Received 22 September 2022, accepted 19 October 2022, date of publication 26 October 2022, date of current version 4 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217656

RESEARCH ARTICLE

Infrared Detection of Small-Moving Targets Using Spatial Local Vector Difference and Temporal Sample Consensus Measures

KAIYUAN LI[1], YUNSHENG ZHANG[2,5], YIQIONG ZHANG[3], YITING CHEN[4], AND CHIHANG ZHAO[5]

[1]Economics and Management School, Wuhan University, Wuhan 430072, China
[2]Research Center of Hubei Logistics, Hubei Cooperative Innovation for Emissions Trading System, Hubei University of Economics, Wuhan 430205, China
[3]School of Computer and Communication Engineering, Northeastern University, Qinhuangdao 066099, China
[4]School of Biomedical and Engineering, South Central University for Nationalities, Wuhan 430074, China
[5]School of Transportation, Southeast University, Nanjing 210096, China

Corresponding authors: Yunsheng Zhang (yunshengzhang@hbue.edu.cn) and Chihang Zhao (chihangzhao@seu.edu.cn)

This work was supported in part by the China Postdoctoral Science Foundation under Grant 2018M631169; in part by the Ministry of Education Project of Humanities and Social Sciences for Young Researchers under Grant 18YJCZH253; in part by the Hubei Provincial Natural Science Foundation under Grant 2019CFB580; in part by the Social Science Fund of Shaanxi Province under Grant 2019S048; in part by the Key Science and Technology Projects of Transportation Industry of Ministry of Communications under Grant 2020-MS5-134; in part by the Transportation Science and Technology Plan Project of Shandong Province under Grant 2020B36; in part by the Science and Technology Project of Shandong High Speed Group Company Ltd., under Grant (2020) SDHS-GSGF-09; and in part by the Scientific Research Project of Hubei Department of Education under Grant 20Y145 and Grant D20202202.

ABSTRACT Detection small moving targets in infrared sequences with accuracy and low computation time is challenging work in infrared search and tracking systems. A frequent approach for this task is to strengthen small weak targets and diminish the clutter of background. However, the background and small targets' pixel values are close to each other and thus only a few background suppressing models are suitable to infrared small target segmentation. To efficiently resolve the problem that most of the classic approaches cannot manage low signal to noise ratio and weak objects without details, a novel spatial-temporal local difference measure is introduced for moving objects segmentation in infrared sequences. First of all, to strengthen targets, a new local vector dissimilarity measure is employed to demonstrate the difference between the weak target and their surrounding backgrounds and calculate the spatial saliency feature figure. Then the local sample consensus of succession images is used to compute the temporal varying feature figure. Afterward, the combined saliency feature map is measured by taking spatial and temporal feature images into account. Finally, the small moving targets are segmented employing an adaptive threshold approach. Abundant quantitative and qualitative experimental results have demonstrated that the introduced approach is remarkable and has a superior accuracy in terms of performance on both public and real datasets in comparison to the state-of-the-art spatial and temporal models.

INDEX TERMS Infrared sequence, small moving target, sample consensus, low signal to noise ratio, spatial and temporal feature model.

I. INTRODUCTION

With rapidly promoting number of infrared search and track surveillance systems, to reliably and rapidly segment foreground targets from low light and weak dark scenario is a pre-requisite procedure for application such as infrared vision understanding and high-level surveillance applications in field of missile precision guidance, and pre-warnings based on military night vision, etc.. Hence this topic has attracted researchers' attention in recent decades [1]. The objects are

The associate editor coordinating the review of this manuscript and approving it for publication was Ravibabu Mulaveesala.

usually weak and buried in complex backgrounds with low signal to noise ratio. They may also be short of texture or structure information owing to complex backgrounds scenes and the long imaging distance. Furthermore, interference caused by sky-sea line, cloud edge, and different weather conditions may result in high false alarms. These special characteristics give rise to an intricate and difficult problem during moving objects segment of infrared scenes [2].

A primary approach that extracts moving objects from infrared scenes is background subtraction. This approach is capable of segmenting the pixel regions with different intensity from the background model. It achieved good performance in visible surveillance videos [3]. To manage the special characteristics of infrared sequences, numerous methods for strengthening targets and restraining complex background have been presented in recent years with different designed philosophies for handing the temporal or spatial features. These related methods treating weak small target without details and low signal to noise ratio infrared scenes can be approximately split into single-frame using spatial saliency features figure and multi-frames using spatial-temporal saliency features figure [4].

Frequently-used single-frame approaches for weak small targets can be divided into two kinds. The first kind pays close attention to background that can be built using predictions or estimations, and foreground target can be segmented employing comparing the difference between the original frame and the predicted background image calculated using background model. The popular background estimations approaches include Bayes estimation, morphology filter, top-hat transform filter, max-median or max mean filter and so on [5]. Nevertheless, these methods based on background estimation are sensitive to noise or depend on the structural element's size and shape.

The second class pays attention to special targets characteristics and segments foreground from the background in order to improve the saliency feature of target. Conventional strengthened techniques include the difference of Gaussian filter, the Laplacian Gaussian filter, high-pass filters, and difference of Gabor filter [6]. Although these convolution filters can effectively deal with low-frequency clutter, they do not manage high-frequency clutter and noise. To improve the ability of suppressing false alarms, human visual system-based methods been have introduced in recent years. They highlight the saliency of target by constructing specific metrics.

Attempts to resolve the dilemma have resulted in the development of strengthening the weak object and alleviate background clutter simultaneously, according to the biological visual mechanism, Chen et al. [7] introduced an effective local contrast measure (LCM) employing a nested window structure including eight directions. The LCM method has been widely developed and numerous improved approaches based on LCM have been presented. To enhance the robustness and computational efficiency of computing local contrast measure in infrared sequences, improved local contrast measure (ILCM) [8] utilized image patches to replace original pixels and the maximum of central cell is replace using mean value.

