the results of other methods with RGB information only, which demonstrates the importance of thermal information and the complementary benefits of RGBT data. In particular, our approach outperforms the top three RGB methods including DSS, MDF, and FCNN with 4.6%, 8.5% and 9.7%, respectively. It proves the effectiveness of our approach for adaptively incorporating thermal information. Furthermore, for the sake of fairness, we compare our method with several

TABLE IV

F-MEASURE BASED ON ATTRIBUTE OF THE PROPOSED DATASET WITH 11 METHODS ON THE VT821 DATASET, INCLUDING RBD [27], CA [25], RRWR [26], MR [7], MSS [32], BL [30], MTMR [12] AND DSS [23], MILPS [33], FCNN [20], SDGL [24]. WHERE THE BEST, THE SECOND AND THE THIRD BEST RESULTS ARE IN RED, GREEN AND BLUE COLORS, RESPECTIVELY

	RBD	CA	RRWR	MR	MSS	BL	MTMR	DSS	MILPS	FCNN	SDGL	Ours
BSO	0.843	0.809	0.756	0.797	0.702	0.753	0.795	0.592	0.772	0.766	0.817	0.771
BW	0.518	0.439	0.431	0.493	0.521	0.430	0.495	0.621	0.503	0.696	0.596	0.558
CB	0.691	0.695	0.712	0.731	0.659	0.634	0.659	0.665	0.729	0.721	0.789	0.782
CIB	0.692	0.597	0.581	0.634	0.624	0.577	0.615	0.645	0.641	0.727	0.699	0.634
IC	0.536	0.538	0.547	0.591	0.551	0.477	0.598	0.58	0.579	0.629	0.689	0.673
LI	0.621	0.659	0.651	0.657	0.623	0.521	0.675	0.618	0.641	0.659	0.723	0.722
MSO	0.628	0.613	0.607	0.642	0.615	0.592	0.673	0.669	0.651	0.690	0.737	0.725
OF	0.658	0.637	0.654	0.689	0.589	0.579	0.677	0.498	0.624	0.655	0.722	0.736
SA	0.552	0.603	0.620	0.587	0.530	0.441	0.648	0.607	0.599	0.596	0.699	0.714
SSO	0.341	0.238	0275	0.328	0.225	0.284	0.411	0.444	0.247	0.259	0.513	0.607
TC	0.592	0.561	0.567	0.608	0.556	0.563	0.642	0.634	0.617	0.655	0.713	0.692
Entire	0.622	0.609	0.628	0.665	0.580	0.556	0.680	0.640	0.644	0.653	0.744	0.739

TABLE V

F-MEASURE BASED ON ATTRIBUTE OF THE PROPOSED DATASET WITH 11 METHODS ON THE VT1000 DATASET, INCLUDING RBD [27], CA [25], RRWR [26], MR [7], MSS [32], BL [30], MTMR [12] AND DSS [23], MILPS [33], FCNN [20], SDGL [24]. WHERE THE BEST, THE SECOND AND THE THIRD BEST RESULTS ARE IN RED, GREEN AND BLUE COLORS, RESPECTIVELY

	RBD	CA	RRWR	MR	MSS	BL	MTMR	DSS	MILPS	FCNN	SDGL	Ours
BSO	0.813	0.796	0.735	0.75	0.752	0.643	0.741	0.613	0.717	0.794	0.771	0.697
CB	0.488	0.444	0.458	0.468	0.444	0.433	0.541	0.609	0.499	0.551	0.627	0.620
CIB	0.606	0.589	0.534	0.572	0.589	0.623	0.565	0.632	0.644	0.675	0.693	0.590
IC	0.460	0.530	0.458	0.506	0.53	0.475	0.520	0.594	0.528	0.591	0.627	0.552
LI	0.646	0.640	0.647	0.626	0.640	0.385	0.648	0.427	0.615	0.680	0.648	0.602
MSO	0.724	0.665	0.681	0.690	0.665	0.582	0.738	0.713	0.732	0.754	0.773	0.739
OF	0.640	0.588	0.645	0.580	0.588	0.446	0.627	0.437	0.609	0.632	0.669	0.621
SA	0.705	0.664	0.686	0.621	0.632	0.423	0.703	0.624	0.700	0.723	0.753	0.724
SSO	0.456	0.312	0.415	0.444	0.352	0.314	0.556	0.603	0.479	0.425	0.681	0.674
TC	0.543	0.513	0.508	0.584	0.540	0.507	0.594	0.573	0.577	0.605	0.670	0.674
Entire	0.650	0.610	0.616	0.667	0.615	0.596	0.673	0.634	0.661	0.671	0.727	0.693

