Grayscale-Thermal Tracking via Inverse Sparse Representation-Based Collaborative Encoding

Bin Kang, Dong Liang, Wan Ding, Huiyu Zhou, and Wei-Ping Zhu, *Senior Member, IEEE*

***Abstract*— Grayscale-thermal tracking has attracted a great deal of attention due to its capability of fusing two different yet complementary target observations. Existing methods often consider extracting the discriminative target information and exploring the target correlation among different images as two separate issues, ignoring their interdependence. This may cause tracking drifts in challenging video pairs. This paper presents a collaborative encoding model called joint correlation and discriminant analysis based inver-sparse representation (JCDA-InvSR) to jointly encode the target candidates in the grayscale and thermal video sequences. In particular, we develop a multi-objective programming to integrate the feature selection and the multi-view correlation analysis into a unified optimization problem in JCDA-InvSR, which can simultaneously highlight the special characters of the grayscale and thermal targets through alternately optimizing two aspects: the target discrimination within a given image and the target correlation across different images. For robust grayscale-thermal tracking, we also incorporate the prior knowledge of target candidate codes into the SVM based target classifier to overcome the overfitting caused by limited training labels. Extensive experiments on GTOT and RGBT234 datasets illustrate the promising performance of our tracking framework.**

***Index Terms*— Grayscale-thermal tracking, inverse sparse representation, discriminant analysis, feature selection.**

I. Introduction

VISUAL tracking plays a very important role in computer vision due to its many applications in video analysis [1], vehicle navigation [2] and human-computer interaction [3]. Despite the significant progress made recently, visual tracking under bad weather, such as rain and smog, remains very challenging because the visible spectrum camera only collects limited light in poor weather, making it difficult to discriminate the foreground target from the background.

(a) grayscale image (b) thermal image

Fig. 1. An example of grayscale-thermal video pairs. The grayscale video sequence has very limited illumination, while the thermal video sequence is robust to the illumination changes. Grayscale-thermal tracking aims to exploit the complementarity property to guarantee the robust tracking performance in both thermal and grayscale sequences.

With the rapid development of multimedia and internet of things, thermal infrared camera has become economically affordable. Such a camera can capture the thermal infrared radiation emitted by the subjects with temperature above absolute zero, and hence is suitable for night surveillance. For this reason, the joint use of visible spectrum camera and thermal infrared camera offers two advantages: 1) Thermal infrared camera is robust to the illumination change, which can provide complementary data to visible spectrum that are captured under poor light condition; 2) The gray feature in visible spectrum camera would contribute to solving the crossover problem in thermal infrared camera based object detection. Therefore, grayscale-thermal tracking with both grayscale and thermal features can effectively tackle the bad weather challenge [4].

In grayscale-thermal tracking, the grayscale and thermal video sequences are obtained in pairs (see Fig. 1). Exploiting the complementarity property of the grayscale and thermal information to enhance the tracking performance is actually a multi-modality fusion problem. Existing fusion methods for grayscale-thermal tracking can be briefly classified into two categories. The first one is the particle fusion based method, in which fusing two particle filter models requires to simultaneously extract robust features from the grayscale and thermal video sequences for estimating the particle weights. To this end, Cvejic *et al.* [5] adopted the color cue and the structural similarity measure, Leykin and Hammoud [6] proposed to use the likelihood of the background extraction result, and Talha and Stolkin *et al.* [7] used color-based particle filter to model the appearance of the grayscale and

Manuscript received May 15, 2019; revised October 5, 2019; accepted December 4, 2019. Date of publication December 24, 2019; date of current version January 30, 2020. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB0802300, in part by the National Natural Science Foundation of China (NSFC) under Grant 61801242, Grant 61571240, Grant 61601223, Grant 61871235, Grant 61876093, and Grant 61802206, the National Science Foundation (NSF) of Jiangsu Province under Grant BK20170915, Grant BK20181393, and Grant BK20161072, and in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mireille Boutin. *(Corresponding author: Dong Liang.)*

B. Kang is with the Department of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210023, China.

D. Liang is with the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: liangdong@nuaa.edu.cn).

W. Ding is with the Department of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China.

H. Zhou is with the Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K.

W.-P. Zhu is with the Department of Electrical and Computer Engineering, Concordia University, Montréal, QC H3G 1M8, Canada.