In order to effectively suppress the single-pixel noise, novel local contrast measure (NLCM) [9] utilized the mean value of the first 3 maximal values among central cell to substitute the mean of central cell. The multi-scale patch-based LCM and multi-scale relative LCM adopt multi-scale technology to detect small targets having different sizes and shapes [10]. In [11], the weighted homogeneity characteristic of the surrounding regions was combined with LCM to enhance the target and the performance was more effective in the field of the signal-to-clutter gain. These improved LCM methods calculate the ratio value between the middle cell and corresponding surroundings cell of nested to promote weak target.

Other enhanced LCM methods have employed tri-layer filter structure or difference-form to eliminate a high brightness background. A method was proposed to detect the foreground target region using a three-layer patch-image model and weak small target enhancement with variance difference [12]. To detect weak targets in low signal-to-clutter ratio scenes, the intensity levels of nine regions are calculated using difference-form to improve the local contrast map algorithm [13]. To solve the problem that small and dim foreground target is of low signal-to-noise ratio and surrounded in a complex and mixed background scenes, local dissimilarity and local brightness difference were utilized in [14] to demonstrate the dissimilarity between targets and corresponding surrounding background region with nested structure. To better highlight target and suppress background interference, a Tri-layer structure is designed in [15] to measure the double-neighborhood gradient.

To enhance true target, both the ratio and difference operations are employed in [16] to compute the local contrast information in Tri-layer structure. A novel small target detection algorithm was introduced in [17] based on absolute directional mean difference using internal and external windows filter structure. LCM-based approaches either decrease the background noise or increase the weak target instead of handling these two aspects at once, these approaches perform well in simple surrounding backgrounds but have poor performance in complex backgrounds infrared sequences. Moreover, the single-frame models use spatial saliency feature based on their detection capability, but neglect the correlation of temporal saliencies feature in removing false alarms.

Due to the fact that the temporal saliency through successive frames can be used as a reinforcement of spatial saliency feature to diminish false alarms, a raising number of investigators have introduced multi-frames model to segment the small infrared targets employing temporal-spatial features. Such methods can be divided into three kinds: low-rank-based model, deep-neural-network-based model, and temporal-spatial local saliency feature. Low-rank-based approaches assume that low-rank and outliers correspond to the background and foreground target, respectively. For example, to handle the short of context information of both the target and corresponding background in spatial-temporal domain, a spatial-temporal tensor model was established in [18] depended on the sparse prior of the weak target and the local correlation of the background. Although low-rank based models that use with recent optimization techniques may enhance the speed of operation, they cannot realize real-time foreground detection.

Researchers are now starting to develop a deep neural network (DNN) to model ideal background sequences. Based on a convolutional neural network model pre-trained on public dataset, a novel multi-scale fully convolutional network architecture was introduced to detect in [19] moving objects of infrared videos. Recently, based on convolutional neural networks and handcrafted feature methods, a novel infrared small target segmentation network was introduced in [20] to handle the low signal-to-noise ratio problem. These neural network-based models outperform unsupervised models and the state-of-the-art deep learning approaches allow adaptation to targets size and shape if an enormous amount of training data-set is offered. The neural network-based approaches work in off-line mode and may lack real-time performance. in addition, when the size of infrared small targets is very small, DNN based feature learning techniques are still weak.

To excavate comprehensive and precise information concerning the target and the background, an increasing number of temporal-spatial local saliency feature models are developed recently. To enhance and detect moving weak objects among infrared video scenes, spatial–temporal local contrast filter (STLCF) was proposed in [21] based on combining a novel spatial-temporal local contrast feature. Then, to describe spatial–temporal local contrast map (STLCM) [22], a spatial local contrast map (SLCM) and temporal local contrast map (TLCM) were multiplied employing the effective multi-direction filters fusion and enhancing time domain difference. Afterwards, to change the pattern of spatial and temporal feature fusion, a novel spatial and temporal local difference measure (STLDM) was introduced in [23] based on 3-D spatial-temporal domain and three fixed block frames. To focus on low-altitude background scenes, a novel spatial and temporal saliency model (NSTSM) was studied in [24] by combining spatial variance saliency feature and temporal gray saliency feature, It achieved better detection performance in low-altitude slow small targets detection. To detect remote flying drones, a robust switching spatial-temporal fusion (STF) segmentation model was proposed in [25] by suppressing the noises or clutters and strengthening the contrast between the target and background simultaneously. Our previous spatial-temporal local vector difference measure (STVDM) [26], which is based on frame difference method [27] and the local vector difference [28], can effectively enhance the targets and alleviate the clutters background.

The successive frames should be registered before object detection using the low-rank-based method and deep neural network-based method. Thus temporal-spatial local saliency feature methods are an active topic in applications. However, there are still some unfavorable regarding temporal-spatial local saliency feature approaches. In the first place, the spatial feature usually calculates the dissimilarity between the central cell and surrounding cells. However, when intensity values of the background pixels and weak objects are similar or the targets are dim, the dissimilarity between them will not be obvious. Second, issue is how many frames are to be selected to obtain the temporal difference and establish a temporal feature mapping.

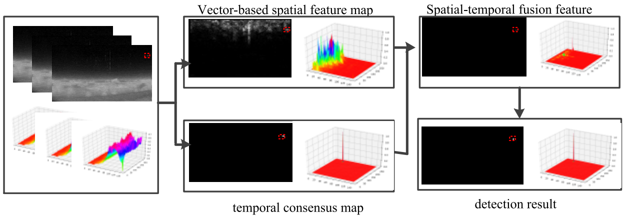
Vector-based spatial feature map
Spatial-temporal fusion feature
temporal consensus map
detection result

FIGURE 1. The flowchart of presented approach.

From the above literature review, one can conclude that moving weak target segmentation is an intractable work in infrared videos. Moreover, combining the advantage of these models will be a promising strategy to improve temporal-spatial local saliency methods. The structure of the foreground targets is dissimilar to background area and the size of the objects may transform according to the object's types, distance, and infrared scenes. To efficiently overcome the issue that most of current models cannot effectively manage the situation where the intensity values of the background pixels and weak targets are similar, a novel spatial and temporal model with spatial local vector difference measure and temporal sample consensus is developed in this paper. The main contributions of this spatial-temporal model can be summarized as the following statements: (1) To use local region vector, we construct a strengthened possible target spatial saliency feature image based on the cosine similarity of the local vector. (2) To use local sample consensus of succession frames and consensus background samples model, we construct an enhanced target temporal varying mapping. (3) to employ spatial and temporal feature map varying simultaneously, we propose novel Spatial-Temporal saliency feature map by multiplying the feature images at the corresponding positions and extract targets using an adaptive segmentation method.

