A Novel Visible-Depth-Thermal Image Dataset of Salient Object Detection for Robotic Visual Perception

Kechen Song, *Member, IEEE*, Jie Wang, Yanqi Bao, Liming Huang, and Yunhui Yan

***Abstract*—Visual perception plays an important role in industrial information field, especially in robotic grasping application. In order to detect the object to be grasped quickly and accurately, salient object detection (SOD) is employed to the above task. Although the existing SOD methods have achieved impressive performance, they still have some limitations in the complex interference environment of practical application. To better deal with the complex interference environment, a novel triple-modal images fusion strategy is proposed to implement SOD for robotic visual perception, namely visible-depth-thermal (VDT) SOD. Meanwhile, we build an image acquisition system under variable lighting scene and construct a novel benchmark dataset for VDT SOD (VDT-2048 dataset). Multiple modal images will be introduced to assist each other to highlight the salient regions. But, inevitably, interference will also be introduced. In order to achieve effective cross-modal feature fusion while suppressing information interference, a hierarchical weighted suppress interference (HWSI) method is proposed. The comprehensive experimental results prove that our method achieves better performance than the state-of-the-art methods.**

***Index Terms*—Industrial visual perceptron, salient object detection (SOD), visible-depth-thermal (VDT) images.**

I. Introduction

AS AN important perceptron in industrial information field, machine vision has been widely used in various robot fields [1], [2], [3], [4], [5], [37]. Especially in service robots, in order to cope with the complex and changeable grasping task, visual perceptron needs to detect the object to be grasped quickly and accurately. Salient object detection (SOD) is one of the key technologies in the above object grasping task. In addition, SOD is often used in visual processing tasks of industrial robots, including industrial image detection and segmentation, which provides preprocessing of the object for these tasks.

This article mainly studies the salient object visual perception ability of home service robot. The service level of the robot is closely related to the health, happiness, and even life of the employer, especially the elderly. To improve the service ability and quality of home service robot, it is necessary to equip the robot with good visual perception system, so that it can have full visual recognition ability like human. Currently, the visual perception system of home service robot equipped with RGB cameras still has the following problems: 1) The objects to be recognized in the home environment are usually small, numerous and dense, and vulnerable to the background interference; 2) When the light is insufficient, the robot's object detection ability is greatly reduced, which greatly affects its service quality, as shown in Fig. 1.

To solve the above problems, we equipped the home service robot with more visual sensors (depth and thermal infrared) other than RGB cameras to further improve the robot's visual perception ability, as shown in Fig. 2. At the same time, multimodal images are necessary for perceptual training of robot control agent. However, the current SOD dataset is mainly from the natural images, and there is no multimodal image dataset suitable for home service robot, especially triple-modal (visible-depth-thermal) dataset.

Although these increased depth and thermal infrared sensors can better detect salient objects, they also bring some interference challenges. Depth sensor [6], [7] can effectively distinguish the distance differences between object and background, but it also introduces some distracting information. As shown in Fig. 3, the background of the depth image without salient object is very messy, which distracts the detection focus of SOD. Moreover, the depth information of salient object is incomplete when there is no distance difference between the salient objects,

Manuscript received 13 April 2022; revised 24 June 2022 and 1 September 2022; accepted 16 October 2022. Date of publication 27 October 2022; date of current version 16 June 2023. Recommended by Technical Editor: J.-H. Han and Senior Editor N. Tsagarakis. This work was supported in part by the National Natural Science Foundation of China under Grant 51805078, in part by the Fundamental Research Funds for the Central Universities under Grant N2103011, in part by the Central Guidance on Local Science and Technology Development Fund under Grant 2022JH6/100100023, and under Grant 111 Project B16009. *(Corresponding authors: Kechen Song; Yunhui Yan.)*

Kechen Song, Jie Wang, Yanqi Bao, and Yunhui Yan are with the School of Mechanical Engineering and Automation, Northeastern University, Shenyang 110819, China, also with the National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang 110819, China, and also with the Key Laboratory of Data Analytics and Optimization for Smart Industry, Ministry of Education, Northeastern University, Shenyang 110819, China (e-mail: songkc@me.neu.edu.cn; 1970193@stu.neu.edu.cn; yanqibao@stumail.neu.edu.cn; yanyh@mail.neu.edu.cn).

Liming Huang is with the Department of Computer Science, University of Exeter, Exeter EX4 4PY, U.K. (e-mail: lh830@exeter.ac.uk).

The dataset is available at: https://github.com/VDT-2048/VDT-Dataset

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMECH.2022.3215909.

Digital Object Identifier 10.1109/TMECH.2022.3215909

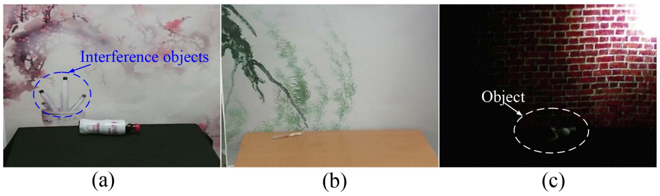

Fig. 1. Difficult challenges of home service robot equipped with RGB. (a) Similar appearance: some interference objects in the background have the similar appearance and seriously affect the detection performance of SOD. (b) Small salient object: some small salient objects also increase the difficulty of SOD. (c) Low illumination: object is very difficult to detect due to lack of adequate illumination. (a) Similar appearance. (b) Small salient object. (c) Low illumination.

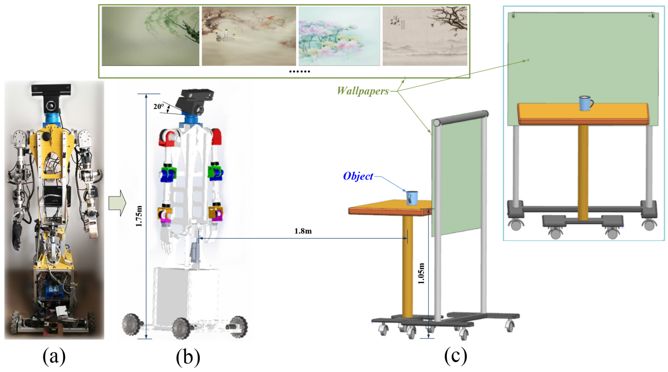

Fig. 2. Overall layout of the image acquisition system in this work. The system mainly consists of three parts: the robot body (entity and 3-D model are shown in (a) and (b), respectively), visible-depth-thermal (VDT) camera component and auxiliary acquisition platform (c). (a) Robot entity. (b) 3D model of robot. (c) Auxiliary acquisition platform.

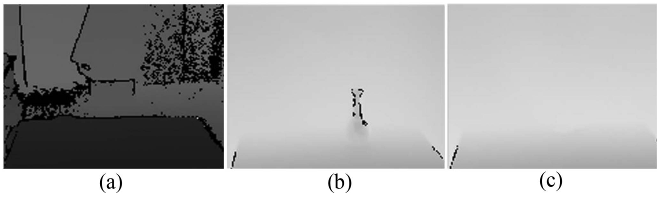

Fig. 3. Difficult challenges of depth image. (a) Background messy: the background without salient object is very messy, which distracts the detection focus of SOD. (b) Depth information incomplete: the depth information of salient object is incomplete when there is no distance difference between the salient objects, or the difference is very small. (c) Small salient objects: the depth image is difficult to distinguish some small salient objects. (a) Background messy. (b) Depth information incomplete. (c) Small salient objects.

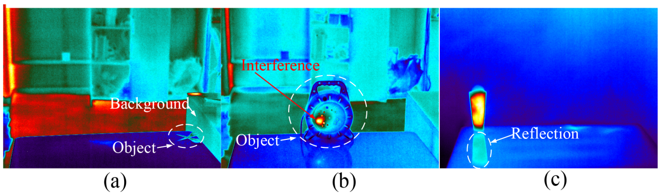

Fig. 4. Difficult challenges of thermal image. (a) Thermal crossover: the temperature of the salient object is the same as that of part of the background, which greatly increases the difficulty of object detection. (b) Thermal radiation dispersion: part of a salient object is more salient than the whole object. (c) Heat reflection: some objects will appear heat reflection phenomena. (a) Thermal crossover. (b) Thermal radiation dispersion. (c) Heat reflection.

or the difference is very small. In addition, depth is still difficult to distinguish some small salient objects. Different from visible and depth images, thermal infrared sensor can sense the slight temperature difference between the salient object and the background even in the low illumination and completely dark environment [8], [38], [39], [40], [41]. However, there are some difficult challenges that need to be addressed for thermal image, such as thermal crossover, thermal radiation dispersion, and heat reflection. As shown in Fig. 4(a), the temperature of the salient object is the same as that of part of the background, namely thermal crossover, which greatly increases the difficulty of object detection. Fig. 4(b) presents an example of thermal radiation dispersion, i.e., part of a salient object is more salient than the whole object, which causes interference to detection. In addition, some objects will appear heat reflection phenomena, as shown in Fig. 4(c), which is also an important interference.

