Fusion of infrared and visible images for remote detection of low-altitude slow-speed small targets.

SUN, H., LIU, Q., WANG, J., REN, J., WU, Y., ZHAO, H. and LI, H.

2021

Fusion of Infrared and Visible Images for Remote Detection of Low-Altitude Slow-Speed Small Targets

Haijiang Sun, Qiaoyuan Liu, Jiacheng Wang, Jinchang Ren, Yanfeng Wu, Huimin Zhao, and Huakang Li

***Abstract*—Detection of the low-altitude and slow-speed small (LSS) targets is one of the most popular research topics in remote sensing. Despite of a few existing approaches, there is still an accuracy gap for satisfying the practical needs. As the LSS targets are too small to extract useful features, deep learning based algorithms can hardly be used. To this end, we propose in this article an effective strategy for determining the region of interest, using a multiscale layered image fusion method to extract the most representative information for LSS-target detection. In addition, an improved self-balanced sensitivity segment model is proposed to detect the fused LSS target, which can further improve both the detection accuracy and the computational efficiency. We conduct extensive ablation studies to validate the efficacy of the proposed LSS-target detection method on three public datasets and three self-collected datasets. The superior performance over the state of the arts has fully demonstrated the efficacy of the proposed approach.**

***Index Terms*—Background subtraction, image fusion, low-altitude and slow-speed small (LSS) target detection, saliency detection.**

I. Introduction

THE "low-altitude and slow-speed small" target (LSS target), such as unmanned aerial vehicles (UAVs), is a general term for small aviation device/equipment with a flight altitude less than 2 km and a flight speed less than 50 km/h. Detection of the LSS targets is a key technology for precision navigation, infrared search, tracking and reconnaissance warning system, which has been widely applied in many practical applications such as air transportation [1], aerial mapping [2], and many tasks in military confrontation [3]. As a result, the LSS-target detection indeed has important research significance in the field of ground air defense. This approach could directly determine the operating distance and detection sensitivity of the corresponding systems. Although great breakthroughs have been made in recent years [4]–[7], [40], [41], including the background subtraction algorithm based on the visible light image [6], detecting the LSS target with the combination of visible light image, infrared image and even hyperspectral image [8], but there are still great challenges for putting this approach into practical applications.

Fig. 1. LSS targets in different aviation scenarios.

As the LSS targets usually appear like a spot (see Fig. 1), the detection can be easily affected by the background objects. In this article, we summarize the difficulties of the LSS-target detection into three aspects. The first is the complex background, and how to accurately detect the LSS targets while overcoming the effects of complex scenes is always a challenge in this context. The second is the weak target, as it is very hard to determine the specific trajectories for LSS targets under low speeds. Due to their small sizes in the field of view, it is difficult to build accurate templates for target detection. The third is the high false alarm rate: considering that there always exist faked objects with similar features to the LSS targets, how to accurately differ the LSS targets from complex scenes to reduce the false alarms is another major issue.

To tackle the aforementioned challenging problems, various methods have been proposed for LSS-target detection. According to the image sources used, these approaches can be divided into two categories, i.e., LSS-target detection in the visible light image and infrared images, respectively. Lou *et al.* [9] introduced the saliency and regional stability for feature extraction in visible-light images. A segmentation threshold was used to distinguish the target for accurate detection, but it was unsuitable for complex scenes. Xie *et al.* [10] proposed the peer group filter to improve the signal-to-noise ratio of the infrared image in LSS-target detection. Although small targets could be extracted

Manuscript received September 12, 2020; revised October 24, 2020, December 3, 2020, January 24, 2021, and February 16, 2021; accepted February 19, 2021. Date of publication February 24, 2021; date of current version March 17, 2021. This work was supported in part by the Key Laboratory of Airborne Optical Imaging and Measurement, Chinese Academy of Sciences and International Cooperation Project of Changchun Institute of Optics, Fine Mechanics and Physics under Grant Y9U933T190, in part by the Dazhi Scholarship of the Guangdong Polytechnic Normal University, National Natural Science Foundation of China under Grant 62072122, and in part by the Education Department of Guangdong Province under Grant 2019KSYS009. *(Haijiang Sun and Qiaoyuan Liu contributed equally to this work.)* (*Corresponding authors: Jinchang Ren; Huimin Zhao*.)

Haijiang Sun, Qiaoyuan Liu, and Jiacheng Wang are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China (e-mail: sunhaijiang@126.com; liuqy@ciomp.ac.cn; wjcforward@126.com).

Jinchang Ren is with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China, and also with the National Subsea Centre, Robert Gordon University, AB10 7AQ Aberdeen, U.K. (e-mail: jinchang.ren@ieee.org).

Yanfeng Wu is with the 28th Research Institute of China Electronics Technology Group, Nanjing 210001, China (e-mail: wuyfciomp@yahoo.com).

Huimin Zhao and Huakang Li are with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China (e-mail: zhaohuimin@gpnu.edu.cn; 865020993@qq.com).

Digital Object Identifier 10.1109/JSTARS.2021.3061496

from the background accurately, it was too time-consuming for real-time tasks. Pan *et al.* [37] proposed a double-layer local contrast measure (DLCM) approach for small target detection in infrared images. A double-layer-diagonal gray difference contrast was used to enhance the visual saliency of the target and alleviate the impact of the background clutter and noise.

In general, the images captured by the visible cameras show clear texture, edge, and high spatial resolution than the infrared ones, which are useful for target detection. However, the quality of the image produced can be affected by the weather conditions and environmental illumination *et al.*, leading to failure of detection at night or during bad weather conditions. However, the infrared imaging can supplement the deficiency of visible images. Although the infrared image suffers from a relatively lower spatial resolution and weaker textures, it can produce particularly better images at night or in bad weather.

In order to enhance the generality of LSS-target detection, the combination of both the visible images and the infrared images has been widely used. Bavirisetti and Dhuli [11] proposed image fusion and secondary decomposition of the image-based LSS-target detection. Qiu *et al.* [12] applied combined strategies on the visible and infrared images for accurate target extraction, in which an inter-frame difference based rough entropy model was used for locating the moving target in the infrared images, along with a local binary pattern model for the visible images and a Gaussian mixture model for background modeling. Although these have improved the accuracy and generality of the algorithm, the information of visible and infrared images was not fully utilized.

To tackle the effect of the complex background in LSS-target detection, various background modeling approaches have been proposed, either in the visible light images or the infrared images. Barnich and Van Droogenbroeck [13] proposed an unordered background modeling algorithm, which stored a sample set for all pixels and randomly selected the neighboring pixels for updating the samples. This method was robust to noise but could hardly deal with color changes such as illumination and ghost effect. The approach was further improved by Hofmann *et al.* in [14], where a pixel level nonparametric background-modeling algorithm was proposed. Although it could adaptively detect small targets in complex background, the algorithm was quite complex and time-consuming.

In this article, combining infrared and visible images for LSS-target detection is also focused. Different from the existing methods, our proposed approach fuses the features and information from visible and infrared images in multilayers. First, a background model is used from the perspective of inter-frame difference from three consecutive frames rather than two, which can help to obtain the accurate motion region while greatly decreasing the impact of background. Second, the weighted moving average (WMA) [15] and weighted moving variance (WMV) [16] are utilized to obtain the target candidate regions from the infrared and visible images, respectively. Third, based on the target candidate regions, a self-balanced sensitivity segmentation algorithm (SuBSENSE) [17] is used to construct a local area background model for refined LSS-target detection. The Toet A.TNO (Image fusion dataset) dataset [18] and the visible-infrared database (VID) dataset [19] are used for evaluating the performance of image fusion and target detection, respectively. Experimental results have validated the efficacy of the proposed model for LSS-target detection in comparison to the state-of-the-art (SOTA) approaches.

II. Proposed Algorithm

The diagram of the proposed LSS-target detection approach is illustrated in Fig. 2, which has two modules for fusion-based feature extraction. The region of interest (ROI) extraction module is designed to predict the candidate target region with multidetector images. Given three frames of the infrared images and visible images, WMA and WMV can output a candidate target region, denoted as rec1 and rec2, respectively. The minimum bounding box of these two regions can be taken as the extracted ROI that carries target information from both infrared and visible images. In the multiscale layered image fusion module, a fused image is generated for LSS-target detection by saliency maps from the extracted base layers and detail layers. The detected ROIs and the fused image are then inputted to an improved SubSENSE module for further refining the locations of the LSS targets.

Specifically, we first apply the background subtraction to the first three frames of the visible and its corresponding infrared image sequences to determine the initial candidate regions of targets. Here, we denote the first three frames of the infrared image and the visible images as $I_{\mathrm{IR}} = \{I_{\mathrm{IR\_1}},\ I_{\mathrm{IR\_2}},\ I_{\mathrm{IR\_3}}\}$ and $I_{\mathrm{Vis}} = \{I_{\mathrm{Vis\_1}},\ I_{\mathrm{Vis\_2}},\ I_{\mathrm{Vis\_3}}\}$, respectively. By applying the WMA [15] and WMV [16] methods on I_{IR} and I_{Vis}, a rough foreground and background segmentation can be obtained, which can provide a good guidance for the subsequent target detection. Specifically, by applying the WMA on I_{IR}, several foreground masks can be obtained, where we use $\mathrm{rec1}$ to denote the bounding box of the targets. Similarly, a bounding box $\mathrm{rec2}$ of detected targets from the visible images can be determined from by applying WMV on the visible images. The final ROI of the target candidate region *R* can be obtained via a union of two sets rec1 and rec2 as follows, and the details are presented in Section III:

$$\mathrm{rec1} = \mathrm{WMA}\left(I_{\mathrm{IR}}\right) \tag{1}$$

$$\mathrm{rec2} = \mathrm{WMV}\left(I_{\mathrm{Vis}}\right) \tag{2}$$

$$R = \mathrm{rec1} \cup \mathrm{rec2}. \tag{3}$$

The target candidate region generated by the background subtraction model will be applied to each subsequent fusion image. In this way, the detection algorithm SubSENSE [17] only needs to be applied to the determined ROI rather than the whole image for improved efficiency.