The overall structure of the presented adaptive approach is demonstrated in Fig. 1. To begin with, a local dissimilarity descriptor is conceived based on the cosine similarity of local region vector to strengthen possible targets region in the spatial domain. Secondly, the local sample consensus of succession frames is used to calculate the temporal varying mapping. Then, the combined saliency feature image is determined by taking spatial and temporal feature images into account. Finally, the weak targets are detected employing an adaptive threshold technique.

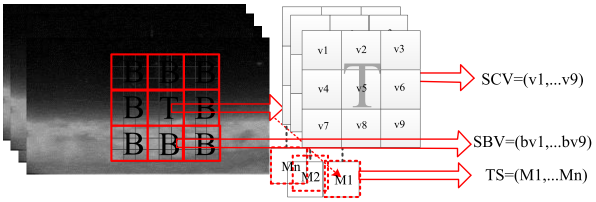

FIGURE 2. **Spatial and temporal vectors samples constructed by the sliding window filter.**

The remainder of article is structured as the following. In the second section, the Spatial-Temporal model is illustrated in detail. The qualitative and quantitative of experimental results are demonstrated in Section 3 by comparing the presented model with the state-of-the-art spatial and temporal approaches. Finally, A brief summary of some of the relevant conclusions and subsequent work are provided in Section 4.

II. PROPOSED METHOD

As mentioned above, current methods cannot cope with low signal to noise ratios and the intensity pixels values of the small-targets and background being similar. To efficiently alleviate these issues, an adaptive temporal-spatial model is developed for small target detection in infrared sequences. In what follows, we present the detailed description of the framework.

A. SPATIAL AND TEMPORAL FEATURE BUILDED OF PIXEL

It is assumed that $f_1, \cdots, f_l$ is an frames sequence of infrared videos, where f_l is the current processed frame. We further presume that the $3p \times 3p$ sliding window is centered at given pixel location (x, y) in a frame and the corresponding window can be classified into nine $p \times p$ cells, and the middle of window can go through all pixels of frame. As demonstrated in Fig.2, the middle cell denoted as T, and B labeled the eight directions of the surrounding background cells in one frame. To decide if the observed pixel location (x, y) is target, the similarity feature between the middle cell and the corresponding background cells must be measured. Suppose each cell has 3×3 pixels and the grayscale values of pixels in each cell are vectorized along the column as a vector in Fig. 2. The middle cell T is vectorized as $SV = (\mathrm{v}1, \mathrm{v}2, \cdots \mathrm{v}9)$ and the vector of each B is constructed employing the same scheme. To find the temporal varying feature map, the past observed intensity sample value M of pixel at location (x, y) can be constructed using n successive frames and the temporal samples model can be denoted as $TS = (\mathrm{M}1, M2, \cdots Mn)$.

B. SPATIAL SALIENCY FEATURE MAPPING

The area where a target located is different from its neighborhood background. That is, despite the fact that intensity values of the small-targets and background are close to each other and the distinguishability is not remarkable, the object region is different in the local spatial vector domain feature. To promote the sensitivity between background and object region, the local spatial vectors cosine similarity is employed according to relation of middle cell and background cells. The similarity between the central cell vector $\overrightarrow{v(x, y)}$ at location (x, y) and one of the corresponding background cells vectors $\overrightarrow{bv(x, y)_j}, j \in \{1, 2, \cdots, 8\}$ is defined as the vector's collinearity. The sensibility of vector's collinearity can overcome the problem of weak-small targets without details and low signal to noise ratio. The measure of collinearity is demonstrated employing the angle of two vectors $\theta\left(\overrightarrow{v(x, y)}, \overrightarrow{bv(x, y)_j}\right)$ and is computed as the following form:

$$\theta\left(\overrightarrow{v(x, y)}, \overrightarrow{bv(x, y)_j}\right) = arc\ cos\left(\frac{\overrightarrow{v(x, y)} \cdot \overrightarrow{bv(x, y)_j}}{\left\|\overrightarrow{v(x, y)}\right\| \times \left\|\overrightarrow{bv(x, y)_j}\right\|}\right), \tag{1}$$

where $\left\|\overrightarrow{v(x, y)}\right\|$ is the magnitude of vector and $\overrightarrow{v(x, y)} \cdot \overrightarrow{bv(x, y)_j}$ is the dot product. For every middle pixel in central cell, the angle $\theta\left(\overrightarrow{v(x, y)}, \overrightarrow{bv(x, y)_j}\right)$ is used to determine if the central cell matches with the neighboring background cell or not. Therefore, to decide if the middle pixel is a target, the dissimilarity between the middle cell and the neighboring background cells is calculated employing angle. When the direction of local vectors is changing, the corresponding values of angle vary from 0 to $\frac{\pi}{2}$. The decreasing value of angle will lead to the increasing similarity. The vectors' collinearity that only employs the direction of vectors and does not utilize vectors' lengths is more sensitive and robust to low signal-to-noise ratio infrared sequences. Consequently, employing the local spatial vectors' cosine similarity, a spatial feature map can be constructed to describe the relation of target and the surrounding background. The spatial local vector difference measure (SLVDM) feature image of one infrared frame is generated as follows:

$$SLVDM\ (i, j) = \max \theta\left(\overrightarrow{v(x, y)}, \overrightarrow{bv(x, y)_j}\right), \quad j \in \{1, 2, \cdots, 8\}, \tag{2}$$

C. TEMPORAL VARYING FEATURE MAPPING

The temporal background samples model $TS(x, y)$ of pixel at location (x, y) is comprised of n samples and denoted as the following:

$$TS(x, y) = \{M_1(x, y), M_2(x, y), \cdots M_n(x, y)\}, \tag{3}$$

Among which $M_j(x, y), (j = 1, \cdot \cdot \cdot, n)$ are the past observed intensity samples value of the same pixel position and n denotes the number of samples in temporal background model. The number of samples depends on manual setting and $n = 15$ in this temporal samples model. Then the pre-n-frame of the infrared frames sequence is employed to initialize the model.