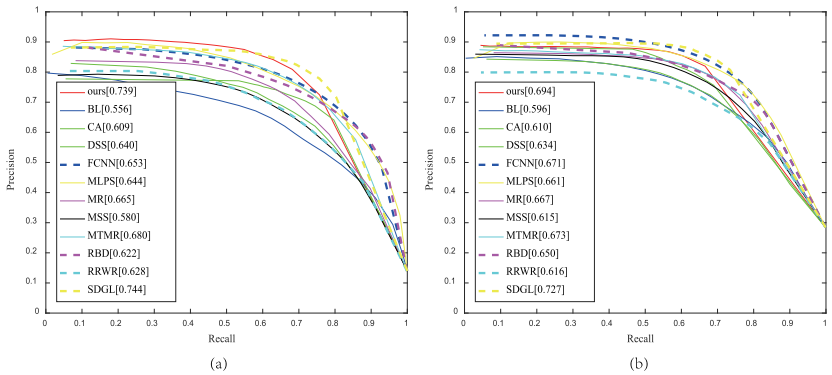


Fig. 6. (a) PR curves of the proposed approach with other baseline methods with RGB-T input on the VT821 dataset. (b) The result on the VT1000 dataset.

extended RGBT methods including MR, CA, RRWR, RBD, BL, DSS, MSS, MILPS, FCNN. We compare with other methods likewise, and achieve the second in F-measure, only 0.5% less than the top method SDGL [24] that is discussed above, and outperform the third best MR [7] with 7.5%. From what has been discussed above, our method achieves remarkable results compared with existing baseline methods. The commendable results are on account of our proposed novel approach based on a cooperative ranking algorithm.

In addition, although our method has low recall, we achieve highest precision in all methods. That is to say, our results contain very few non-salient regions but miss some salient regions. We could use some techniques like interactive and semi-supervised segmentation to refine our results and the results with high recall are easily obtained.

Fig. 7 shows some sample results of our approach against other baseline methods with different inputs. The evaluation results show that the proposed approach can detect the salient objects more accurate than other methods by adaptively and collaboratively leveraging RGBT data. It is worth noting that some results using single modality are better than using RGBT data. It is because that the redundant information introduced by the noisy or malfunction modality sometimes affects the fusion results in bad way.

2) Challenge-Sensitive Performance: For analyzing the attribute-sensitive performance of our approach against other methods, we present the F-measure in Table IV. We evaluate our method with 11 attributes (i.e., BSO, BW, CB, CIB, IC, LI, MSO, OF, SA, SSO, TC) on the VT821 dataset. From Table IV, we can observe that our approach outperforms