From the above analysis, it can be seen that any single modal image has both merits and disadvantages. To overcome the disadvantages of single modal image, researchers proposed dual-modal images SOD methods using depth and thermal images, namely RGB-D SOD [6], [7] and RGB-T SOD [8], [38], [39], [40], [41]. RGB-D SOD effectively overcomes the problem that RGB is susceptible to the interference of distance perception, i.e., some interference objects in the background have the similar appearance, as shown in Fig. 1(a). RGB-T SOD solves the problem that RGB is susceptible to the interference from illumination changes. Although dual-modal images solve the problems existing in single-modal image to some extent, the detection performance of RGB-D and RGB-T will be greatly affected in complex scenes with multiple interference factors, as shown in the subsequent challenging scenes in this article. Especially, when both modalities are subject to certain challenges, the detection effect of dual-modal methods will decrease sharply, as shown in Fig. 15.

In order to solve the problems of single modal and dual-modal images, this article introduces the idea of fusing triple-modal images together for detection. Actually, in the past decade, visible-depth-thermal (VDT) image analysis has achieved a wide range of interest and has been applied in many fields to perform specific tasks, including pedestrian detection [22], person tracking [23], mobile robots [24], medical diagnostics [25]–[26], human body segmentation [27], gas leak inspection [28], 3-D thermal mapping reconstruction [29], etc. These studies also further verify that the idea of fusing triple-modal images together for detection can adapt to more complex interference environment. However, the VDT image analysis technique has not been applied to SOD for robotic visual perception. Therefore, to better deal with the complex interference environment, a new strategy fused triple-modal images is proposed to implement SOD task, namely VDT SOD. Our motivation is to take advantage of the strengths of the triple modal images to make up for their respective weaknesses. VDT SOD can not only effectively solve the difficult challenges for single modal image, but also overcome the limitations of dual-modal images. Meanwhile, we build an image acquisition system under variable lighting indoor scene and construct a novel benchmark dataset for VDT SOD, namely VDT-2048 dataset.

Since the existing dual-modal SOD methods are verified on two different types of datasets separately (i.e., RGB-D datasets and RGB-T datasets), and there is a lack of a unified dataset to simultaneously verify the merits and disadvantages of these methods. The proposed VDT-2048 dataset provides fourteen challenging scenes, which can be used as a unified dataset to simultaneously verify the performance of dual-modal methods.

It is inevitable that multiple modal images will assist each other to highlight the salient region but also introducing interference. In order to achieve the complementary fusion of cross-modal information and the purpose of effectively suppress interference, a hierarchical weighted suppress interference (HWSI) method is proposed.

Our contributions are summed up as follows:

1) A novel triple-modal images fusion strategy is proposed to implement the SOD for robotic visual perception. To the best of our knowledge, it is the first work to investigate the SOD from VDT images, i.e., VDT SOD.
2) We construct a novel benchmark dataset for VDT SOD, namely VDT-2048 dataset. It collects 34 household items in the seven most common household scenes, including 2048 image groups of triple-modal images.
3) The proposed VDT-2048 dataset provides fourteen challenging scenes, which can be used as a unified dataset platform to simultaneously verify the cross-model generalization performance of dual-modal SOD methods.
4) We propose a new HWSI method. HWSI mainly includes three modules, i.e., dual-modal attention fusion module (DMAFM), triple-modal interactive weighting module (TMIWM) and global attention-weighted fusion module (GAWFM). Each module is designed to use cross-modal information weighting to highlight salient regions and effectively suppress interference.

II. Image Acquisition System and Dataset

A. Image Acquisition System

To realize the SOD for the robotic grasping task in a family environment, we employ a robot developed by our lab to build an image acquisition system under variable lighting scene. The image acquisition system mainly consists of three parts: the robot body, VDT camera component and auxiliary acquisition platform. The overall layout of the image acquisition system is shown in Fig. 2.

VDT camera component is constructed to capture multimodal images. This camera component consists of two cameras: a motion-sensing camera (Microsoft Kinect v2, including visible and depth sensors) and a thermal camera (FLIR-A655sc, focal length: 25 mm, FOV: 25° × 19°, spectrum: 7.5–14 μm). Image resolution of these cameras is different, Visible (V): 1920 × 1080, Depth (D):512 × 424, Thermal (T):640 × 480. These cameras are mounted vertically to a rotating platform as the head of the robot.

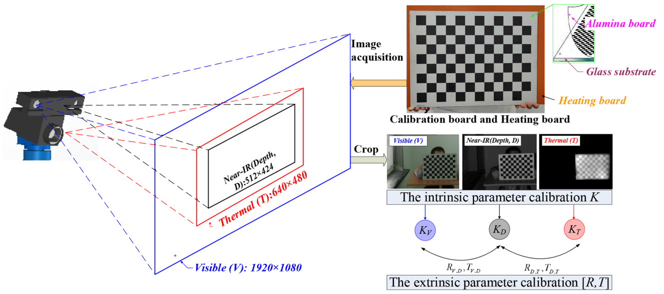

Fig. 5. Calibration of the VDT camera component.

The auxiliary acquisition platform is mainly composed of a table and a wallpaper shelf. To increase the challenge of the image dataset, the following different experimental settings are carried out for the image acquisition process. 1) Changing size, type, and number of objects in the table. 2) Changing different wallpapers to enhance the background interference challenges for visible image. 3) With or without the wallpaper shelf to enhance the background interference challenges for depth image. 4) Changing the intensity and location of the light.

B. Calibration, Registration, and Annotation

Multimodal SOD usually requires the aligned images. But in this work, the constructed VDT cameras have different FOV and image resolution. In order to obtain the aligned images, the constructed VDT cameras should be calibrated firstly. The calibration parameters are then used to construct the final aligned images. For the multimodal image calibration, the existing studies of VDT image analysis give us inspiration and reference [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Different from the existing calibration patterns [22], [30], [31], a checkerboard pattern is used, as shown in Fig. 5.

Kinect v2 includes two image sensors, visible sensor and Near-infrared (Near-IR) sensor. Near-IR sensor is commonly used to generate the depth image and is often referred to as depth sensor. Although the principles of these two sensors are different, they have similar spectral ranges. Therefore, these two sensors can be calibrated using the same calibration board, e.g., printed checkerboard using paper.

Different from the Kinect v2, the calibration difficulty of the thermal camera is how to obtain high quality images of calibration board pattern. To solve this problem, we designed and manufactured a calibration board composed of two materials, as shown in Fig. 5. The pattern of our calibration board is a checkerboard with 12×9 (30 mm for every square grid). This pattern is printed onto an alumina plate, which is then mounted on a glass substrate. Moreover, a heating board is used to heat the calibration board. From Fig. 5, it can be seen that our calibration board can meet the calibration requirements of multimodal cameras, that is, every camera can obtain high quality pattern images including clear edges and high contrast.

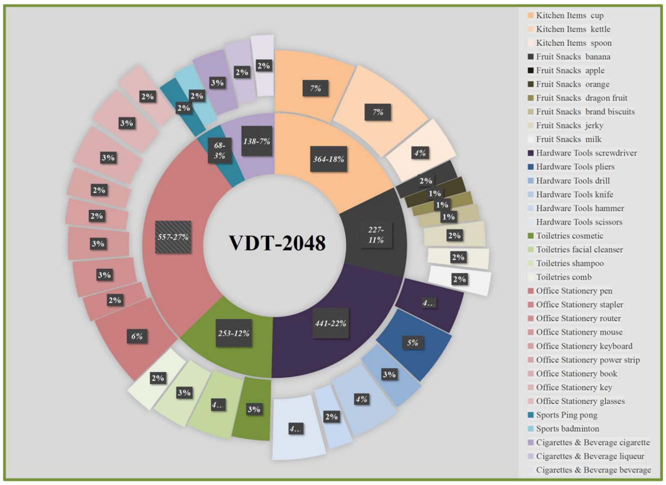

Fig. 6. Proportion of each scene and item category of VDT-2048 dataset. 34 household items in seven most common household scenes: kitchen items (cup, kettle, spoon), fruit snacks (banana, apple, orange, dragon fruit, brand biscuits, jerky, milk), hardware tools (screwdriver, pliers, drill, knife, hammer, scissors), Toiletries (cosmetic, facial cleanser, shampoo, comb), office stationery (pen, stapler, router, mouse, keyboard, power strip, book, key, glasses), sports (ping pong, badminton), and cigarettes and beverage (cigarette, liqueur, beverage).