After building of the initial temporal background samples model, small moving targets segmentation process is used to segment objects from incoming infrared sequences and define the temporal varying feature map. At segmentation phase, the presented temporal background model gets the varying information by comparing the incoming intensity value at each pixel location with samples of constructed model. For incoming pixels, the distance $d_j(x, y)$ is calculated between intensity value $\mathrm{i}(\mathrm{x}, \mathrm{y})$ of pixel $p(x, y)$ and sample value $\mathrm{M_j}(\mathrm{x}, \mathrm{y})$ $(\mathrm{j} = 1, \cdots, \mathrm{n})$ in its background model $TS(x, y)$. The varying criterion for pixel $p(x, y)$ is defined as the followings equations and the same is implemented for all pixels in the frame.

$$\mathrm{d_j}(\mathrm{x}, \mathrm{y}) = \begin{cases} 1 & \text{if } \left|\mathrm{i}(\mathrm{x}, \mathrm{y}) - \mathrm{M_j}(\mathrm{x}, \mathrm{y})\right| < \varepsilon, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

$$s(x, y) = \sum_{j=1}^{n} d_j(x, y), \tag{5}$$

$$TLVM(x, y) = \begin{cases} 1 & s(x, y) < Th, \\ 0 & otherwise. \end{cases} \tag{6}$$

ε is a distance threshold and is used to compare a new pixel value to pixel samples, if $\left|\mathrm{i}(\mathrm{x}, \mathrm{y}) - \mathrm{M_j}(\mathrm{x}, \mathrm{y})\right| < \varepsilon$, it implies that $i(x, y)$ and $M_j(x, y)$, $(j = 1, \cdots, n)$ is identical and matched, and the possibility of stable pixel will increase. The values of ε will influence the performance. We select it to be at the beginning range of the plateau, that is $\varepsilon = 15$. $s(x, y)$ is employed to denote the times of match between current pixel and temporal background samples model. $TLVM(x, y)$ denotes the temporal local varying measure feature (TLVM) result. If the match times is less then Th, the current pixel is varying and it may be a part of targets. Otherwise, it is detected as background pixel. The fixed predefined threshold Th is the minimum match number of current pixel to classify the pixel as background. Increasing the value of Th is likely to improve the computational cost, and it depends on number of samples in background and the trade-off between accuracy results and computational complexity. Based on our experiments, we set the optimal value of Th as $Th = 2$ in this paper.

The updating of temporal background model is a procedure where the timeworn samples of the background model is substituted employing the current background pixel to adapt to the infrared scenes change. In order to guarantee that the "right" representation samples can be maintained in the model, the random subsampling and conservative update scheme are employed. At the end of segmentation, the samples of temporal background model are renovated as the following process: in the first place, if current pixel is detected as background, that is $TLVM(x, y) = 1$, the corresponding intensity value replace one sample of background model $TS(x, y)$ with $1/\theta$ ratio, where θ is a subsampling factor and similar to ViBe method [29]. There is one more point, one of its neighbor pixels in small region is randomly chosen and the current pixel replaces the selected samples of neighbor background $\mathrm{TS}(\mathrm{x}', \mathrm{y}')$ with the $1/\theta$ ratio. The neighbor updating improves the ghost area to be automatically and quickly infused into the model at the end of detection. We set $\theta = 16$ in this paper.

D. FOREGROUND OBJECT DETECTION

To detect the objects from the complex infrared environment, SLVDM is fused with TLVM by multiplication fusion scheme as follows:

$$STLVDM(x, y) = SLVDM(x, y) \times TLVM(x, y), \tag{7}$$

in which

$$SLVDM(x, y) = \frac{SLVDM(x, y)}{\max_{x,y}\{SLVDM(x, y)\}}, \tag{8}$$

Fig. 1 demonstrates the STLVDM graphic and corresponding 3-D image obtained based on our approach, where the real foreground object has the maximum gray value and this value is the most obvious in the image. The most salient region in the STLVDM image is believed to be the target. Employing this analysis, the conventional adaptive threshold technique is selected to segment the object, and the threshold value of STLVDM image is defined as follows:

$$Th = \mu_{STLVDM} + k \times \delta_{STLVDM}, \tag{9}$$

in which μ_{STLVDM} and δ_{STLVDM} denote the mean and standard deviation of the STLVDM, respectively, and the parameter k is an adjustable parameter. The optimizing range of k is from 5 to 20. When the intensity value of STLVDM image is bigger than the adaptive threshold, it is detected as a target.

III. EXPERIMENTS RESULTS

A. EXPERIMENTAL DATA-SET AND SETTING PARAMETERS

The comparative experiments of the presented spatial and temporal local vector difference feature measure method are evaluated on four groups of true infrared image sequences based on qualitative and quantitative performance. Seq.1 describes the scene about the weak small plane of thermal video and was selected from one of public research infrared database named Terravic Motion IR Database [30]. This well-known dataset can be downloaded from https://vciplokstate.org/pbvs/bench/Data/05/download.html. Seq.2, Seq.3, and Seq.4 are selected the scene about weak small "planes" of thermal video in recent public IR Database [17]. It is a recent public research infrared database, and some compared model and images can be downloaded from https://github.com/moradisaed/IRimages/tree/main/Images. The details of the four sequences dataset are demonstrated in table 1. To assess the detection performance of the presented method, presented approach was compared with several the state-of-the-art spatial and temporal local model including STLCF [21], STLDM [23], NSTSM [24], STF [25] and STVDM [26]. All the methods have been rigorously tested on typical thermal sequences with small moving objects. Due to training offline, supervised change detection approaches are not included in the compared methods.

TABLE 1. Details of four infrared scenes.