Digital Object Identifier 10.1109/TIP.2019.2959912

thermal targets. The particle weight estimation is sensitive to the occlusion, thus causing a bottleneck for the particle fusion based grayscale-thermal tracking methods in challenging video pairs.

The second category of multi-modality fusion relies on sparse representation to effectively overcome the occlusion [8], which is our focus in this paper. Based on the ways of exploiting the complementarity property, the sparse representation based grayscale-thermal tracking can further be categorized into two kinds: the first one is to concatenate the grayscale and the thermal images patches into a vector to sparsely model the moving target [9], [10]. Those works usually assume that the target patches extracted from the grayscale and the thermal video sequences can all work well. However, they would yield poor tracking performance in challenging video sequences where such an assumption does not hold. In fact, there exist not only the potential similarity but also a large gap between the grayscale and the thermal video sequences. Hence, the reliability of the extracted target patches cannot be guaranteed. Considering this fact, the second one integrates the multi-modality fusion and the modality reliability estimation into a unified optimization problem [11], [12], which can adaptively make the grayscale and thermal information complement with each other.

Generally speaking, the previous sparse representation based methods require the corresponding targets in the grayscale and thermal video sequences to yield a similar sparse representation. However, the state-of-the-art methods in [11] and [12] could not meet this requirement in some practical cases such as that in Fig. 1, where the dissimilarity between the target and the background in Fig. 1(a) and (b) is significant. Especially in the grayscale image, the dog is immersed in the darkness, and only a little useful information can be used to represent the appearance of the dog. The reason why it is difficult to achieve visual tracking in Fig. 1 is that there exist a chicken-and-egg problem: without exploring the correlation between the targets in the grayscale and thermal images, it is hard to directly extract discriminative information from the grayscale target. On the other hand, if we cannot use discriminative feature to represent the target, it may involve corruption in target correlation analysis. Existing methods often consider the feature selection and the target correlation as two separate issues, for example those in [9]–[12], ignoring the interdependence between them. Moreover, in contrast to grayscale-thermal tracking, the existing spectrum camera based tracking methods only pay attention to the grayscale information, and hence could not solve the dilemma in Fig. 1 either. The aforementioned observations motivate our work in this paper.

In this paper, we propose an inverse sparse representation based framework (see Fig. 2) to address the challenge in grayscale-thermal tracking, in which the inverse sparse representation, the feature selection and the multi-view correlation analysis are firstly integrated into a unified optimization problem (JCDA-InvSR model) for collaborative target candidate encoding. Secondly, the target candidate codes are used to achieve SVM based target classification. Introducing the unified optimization based target candidate encoding in the tracking framework can overcome the dilemma in challenging video pairs due to two reasons: 1) Feature selection can highlight the discriminative information in a certain image, while correlation analysis can enforce the strong target (the target can be discriminated from the background) to give more compensation to the weak target (the target is difficult to be discriminated from the background) through exploring the correlation between different kinds of images. Integrating the feature selection and the multi-view correlation analysis into the joint optimization model can alternately optimize both the target discrimination within a certain image and the target correlation in different images, thereby making full use of the complementarity property. 2) Inverse sparse representation is an extension of traditional sparse representation. It has been proved in [13] that if inverse sparse representation is considered as the target encoder, the target codes are robust to the illumination change and the occlusion. The main contributions of this paper are summarized as follows:

- We propose a collaborative encoding model called joint correlation and discriminant analysis based inver-sparse representation (JCDA-InvSR) to jointly encode the target candidates in the grayscale and thermal video sequences. Since JCDA-InvSR can achieve discriminative feature selection in a common subspace, it can simultaneously enhance the discrimination and robustness of target candidate codes in the grayscale and thermal video sequences.
- The proposed JCDA-InvSR model involves a joint optimization problem that not only minimizes the inverse sparse coding error but also maximizes the correlation between the grayscale and thermal observations. For practical applications, we propose an alternative reconstruction method to solve this optimization problem.
- We design a visual tracking framework based on JCDA-InvSR, which incorporates the prior knowledge of target candidate codes into the SVM based optimization problem as a regularizer. This can effectively avoid the overfitting caused by the insufficient training samples.
- Extensive experiments on GTOT, RGBT234 and TU-VDN datasets show that our JCDA-InvSR model not only can collaboratively encode the grayscale and thermal target candidates, but also can be used for multi-view observation encoding.

To the best of our knowledge, only a few works focus on the inverse sparse representation based grayscale-thermal tracking. Note that our previous work [14] also introduces inverse sparse representation in grayscale-thermal tracking. The main differences between this paper and the work in [14] are summarized as follows: 1) The inverse sparse representation model in [14] only explores the correlation between the grayscale and the thermal targets, but it cannot explore the inter and intra class similarity in the grayscale and the thermal sequences, respectively. The JCDA-InvSR proposed in this paper uses multi-objective programming to integrate the inver-sparse representation, the feature selection and the multi-view correlation analysis into a unified optimization problem, which cannot only explore target correlation between different image domains, but also can exploit the class similarity within certain image. 2) The inverse sparse representation

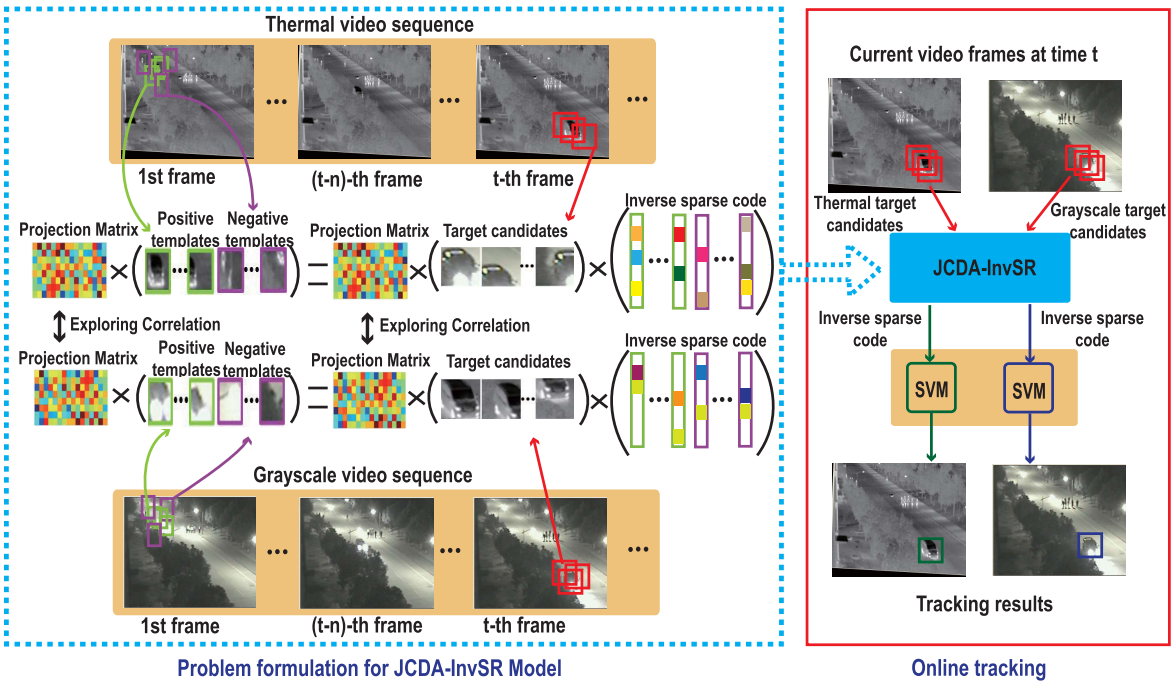


Fig. 2. The proposed inverse sparse representation based tracking framework. The JCDA-InvSR model aims to build a joint optimization to estimate the projection matrices and inverse sparse representation codes for grayscale-thermal target candidates. The projection matrices adopt multi-view correlation analysis and feature selection to exploit the correlation between grayscale and thermal targets for extracting the useful information in grayscale-thermal targets.