In our approach, image fusion is the key to make the full use of infrared and visible images. First, the rolling guide filter (RGF) is applied to the infrared and visible images, separately, in order to decompose the images and enhance the details. This is followed by the visual saliency to fuse the decomposed image

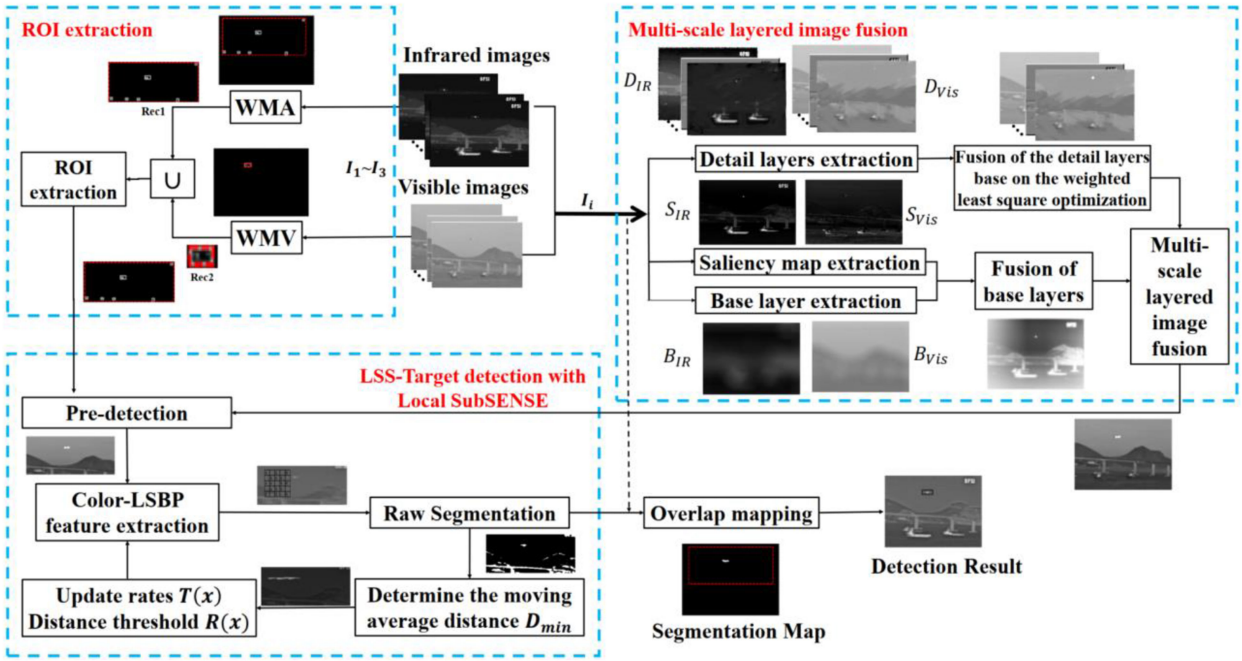


Fig. 2. Flowchart of the proposed LSS-target detection approach.

layer as the input of SubSENSE [17] detection method

$$\{D_{\mathrm{IR}}, B_{\mathrm{IR}}\} = \mathrm{RGF}\left(I_{\mathrm{IR\_}i}\right),\ D_{\mathrm{IR}} = \left\{d_{\mathrm{IR}}^{1}, d_{\mathrm{IR}}^{2}, \ldots d_{\mathrm{IR}}^{N}\right\} \tag{4}$$

$$\{D_{\mathrm{Vis}}, B_{\mathrm{Vis}}\} = \mathrm{RGF}\left(I_{\mathrm{Vis\_}i}\right),$$

$$D_{\mathrm{Vis}} = \left\{d_{\mathrm{Vis}}^{1}, d_{\mathrm{Vis}}^{2}, \ldots d_{\mathrm{Vis}}^{N}\right\} \tag{5}$$

$$\{S_{\mathrm{IR\_}i}, S_{\mathrm{Vis\_}i}\} = \mathrm{Saliency}\left(I_{\mathrm{IR\_}i}, I_{\mathrm{Vis\_}i}\right) \tag{6}$$

where D_{IR} and B_{IR} denote, respectively, the extracted N detailed layers from the infrared image and base layer; D_{Vis} and B_{Vis} denote, respectively, the extracted N detailed layers and base layer from the visible image. As shown in Fig. 2, all the extracted detail layers D_{IR} and D_{Vis} are fused according to the weighted least square optimization, and all the extracted base layers B_{IR} and B_{Vis} combined with the corresponding saliency map S_{IR} and S_{Vis} are fused. Combining all the fused images as shown in (25), we can get the final fused image F_i.

Based on the background subtraction model R and the fused image F_i, the position of the LSS target in the current frame can be predetected, which is then further refined by an improved SubSENSE [17] method as detailed in Section V

$$l = \mathrm{Sub}\left(R, F_i\right). \tag{7}$$

Finally, the refined mask l is mapped to the original image to produce the segmentation map and the final detection results.

III. Extraction of Candidate Target Regions

For accurate detection of the moving targets, we first introduce a three-frame image difference algorithm to extract the candidate regions of the targets, in which the adjacent three frames are taken as a group for calculating the image difference for robustness. Considering the low speed of the LSS target, only the initial three frames are used for background modeling.

A. ROI Extraction in Infrared Images

For the infrared images, the WMA algorithm is employed for candidate target region extraction, which is more suitable for processing infrared images due to its fast calculation speed and sensitivity to illumination changes. WMA uses the weighted average of the initial three infrared frames to build a background subtraction model by replaying more weights on the observation values closer to the predicted time

$$I_{\mathrm{mean}} = \sum_{i=1}^{3} w_i * I_i \tag{8}$$

where I_{mean} is the weighted average of the image pixels, w_i is the corresponding weight with $\sum_{i=1}^{3} w_i = 1$; in our approach, we set $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$ as an example.

As infrared images conform to the laws of thermodynamics, the 1-D information entropy is used as a threshold to distinguish the foreground from the background as follows:

$$H_b = -\sum_{i=0}^{\delta} \frac{p_i}{W_1(\delta)} \times \log\left(\frac{p_i}{W_1(\delta)}\right) \tag{9}$$

$$H_f = -\sum_{i\,=\,\delta+1}^{L-1} \frac{p_i}{W_2(\delta)} \times \log\left(\frac{p_i}{W_2(\delta)}\right) \tag{10}$$

where δ is the threshold to distinguish the gray values of the background before and after; W_1 and W_2 are the gray value probability of the background and the target, respectively, and L is the number of gray levels of the infrared image. p_i is

the probability distribution of image gray levels, which can be computed by $p_i = \frac{n_i}{M \times N}$, where $M \times N$ is the total number of pixels, and n_i indicates the number of pixels with gray level i.

The segmentation threshold can then be determined by

$$\begin{cases} H_t\ (\delta) = H_b\ + H_f \\ \delta^* = \ \mathrm{argmax} H_t\,(\delta) \end{cases} \tag{11}$$

where $H_t(\delta)$ is the total information entropy of the image, and δ^* is the determined segmentation threshold by maximizing the 1-D information entropy of the image [38].

For detecting LSS targets, first pixel-based difference between the current frame I_i and the corresponding weighted average image I_{mean} is obtained as I_{sub}. By comparing I_{sub} against the threshold δ^*, the foreground can be determined by

$$B_t = \begin{cases} 1,\ I_{\mathrm{sub}} > \delta^* \\ 0,\ I_{\mathrm{sub}} \le \delta^* \end{cases},\ \ I_{\mathrm{sub}} = |I_i - I_{\mathrm{mean}}|. \tag{12}$$

The algorithm can then dynamically update the background model based on the weighted background model and the current frame to adapt the model to scene changes as follows:

$$B_{t+1} = (1-\alpha)\ B_t + \alpha I_t \tag{13}$$

where α is the update learning rate, and I_t is the tth frame; B_t and B_{t+1} denote the background model at time t and $t+1$, respectively.

Although the foreground can be extracted from the infrared image above, it is relatively too rough to estimate the final target location. However, it helps to provide a coarse location of the target for efficiency. As shown in Fig. 2, the binary image outputted by WMA, the white area represents the extracted foreground, and the remaining represents the background. We first determine the bounding box for all the white areas as rec1, which is the candidate target region of the infrared image corresponding to the current background model.