Scene	Frames	Background state	Target state
1	200	Gloomy sky with clutter and many bright spot noises	Size: varies from 7×5 to 4×3
2	50	Strong bright spot noise and Strong bright background	low local contrast Size： 4×3
3	50	Heavy clutter and jump changed background	low local contrast Size： 4×5
4	50	Heavy cloud and sky-cloud edge noise	Tiny Size： 2×2

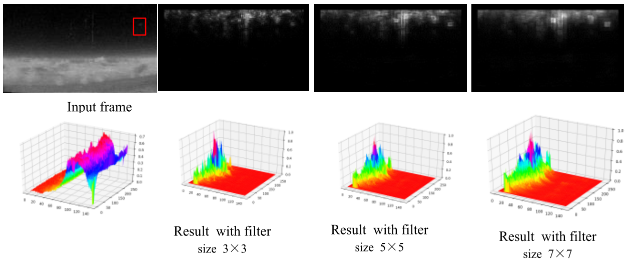

FIGURE 3. Results of spatial image with different cell size.

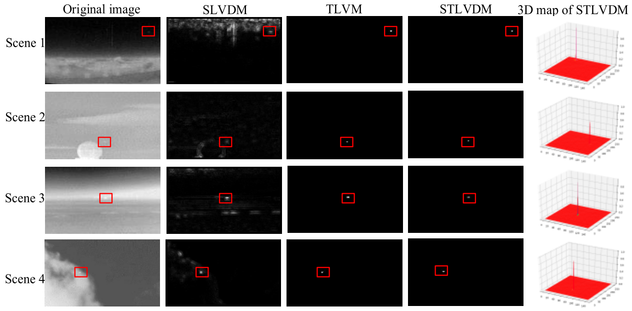

FIGURE 4. Spatial, temporal, spatial-temporal map and 3-D image obtained by the proposed model. The target is displayed in red rectangle.

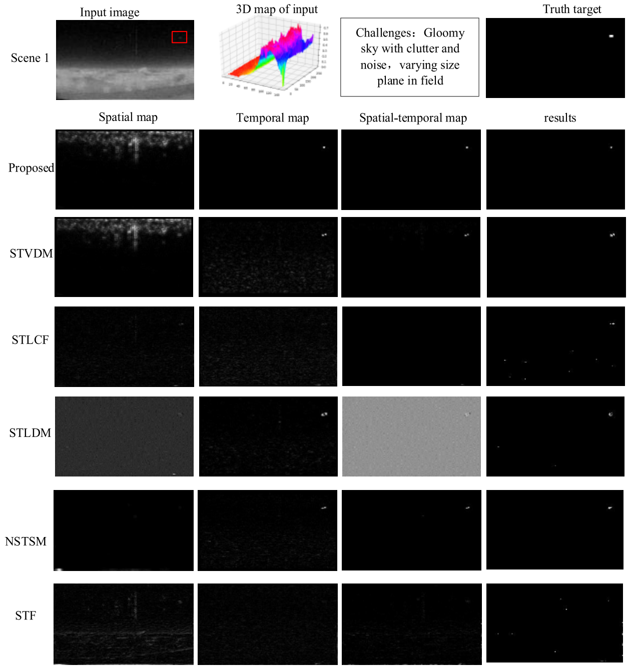

FIGURE 5. The spatial, temporal, spatial-temporal map and final result compared in scene 1.

In order to measure the effect of crucial parameters in enhancing or diminishing the performance of the proposed approach in different infrared environments, the size p of cell in window was described in this subsection. The different values of p were tested on data of "helicopter" in VOT2015. The sizes of the cells modify the spatial feature map and the ultimate results for different cell sizes is illustrated in Fig.3. According to the figure provided by our model, although the 3-D images of corresponding spatial feature map are close to each other, we can see clearly that the cluttered cloud noise has been alleviated remarkably and the objects have been enhanced in the spatial feature image with different sizes of cells.

A perfect size p is important to balancing the sensitivity, detection precision and computational complexity in the method. smaller filters will decrease the computational complexity, but it also brings down the precision of foreground detection. With these Considerations, the parameter of p is set to 3 for all videos. In addition, the STLCF, STLDM, STVDM and STLCF calculate the spatial feature map employing the use of l forward and l backward frames with $l = 4$ in the all compared approaches.

B. QUALITATIVE EVALUATIONS

The qualitative evaluation is a subjective assessment of detection performance employing visual measurement of detected binary objects' masks between the compared methods and real target' mask in the same frames. The segmentation results of spatial, temporal, spatial-temporal map and 3-D figure in four real infrared sequences are presented in Fig.4, in which the real small object is promoted and the background noise is alleviated in spatial-temporal figure. To improve object information, the spatial feature image of the proposed approach improves the issue of low local contrast and eliminate a remarkable number of background interferences and violent bright spot noises. In addition, ambiguity foreground targets are further minimized in temporal feature map. The relatively completed saliency target is acquired in spatial-temporal image by combing spatial and temporal maps. It is obvious that almost no residual background clutter exists in 3-D figure of scene 1-4 and the all the targets area have the remarkable intensity value. Obviously, although a little amount of residual bad noise can be noticed among spatial feature map, the fusing scheme and adaptive detection threshold can easily detect the target region intensity value and effectively deal with residual noise.

The qualitative comparisons result of five similar state-of-the-art spatial-temporal approaches–STLCF [21], STLDM [23], NSTSM [24], STF [25], STVDM [26] and proposed model based on public typical thermal videos dataset are demonstrated in Figs.5-8. The first column of each

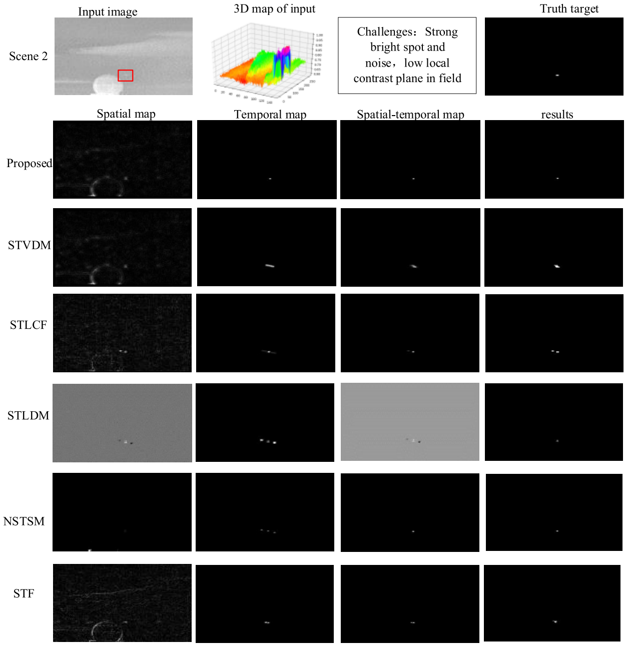

FIGURE 6. The spatial, temporal, spatial-temporal map and final result compared in scene 2.