model in [14] is based on a single objective function. By contrast, the JCDA-InvSR requires to simultaneously optimize two objective functions. Since JCDA-InvSR is much more sophisticated than that reported in [14], we propose a practical reconstruction method to accelerate the tracking speed. 3) We have revised the training process of SVM to enhance the target classification performance.

This paper is organized as follows. In Section II, we discuss the background and the related works. Section III illustrates our proposed inverse sparse representation based collaborative encoding model in detail. Section IV uses the proposed encoding model to achieve visual tracking. Experiment results and related discussions are given in section V, and finally conclusions are presented in Sections VI.

II. Background and Related Works

A. Sparse Representation Based Visual Tracking

1) Sparse Representation in Visible Spectrum Tracking: In traditional sparse representation based visual tracking, the target observation matrix (target candidate matrix) $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$ is firstly obtained from particle filter method, then the observation likelihood is estimated by solving the following problem [15]

$$arg \min_{\mathbf{\Theta}} \|\mathbf{Y} - \mathbf{D}\mathbf{\Theta}\|_F^2 + \lambda \|\mathbf{\Theta}\|_{2,1}, \tag{1}$$

where $\mathbf{D} = [\mathbf{D}_P, \mathbf{D}_N]$ is the target dictionary, where $\mathbf{D}_P$ and $\mathbf{D}_N$ are the positive and the negative sub-matrices (the foreground and the background templates), $\mathbf{\Theta}$ is the sparse representation matrix of observation matrix $\mathbf{Y}$. Problem (1) for traditional visual tracking is often considered as a classifier, which is to calculate the importance of observation vectors $\mathbf{y}_i (i = 1, 2, \ldots, n)$ according to the reconstruction error $\|\mathbf{y}_i - \mathbf{D}\boldsymbol{\theta}_i\|_2^2$. The observation vector with the minimum reconstruction error is the final tracking result for the current frame. Since the dimension of $\mathbf{Y}$ is often very large, solving problem (1) incurs high computational complexity. Inspired by Eq. (1), extensive works [16]–[20] have been done to enhance the robustness and reduce the computational complexity of the sparse representation based visual tracking. However, those methods cannot give good tracking performance in poor light condition and severe occlusion because the discriminative target information extracted from the grayscale video sequences is very limited.

2) Inverse Sparse Representation in Visible Spectrum Tracking: Compared with Eq. (1), the inverse sparse representation model is originally proposed in [21] for visual tracking. It is in general written as

$$arg \min_{\mathbf{U}} \|\mathbf{D} - \mathbf{Y}\mathbf{U}\|_F^2 + \lambda \|\mathbf{U}\|_1. \tag{2}$$

In this problem, the target dictionary $\mathbf{D}$ is inversely represented by the observation matrix $\mathbf{Y}$, and $\mathbf{U}$ is the inverse sparse representation matric. Here, the dimension of $\mathbf{D}$ is far less than that of $\mathbf{Y}$, hence, problem (2) has obviously lower computational complexity than problem (1). Since the target dictionary is composed of the positive and the negative templates, using target candidates to inversely represent the target dictionary can indicate the likelihood of target candidates belonging to the foreground and background. This is actually a new method to make the target candidate more class discriminative. Based on this observation, problem (2) can be regarded as a target candidate encoder in traditional visual tracking. The state-of-the-art inverse sparse representation based tracking methods include [13], [22]–[25]. However, those methods only use grayscale information to achieve visual tracking, and thus may yield tracking drift in severe background clutter.