B. ROI Extraction in Visible Images

For the visible images, based on the original WMA background modeling, the mean I_{mean} and the standard deviation is calculated. As the visible images always contain rich texture features, this is namely WMV background modeling

$$\mathrm{var}\ (I_i, w_i) = w_i\ * (I_i - I_{\mathrm{mean}})^2 \tag{14}$$

$$\mathrm{std}\ = \sqrt{\sum\nolimits_{i\in[13]} \mathrm{var}\,(I_i, w_i)\,/3} \tag{15}$$

where $\mathrm{var}(.)$ and std are the weighted variance and weighted standard deviation of the image pixels, respectively. Besides, other background modeling processes are as same as the infrared images, and w_i follows the same distribution as the WMA does. Therefore, the candidate target region rec2 can also be generated from the visible image, where the final target candidate region could be obtained using (3).

IV. Multilayer Image Fusion Based on Rolling Guided Filtering

Considering that the infrared images and the visible images contain different but supplementary information, our approach proposes to fuse them in multilayers to make the full use of their information. In our approach, the RGF [20] is used for multiscale decomposition and detail enhancement. Specifically, the visible and infrared images are first decomposed into the corresponding base layer and detail layers, which are then fused separately. Accordingly, this could not only retain and enhance more useful image details but also enable a favorable detection of LSS targets.

A. Extraction and Fusion for the Base Layers

The RGF-based decomposition is an iterative process [39], which can decompose an image into *N* layers, where the filtered detail image of the *j*th layer is given by

$$\mu^j = \ \mathrm{RGF}\left(\mu^{j-1}, \sigma_s^{j-1}, \sigma_r, T\right) \tag{16}$$

$$d^j = \mu^{j-1}\ - \mu^j \tag{17}$$

where μ^j is the image after the *j*th layer filtering, d^j is the decomposing images of the *j*th layer, σ_s^{j-1} is the scale parameter, σ_r is the weight range parameter, and *T* is the number of iterations.

Nevertheless, a Gaussian filter rather than RGF is applied in our approach to obtain the base-layer image. The RGF algorithm utilized in our approach lies in the advantages of its scale-aware and edge-preserving properties. On the contrary, the base layer is the coarsest version of the source image, which is used to control the global contrast and appearance of the fused image. As a result, it is unnecessary to apply RGF to the base layer as there is no need to preserve edge or detail information for this layer. For efficiency, the Gaussian filtering with a larger standard deviation is chosen in our method to get the base layer in the last filter layer with $j = N$ as follows:

$$\mu^N = \ \mathrm{Gaussian}\left(\mu^{N-1}, \sigma_s^{N-1}, \sigma_r, T\right) \tag{18}$$

$$d^N = \mu^{N-1}\ - \mu^N. \tag{19}$$

By setting $\sigma_s^N = \ 2\sigma_s^{N-1}$, the infrared base-layer image B_{IR} and the visible base-layer image B_{Vis} can be obtained. For fusion of the base layer, the fusion rules are given as follows:

$$\begin{cases} B_F = W_b\ B_{\mathrm{IR}} + (1 - W_b)\ B_{\mathrm{Vis}} \\ W_b = \frac{1}{2}\ (1 + S_{\mathrm{IR}} + S_{\mathrm{Vis}}) \end{cases} \tag{20}$$

where S_{IR} and S_{Vis} are the normalized saliency images generated from I_{IR} and I_{Vis}, respectively, using the widely used saliency map extraction method visual saliency map [39]. With the fusion weight W_b calculated by S_{IR} and S_{Vis}, the base layer fusion image can be finally determined.

B. Extraction and Fusion for the Detail Layers

For the fusion of the detail layers, the fusion rules are as follows. First, the initial fusion detail layer M^j is obtained as

$$M^j = W^j\ d_{\mathrm{IR}}^j + \left(1 - W^j\right) d_{\mathrm{Vis}}^j \tag{21}$$

where the fusion weight can be obtained using

$$W^j = \begin{cases} 1\ \left| d_{\mathrm{IR}}^j > d_{\mathrm{Vis}}^j \right| \\ 0\ \text{otherwise.} \end{cases} \tag{22}$$

The weighted least square method is used to obtain the final detail fusion layer D_i^j as follows:

$$\min \sum_n \left[\left(D_n^j - M_n^j\right)^2 + \lambda \cdot \Lambda \cdot \left(D_n^j - \left(d_{\mathrm{Vis}}^j\right)_n\right)^2 \right] \quad (23)$$

where $(D_n^{\mathrm{j}} - M_n^{\mathrm{j}})^2$ is the Euclidean distance between the fusion detail layer D_n^{j} and the initial fusion detail layer M_n^{j}, n denotes the spatial location of a pixel, λ is a balance control parameter, $\Lambda = (|\sum_{\omega_n} (d_{\mathrm{IR}}^j)_n| + 0.0001)^{-1}$ is the spatial-varying weight parameter, and ω_n represents the square sliding window centered at pixel *n*.

A larger window would not only blur the fused image but also generate a higher dimensional spatial-varying weight, resulting in increased computational cost. On the other hand, a smaller window cannot effectively remove the noise and irrelevant IR details. After testing, the size of 7×7 is used, as it produces satisfactory results, and (23) can be rewritten as

$$\left(1 + \lambda A^j\right)\ D^j = M^j\ A^j d^j \quad (24)$$

where A^j is a diagonal matrix containing the weights of all pixels. Finally, the fused image *F* is obtained by

$$F\ = B_F\ + D^1 + D^2 + \cdots + D^N. \quad (25)$$

V. LSS-Target Detection With Local SuBSENSE

To extract LSS targets from the candidate target regions of the fused image, we integrate an improved SuBSENSE algorithm [17] with a proposed background modeling in local regions. The input of the background model is the fused image, and the modeling area is the comprehensive target candidate areas as determined in Section III. The entire detection process consists of five steps, i.e., background model initialization, recursive moving average distance calculation, noise suppression, local distance threshold control, and background model update. Besides, a false alarm eliminating algorithm is also introduced for further improving the detection accuracy.

A. Improved SubSENSE Detection Algorithm

Step 1: Based on the spatio–temporal binary similarity and the Color-LBSP descriptor [22], SuBSENSE could characterize the pixel representation with a nonparameter model. We initialize the background model by

$$S_t\ (x) = \begin{cases} 1, & if\ \{\mathrm{dist}\,(F_t\,(x)\,, B\,(x)) < R_{\max}\} < m \\ 0, & \text{otherwise} \end{cases} \quad (26)$$

where $F_t(x)$ is the fused image at time *t*, $B(x)$ is the historical sample, the element in the background model $S_t(x)$ actually is the segmentation result, and $R_{\max}$ is the maximum distance threshold; m is the requested minimum number of matches for background classification; $\mathrm{dist}(F_t(x), B(x))$ is the distance between the current observation and the given background. By reasonably selecting the maximum distance threshold, we can better resist unrelated changes of the model.

Step 2: Once the background model is determined, the recursive moving average distance between the current pixel and the pixel in the sample set can be calculated as follows:

$$D_{\min}\,(x) = D_{\min}\,(x) \cdot (1 - \alpha_2) + d_t\,(x) \cdot \alpha_2 \quad (27)$$

where α_2 is the learning rate and $d_t(x)$ is the minimum normalized Color-LBSP between all samples in $F_t(x)$ and $B(x)$. Since $D_{\min}(\mathrm{x})$ is bound to [0, 1], an entirely static background region would have $D_{\min}(x) \approx 0$, and a dynamic region to which the model cannot adapt to would have $D_{\min}(x) \approx 1$. Areas with foreground objects show higher $D_{\min}$ values, because foreground detection is defined by the difference between the pixel model and local observations.

Step 3: Before applying local distance threshold control, we define a 2-D map of pixel-level accumulators v to improve the detection accuracy. For every new segmented frame, we can obtain the binary map of all blinking pixels at time *t* from the background model S_t. Finally, we update v as follows:

$$v\ (x) = \begin{cases} v\,(x) + 1, & \text{if } S_t\,(x) = 1 \\ v\,(x) - 0.1, & \text{otherwise.} \end{cases} \quad (28)$$

For regions with little labeling noise, we will typically have $v(x)\ \approx\ 0$, whereas for regions with unstable labeling, we have large positive $v(x)$ values. This method can guide the dynamic motion background to trigger the feedback mechanism, whereas conventional models cannot provide such feedback.

Step 4: In order to dynamically select the suitable thresholds, the local distance in the SuBSENSE [17] background model is used. The local distance thresholds R can be recursively adjusted for each new frame as shown in the following equation:

$$R\ (x) = \begin{cases} R\,(x) + v\,(x)\,, & \text{if } R\,(x) < (1 + 2 \cdot D_{\min}\,(x))^2 \\ R\,(x) - \frac{1}{v(x)}, & \text{otherwise} \end{cases} \quad (29)$$

where $R(x)$ are a set of continuous values; the exponential relation between $R(x)$ and $D_{\min}(x)$ is chosen over a linear relation as it favors sensitive behavior in static regions (and thus helps to generate sparse segmentation noise), but also provides robust and rapidly scaling thresholds elsewhere.

Here, the segmentation noise indicator $v(x)$ is used as a factor that, in dynamic regions, allows faster threshold increments and can even freeze $R(x)$ in place when $D_{\min}(x)$ recedes to lower values. This parameter control method is conducive to generating sparse segmentation noise and helps to provide reliable and rapidly expanding thresholds elsewhere.