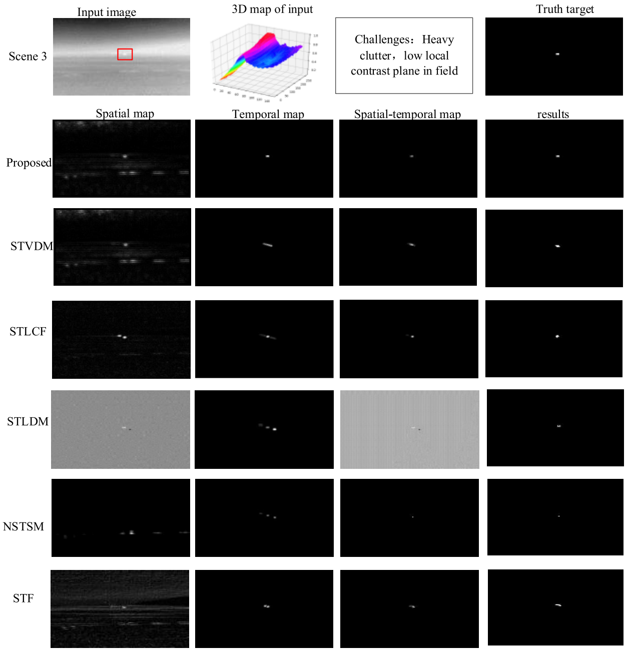

FIGURE 7. The spatial, temporal, spatial-temporal map and final result compared in scene 3.

figure demonstrates the representative frames, corresponding 3D-map, challenges and the corresponding ground-truth, respectively. The column 2-7 are comparison segmentation results of spatial, temporal, spatial-temporal feature maps and detection masks. The detection masks of compared approaches are pixel-based, and the experimental mask of each method are original segmentation results without morphological post-processing. The major difficulties of environment 1 are dim sky with clutter, varying target size, and ambiguity noise. As illustrated in Fig.5, the proposed model, STVDM, STLDM, and STF enhance the target in the spatial map and STLCM, STVDM, NSTSM, and STLCF have a ghost in temporal map. The target of spatial-temporal map of proposed model, STLDM and NSTSM is obvious. The proposed model, STLVDM and NSTSM obtain better detection masks results. The vector based spatial map enhances the gloomy sky region well and depresses the bright cloud region without affecting the real bright plane.

The major challenges of scenes 2 are strong bright spot and heavy clutter cloud. As illustrated in Fig.6, the proposed model, STVDM and STF are better at improving the object in the spatial map compared with other approaches, and ghosts have been generated by STLCF and STLDM in the spatial map. Because the target moves quickly, the STLCF, STLDM, STVDM, NSTSM and STF cannot deal with the ghost well in the temporal map. The proposed model and NSTSM obtain a complete target in the spatial-temporal map. Therefore, the proposed model and NSTSM obtain better detection results. The vector based spatial map can enhance the low local contrast target and depress the noise in temporal map.

The heavy clutter background of scene 3 is illustrated in Fig.7 and it is difficult to segment the target from background due to the pixel values of weak small targets and the background scene being close to each other. It can be seen that the NSTSM cannot manage this scene well and the other methods cannot effectively deal with the ghost in the temporal map. The heavy clutter cloud of scene 4 is illustrated in Fig.8 and it is difficult to manage the tiny and low local contrast target. The NSTSM cannot cope with this scene but our method has perfect performance by enhancing the target and decreasing the clutter background in these scenes.

C. QUANTITATIVE EVALUATIONS

To further illustrate the robustness and effectiveness of proposed vector-based spatial-temporal method, the quantitative assessment among compared algorithms are used to evaluate the model with the corresponding ground-truths. To quantify the segmentation results of methods under complex infrared sequences at pixel level, Recall, Precision, and F-measure [31] are employed, where the higher these metrics denotes better detection performance. These corresponding metrics are defined as the following equations:

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{10}$$

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$F - measure = \frac{2 \cdot \text{Recall} \cdot Precision}{\text{Recall} + Precision}, \tag{12}$$

where true positive (TP) denotes the number of correctly classed as object pixels, false negative (FN) denotes the

TABLE 2. Three metrics based on four public scenes.

SEQUENCES	METRICS	STLCF	STLDM	NSTSM	STF	STVDM	PROPOSED
SCENE 1	RECALL	0.2135	0.4270	0.4978	0.0355	0.4982	**0.6056**
	PRECISION	0.2298	0.6629	**0.9929**	0.0414	0.7734	0.9890
	F-MEASURE	0.1806	0.4758	0.6218	0.0166	0.5618	**0.6698**
SCENE 2	RECALL	**0.9897**	0.3225	0.5940	0.8910	0.9880	0.7910
	PRECISION	0.5524	0.4761	**0.9836**	0.6382	0.3690	0.6763
	F-MEASURE	0.6660	0.3418	0.6965	**0.6981**	0.5007	0.5986
SCENE 3	RECALL	0.8799	0.2762	0.3973	0.6077	**0.8839**	0.8286
	PRECISION	0.6130	0.4504	0.3314	0.4382	0.8376	**0.9152**
	F-MEASURE	0.6786	0.3010	0.3178	0.4648	0.8129	**0.8179**
SCENE 4	RECALL	0.8175	0.6815	0.2477	0.5572	0.6818	**0.8798**
	PRECISION	0.4025	0.2893	0.1368	**0.8148**	0.2312	0.6586
	F-MEASURE	0.4831	0.3068	0.1231	0.5741	0.3119	**0.6444**

TABLE 3. The average computing speed in terms of frames/second(fps).