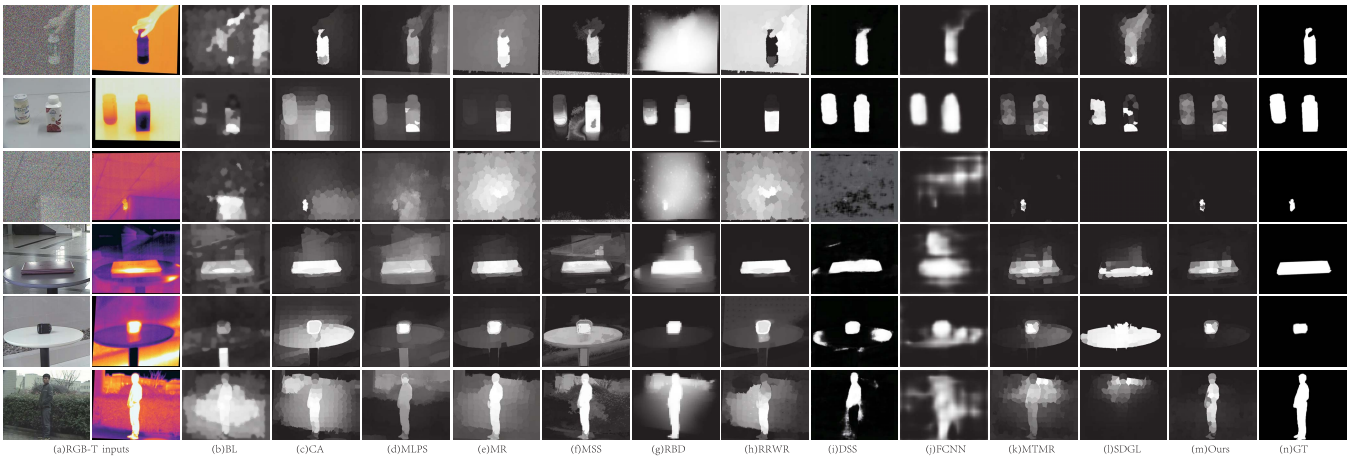


Fig. 7. Sample results of the proposed approach and other baseline methods with different modality inputs. (a) Input RGB and thermal image pair. (b-l) The results of the baseline methods which generated by the extended RGB-T approaches. (m) The results of our proposed approach. (n) The ground truth.