According to the pattern images of multimodal cameras, the intrinsic parameters (K_V, K_D, K_T) and the extrinsic parameters ($[R_{V,\ D}, T_{V,\ D}]$, $[R_{D,\ T}, T_{D,\ T}]$) can be calculated using the camera calibration toolbox [9]. The calculated calibration parameters are then used to register the acquired multimodal images and construct the final aligned images.

For the aligned multimodal images, we manually annotated the salient objects using an annotation tool *Photoshop*. The annotation implements a pixel-level mask for salient object and obtains the ground truth for each group of the multimodal images. Most of the annotation can be realized using the visible image. For some objects that could not be seen clearly in the visible image, we annotated them using the thermal image.

C. VDT-2048 Dataset Analysis

Using the above image acquisition system, we construct a novel benchmark dataset for VDT SOD, namely VDT-2048 Dataset. This dataset contains 2048 image groups, and each group contains triple-modal images (i.e., visible image, depth image, and thermal image). All of the images have the same resolution of 640×480. This dataset collected 34 household items in the seven most common household scenes. The proportion of each scene and item category is shown in Fig. 6. The distribution of challenging scenes is shown in Fig. 7. Details are shown in Figs. 8 –10.

According to the four different experimental settings in Section II Part A, a reasonable and realistic challenging scene is created for each modal when constructing the dataset. Therefore, each modal in the dataset is further divided into several challenging scenes. Then, 14 challenging subdatasets are obtained in VDT-2048.

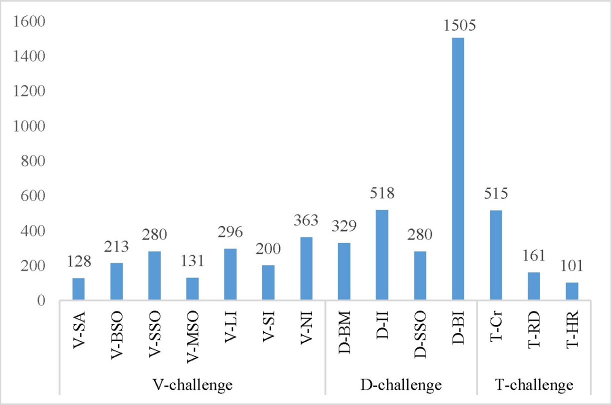

Fig. 7. Distribution of challenging scenes in VDT-2048.

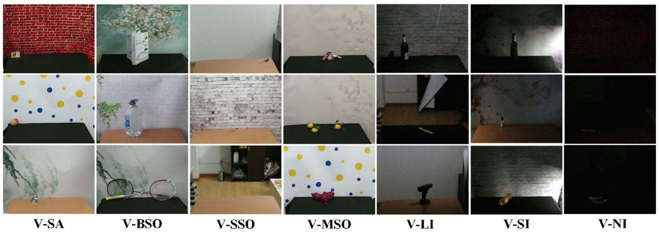

Fig. 8. Sample display of V challenging scenes.

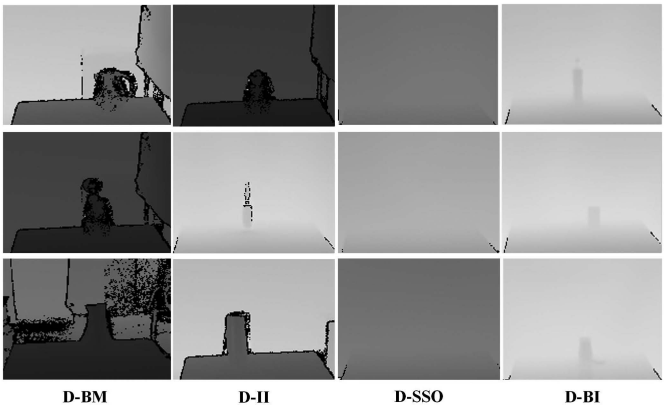

Fig. 9. Sample display of D challenging scenes.

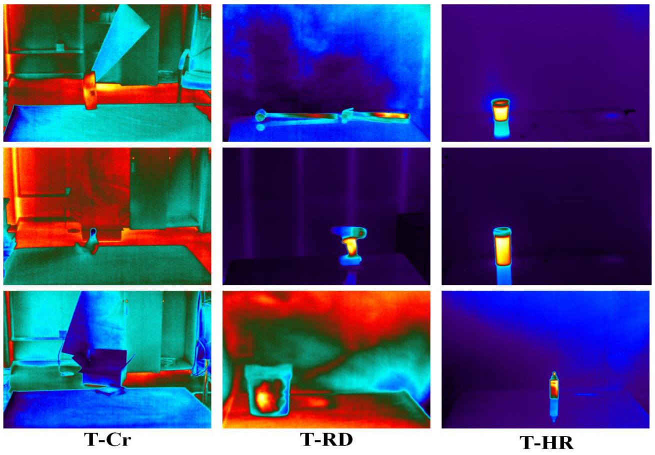

Fig. 10. Sample display of T challenging scenes.

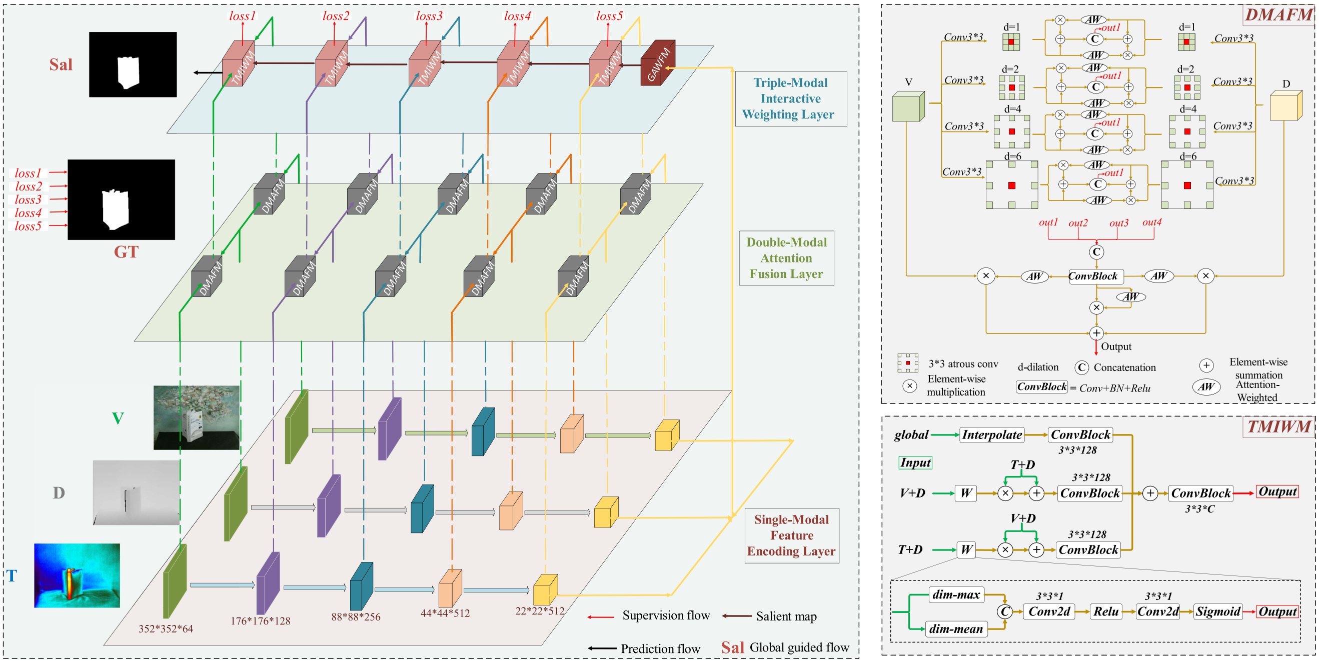

Fig. 11. Overall architecture of the proposed HWSI method and two main modules. The left is the overall architecture of the proposed HWSI method. The upper right is the dual-modal attention fusion module (DMAFM), V+D fusion module is taken as an example to show the specific module details: four atrous convolutions with different dilation rates are used to explore multi-scale and complementary information, the mutual attention-weighted (AW) module is used to suppress interference and achieve cross-modal fusion. The lower right is the triple-modal interactive weighting module (TMIWM), TMIWM weights the two outputs of the DMAFM module to achieve the complementary fusion of triple-modal information.