METHODS	STLCF	STLDM	NSTSM	STF	STVDM	PROPOSED
FPS	27.9	24.2	22.1	**34.0**	17.1	12.6

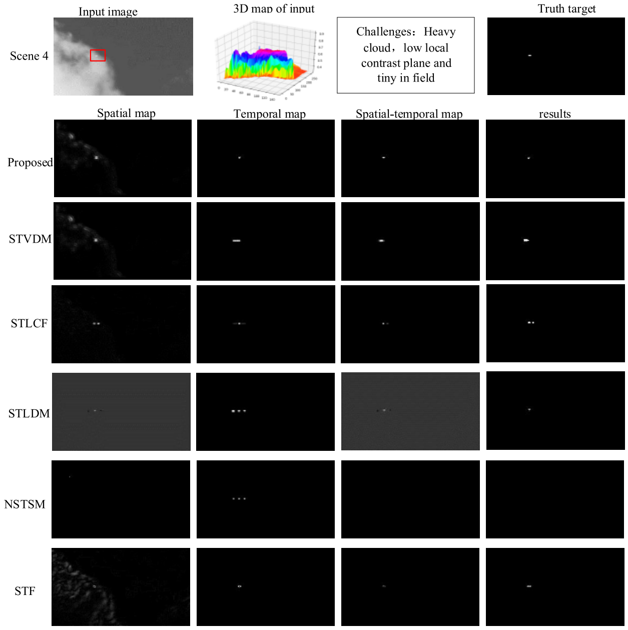

FIGURE 8. The spatial, temporal, spatial-temporal map and final result compared in scene 4.

number of objects incorrectly classed as background, and false positive (*FP*) is the total number of false positive pixels that are incorrectly classed as object. Others metrics such as percentage of correct classification and Similarity which are directly correlated to Recall, Precision, F-measure are also based on *TP*, *FP* and *FN*. Therefore, the three main metrics are commonly used in targets detection. In addition, F-measure reconciles the accuracy measurements of precision and recall using fairly weighting their harmonic balances and is employed as a frequently-used indicator of the comprehensive performance in target detection approaches. All the compared approaches were operated in the VS2012+ opencv2.4.2 environment and on a PC equipped with 2.66 GHz Pentium(R) Dual-Core CPU E5300 and 2GB RAM.

The average compared results of Recall, Precision, F-Measure under four sequences is demonstrated in Table 2. Typical sample images of scene1—scene 4 were selected to computed three metrics based on the ground-truth references, respectively. The bold value of each column demonstrates the best measurements performance and underlined values of each column denote the second-best measurements performance. As illustrated in table 3, the performance of NSTSM performs is better in terms of precision and F-Measure in scene 1 and scene 2, but it has undesirable performance in scene 4. STLCF, STLDM and STF perform well in scene 2 and scene 3, but there are a little unfavorable when they compared with our proposed method. The results of STLCF and STF are slightly adverse in scene 1. Our algorithm obtains perfect performance in average metrics in all compared datasets. According to the compared results provided by table 2, it can be seen that the presented method provides desirable F-measure results in comparison with other methods under the same infrared video sequences. The developed spatial and temporal vectors-based model enhances the spatial low local contrast features and further alleviates ambiguity targets in temporal feature map.

Finally, the average operating speed of six compared methods is calculated using the image size of 256 × 144. As can be seen from the data in table 3, the frames/second (fps) of our algorithm and STF model have the maximum and minimum operation time, respectively. Our method takes a spatial-temporal sliding convolution filters that traverses the image and video to acquire spatial vectors feature and temporal feature consensus background model. In addition, the process of forming spatial vectors and background model construction are important factors affecting computational speed. Although our method takes the most operation time, it can meet the requirement of real-time infrared surveillance environment by employing the parallel acceleration operation.

IV. CONCLUSION

In this study, a robust vector-based spatial-temporal object segmentation approach is developed to detect small moving objects in complex infrared scenes. To strengthen targets and calculate the dissimilarity between the small object and their surrounding background, a new local vector dissimilarity measure is introduced to calculate the spatial saliency feature mapping. To compute the temporal varying feature, temporal feature consensus background model is introduced based on the local mean of succession frames and the temporal saliency feature mapping is calculated employing subtraction background model. The final fusion saliency feature map is determined taking spatial and temporal feature images into account. Then the weak targets are detected employing an adaptive detection technique. Extensive experimental data have demonstrated that the presented approach is more resultful and has a remarkable accuracy performance on public data sets. However, our method is the most time-consuming and does not deal with open real infrared images. Hence, we will attempt to generate simulated infrared images by employing deep learning to test our algorithm and also pour attention into diminishing the influence of these parameters in the subsequent work.

REFERENCES

[1] X. Kong, C. Yang, S. Cao, C. Li, and Z. Peng, "Infrared small target detection via nonconvex tensor fibered rank approximation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–21, 2022, doi: 10.1109/TGRS.2021.3068465.

[2] Q. Li, J. Nie, and S. Qu, "A small target detection algorithm in infrared image by combining multi-response fusion and local contrast enhancement," *Optik*, vol. 241, Sep. 2021, Art. no. 166919.

[3] B. Garcia-Garcia, T. Bouwmans, and A. J. Rosales Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Comput. Sci. Rev.*, vol. 35, Feb. 2020, Art. no. 100204.

[4] J. Gao, Z. Lin, and W. An, "Infrared small target detection using a temporal variance and spatial patch contrast filter," *IEEE Access*, vol. 7, pp. 32217–32226, 2019.

[5] C. Wang and L. Wang, "Multidirectional ring top-hat transformation for infrared small target detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8077–8088, 2021.

[6] Y. Bi, X. Bai, T. Jin, and S. Guo, "Multiple feature analysis for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1333–1337, Aug. 2017.

[7] C. L. P. Chen, H. Li, Y. Wei, T. Xia, and Y. Y. Tang, "A local contrast method for small infrared target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 574–581, Jan. 2014, doi: 10.1109/TGRS.2013.2242477.