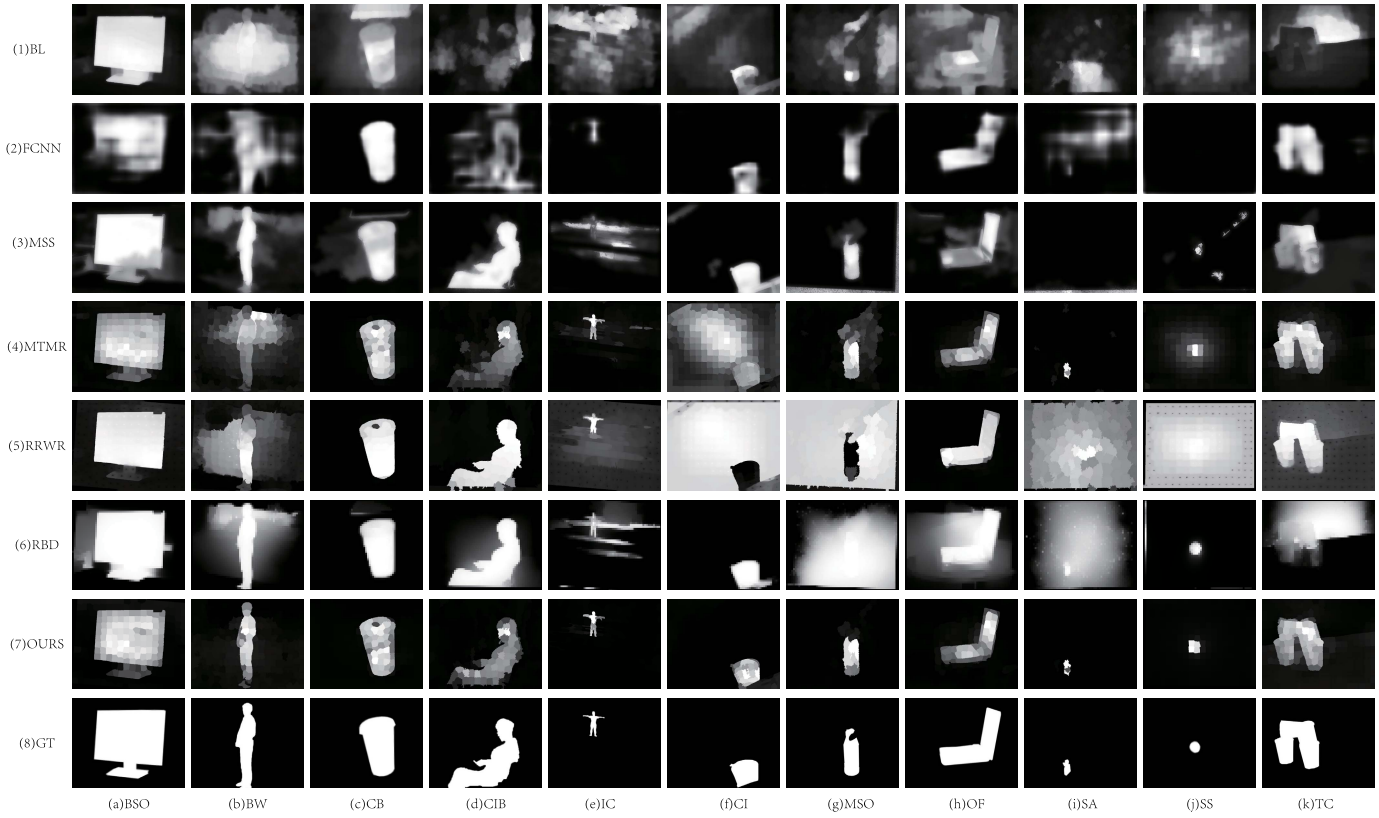


Fig. 8. Subjective sample results of the challenge-sensitive performance. (a-k) represents different annotated challenges of our RGB-T dataset. (1-6) The results of the baseline methods which generated by the extended RGBT approaches. (7) The results of our proposed approach. (8) The ground truth.

other RGBT methods on most of the challenges except BSO, BW and CIB subsets. Especially in OF, SA and SSO, our approach achieves the best results compared with other methods. For BSO, RBD [27] achieves a superior performance over ours, as the background connectivity evaluation of RBD can discover big salient objects by calculating the associations between objects and boundaries. Consider both of these aspects, we can explore more relations among superpixels in graph construction and add some background association information to handle the challenge of BSO. For BW and CIB, most of methods have bad performance, but FCNN [20] obtains the highest F-measure, as it can be trained with various data to handle several challenges.

Furthermore, we present some subjective results comparing to other six methods including BL [30], FCNN [20], MSS [32], MTMR [12], RRWR [26] and RBD [27] in Fig. 8. From the results, we achieve better results in most of the subsets such as BW , IC, CI, MSO, SA, SS and TC. However, in CB, CIB and OF, the outlines of our results are clear, but the contents are not uniformly highlighted. The unsatisfactory results might be due to the weak representations of color features in complex scenarios.

TABLE VI
AVERAGE RUNTIME COMPARISON ON THE VT821 DATASET

Method	MR [7]	RBD [27]	CA [25]	RRWR [26]	MILPS [33]	FCNN [20]	MSS [32]	BL [30]	DSS [23]	MTMR [12]	SGDL [24]	Ours
Code Type	M	M	M	C++	M	Caffe	M	M	Caffe	M	M	M
Runtime(s)	0.55	3.78	0.79	1.57	93.2	0.13	7.35	21.03	0.06	1.89	2.23	1.41

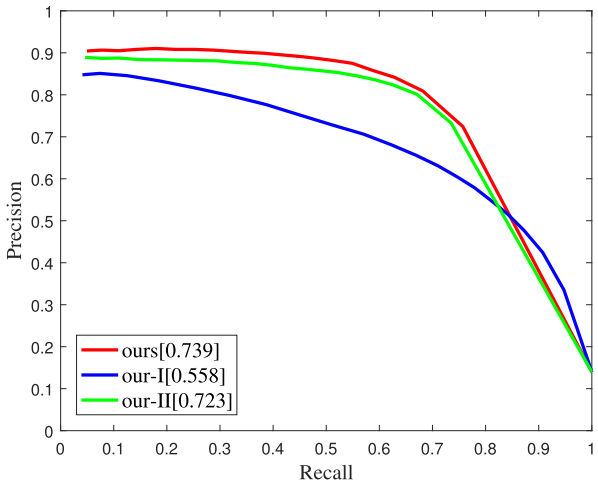


Fig. 9. Evaluation results of the proposed approach with its variants on the VT821 dataset.

C. Evaluation on VT1000

To further testify the effectiveness of our approach, we conduct more experiments on the recently proposed RGBT dataset, i.e., VT1000 [24]. The performance is reported in Table III and Fig. 6. As shown in the Table III and Fig. 6, it is not hard to observe that our approach achieves the outstanding results, which achieve the second best in F-measure, less than the top method SDGL [24] with 3.4% and outperforming the third best MTMR [12] with 2%. These observations are similar to VT821. Furthermore, we also perform evaluation on different attributes (i.e., BSO, SSO, MSO, LI, CB, CIB, SA, TC, IC,OF) on the VT1000 dataset, which is demonstrated in Table V. We can see that our approach outperforms other RGBT methods on most of the challenges, but in the part of challenges.