As shown in Fig. 8, V information mainly has seven challenging scenes. V-SA (similar appearance): the salient object has a similar color or shape to the background. V-BSO (big salient object): the ratio of the sum of salient pixels to the total pixel sum of the entire image is greater than 0.08. V-SSO (small salient object): the ratio of the sum of salient pixels to the total pixel sum of the entire image is less than 0.007. V-MSO (multiple salient objects): the number of salient objects is more than one. V-LI (low illumination): images are collected under low illumination, and objects are not easier to identify visually. V-SI (side illumination): illumination is given from the side of salient objects, and the brightness of salient objects is uneven. V-NI (no illumination): the image is collected under no illumination, and objects are visually difficult to identify.

As shown in Fig. 9, D information mainly has four challenging scenes. D-BM (background messy): background messy when there is no wallpaper. D-II (information incomplete): partial lack of D information leads to incomplete information of salient objects. D-SSO (small salient objects): the ratio of the sum of salient pixels to the total pixel sum of the entire image is less than 0.007. D-BI (background interference): using wallpaper as a background to interfere with D information.

As shown in Fig. 10, T information mainly has three challenging scenes. T-Cr (crossover): the salient object has a similar temperature to the surrounding or other objects. T-RD (radiation dispersion): part of a salient object is more salient than the whole object. T-HR (heat reflection): the heat radiation of the salient object is reflected.

A detailed description and dataset download link are available at:[1]

III. Proposed V-D-T SOD Method

A. Architecture Overview

The proposed method uses the feature extraction part of the classification network VGG16 as the backbone. A three-stream encoding network is constructed to encode the images of the triple-modal separately and extract five-level features with different resolutions. The five-level features of the triple-modal extracted are, respectively, recorded as: visible feature V ($V_1 \sim V_5$), depth feature D ($D_1 \sim D_5$), and thermal infrared feature T ($T_1 \sim T_5$). We created a hierarchical network architecture to deal with the challenging scenes in the dataset and overcome the limitations of each of the triple-modal images. The feature information between different modalities is weighted to suppress the interference of a single modal information and make up for the possible lack of information in a certain modal. The overall architecture of the proposed HWSI method is presented in Fig. 11.

From the vertical perspective, the network is mainly divided into three layers: the single-modal feature encoding layer in the first layer (i.e., bottom layer), the dual-modal attention fusion layer in the second layer (i.e., middle layer), and the triple-modal interactive weighting layer in the third layer (i.e., top layer). The

[1] [Online]: Available https://github.com/VDT-2048/VDT-Dataset

bottom layer is mainly used to extract the multiscale features by the encoding process. The middle layer is used as a ladder to process the features extracted from different modalities. This layer uses two DMAFM modals to process V+D and D+T, respectively, and achieves cross-modal attention fusion of two dual-modal features. The top layer uses TMIWM to interactively weight the dual-modal features after fusion in the middle layer to achieve the purpose of suppressing interference and complementary fusion.

B. Proposed Dual-Modal Attention Fusion Module

As mentioned in the introduction, any single modal image has its own advantages and disadvantages. Therefore, modal fusion can effectively compensate for their disadvantages. The fusion module of cross-modal features is the key factor to obtain excellent performance. Most of the existing dual-modal SOD methods [10], [11] adopt simple addition, multiplication, and concatenation operations to fuse cross-modal features. These fusion operations are insufficient for the exploration of complementary information and the fusion of cross-modal features. More importantly, these operations do not effectively suppress the interference information in the cross-modal features. Especially in the VDT SOD task, the interference information will be more serious. To better solve the above problems, D information is employed as a triple-modal information communication bridge to achieve better fusion performance for VDT SOD. Moreover, DMAFM is proposed to achieve cross-modal V+D and D+T complementary fusion and mutual weighting to eliminate interference.

Fig. 11 illustrates the proposed DMAFM, i.e., V+D fusion module is taken as an example to show the specific module details. Four atrous convolutions with different dilation rates are used to explore multiscale and complementary information. For the two modal features of four different scales, the mutual attention-weighted (AW) module is used to suppress interference and achieve cross-modal fusion.

Four atrous convolutions with different dilation rates are first used to explore features of different scales (the operations on the obtained four scale information are similar). x^v and x^d represent the input V and D features, respectively. Here, taking the feature $\phi_1(x^v)$ processed by the atrous convolution of $d = 1$ as an example. The *AW* module is used to generate the attention-weighted feature map of $\phi_1(x^v)$. The attention-weighted feature map is multiplied by $\phi_1(x^d)$ to achieve cross-modal feature attention. Then, the result after cross-modal feature attention is added into $\phi_1(x^d)$ to realize the weighting of attention of V to D

$$f_i^d = f_i^d\left(x^d, x^v\right) = \varphi\left(\phi_n\left(x^v\right)\right) \times \phi_n\left(x^d\right) + \phi_n\left(x^d\right) \quad (1)$$

where $\phi_n(\bullet)$ represents 3∗3 atrous convolution operation with a dilation rate of *n*, $\varphi(\bullet)$ represents the attention-weighted module, $i = 1,\ldots,4$; $n = 12{,}46$. First, the channel attention mechanism is used to weight the feature channels, and then the spatial attention is explored to generate feature weighted attention maps. The same principle can be used for V

$$f_i^v = f_i^v\left(x^v, x^d\right) = \varphi\left(\phi_n\left(x^d\right)\right) \times \phi_n\left(x^v\right) + \phi_n\left(x^v\right). \quad (2)$$

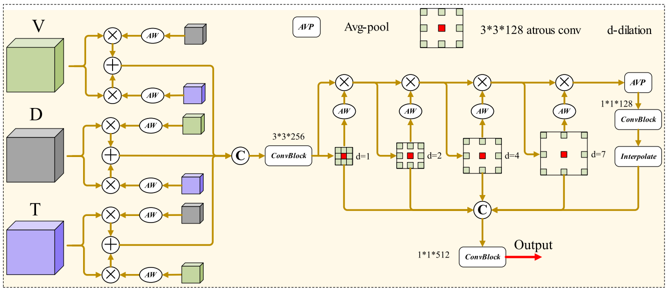

Fig. 12. Proposed global attention weighted fusion module (GAWFM).

Finally, the two features weighted by mutual attention are concatenated to obtain the output of each branch

$$f'_i = CAT\left(f_i^v, f_i^d\right) \quad (3)$$

where *CAT* indicates the concatenate operation. The output of the obtained four branches is fully fused through concatenate and convolution block operation. The fused features are weighted by attention to V, D, and itself, respectively. Finally, the final output is obtained by adding. The output of the dual-modal feature V+D is expressed as

$$f_j^{vd} = \sigma\left\{\beta\left[Conv_{3*3}\left(CAT\left(f'_1, \cdots, f'_4\right)\right)\right]\right\}, (j = 1, \ldots, 5) \quad (4)$$

$$Y_j^{vd} = f_j^{vd} \times \varphi\left(f_j^{vd}\right) + x_j^v \times \varphi\left(f_j^{vd}\right) + x_j^d \times \varphi\left(f_j^{vd}\right) \quad (5)$$

where $Conv_{n*n}$ represents the convolution of the convolution kernel *n*∗*n*, $\beta[\bullet]$ represents *BN* processing, and $\sigma\{\bullet\}$ represents *Relu* activation. The output of the dual-modal feature T+D is expressed as

$$Y_j^{td} = f_j^{td} \times \varphi\left(f_j^{td}\right) + x_j^t \times \varphi\left(f_j^{td}\right) + x_j^d \times \varphi\left(f_j^{td}\right). \quad (6)$$

C. Proposed Global Attention-Weighted Fusion Module

The high-level features with the smallest resolution contain rich semantic information. These semantic features can better achieve cross-modal information complementary fusion and play an important guiding role in the subsequent decoding process. Recently, ASPP is employed to explore global semantic and to guide the decoding process for SOD [12], [13]. But, the direct information aggregation at multiple scales will weaken its representation ability due to some interference. In this work, GAWFM is proposed to suppress the interference information of the triple-modal through mutual weighting, and realize the mining and fusion of the complementary information.