[8] J. Han, Y. Ma, B. Zhou, F. Fan, K. Liang, and Y. Fang, "A robust infrared small target detection algorithm based on human visual system," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2168–2172, Dec. 2014.

[9] Y. Qin and B. Li, "Effective infrared small target detection utilizing a novel local contrast method," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1890–1894, Dec. 2016.

[10] J. Han, K. Liang, B. Zhou, X. Zhu, J. Zhao, and L. Zhao, "Infrared small target detection utilizing the multiscale relative local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 612–616, Apr. 2018, doi: 10.1109/LGRS.2018.2790909.

[11] P. Du and A. Hamdulla, "Infrared small target detection using homogeneity-weighted local contrast measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 514–518, Mar. 2020, doi: 10.1109/LGRS.2019.2922347.

[12] M. Nasiri and S. Chehresa, "Infrared small target enhancement based on variance difference," *Infr. Phys. Technol.*, vol. 82, pp. 107–119, May 2017.

[13] P. Lv, S. Sun, C. Lin, and G. Liu, "A method for weak target detection based on human visual contrast mechanism," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 261–265, Feb. 2019, doi: 10.1109/LGRS.2018.2866154.

[14] C. Xia, X. Li, L. Zhao, and R. Shu, "Infrared small target detection based on multiscale local contrast measure using local energy factor," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 157–161, Jan. 2020, doi: 10.1109/LGRS.2019.2914432.

[15] L. Wu, Y. Ma, F. Fan, M. Wu, and J. Huang, "A double-neighborhood gradient method for infrared small target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 8, pp. 1476–1480, Aug. 2021, doi: 10.1109/LGRS.2020.3003267.

[16] J. Han, S. Moradi, I. Faramarzi, C. Liu, H. Zhang, and Q. Zhao, "A local contrast method for infrared small-target detection utilizing a tri-layer window," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1822–1826, Oct. 2020, doi: 10.1109/LGRS.2019.2954578.

[17] S. Moradi, P. Moallem, and M. F. Sabahi, "Fast and robust small infrared target detection using absolute directional mean difference algorithm," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107727.

[18] H. Liu, L. Zhang, and H. Huang, "Small target detection in infrared videos based on spatio-temporal tensor model," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8689–8700, Dec. 2020.

[19] D. Zeng and M. Zhu, "Multiscale fully convolutional network for foreground object detection in infrared videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 617–621, Apr. 2018.

[20] Q. Hou, Z. Wang, F. Tan, Y. Zhao, H. Zheng, and W. Zhang, "RISTDnet: Robust infrared small target detection network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3050828.

[21] L. Z. Deng, H. Zhu, C. Tao, and Y. T. Wei, "Infrared moving point target detection based on spatial–temporal local contrast filter," *Infr. Phys. Technol.*, vol. 76, pp. 168–173, May 2016.

[22] B. Zhao, S. Xiao, H. Lu, and D. Wu, "Spatial–temporal local contrast for moving point target detection in space-based infrared imaging system," *Infr. Phys. Technol.*, vol. 95, pp. 53–60, Dec. 2018.

[23] P. Du and A. Hamdulla, "Infrared moving small-target detection using spatial-temporal local difference measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1817–1821, Oct. 2020, doi: 10.1109/LGRS.2019.2954715.

[24] D. Pang, T. Shan, P. Ma, W. Li, S. Liu, and R. Tao, "A novel spatiotemporal saliency method for low-altitude slow small infrared target detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2020.3048199.

[25] J. Xie, J. Yu, J. Wu, Z. Shi, and J. Chen, "Adaptive switching spatial-temporal fusion detection for remote flying drones," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 6964–6976, Jul. 2020, doi: 10.1109/TVT.2020.2993863.

[26] Y. Zhang, K. Leng, and K. -S. Park, "Infrared detection of small moving target using spatial-temporal local vector difference measure," *IEEE Geosci. Remote Sens. Lett.*, doi: 10.1109/LGRS.2022.3157978.

[27] Y. Zhang, W. Zheng, K. Leng, and H. Li, "Background subtraction using an adaptive local median texture feature in illumination changes urban traffic scenes," *IEEE Access*, vol. 8, pp. 130367–130378, 2020, doi: 10.1109/ACCESS.2020.3009104.

[28] Y. Zhang, K. Leng, and K. Park, "Adaptive vector-based sample consensus model for moving target detection in infrared video," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7506505, doi: 10.1109/LGRS.2022.3150760.

[29] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011, doi: 10.1109/TIP.2010.2101613.

[30] (Jul. 2016). *IEEE OTCBVS WS Series Bench; Roland Miezianko Terravic Research Infrared Database*. [Online]. Available: http://vcipl-okstate.org/pbvs/bench/index.html

[31] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 393–400.

KAIYUAN LI received the master's degree in human resources management and organizational behavior from Lingnan University, Hong Kong, in 2017. She is currently pursuing the Ph.D. degree with Wuhan University, China. Her research interests include international finance, energy finance, and big data.

YUNSHENG ZHANG received the Ph.D. degree in transportation engineering from Southeast University, Nanjing, China, in 2017.

From April 2020 to September 2022, he worked as an Assistant Professor with the School of Mechanical, Gachon University, South Korea. He is currently an Associate Professor with the Hubei University of Economics and Southeast University. His research interests include image processing and intelligent transportation systems.

YIQIONG ZHANG received the B.E. degree from Northeastern University, Qinhuangdao, China, in 2021. His research interests include data mining and information security.

YITING CHEN is currently pursuing the B.S. degree with the School of Biomedical Engineering, South Central University for Nationalities. Her current research interests include information fusion and big data.

CHIHANG ZHAO received the Ph.D. degree in precision mechanics from Southeast University, Nanjing, China, in 2004. From 2006 to 2007, he was a Research Fellow with Korean University, Seoul, South Korea. From 2012 to 2013, he was a Research Professor with Griffith University, Brisbane, Australia. He is currently a Professor with the School of Transportation, Southeast University, Nanjing. His research interests include intelligent transportation and pattern recognition.