3) Sparse Representation in Grayscale-Thermal Tracking: Using sparse representation to model the grayscale-thermal video pairs, the key point is to adaptively evaluate the contributions of different sparse representation models. To solve this problem, Li *et.al.* proposed a collaborative sparse representation, which is formulated as [12]

$$arg\min_{\mathbf{\Theta},\alpha}\sum_{k=1}^{2}\frac{(\alpha^k)^s}{2}\|\mathbf{Y}^k-\mathbf{D}^k\mathbf{\Theta}^k\|_F^2 + \lambda\|\mathbf{\Theta}\|_{2,1}+\sum_{k=1}^{2}\left(\phi^k(\alpha^k)^s+(1-\alpha^k)^s\right), \quad (3)$$

where $\mathbf{Y}^1$ and $\mathbf{Y}^2$ denote the observation matrices of the grayscale and thermal video sequences, respectively. $\mathbf{\Theta} = [\mathbf{\Theta}^1, \mathbf{\Theta}^2]$ is the sparse representation matrix of $\mathbf{Y} = [\mathbf{Y}^1, \mathbf{Y}^2]$, and α^k is the reliable weight for the sparse representation error of $\mathbf{Y}^k$. Problem (3) is actually a multi-modality based sparse representation, in which α^k can be online updated for adaptive multi-modality fusion. In fact, there exist not only the potential similarity but also a large gap between grayscale and thermal target candidates, which means that only partial information in $\mathbf{Y}_i$ is good for multi-modality fusion. Moreover, problem (3) can not highlight useful information in $\mathbf{Y}^k$, and thus can not guarantee that the same target in the grayscale and the thermal images receives similar sparse representation results in challenging scenarios. In addition, problem (3) uses sparse representation to achieve multi-modality fusion, and its computational complexity is in general very high.

B. Other Related Works

1) Deep Learning in Visible Spectrum Camera Based Visual Tracking: With the rapid development of artificial intelligence, deep learning has become a useful tool in visible spectrum camera based visual tracking. According to different network structures, the state-of-the-art methods can be categorized into CNN based trackers [26]–[29], RNN based trackers [30], [31], Siamese network based trackers [32]–[35], *etc*. Different from visible spectrum camera based visual tracking, the video pairs in grayscale-thermal tracking often contain target discrimination bias (the same target has significantly different target discrimination in different image domain), which would inevitably incur label noise in appearance training. In this case, it is hard for deep learning to unleash its potential in grayscale-thermal tracking because the classification performance of deep learning is sensitive to the choice of training samples and tends to be overfitting in the presence of label noise. Due to this fact, the non-supervised appearance model, such as sparse representation, has become a priority research topic in grayscale-thermal tracking.

2) Correlation Filter in Visible Spectrum Camera Based Visual Tracking: Correlation filter based tracking methods have attracted a great deal of attention because it can convert the spatial correlation to the element-wise multiplication in the frequency domain, thereby having the advantage of being computationally efficient for real-time tracking. MOSSE [36] is the first one that learns correlation filter with few samples in the frequency domain. After this work, notable improvements have been made through introducing the kernel trick [37], the deep feature [38]–[40], the context information [41] or the spatial/temporal regularization [42], [43] *etc*, in the ridge regression model. Those methods cannot be directly applicable to grayscale-thermal tracking because they only focus on the grayscale information. If modifying aforementioned methods through adopting multiple correlation filters to simultaneously enhance the grayscale and thermal tracking performance, target discrimination bias may break the consistency of circulant samples in two video sequences, causing unstable regression estimation results.