Fig. 12 presents the proposed GAWFM. For the features of the triple-modal (x_5^v, x_5^d, x_5^t), each modal is weighted by the attention of the other two modal. In this way, the attention weighting between the triple-modal can effectively highlight the salient region and suppress the interference. The mutually weighted output uses concatenate and convolution block to fuse the features and to reduce the number of channels

$$x^{vdt} = \sigma\left\{\beta\left[Conv_{3*3}\left(CAT\left(x_5^v \times \varphi\left(x_5^d\right) + x_5^v \times \varphi\left(x_5^t\right), x_5^d\right.\right.\right.\right.$$

$$\times\varphi\left(x_5^v\right)+x_5^d\times\varphi\left(x_5^t\right), x_5^t\times\varphi\left(x_5^d\right)+x_5^t\times\varphi\left(x_5^v\right)\right)\right)\right]\Big\}. \tag{7}$$

Then, four attention-weighted residual atrous convolutions and a global pooling are used to further mine multiscale salient features

$$\begin{aligned} f_1 &= \phi_1(x^{vdt}) \\ f_2 &= \phi_2(\varphi(f_1)\times x^{vdt}) \\ f_3 &= \phi_3(\varphi(f_2)\times\varphi(f_1)\times x^{vdt}) \\ f_4 &= \phi_4(\varphi(f_3)\times\varphi(f_2)\times\varphi(f_1)\times x^{vdt}) \\ f_5 &= upsample\{\sigma\{\beta[\text{Conv}_{3*3}(\text{Avp}(\varphi(f_4) \\ &\quad \times\varphi(f_3)\times\varphi(f_2)\times\varphi(f_1)\times x^{vdt}))]\}\} \end{aligned} \tag{8}$$

where $upsample\{\bullet\}$ represents interpolation up-sampling, Avp represents adaptive average pooling.

Finally, the multiscale information is fused using concatenate, and the number of channels is restored through the convolution block to obtain the global semantic guidance feature. The output of this module can be expressed as

$$G^{vdt} = \sigma\{\beta[Conv_{3*3}(CAT(f_1, f_2, f_3, f_4, f_5))]\}. \tag{9}$$

D. Proposed Triple-Modal Interactive Weighting Module

The DMAFM module serves as a bridge to effectively communicate the triple-modal information, while TMIWM plays an important role in achieving the complementary integration of the triple-modal. This module mainly has two functions: 1) On the basis of inheriting the decoding features of the previous level, TMIWM supplements the detailed features of the current level, so that the entire decoding process is continuously enriched with information. 2) TMIWM further weights the two outputs of the DMAFM module to achieve the complementary fusion of triple-modal information. This module will highlight the significant region and suppress interference information. Fig. 11 shows the details of TMIWM.

Through interpolation up-sampling and a convolution block processing $global$ (the output of the previous stage TMIWM), the resolution is consistent with the current stage, and the number of channels is reduced to 128. This simple inheritance guarantees the proportion of the previous-level feature information in the current processing module. The processed result is used as one of the three branches of this module. This branch can be expressed as

$$\begin{aligned} F_j^{\text{glo}} &= \sigma\{\beta[\text{Conv}_{3*3}(\text{upsample}(F_{j-1}))]\} \\ &\quad \times\left(j=1,\ldots,5, F_0=G^{\text{vdt}}\right). \end{aligned} \tag{10}$$

The DMAFM module has used D information as a bridge to obtain the attention-weighted fusion output between the V+D and D+T modalities. Therefore, we design W to generate a weighted feature map at the spatial level, and use interactive multiplication to achieve weighted fusion between the triple-modal. The fusion results are added to the unweighted output. Finally, a convolution block is used to unify the number of channels to 128

$$F_j^{\text{tdv}} = \sigma\left\{\beta\left[\text{Conv}_{3*3}\left(Y_j^{td}+Y_j^{td}\times W\left(Y_j^{vd}\right)\right)\right]\right\} \tag{11}$$

$$F_j^{\text{vdt}} = \sigma\left\{\beta\left[\text{Conv}_{3*3}\left(Y_j^{vd}+Y_j^{vd}\times W\left(Y_j^{td}\right)\right)\right]\right\}. \tag{12}$$

W explores the maximum value and average value of each feature map in the channel dimension to obtain two weighted feature maps. We use concatenate, convolution, and activation functions to obtain a weighted output of the input feature. The output of TMIWM is expressed as

$$F_j = \sigma\left\{\beta\left[\text{Conv}_{3*3}\left(F_j^{glo}+F_j^{tdv}+F_j^{vdt}\right)\right]\right\}. \tag{13}$$

TMIWM realizes the interactive fusion of triple-modal information, highlights the salient region, and effectively suppresses the interference. It facilitates the supplementation of valuable detailed information in the decoding process. We use two losses (i.e., binary cross entropy and IOU) to supervise the output of each level of TMIWM. For the first loss, a 3∗3 convolution is used to reduce the number of channels to 1, and then binary cross entropy is used to calculate the loss. For the second loss, the same applies to convolution to reduce the number of channels to 1. And we calculate the IOU loss after *sigmoid* is activated. The two calculated losses are added together as the losses of each level of TMIWM

$$L_j = L_j^{\text{bce}} + L_j^{\text{iou}} \tag{14}$$

where L_j represents the loss of the jth level TMIWM, and L_j^{bce} and L_j^{iou} represent the corresponding binary cross entropy loss and IOU loss, respectively.

E. Joint Learning and Saliency Map Prediction

We use joint learning to train our network. Each level of TMIWM produces a loss L_j. Since the output of the first stage needs to generate the final salient map during the test stage, a boundary loss is added to this stage [15]. It uses the Laplacian operator to generate the boundary of the saliency map and the ground truth map, and then only uses the cross-entropy loss supervision to generate the boundary, denoted as L_s.

The overall loss in the training phase can be expressed as

$$L_{\text{total}} = L_1 + L_2 + \chi L_3 + \gamma L_4 + \lambda L_5 + \kappa L_s \tag{15}$$

where the initial value of all loss coefficients is set to 1. Through experiments, the coefficients are adjusted through experiments: χ is adjusted to 0.9, γ , and κ are adjusted to 0.8, and λ is adjusted to 0.7.

IV. Experiments

A. Datasets and Implementation Details

For the VDT-2048 dataset, we randomly sampled 1048 image groups as the training set, and the rest as the test set. We divided the challenging scenes in the test set, and used the trained model to test the challenging scenes. The proposed method is built using the Pytorch framework, and a single NVIDIA GTX 2080 Ti GPU is used for all training and testing. The batchsize is set to 2 during training, and 60 epochs are trained at the same time. The stochastic gradient descent is used to optimize the parameters.

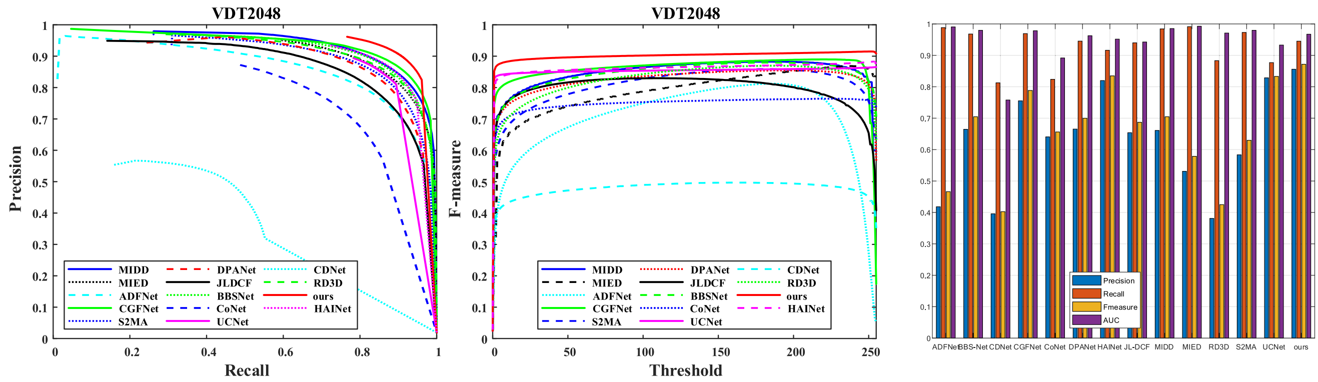

Fig. 13. Quantitative comparison results between our proposed method and the latest methods. The first column is precision-recall curves, the second column shows the measure-threshold curves, and the third column presents the average precision, recall, measure, and AUC score of different methods.