Step 5: To overcome the influence of lighting, shadows, and moving targets on the detection results, the SuBSENSE background model is further updated as follows, where $T(x)$ represents the model update rate:

$$T\,(x) = \begin{cases} T\,(x) + (v\,(x) \cdot D_{\min}\,(x))^{-1}, & S_t\,(x) = 1 \\ T\,(x) - v\,(x)\,(D_{\min}\,(x))^{-1}, & S_t\,(x) = 0. \end{cases} \quad (30)$$

If the current pixel is classified as a background, then any random pixel in the background sample has $1/T$ probability of being replaced by the current pixel, and its neighboring pixels have $1/T$ probability of being randomly replaced by any value in its neighborhood sample. Random replacement of samples and random updates of pixels can prevent static foreground objects from being quickly absorbed and ensure the authenticity

of short-term and long-term background representations in the model. The spatial consistency of the background model can suppress the impact of camera shake on LSS-target detection.

B. False Alarm Eliminating Algorithm

In our approach, four evaluation indexes are used to reduce the false alarms, including the area ratio change, the centroid distance, the filling rate difference, and the aspect ratio difference, as detailed in the following:

1) to automatically obtain the smallest bounding boxes of all candidate targets;
2) calculate the target area ratio

$$\Delta s = \left| \frac{|t|}{K(x,y)} \right| \tag{31}$$

3) calculate the centroid distance

$$\Delta c = \left\| \frac{c_x}{b_w} + \frac{c_y}{b_n} \right\|^2 \tag{32}$$

4) calculate the aspect ratio

$$\Delta a = b_w / b_n \tag{33}$$

where $|t| = \sum K(x,y)$ denotes the number of pixels of the target in the image; $K(x,y)$ represents the area of the bounding box of the candidate target; and b_w and b_n denote, respectively, the width and height of that bounding box.

Specifically, the aspect ratio reflects the morphological state of the target to some extent as the ideal LSS target appears like a spot, i.e., the aspect ratio is close to 1. If the width or height are much disparity to each other, the suspected area becomes a long strip, which will unlikely be determined as an LSS target even if it meets other constraints. Therefore, it will be classified as a false target to be eliminated. In our approach, the aspect ratio that can tolerate false targets is set to 0.35. Assume the set R contains m candidate targets $\{r_1, r_2, \ldots r_m\}$, and T_f is the target set after removing the false alarms. The specific process of eliminating the false targets is given as follows.

False Alarms Eliminating Algorithm

Input: Candidate target set R
1. $t_i = \{r_i\}, T_i = \{\emptyset\}$ Δ Initialization
2. For i $= 1, r_i \in R$
3. If $\forall S_i : \begin{cases} 0.8 \le \Delta s < 1 \\ 0.75 \le \Delta c < 1 \\ \Delta a \ge 0.35 \end{cases}$ Then
4. $T_{new} = T_{exist} \cup t_i$ Δ Update target Set
5. Else $T_{exist} = T_f$ Δ Eliminate false alarms
End for
Output: Target set T_f

VI. Experimental Results and Analysis

In this article, three groups of multisource images are selected for performance assessment. Video 1 is the "Take 1 sequence" of the "Guanabara Bay: Outdoor" in the Toet A.TNO (Image fusion dataset) dataset [18], which is referred to as Guanabara Bay in this article. Video 2 is the "fire_sequence" of the "DHV" in the Toet A.TNO dataset, denoted as DHV in this article. Video 3 is the "Duine_Image" of "FEL_Image" in the VID dataset [19], which is referred to as FEL in this article. The specific information of the test datasets in terms of the camera scene and image resolution is displayed in Table I.

TABLE I
Test Video Information

Test Video Set	Scenes	Frame size
Guanabara Bay	Forest fire	460×380
DHV	Field surveillance	360×270
FEL	Port surveillance	720×480

TABLE II
Self-Acquired Video Information

Test Video Set	V1	V2	V3
Frame Size	320×240	480×320	480×640

Due to the particularity of the LSS-target detection problem, there are rare suitable datasets publicly available. In order to enhance the persuasiveness of the proposed method, three self-acquired datasets are added, which are named as V1, V2, and V3 in this article. All the self-acquired videos are focused on the UVA surveillance, and the specific information in terms of image resolution and image resolution is listed in Table II.

For verifying the efficacy and robustness of the proposed approach, we have collected our own datasets from UAVs under various challenging scenarios, such as small-scale, fast-moving, similar background, dark illumination, complex background, etc. Among them, the dataset V1 contains small-scale targets, and V2 is for targets of similar background and fast-moving. As can be seen, the drone in V2 has similar intensity values to the surrounding pixels, whereas the target is blurred due to the rapid rotational movement of its propeller, leading to difficulty for target detection. As for the V3 dataset, it is mainly for dark illumination and complex background, where the trees and the monitoring machines in the background may introduce false targets and extra difficulty for accurate target detection.

The parameter settings of our method are given below. The number of decomposition levels is typically set as $N = 4$, which is the same as most of other fusion methods [39] and has produced satisfactory fusion results. The initial spatial weight is set as $\sigma_s^0 = 2$. Generally, the value of λ is in the range of [0.005, 0.02], and we set $k = 0.01$ in this article.

The experimental results of the extracted candidate targets are shown in Fig. 3. As shown in Fig. 3(a), the target is obviously visible in the visible image but invisible in the infrared image. In Fig. 3(b), the visibility of the target is lower in the visible image but more obvious in the infrared image. In Fig. 3(c), the target has a rough outline in the visible image and more obvious in the infrared image. However, only aircraft engines with high heat can be observed in infrared images, and image information of other parts seems missing. The corresponding results of the self-acquired result are shown in Fig. 3(d)–(f). The

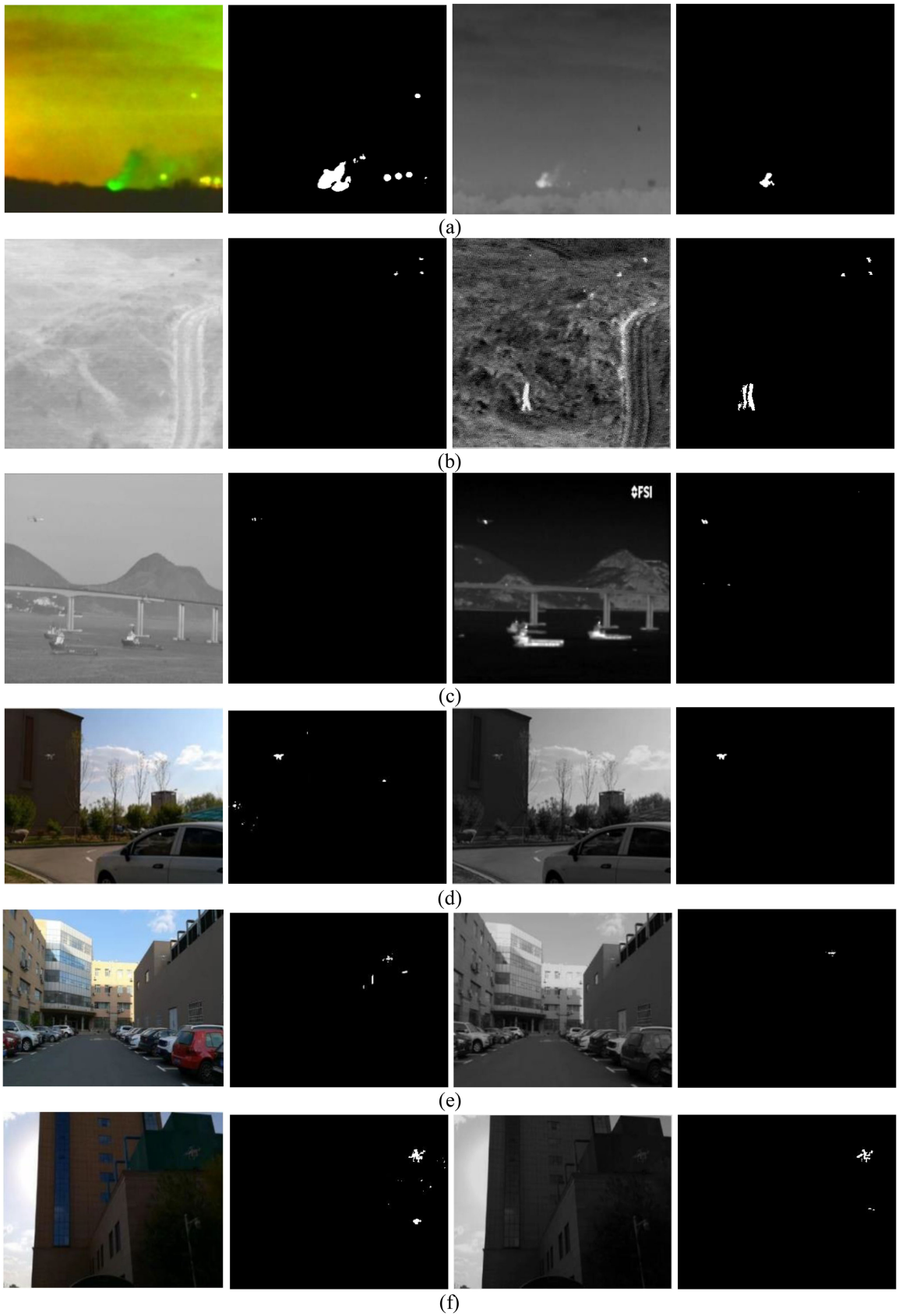


Fig. 3. Visual results of LSS-target detection from visible and infrared images with a poor resolution, separately. The first column is the original visible images, the second column is the extracted candidate regions from visible images, the third column is the original infrared image, and the fourth column is the extracted candidate regions from infrared images.