D. Ablation Study

To justify the significance of the main components of the proposed approach, we implement two special versions for comparative analysis, including: 1) Ours-I, that removes the modality weights in the proposed ranking model, i.e., fixes $r^k = \frac{1}{K}$ in (5), and 2) Ours-II, that removes the cross-modality consistent constraints in the proposed ranking model, i.e., sets $\lambda = 0$ in (5). The results are presented in Fig. 9, and we can make the following observations and conclusions. 1) Our method substantially outperforms Ours-I. It proves the significance of the introduced weighted variables to achieve adaptive fusion of different inputs. 2) The complete algorithm achieves superior performance than Ours-II, verifying the robustness of the cross-modality consistent constraints.

E. Runtime Comparison

In addition, we calculate the runtime of our proposed approach with other baseline methods on VT821 dataset. All results are evaluated on the platform of MATLAB 2016b, with the Intel i7 4.0GHz CPU and 32GB RAM, expect for two deep learning metohds (DSS [23] and FCNN [20]), which are run in the Caffe environments with Python 3.6. As showed in Table VI, the proposed approach costs an average of 1.41 second per RGBT image pair with the size of 480 $\times$ 640. Note that the most time consuming procedure is the inversion operation of a matrix at each iteration in the **S**-subproblem 12. Therefore, we will employ a linearized operation [40] to avoid inverse operations in the future work.

VII. INSIGHTS AND POTENTIAL DIRECTIONS

In this work, we propose a comprehensive image benchmark for RGBT SOD purpose, includes a dataset with statistic analysis, fifteen baseline methods with different inputs and four evaluation metrics. With this benchmark, we propose an effective approach based on a cooperative ranking algorithm for RGBT SOD. We conduct a large scale experiments to evaluate the performance of our approach with other baseline methods on the proposed dataset VT821 and recently presented dataset VT1000. The results show that by fusing RGB and thermal data adaptively, we can achieve favorable performance against other methods. We think that our work could provide some new insights to the field of RGBT saliency detection. Based on these insights, we also discuss some potential research directions for RGBT SOD.

First of all, it can be observed from experimental results that taking the account of both RGB and thermal cues is critical to boost the performance of SOD. This is because the imaging procedure of thermal cameras is insensitivity to lighting and weather conditions and could also suppress background clutter to some extent. On the contrary, visible spectrum could distinguish objects from background when thermal spectrum is disturbed by thermal crossover or reflection likewise.

Furthermore, we conclude some insights from the experimental results as follow. First, adaptive fusion of RGB and thermal modalities is effective for RGBT SOD. Through integrating different modalities with reliability weights, the noisy information introduced by less reliable modality is suppressed and thus the performance is boosted (*e.g.*, MTMR [12], Ours). Second, collaborative learning of different modalities is significant for robust RGBT SOD. It could be observed from the performance gains of collaboratively computing the ranking functions of different modalities (*e.g.*, Ours) or jointly learning the graph using different modality graphs (*e.g.*, SDGL [24]). Third, taking the heterogeneity of different modalities into account is useful to RGBT SOD. In addition to the collaboration, the modeling of the heterogeneity could

make use of different properties of multiple modalities for better results of RGBT SOD (*e.g.*, Ours). Last but not least, powerful feature representations are important for high performance RGBT SOD. It can be observed from the outstanding performance of the incorporation deep learning features (*e.g.*, SDGL [24]).

Based on these results and insights, we present some potential research directions of RGBT SOD, including adaptive fusion of different modalities, effective deep representation with both collaboration and heterogeneity considerations, aggregation of hierarchical deep features and deep mutual learning for the interactions of different modalities.

REFERENCES

[1] J. Guo, T. Ren, J. Bei, and Y. Zhu, "Salient object detection in RGB-D image based on saliency fusion and propagation," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2015, Art. no. 59.

[2] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2017.

[3] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weld: Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 2574–2585, 2018.