TABLE I
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT MODEL METHODS (RED REPRSENTS THE BEST, BLUE REPRESENTS THE SECOND BEST)

Models	Test-set				
	$E_m\uparrow$	$S_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$
MIDD$_{2021}$	0.880	0.914	0.817	0.704	0.004
MIED$_{2020}$	0.783	0.889	0.750	0.578	0.005
ADFNet$_{2020}$	0.686	0.867	0.636	0.466	0.009
CGFNet$_{2021}$	0.936	0.921	0.850	0.788	0.004
S2MA$_{2020}$	0.829	0.898	0.768	0.629	0.005
DPANet$_{2020}$	0.876	0.889	0.783	0.700	0.005
JL-DCF$_{2020}$	0.869	0.854	0.723	0.687	0.006
BBS-Net$_{2020}$	0.881	0.912	0.817	0.704	0.005
CoNet$_{2020}$	0.876	0.808	0.678	0.656	0.008
UC-Net$_{2020}$	0.968	0.891	0.838	0.833	0.004
(CDNet$_{2021}$)	0.649	0.704	0.415	0.402	0.011
RD3D$_{2021}$	0.839	0.909	0.790	0.653	0.004
HAINet$_{2021}$	0.965	0.910	0.853	0.827	0.004
Ours	0.981	0.932	0.897	0.872	0.003

The initial weight decay is set to 5e-4, the momentum is 0.9, and the learning rate is 1e-3. When the training reaches the 18th, 34th, 43th, and 53th epoch, the learning rate is reduced to 1/10 of the original, and the image resolution is resized to 352×352 during training and testing.

B. Evaluation Metrics

We use five evaluation metrics consistent with [14], [15], which are widely used in SOD. They are E-measure (E_m), S-measure (S_m), Weighted F-measure (W_F) [16], *MAE*, and F-measure (F_m), respectively.

C. Comparison With the State-of-the-Art SOD Methods

Currently, there is no publicly available triple-modal SOD methods. For more comprehensively verify the effectiveness of the constructed dataset VDT-2048 and the proposed method, we compared thirteen state-of-the-art dual-modal SOD methods based on deep learning in the past two years, as shown in Table I. MIDD [14], MIED [32], ADFNet [33], and CGFNet [15] are RGB-T salient detection methods (green in Table I). The rest of the methods are RGB-D salient detection methods (orange in Table I), DPANet [17], JL-DCF [10], BBSNet [18], CoNet [19], UC-Net [20], HAINet [21], S2MA [34], CDNet [35], and RD3D [36]. All compared methods use the image groups of the corresponding modal in the VDT-2048 dataset to train and test. All methods are uniformly trained and tested on the same device.

Quantitative performance comparison: Table I shows the quantitative indicators of the latest thirteen salient detection methods and our proposed method on VDT-2048 dataset. It can be seen that from the Table I and Fig. 13, our proposed HWSINet has achieved better performance. Our method has 1.3% and 1.2% improvement in E_{m} and S_m indicators, and has 4.7% and 5% improvement in F_m and W_F indicators compared to the suboptimal method. The quantitative comparison indicators also prove that the RGB-D and RGB-T methods have certain limitations in dealing with challenging scenes. Tables II–IV, respectively, show the comparison of test indicators between the proposed method and the latest salient detection methods in challenging scenes with different modalities. Comparing our method with the suboptimal method: 1) It can be seen from Table II, in the challenging scenes of V-NI, V-SI, and V-LI, 13.7%, 4.5%, and 4.2% are improved on the F_m indicator, 7.1%, 4.4%, and 4% are improved on the W_F indicator. 2) It can be seen from Table III, in the challenging scenes of D-BM, 5% is improved on the F_m indicator, 6% is improved on the W_F indicator. 3) It can be seen from Table IV, in the challenging scenes of T-Cr, 6.4% is improved on the E_{m} indicator, 13.5% is improved on the F_m indicator, 7.1% is improved on the W_F indicator. The comparison results also reveal some of the drawbacks of dual-modal salient detection methods when dealing with challenging scenes. Comparison of visualization results: Fig. 14 shows the visualization results of the proposed method and the latest methods in the challenging scenario of the VDT-2048 dataset.

Overall, we can see that our method has higher robustness and superior performance in fourteen challenging scenes. It is undeniable that some dual-modal methods have also achieved competitive results when dealing with a small number of simple

TABLE II
QUANTITATIVE COMPARISON RESULTS OF RGB-D AND RGB-T SOD METHODS IN V CHALLENGING SCENES (RED REPRSENTS THE BEST, BLUE REPRESENTS THE SECOND BEST)

Models	**V-** Challenging Scenes																				
	V-SA (Similar appearance)			V-BSO (Big Salient Object)			V-SSO (Small Salient Object)			V-MSO Multiple Salient Object			V-LI (Low Illumination)			V-SI (Side Illumination)			V-NI (No Illumination)		
	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$
MIDD$_{2021}$	0.896	0.812	0.711	0.986	0.920	0.913	0.675	0.670	0.434	0.913	0.823	0.744	0.889	0.795	0.702	0.893	0.802	0.709	0.793	0.718	0.590
MIED$_{2020}$	0.828	0.756	0.612	0.971	0.888	0.857	0.412	0.535	0.173	0.854	0.770	0.655	0.810	0.739	0.595	0.807	0.745	0.596	0.698	0.669	0.477
ADFNet$_{2020}$	0.674	0.631	0.440	0.957	0.788	0.811	0.374	0.431	0.138	0.751	0.651	0.529	0.702	0.618	0.468	0.696	0.612	0.469	0.590	0.529	0.364
CGFNet$_{2021}$	0.937	0.838	0.782	0.990	0.927	0.933	0.785	0.745	0.552	0.948	0.847	0.807	0.942	0.830	0.783	0.939	0.835	0.781	0.885	0.786	0.702
JL-DCF$_{2020}$	0.896	0.733	0.714	0.985	0.890	0.921	0.599	0.577	0.354	0.919	0.755	0.741	0.891	0.696	0.704	0.888	0.693	0.705	0.773	0.485	0.532
S2MA$_{2020}$	0.839	0.763	0.629	0.984	0.914	0.898	0.596	0.636	0.345	0.882	0.789	0.692	0.820	0.723	0.602	0.825	0.732	0.605	0.683	0.577	0.448
DPANet$_{2020}$	0.924	0.801	0.748	0.989	0.929	0.925	0.686	0.654	0.441	0.899	0.784	0.723	0.875	0.760	0.686	0.897	0.765	0.714	0.728	0.583	0.492
BBS-Net$_{2020}$	0.903	0.830	0.717	0.990	0.937	0.930	0.660	0.693	0.411	0.921	0.829	0.750	0.879	0.786	0.688	0.874	0.777	0.687	0.774	0.648	0.542
CoNet$_{2020}$	0.915	0.703	0.687	0.931	0.776	0.848	0.659	0.499	0.369	0.907	0.710	0.713	0.868	0.643	0.630	0.837	0.577	0.577	0.792	0.489	0.488
UC-Net$_{2020}$	0.978	0.853	0.844	0.988	0.932	0.941	0.914	0.735	0.680	0.969	0.841	0.832	0.964	0.808	0.806	0.965	0.806	0.814	0.923	0.691	0.706
CDNet$_{2021}$	0.667	0.381	0.374	0.878	0.719	0.768	0.407	0.125	0.089	0.809	0.580	0.573	0.641	0.335	0.356	0.607	0.302	0.326	0.353	0.063	0.091
RD3D$_{2021}$	0.863	0.801	0.666	0.986	0.926	0.912	0.577	0.664	0.338	0.893	0.819	0.716	0.838	0.752	0.653	0.843	0.752	0.641	0.708	0.608	0.484
HAINet$_{2021}$	0.972	0.872	0.844	0.991	0.939	0.940	0.902	0.782	0.683	0.963	0.857	0.838	0.962	0.813	0.791	0.958	0.809	0.791	0.908	0.683	0.667
Ours	0.979	0.878	0.856	0.993	0.949	0.946	0.936	0.844	0.751	0.976	0.889	0.866	0.974	0.864	0.840	0.978	0.872	0.851	0.959	0.842	0.803

TABLE III
QUANTITATIVE COMPARISON RESULTS OF RGB-D SOD METHODS IN D CHALLENGING SCENES (RED REPRSENTS THE BEST, BLUE REPRESENTS THE SECOND BEST)