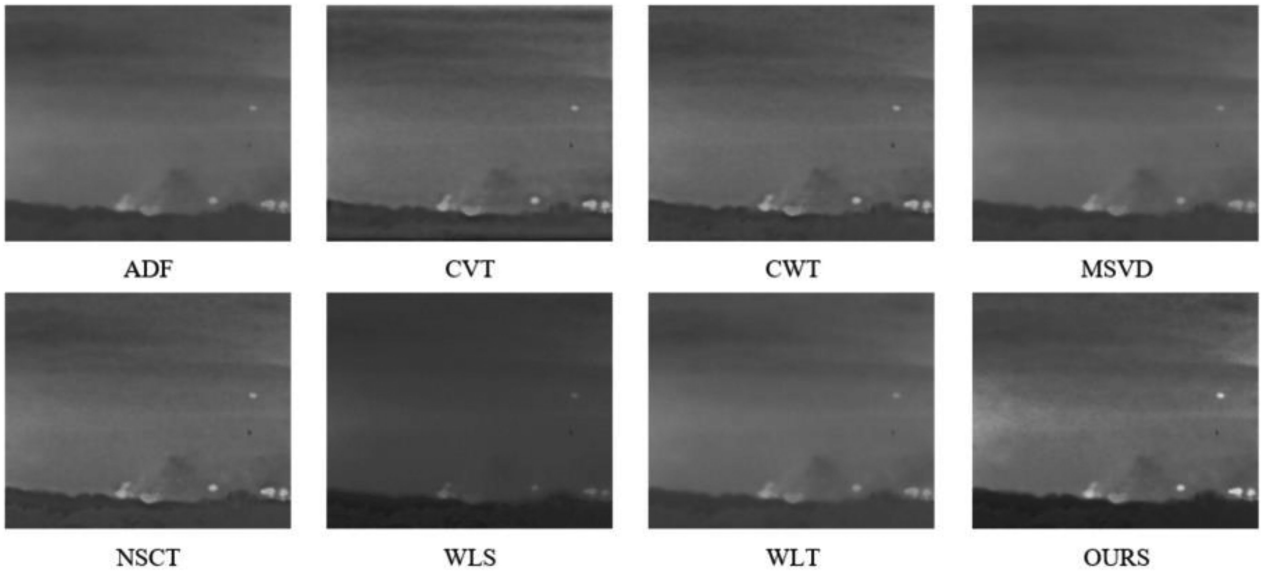


Fig. 4. Fusion comparison on the frames of the Guanabara Bay sequence.

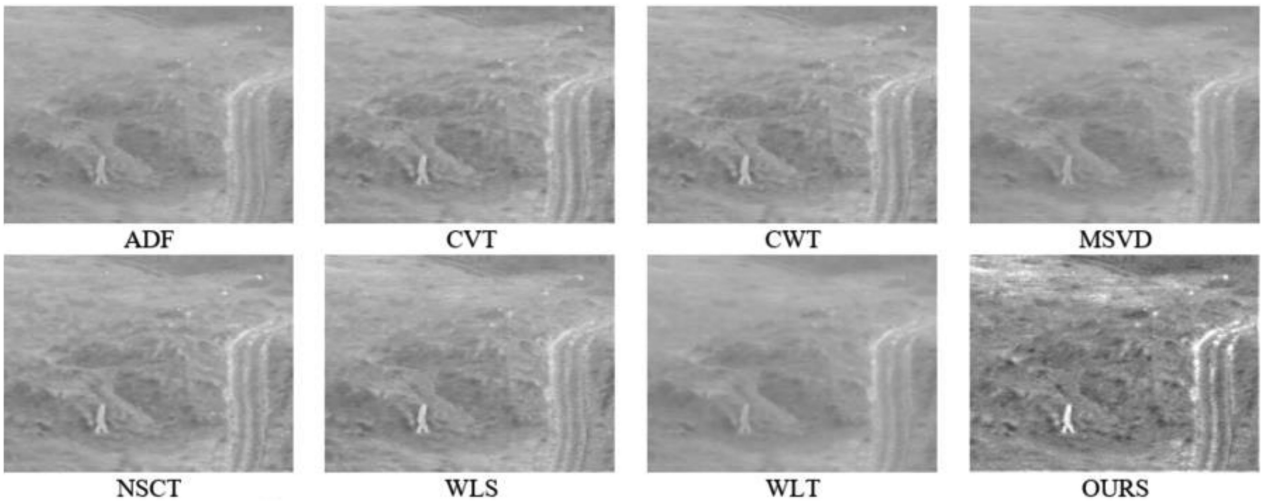


Fig. 5. Fusion comparison on the frames of the DHV sequence.

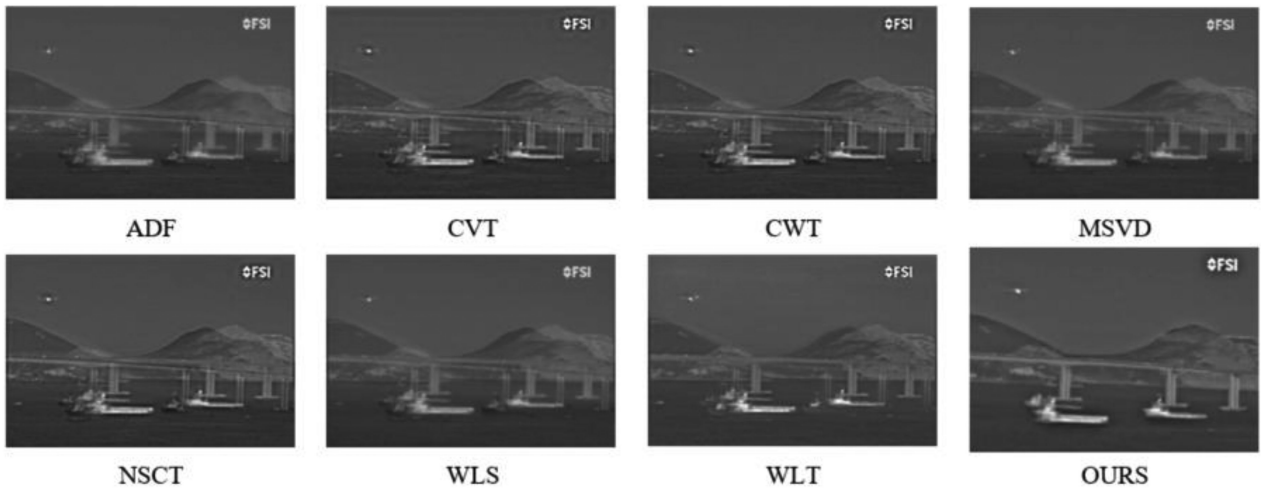


Fig. 6. Fusion comparison on the frames of the FEL sequence.

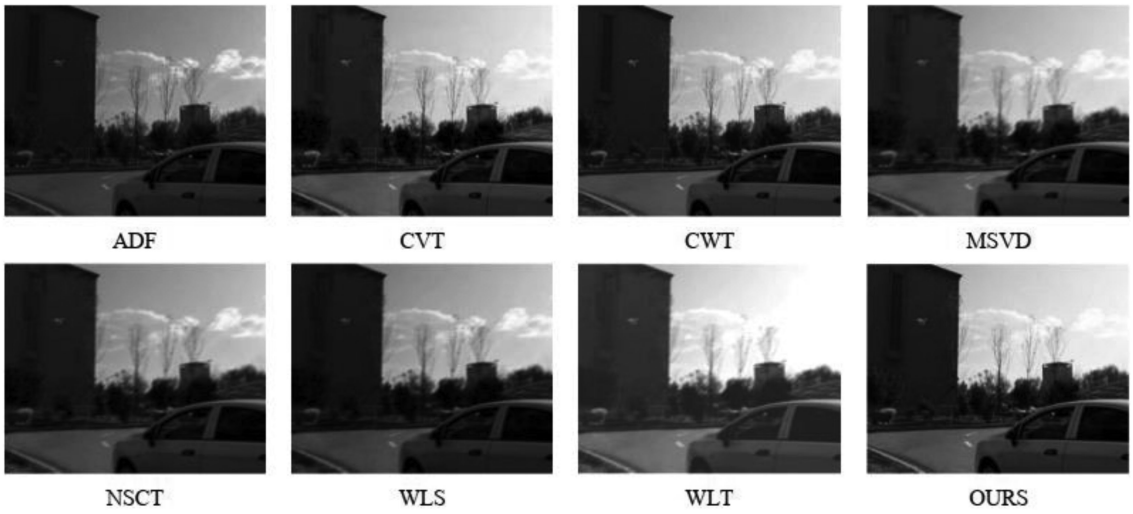


Fig. 7. Fusion comparison on the frames of video V1.