[4] S. Yang, B. Luo, C. Li, G. Wang, and J. Tang, "Fast grayscale-thermal foreground detection with collaborative low-rank decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2574–2585, Oct. 2018.

[5] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.

[6] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 69, p. 106977, Dec. 2019.

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

[8] Y. Ji, H. Zhang, K.-K. Tseng, T. W. S. Chow, and Q. M. J. Wu, "Graph model-based salient object detection using objectness and multiple saliency cues," *Neurocomputing*, vol. 323, pp. 188–202, Jan. 2019.

[9] J. Y. Wang, H. Bensmail, and X. Gao, "Multiple graph regularized protein domain ranking," *BMC Bioinformat.*, vol. 13, no. 1, p. 307, 2012.

[10] C. Li, Z. Nan, Y. Lu, C. Zhu, and T. Jin, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1856–1864.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[12] C. Li, G. Wang, Y. Ma, A. Zheng, B. Luo, and J. Tang, "A unified RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Proc. Chin. Conf. Image Graph. Technol.*, 2018, pp. 494–505.

[13] K. Gu, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 9, pp. 1480–1494, Sep. 2015.

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

[15] S. Wang, M. Wang, S. Yang, and K. Zhang, "Salient region detection via discriminative dictionary learning and joint Bayesian inference," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1116–1129, May 2017.

[16] B. Jiang, L. Zhang, H. Lu, C. Yang, and M. H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1665–1672.

[17] L. Zhu, H. Ling, J. Wu, H. Deng, and J. Liu, "Saliency pattern detection by ranking structured trees," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5468–5477.

[18] Q. Wang, W. Zheng, and R. Piramuthu, "GraB: Visual saliency via novel graph model and background priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 535–543.

[19] B. Li, Q. Liu, X. Shi, and Y. Yang, "Graph-based saliency fusion with superpixel-level belief propagation for 3D fixation prediction," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 2321–2325.

[20] X. Li *et al.*, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[21] L. Wang, L. Wang, H. Lu, P. Zhang, and R. Xiang, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2018.

[22] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 714–722.

[23] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3203–3212.

[24] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," 2019, *arxiv:1905.06741*. [Online]. Available: https://arxiv.org/abs/1905.06741

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

[26] C. Li, Y. Yuan, W. Cai, and Y. Xia, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2710–2717.

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

[28] W. C. Tu, S. He, Q. Yang, and S. Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2334–2342.

[29] J. Wang, H. Lu, X. Li, N. Tong, and W. Liu, "Saliency detection via background and foreground seed selection," *Neurocomputing*, vol. 15, pp. 359–368, Mar. 2015.

[30] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1884–1892.

[31] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5455–5463.

[32] N. Tong, H. Lu, L. Zhang, and X. Ruan, "Saliency detection with multi-scale superpixels," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1035–1039, Sep. 2014.

[33] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1911–1922, Apr. 2017.

[34] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

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

[36] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Proc. Neural Inf. Process. Syst.*, 2004, pp. 169–176.

[37] Y. Xu and W. Yin, "A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1758–1789, 2015.

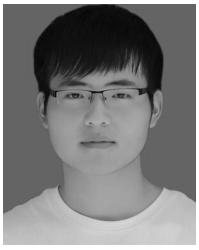
[38] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.

[39] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[40] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 612–620.

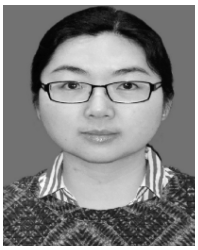
Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively.

He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.

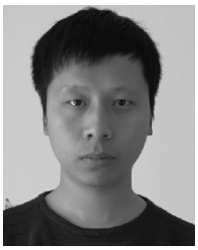


Dongzhe Fan received the B.S degree from Anhui Normal University, Wuhu, China, in 2016. He is currently pursuing the M.S. degree with Anhui University, Hefei, China. His research interests include computer vision and machine learning.

Xiaoxiao Wang received the B.S. degree from the Wannan Medical College, Wuhu, China, in 2018. She is currently pursuing the M.S. degree with Anhui University, Hefei, China. Her current research interests include computer vision and deep learning.



Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision, pattern recognition, and deep learning.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.