Models	**D-** Challenging Scenes															
	D-BM (Background Messy)				D-II (Information Incomplete)				D-SSO (Small Salient Object)				D-BI (Background Interference)			
	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$
JL-DCF$_{2020}$	0.848	0.686	0.648	0.006	0.908	0.767	0.754	0.007	0.599	0.577	0.354	0.002	0.856	0.710	0.666	0.005
S2MA$_{2020}$	0.830	0.757	0.633	0.005	0.881	0.811	0.714	0.006	0.596	0.636	0.345	0.002	0.811	0.754	0.601	0.005
DPANet$_{2020}$	0.745	0.691	0.546	0.007	0.836	0.776	0.686	0.007	0.687	0.654	0.441	0.002	0.891	0.786	0.706	0.004
BBS-Net$_{2020}$	0.868	0.799	0.688	0.004	0.920	0.854	0.779	0.004	0.660	0.693	0.411	0.006	0.869	0.805	0.681	0.005
CoNet$_{2020}$	0.878	0.677	0.651	0.008	0.907	0.715	0.721	0.011	0.659	0.499	0.367	0.002	0.865	0.665	0.634	0.007
UC-Net$_{2020}$	0.969	0.827	0.818	0.004	0.979	0.869	0.868	0.005	0.913	0.736	0.681	0.002	0.964	0.828	0.822	0.004
(CDNet$_{2021}$)	0.558	0.272	0.291	0.015	0.690	0.457	0.483	0.014	0.407	0.125	0.089	0.005	0.635	0.403	0.375	0.010
RD3D$_{2021}$	0.826	0.758	0.635	0.005	0.889	0.820	0.733	0.006	0.577	0.664	0.338	0.002	0.824	0.781	0.627	0.004
HAINet$_{2021}$	0.964	0.836	0.811	0.004	0.975	0.875	0.861	0.004	0.902	0.782	0.683	0.001	0.961	0.847	0.817	0.003
Ours	0.977	0.889	0.859	0.003	0.984	0.912	0.892	0.003	0.936	0.844	0.751	0.001	0.980	0.893	0.865	0.002

TABLE IV
QUANTITATIVE COMPARISON RESULTS OF RGB-T SOD METHODS IN T CHALLENGING SCENES (RED REPRSENTS THE BEST, BLUE REPRESENTS THE SECOND BEST)

Models	**T-** Challenging Scenes											
	T-Cr (Crossover)				T-RD (Radiation Dispersion)				T-HR (Heat Reflection)			
	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$
MIDD$_{2021}$	0.851	0.780	0.656	0.004	0.954	0.850	0.800	0.006	0.923	0.856	0.742	0.004
MIED$_{2020}$	0.734	0.694	0.512	0.006	0.902	0.803	0.711	0.007	0.839	0.769	0.622	0.006
ADFNet$_{2020}$	0.609	0.561	0.376	0.010	0.831	0.704	0.616	0.011	0.719	0.667	0.494	0.008
CGFNet$_{2021}$	0.915	0.825	0.748	0.003	0.979	0.871	0.854	0.005	0.968	0.875	0.829	0.003
Ours	0.974	0.884	0.849	0.002	0.990	0.899	0.891	0.004	0.990	0.913	0.890	0.003

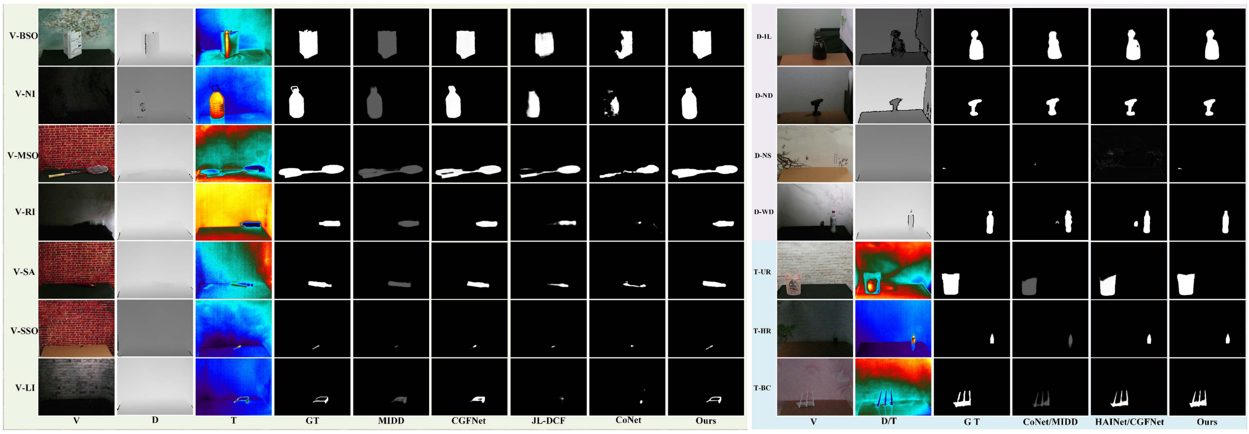

Fig. 14. Comparison of the salient map visualization results of the proposed method and the latest methods in dealing with different challenging scenes.

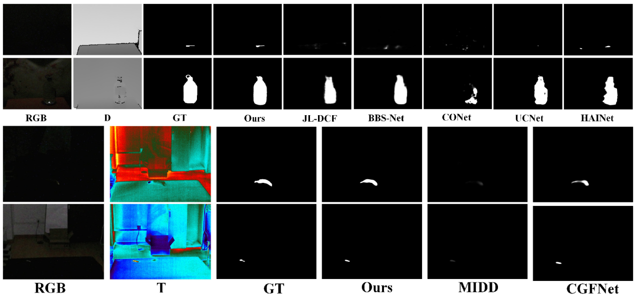

Fig. 15. Visual comparison results of two modalities are disturbed.

challenges. This is because the single-modal image is challenged, and the other modality will provide the necessary supplementary information. But, when both modalities are subject to certain challenges, the dual-modal methods will drop significantly, as shown in Fig. 15. It verifies the effectiveness of our proposed method and the constructed VDT-2048 dataset, and also confirms the importance and necessity of introducing another modality to supplement information and auxiliary detection.

D. Communication Bridge Analysis

In our proposed method, D information is employed as a triple-modal information communication bridge for VDT SOD, i.e., D-V and D-T as the input of DMAFM. In fact, both V and T images can also be used as the communication bridges. As shown in Table V, we designed experiments to verify the use of V and T as bridges and the three inputs of DMAFM without distinguishing triple-modal. It can be seen from the results of the index comparison that using different information as a bridge or constructing three DMAFMs have all achieved competitive performance. In other words, any one of the three modalities can be used as a communication bridge, which verifies the robustness of our proposed method.

TABLE V
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT CONFIGURATIONS OF DMAFM

DMAFM	Test-set				
	$E_m\uparrow$	$S_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$
V-D&V-T	0.981	0.933	0.898	0.869	0.003
T-V&T-D	0.980	0.932	0.895	0.867	0.003
V-D&V-T&T-D	0.981	0.932	0.896	0.868	0.003
D-V&D-T	0.981	0.932	0.897	0.872	0.003

TABLE VI
QUANTITATIVE COMPARISON RESULTS OF PROPOSED MODULES

Modules			Test-set				
DMAFM	TMIWM	GAWFM	$E_m\uparrow$	$S_m\uparrow$	$W\_F\uparrow$	$F_m\uparrow$	$MAE\downarrow$
			0.597	0.580	0.498	0.436	0.030
√			0.966	0.926	0.885	0.845	0.003
	√		0.968	0.930	0.890	0.856	0.003
√		√	0.969	0.927	0.886	0.840	0.003
√	√		0.968	0.931	0.891	0.851	0.004
	√	√	0.977	0.931	0.891	0.862	0.003
√	√	√	0.981	0.932	0.897	0.872	0.003

E. Ablation Analysis

In this section, we will design experiments to verify the effectiveness of the proposed modules (i.e., DMAFM, GAWEM, and TMIWM) and the contribution of different modules to network performance. √ means that the proposed method contains corresponding modules. Table VI shows the comparison results of specific indicators. Obviously, our module has significantly improved the performance of the proposed method. It also verified that our proposed method achieves the complementarity of cross-modal information and effective suppression of interference.