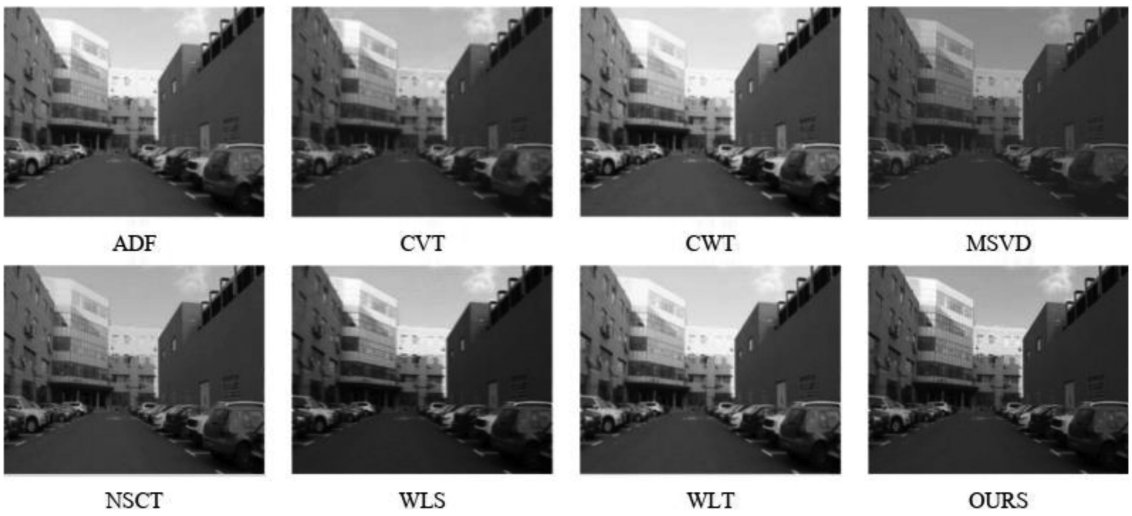


Fig. 8. Fusion comparison on the frames of video V2.

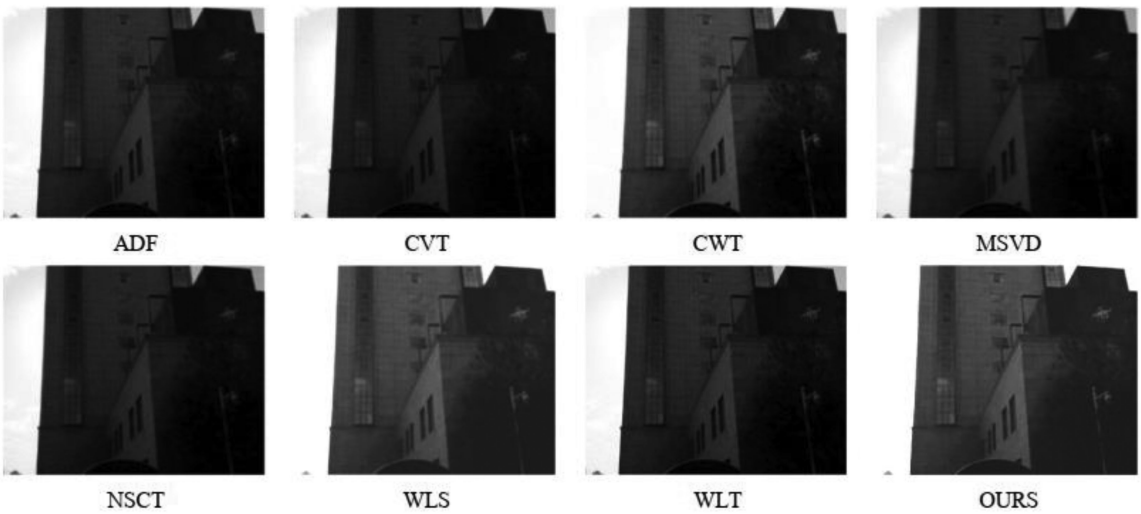


Fig. 9. Fusion comparison on the frames of video V3.

TABLE III
FUSION COMPARISON OF GUANABARA BAY

Video1	AVG	IE	Q^E	SF	SD	MI	f/s
WLT	1.39	5.67	0.35	5.38	18.86	11.35	0.75
CWT	2.56	5.92	0.45	9.91	22.33	11.83	1.13
CVT	2.59	**5.99**	0.43	9.86	22.27	**11.99**	1.90
NSCT	2.58	5.94	0.40	9.925	22.67	11.88	4.39
ADF	1.73	5.77	0.39	6.180	19.09	11.54	1.02
MSVD	2.10	5.75	0.41	8.109	19.33	11.51	0.47
WLS	2.31	5.61	0.46	10.00	20.49	11.22	3.05
OUR	**2.69**	5.86	**0.46**	**10.03**	**22.77**	11.76	**4.94**

The best results in each group highlighted in bold.

TABLE IV
FUSION COMPARISON OF DHV

Video2	AVG	IE	QE	SF	SD	MI	f/s
WLT	0.74	0.34	0.44	1.65	15.99	11.73	0.46
CWT	1.20	5.94	0.58	2.55	16.62	11.88	0.82
CVT	1.41	6.08	0.51	**3.29**	17.96	12.16	1.30
NSCT	1.21	5.96	0.59	2.56	16.78	11.92	2.40
ADF	0.94	5.88	0.55	1.94	16.16	11.76	0.58
MSVD	1.14	5.89	0.48	2.42	16.14	11.77	0.23
WLS	1.42	**6.30**	0.47	3.02	25.02	12.60	1.92
OUR	**1.44**	6.28	**0.60**	3.16	**25.55**	**12.68**	**2.30**

The best results in each group highlighted in bold.

experimental results show that integrating the candidate target regions of infrared and visible images can reduce the missed detection from poor images with a single sensor.

In order to demonstrate the effectiveness of the proposed fusion method, seven classic fusion algorithms including WLT [23], CWT [24], CVT [25], NSCT [26], ADF [27], MSVD [28], and WLS [29] are compared. The fusion results are shown in Figs. 4–9, in which our method can extract more texture and intensity features in the fused image for more effective target detection. After applying the fusion algorithm, almost all the targets can be identified in the six test sequences, thanks to the enhanced saliency measurement, and the amount of image information can also be significantly enriched after fusion.

The detailed results are listed, respectively, in Tables III–VIII, in which the best results are all bolded. As can be seen, the proposed fusion algorithm could perform well and obviously better than other algorithms in combination.

The idea of detecting the LSS target on the fused image and performing background modeling in the candidate target region was also validated as it could greatly increase the detection speed and reduce the false alarm rate. The comparison of the detection result with different SOTA algorithms is shown in Figs. 10 and 11, respectively, where (a) is the input image, (b) is the detection result of the LOBSTER algorithm [36], (c) is the detection result of the PBAS algorithm [14], (d) is the detection result of the (DLCM) algorithm [37], and (e) is the detection result of the

TABLE V
FUSION COMPARISON OF FEL

Video3	AVG	IE	QE	SF	SD	MI	f/s
WLT	2.59	5.75	0.43	4.39	13.22	11.50	0.43
CWT	4.50	5.91	0.52	7.65	14.68	11.82	0.79
CVT	4.58	5.95	0.49	7.75	15.06	11.90	1.09
NSCT	4.53	5.95	0.54	7.70	15.05	11.89	1.35
ADF	4.18	5.82	**0.57**	7.14	13.87	11.65	0.71
MSVD	4.68	5.82	0.48	7.89	13.86	11.65	0.14
WLS	5.01	6.09	0.52	8.69	16.58	12.18	1.07
OUR	**5.03**	**6.10**	0.55	**8.75**	**16.61**	**12.97**	**1.43**

The best results in each group highlighted in bold.

TABLE VI
FUSION COMPARISON OF V1

Video3	AVG	IE	QE	SF	SD	MI	f/s
WLT	1.59	4.75	0.33	3.39	12.22	10.50	0.33
CWT	3.50	4.91	0.42	6.65	13.68	10.82	0.69
CVT	3.58	4.95	0.39	6.75	14.06	10.90	1.09
NSCT	3.53	4.95	0.44	6.70	14.05	10.89	1.25
ADF	3.18	4.82	**0.47**	6.14	12.87	10.65	0.61
MSVD	3.68	4.82	0.35	6.89	12.86	10.65	0.18
WLS	4.01	5.09	0.42	7.69	15.58	11.18	0.95
OUR	**4.03**	**5.10**	0.45	**7.75**	**15.61**	**11.97**	**1.33**

The best results in each group highlighted in bold.

TABLE VII
FUSION COMPARISON OF V2

Video3	AVG	IE	QE	SF	SD	MI	f/s
WLT	0.84	0.44	0.52	1.72	16.01	11.85	0.51
CWT	1.32	6.08	0.63	2.61	16.72	11.92	0.93
CVT	1.50	6.13	0.60	**3.49**	18.06	12.26	1.42
NSCT	1.36	6.12	0.59	2.63	16.88	12.05	2.33
ADF	1.25	5.95	0.61	2.02	16.26	11.86	0.69
MSVD	1.28	5.84	0.53	2.56	16.24	11.81	0.38
WLS	1.51	6.22	0.51	3.22	25.12	12.73	2.07
OUR	**1.55**	**6.38**	**0.73**	3.26	**25.65**	**12.79**	**2.41**

The best results in each group highlighted in bold.

TABLE VIII
FUSION COMPARISON OF V3

Video3	AVG	IE	QE	SF	SD	MI	f/s
WLT	0.94	0.54	0.44	1.95	15.89	11.62	0.36
CWT	2.12	5.89	0.49	2.40	16.52	11.79	0.72
CVT	1.39	6.08	0.53	2.49	17.86	12.05	1.20
NSCT	1.15	5.93	0.62	2.41	16.68	11.83	**2.30**
ADF	0.96	5.66	0.75	1.94	16.06	11.56	0.48
MSVD	2.04	6.79	0.48	2.35	16.04	11.68	0.13
WLS	1.72	6.27	0.73	2.53	24.92	12.40	1.82
OUR	**2.34**	**7.18**	**1.50**	**2.92**	**26.45**	**13.51**	2.20

The best results in each group highlighted in bold.

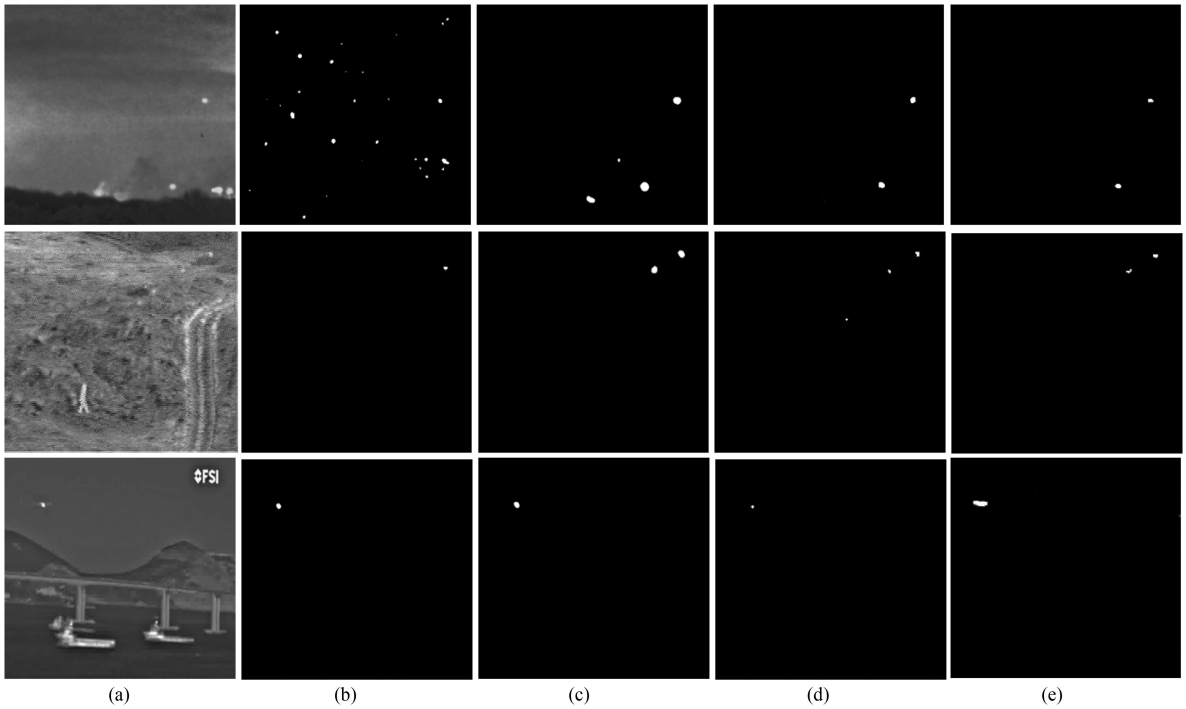


Fig. 10. Comparison of the final detection results of the public datasets: Guanabara Bay, DHV and FEL. From top to bottom, the five rows are the input images and results from the LOBSTER, PBAS, DLCM, and our proposed approach, respectively. (a) Input image. (b) Detection result of LOBSTER. (c) Detection result of PBAS. (d) Detection result of DLCM. (e) Detection result of the proposed method.

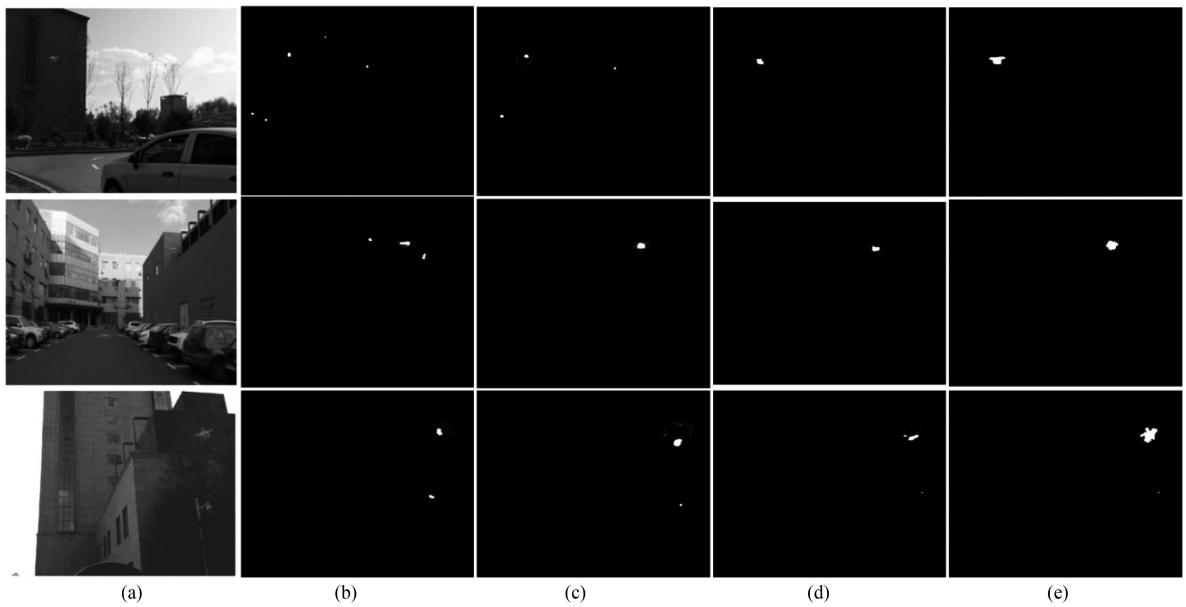


Fig. 11. Comparison of the final detection results of the self-collected datasets: V1, V2 and V3. From top to bottom, the five rows are the input images and results from the LOBSTER, PBAS, DLCM, and our proposed approach, respectively. (a) Input image. (b) Detection result of LOBSTER. (c) Detection result of PBAS. (d) Detection result of DLCM. (e) Detection result of the proposed method.

proposed algorithm. As can be seen, the proposed algorithm can detect the complete LSS target more accurately.

In the process of detection, multiple targets may be detected; however, not every target detected could be considered as LSS target. So, for quantitative analysis, the accuracy and F-value are used to evaluate the detection performance of the approach proposed. The accuracy is the proportion of the LSS targets among all detected targets. The recall rate is the proportion of detected LSS targets in the global samples. The F-value is the weighted harmonic mean of the accuracy and recall rate. As shown in Table IX, it can be seen from the results that the algorithm is still robust even when one of the sensors has poor imaging and achieves a higher detection accuracy than the traditional background-modeling algorithm.

TABLE IX
DETECTION ACCURACY (DA) AND F-MEASURE (F-M)

Test Video	Guanabara Bay		DHV		FEL		V1		V2		V3	
	Da	F-M	Da	F-M	Da	F-M	Da	F-M	Da	F-M	Da	F-M
LOBSTER	0.694	0.687	0.864	0.871	0.9535	0.953	0.9535	0.953	0.9535	0.953	0.9535	0.953
PBAS	0.847	0.855	0.843	0.834	0.966	0.968	0.966	0.968	0.966	0.968	0.966	0.968
DLCM	0.956	0.893	0.915	0.887	0.968	0.969	0.968	0.969	0.968	0.969	0.968	0.969
Ours	**0.962**	**0.973**	**0.964**	**0.978**	**0.975**	**0.975**	**0.975**	**0.975**	**0.975**	**0.975**	**0.975**	**0.975**

The best results in each group highlighted in bold.

TABLE X
COMPARISON OF THE TIME FOR LSS-TARGET DETECTION IN EACH FRAME (TIME IN MILLISECOND FOR EACH FRAME)

	Guanabara Bay	DHV	FEL	V1	V2	V3
LOBSTER	147	81	255	121	93	216
PBAS	89	58	166	73	62	149
DLCM	36	29	45	24	34	53
Ours	93	56	112	65	51	137

In addition, the detection speed of the proposed method is also compared with others in Table X, where the time needed for processing each frame is given, according to two publicly available datasets and three our own datasets. Thanks to the background modeling of the candidate target region, which has greatly increased the detection speed. As can be seen, our method can achieve a high detection accuracy with a very competitive running time in comparison to SOTA approaches.

VII. CONCLUSION

In this article, we have presented a novel algorithm for the detection of LSS targets through the effective fusion of visible and infrared images. First, an ROI extraction module based on 1-D information entropy and weighted average is proposed to reduce the overall calculation time and alleviate the interference from the unnecessary background information. Second, an efficient and effective multiscale layered image fusion module is introduced, with a saliency map for enhancing the image details of the fused image. Finally, accurate LSS-target detection is completed by local background modeling with the local SuBSENSE method. The experimental results also illustrate the efficacy of the two fusion modules, as our proposed approach outperforms the SOTA. More importantly, the proposed approach can be effectively used to detect LSS targets with a high accuracy and robustness even when a single sensor has poor images.

REFERENCES

[1] S. A. Musa *et al.*, "Low-slow-small (LSS) target detection based on micro Doppler analysis in forward scattering radar geometry," *Sensors*, vol. 19, no. 15, 2019, Art. no. 3332.