TABLE VII
PROPOSED METHOD IS COMPARED WITH THE COMPARATIVE METHODS IN TERMS OF RUNNING TIME AND MODEL PARAMETERS

Models	DPANet	CGFNet	JL-DCF	MIDD	BBS-Net	HAINet	Ours
Runtime (FPS)	24.7	10.98	18.26	12.89	18.61	8.41	3.59
Model Size (MB)	370.2	265.8	574.8	209.8	199.8	239.7	403.4
Params(M)	92.4	66.38	143.52	52.43	49.77	59.82	100.77
Flops (G)	58.9	345.18	861.24	216.74	31.14	181.4	357.69

F. Running Time Analysis

Table VII shows the comparison between the proposed method and the comparison methods in terms of model parameters and running time. From the experimental results, it can be seen that the running time of the proposed method is slow with the addition of multiple modules. Although the running time of the proposed method is slower than other methods, it is still competitive in terms of model size and parameter number. In addition, it should be noted that using the information from the three modalities increases the running time but improves the ability to fight challenging scenarios.

V. CONCLUSION

In this work, we investigated the SOD for the robotic grasping task from VDT images. We constructed a novel benchmark dataset, namely VDT-2048 dataset. Our dataset provided 14 challenging scenes, which can be used as a unified dataset platform to simultaneously verify the generalization performance of dual-modal SOD methods. We also proposed a new VDT SOD method, i.e., HWSI. The main motivation for the proposed method is how to effectively suppress interference from the triple-modal images. We designed the method using the idea of hierarchical and proposed three modules to use cross-modal information weighting and to suppress interference. There is no publicly available triple-modal SOD methods, so we compared with the state-of-the-art dual-modal SOD methods on VDT-2048 datasets. The proposed method achieves the best performance in quantitative and qualitative evaluations.

REFERENCES

[1] R. Li and H. Qiao, "A survey of methods and strategies for high-precision robotic grasping and assembly tasks—Some new trends," *IEEE/ASME Trans. Mechatronics*, vol. 24, no. 6, pp. 2718–2732, Dec. 2019.

[2] M. Ilyas, H. Y. Khaw, N. M. Selvaraj, Y. Jin, X. Zhao, and C. C. Cheah, "Robot-assisted object detection for construction automation: Data and information-driven approach," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 6, pp. 2845–2856, Dec. 2021.

[3] L. Chen, P. Huang, Y. Li, and Z. Meng, "Edge-dependent efficient grasp rectangle search in robotic grasp detection," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 6, pp. 2922–2931, Dec. 2021.

[4] H. Wang, B. Yang, Y. Liu, W. Chen, X. Liang, and R. Pfeifer, "Visual servoing of soft robot manipulator in constrained environments with an adaptive controller," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 1, pp. 41–50, Feb. 2017.

[5] X. Liu, W. Chen, H. Madhusudanan, L. Du, and Y. Sun, "Camera orientation optimization in stereo vision systems for low measurement error," *IEEE/ASME Trans. Mechatronics*, vol. 26, no. 2, pp. 1178–1182, Apr. 2021.

[6] D. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[7] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[8] L. Huang, K. Song, J. Wang, M. Niu, and Y. Yan, "Multi-graph fusion and learning for RGBT image saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1366–1377, Mar. 2022.

[9] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[10] K. Fu, D. Fan, G. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection," in *Proc. CVPR*, 2020, pp. 3049–3059.

[11] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGBT salient object detection via fusing multi-level cnn features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.

[12] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. CVPR*, 2019, pp. 3085–3094.

[13] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, May 2021.

[14] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for RGB-thermal salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5678–5691, Jun. 2021.

[15] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-guided fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2949–2961, May 2022.

[16] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. CVPR*, 2014, pp. 248–255.

[17] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection," *IEEE Trans. Image Process .*, vol. 30, pp. 7012–7024, 2021.

[18] D. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. ECCV*, 2020, pp. 275–292.

[19] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proc. ECCV*, 2020, pp. 52–69.

[20] J. Zhang et al., "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. CVPR*, 2020, pp. 8579–8588.

[21] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, 2021.

[22] Y. Choi et al., "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.

[23] I. R. Spremolla et al., "RGB-D and thermal sensor fusion-application in person tracking," in *Proc. 11th Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2016, pp. 610–617.

[24] L. Susperregi et al., "RGB-D, laser and thermal sensor fusion for people following in a mobile robot," *Int. J. Adv. Robot. Syst.*, vol. 10, no. 271, pp. 1–9, 2013.

[25] K. Skala, T. Lipic, I. Sovic, L. Gjenero, and I. Grubisˇic, "4d thermal imaging system for medical applications," *Periodicum Biologorum*, vol. 113, no. 4, pp. 407–416, 2011.

[26] R. Irani et al., "Spatiotemporal analysis of RGB-D-T facial images for multimodal pain level recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 88–95.

[27] J. Rangel, J. Garzón, J. Sofrony, and A. Kroll, "Gas leak inspection using thermal, visual and depth images and a depth-enhanced gas detection strategy," *Revista de Ingeniería*, no. 42, pp. 8–15, 2015.

[28] C. Palmero, A. Clapés, C. Bahnsen, A. Møgelmose, T. B. Moeslund, and S. Escalera, "Multi-modal RGB–depth–thermal human body segmentation," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 217–239, 2016.

[29] Y. Cao et al., "Multi-sensor spatial augmented reality for visualizing the invisible thermal information of 3D objects," *Opt. Lasers Eng.*, vol. 145, pp. 106634–106641, 2021.

[30] C. Bahnsen, "Thermal-visible-depth image registration," M.S. thesis, Aalborg Univ., Aalborg, Denmark, 2013.

[31] N. Kim et al., "Geometrical calibration of multispectral calibration," in *Proc. 12th Int. Conf. Ubiquitous Robots Ambient Intell.*, 2015, pp. 384–385.

[32] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive encoder-decoder network for RGBT salient object detection," 2020, *arXiv:2005.02315vl*.

[33] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT salient object detection: A large-scale dataset and benchmark," *IEEE Trans. Multimedia*, 2022, doi: 10.1109/TMM.2022.3171688.

[34] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13756–13756.

[35] W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "CDNet: Complementary depth network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3376–3390, Mar. 2021.

[36] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, doi: 10.1609/aaai.v35i2.16191.

[37] H. Tian, K. Song, S. Li, S. Ma, and Y. Yan, "Light-weight Pixel-wise generative robot grasping detection based on RGB-D dense fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, Aug. 2022, Art. no. 5017912.

[38] J. Wang, K. Song, Y. Bao, Y. Yan, and Y. Han, "Unidirectional RGB-T salient object detection with intertwined driving of encoding and fusion," *Eng. Appl. Artif. Intell.*, vol. 114, 2022, Art. no. 105162.

[39] S. Ma, K. Song, H. Dong, H. Tian, and Y. Yan, "Modal complementary fusion network for RGB-T salient object detection," *Appl. Intell.*, 2022, doi: 10.1007/s10489-022-03950-1.

[40] W. Zhou, Q. Guo, J. Lei, L. Yu, and J. -N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.

[41] F. Huo, X. Zhu, Q. Zhang, Z. Liu, and W. Yu, "Real-Time one-stream semantic-guided refinement network for RGB-thermal salient object detection," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2512512.

Kechen Song (Member, IEEE) received the B.S. degree in mechanical manufacturing and automation, the M.S. degree in mechanical design and theory, and the Ph.D. degree in mechanical design and theory from the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, in 2009, 2011, and 2014, respectively.

During 2018–2019, he was an Academic Visitor with the Department of Computer Science, Loughborough University, U.K. He is currently an Associate Professor with the School of Mechanical Engineering and Automation, Northeastern University.

Jie Wang received the B.S. degree in mechanical engineering from the School of the North University of China, Taiyuan, China, in 2019, and the M.S. degree in mechanical engineering from the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, in 2022. He is currently working toward the Ph.D. degree in computer science, Tianjin University, Tianjin, China.

Yanqi Bao received the B.S. degree in mechanical manufacturing and automation from the School of Mechanical Engineering and Automation, North University of China, Taiyuan, China, in 2019, and the M.S. degree in mechanical engineering from the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, in 2022. He is currently working toward the Ph.D. degree in computer science, Nanjing University, Nanjing, China.

Liming Huang received the B.S. degree in mechatronic engineering from the Shandong University of Science and Technology, Qingdao, China, in 2017, and the M.S. degree in mechanical engineering from the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, in 2021. He is currently working toward the Ph.D. degree in computer science, University of Exeter, Exeter, U.K.

Yunhui Yan received the B.S. degree in mechanical manufacturing and automation, the M.S. degree in mechanical design and theory, and the Ph.D. degree in mechanical design and theory from the School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China, in 1981, 1985, and 1997, respectively.

Since 1982, he has been a Teacher with Northeastern University, and became as a Professor in 1997. From 1993 to 1994, he stayed as a Visiting Scholar with the Tohoku National Industrial Research Institute, Sendai, Japan.