[2] J. Li *et al.*, "Visual detail augmented mapping for small aerial target detection," *Remote Sens.*, vol. 11, no. 1, 2019, Art. no. 14.

[3] A. Khashman, "Neural system for military target detection," in *Proc. 8th IEEE Int. Conf. Electron., Circuits Syst.*, Malta, 2001, vol. 3, pp. 1359–1362.

[4] G. Pedro *et al.*, "Continuous variance estimation in video surveillance sequences with high illumination changes," *Signal Process.*, vol. 89, no. 7, pp. 1412–1416, 2009.

[5] Q. Shao, Z. Tang, and S. Han, "Hierarchical codebook for background subtraction in MRF," *Infrared Phys. Technol.*, vol. 61, no. 1, pp. 259–264, 2013.

[6] B. Wang *et al.*, "Background subtraction using spatiotemporal condition information," *Optik, Int. J. Light Electron Opt.*, vol. 125, no. 3, pp. 1406–1411, 2014.

[7] Y. Lv, S. Sun, Q. Lin, and R. Liu, "Space moving target detection and tracking method in complex background," *Infrared Phys. Technol.*, vol. 91, no. 17, pp. 107–118, 2018.

[8] S. Chiang, C. Chang, and I. W. Ginsberg, "Unsupervised target detection in hyperspectral images using projection pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1380–1391, Jul. 2001.

[9] J. Lou *et al.*, "Small target detection combining regional stability and saliency in a color image," *Multimedia Tools Appl.*, vol. 76, no. 13, pp. 14781–14798, 2017.

[10] T. Xie, Z. Chen, and R. Ma, "A novel infrared small target detection method based on PGF, BEMD and local inverse entropy," *Acta Infrared Millimeter Wave*, vol. 36, no. 1, pp. 92–101, 2017.

[11] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Phys. Technol.*, vol. 76, no. 6, pp. 52–64, 2016.

[12] S. Qiu *et al.*, "A moving target extraction algorithm based on the fusion of infrared and visible images," *Infrared Phys. Technol.*, vol. 98, no. 27, pp. 285–291, 2019.

[13] O. Barnich and M. Van Droogenbroeck, "ViBE: A powerful random technique to estimate the background in video sequences," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 945–948.

[14] M. Hofmann, P. Tiefenbacher, and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Providence, RI, USA, 2012, pp. 38–43.

[15] B. P. Marcus, "The weighted moving average technique," in *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken, NJ, USA: Wiley, 2010.

[16] J. F. Macgregor and T. J. Harris, "The exponentially weighted moving variance," *J. Qual. Technol.*, vol. 25, no. 2, pp. 106–118, 1993.

[17] P. St-Charles, G. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[18] A. Toet, TNO Image Fusion Dataset, 2014. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

[19] SMT/COPPE/Poli/UFRJ and IME-Instituto Militar de Engenharia, Visible-Infrared Data, 2016. [Online]. Available: http://www02.smt.ufrj.br/∼fusion/

[20] S. Liu, J. Zhao, and M. Shi, "Medical image fusion based on rolling guidance filter and spiking cortical model," *Comput. Math. Methods Med.*, vol. 2015, 2015, Art. no. 156043.

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

[22] L. J. Chipman, T. M. Orr, and L. N. Graham, "Wavelets and image fusion," in *Proc. Int. Conf. Image Process.*, 1995, pp. 248–251.

[23] J. J. Lewis *et al.*, "Pixel-and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[24] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[25] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, no. 12, pp. 147–164, 2015.

[26] D. P. Bavirisetti and R. Dhuli, "Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform," *IEEE Sensors J.*, vol. 16, no. 1, pp. 203–209, Jan. 2016.
[27] V. Naidu, "Image fusion technique using multi-resolution singular value decomposition," *Defence Sci. J.*, vol. 61, no. 5, pp. 479–484, 2011.
[28] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, no. 9, pp. 100–109, 2016.
[29] Z. Liu *et al.*, "MRI and PET image fusion using the nonparametric density model and the theory of variable-weight," *Comput. Methods Programs Biomed.*, vol. 175, no. 11, pp. 73–82, 2019.
[30] X. Bai, Y. Zhang, F. Zhou, and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," *Inf. Fusion*, vol. 22, no. 10, pp. 105–118, 2015.
[31] M. Yin, P. Duan, W. Liu, and X. Liang, "A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation," *Neurocomputing*, vol. 226, no. 19, pp. 182–191, 2017.
[32] X. Kong, L. Liu, Y. Qian, and Y. Wang, "Infrared and visible image fusion using structure-transferring fusion method," *Infrared Phys. Technol.*, vol. 98, no. 16, pp. 161–173, 2019.
[33] M. Shahid Farid, A. Mahmood, and S. A. Al-Maadeed, "Multi-focus image fusion using content adaptive blurring," *Inf. Fusion*, vol. 45, no. 9, pp. 96–112, 2019.
[34] T. Ma *et al.*, "Multi-scale decomposition based fusion of infrared and visible image via total variation and saliency analysis," *Infrared Phys. Technol.*, vol. 92, no. 23, pp. 154–162, 2018.
[35] G. Bilodeau, J. Jodoin, and N. Saunier, "Change detection in feature space using local binary similarity patterns," in *Proc. Int. Conf. Comput. Robot Vis.*, Regina, SK, Canada, 2013, pp. 106–112.
[36] C. Shu and D. Baokuo, "Foreground detection of the adaptive LOBSTER algorithm in a dynamic background," *J. Image Graph.*, vol. 22, no. 2, pp. 161–169, 2017.
[37] S. Pan, S. Zhang, M. Zhao, and B. An, "Infrared small target detection based on double-layer local contrast measure," *Acta Photon. Sin.*, vol. 49, no. 1, 2020, Art. no. 0110003.
[38] Y. Li and Y. Wang, "Video image motion background subtraction based on one-dimensional maximum entropy," *Comput. Modernization*, vol. 3, no. 4, pp. 44–53, 2018.
[39] J. Ma *et al.*, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Phys. Technol.*, vol. 82, pp. 8–17, 2017.
[40] J. Zabalza *et al.*, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4418–4433, Aug. 2015.
[41] Q. Liu *et al.*, "Decontaminate feature for tracking: Adaptive tracking via evolutionary feature subset," *J. Electron. Imag.*, vol. 26, no. 6, 2017, Art. no. 063025.
[42] Y. Yan *et al.*, "Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognit.*, vol. 79, pp. 65–78, 2018.
[43] Z. Wang *et al.*, "A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos," *Neurocomputing*, vol. 287, pp. 768–783, 2018.
[44] G. Sun *et al.*, "Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 91, 2020, Art. no. 102157.
[45] Q. Liu *et al.*, "EACOFT: An energy-aware correlation filter for visual tracking," *Pattern Recognit.*, vol. 112, 2021, Art. no. 107766.

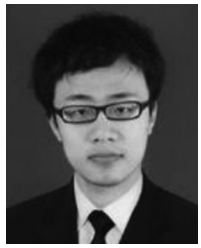
Haijiang Sun received the Ph.D. degree in electronic engineering from Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, China, in 2012.

He is currently a Professor with the Perception and Display Laboratory, CIOMP. His research interests include target recognition and tracking, and high-definition video enhancement and display.

Dr. Sun is an Associate Editor for the *Liquid Crystal and Reality* and *Optical Precision Engineering*.

Qiaoyuan Liu received the master's degree in software engineering, in 2016 and Ph.D. degree in analysis and planning of intelligent environment, in 2019, both from Northeast Normal University, Changchun, China.

She is currently an Assistant Professor with the Changchun Institute of Optics, Precision Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. Her research interests include visual tracking and image analysis.



Jiacheng Wang received the master's degree in circuits and systems from Xidian University, Xi'an, China, in 2014.

From 2014 to 2016, he was a Research Intern with Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, China. Since 2016, he has been a Research Assistant with the Image Processing Laboratory, CIOMP. His research interests include target tracking and embedded software and hardware design.

Jinchang Ren received the Eng.D. in computer vision, in 2000 from Department of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China and the second Ph.D. degree in electronic imaging and media communication from the University of Bradford, Bradford, U.K., in 2009.

He is currently a Professor of Computing Science with the Robert Gordon University, Aberdeen, U.K. His research interests include hyperspectral imaging, computer vision, machine learning, and big data analytics and applications.

Yanfeng Wu was born in Jilin, China, in 1992. He received the bachelor's degree in applied physics in 2015 from Jilin University, Changchun, China, and the M.S. degree in optical engineering from University of Chinese Academy of Sciences, Beijing, China, in 2017.He received the Ph.D degree in electronic engineering from Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, China, in 2020. He is now an assistant professor with the 28th Research Institute of China Electronics Technology Group, Nanjing, China.

His research interests include image enhancement and moving target detection.

Huimin Zhao received the B.Sc. and M.Sc. degrees in signal processing from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1997, respectively, and the Ph.D. degree in electrical engineering from the Sun Yat-sen University, Guangzhou, China, in 2001.

He is currently a Professor and the Dean of the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. His research interests include image/video processing, and information security technology and applications.

Huakang Li received the B.E. degree in electrical and computer engineering, Sun Yat-sen University Southern College, Guangzhou, China, in 2019. He is working toward the master.s degree with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China.

His research interests include deep learning, gait recognition, and image/video processing.