III. Joint Correlation and Discriminant Analysis Based Inver-Sparse Representation

A. Problem Formulation

As stated in the previous section, in challenging video pairs, the target appearance may be disturbed by the adverse factors such as poor light condition, occlusion, *etc*. In this case, only partial information of the targets is useful for multi-modality fusion. This observation motivates us to extract useful information for multi-modality fusion. To this end, we first propose to build a feature selection based inverse sparse representation model to effectively encode the grayscale and the thermal target observations. Let us define $\mathbf{Y}^1 = [\mathbf{y}_1^1, \mathbf{y}_2^1, \ldots, \mathbf{y}_n^1]$ and $\mathbf{Y}^2 = [\mathbf{y}_1^2, \mathbf{y}_2^2, \ldots, \mathbf{y}_n^2]$ to be the grayscale and thermal observation matrices, and $\mathbf{D}^1$ and $\mathbf{D}^2$ are the target dictionaries for the grayscale and thermal video sequences. The proposed inverse sparse representation model is then formulated as

$$\min_{\mathbf{U},\mathbf{W}}\sum_{k=1}^{2}\Big\{\|(\mathbf{W}^k)^T\mathbf{D}^k-\left((\mathbf{W}^k)^T\mathbf{Y}^k\right)\mathbf{U}^k\|_F^2+\lambda\|\mathbf{U}^k\|_1 + Tr\left((\mathbf{W}^k)^T(\mathbf{S}_w^k-\mathbf{S}_b^k)\mathbf{W}^k\right)\Big\}, \quad (4)$$

where $\mathbf{W}^k (k = 1, 2)$ is the projection matrix for extracting the important elements in $\mathbf{Y}^k$ and $\mathbf{D}^k$, $\mathbf{U}^k$ is the inverse sparse representation result of observation projection $(\mathbf{W}^k)^T\mathbf{Y}^k$ and $Tr(\cdot)$ denotes the trace. It is worth mentioning that problem (4) combines the inverse sparse representation and an unsupervised feature selection method called the Maximum Margin Criterion (MMC) [44]. MMC aims at reducing the dimension of the observation vector through using a trained projection matrix to extract some important elements in the observation vector. Since the projection matrix $\mathbf{W}^k$ in MMC is trained through minimizing the intra-class similarity and maximizing the inter-class similarity, it can extract the discriminative elements in the observation vector to make observation projection more class discriminative. Due to the advantage of MMC, in Eq. (4), we do not directly use the grayscale and the thermal observation matrices to build the inverse sparse representation model. Instead, we make use of observation projection $(\mathbf{W}^k)^T\mathbf{Y}^k = [(\mathbf{W}^k)^T\mathbf{y}_1^k, (\mathbf{W}^k)^T\mathbf{y}_2^k, \ldots, (\mathbf{W}^k)^T\mathbf{y}_n^k]$ to stress the discriminative information in the observation matrices for the multi-modality fusion and to enhance the sparsity in $\mathbf{U}^k$. The projection matrix $\mathbf{W}^k$ is online updated through minimizing $Tr((\mathbf{W}^k)^T(\mathbf{S}_w^k-\mathbf{S}_b^k)\mathbf{W}^k)$, where $\mathbf{S}_w^k$ and

$\mathbf{S}_b^k$ are two parameter matrices in the k-th image domain, which are used to calculate the within-class and the between-class variations, respectively.

Since Eq. (4) does not exploit the target correlation between the grayscale and the thermal sequences, it may make the sparsity of the inverse sparse representation results in $\mathbf{U}^1$ different from that in $\mathbf{U}^2$ in challenging scenarios. To overcome this limitation, we introduce Canonical Correlation Analysis (CCA) [45] in Eq. (4) to enhance the difference between the positive and negative templates in the target dictionary. The CCA model is described as

$$\begin{aligned} &\max_{\mathbf{P}^1,\mathbf{P}^2} \quad Tr\big((\mathbf{P}^1)^T\mathbf{D}^1(\mathbf{D}^2)^T\mathbf{P}^2\big), \\ &s.t. \quad \sum_{k=1}^{2}(\mathbf{P}^k)^T(\mathbf{D}^k(\mathbf{D}^k)^T)\mathbf{P}^k = \mathbf{I} \end{aligned} \tag{5}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{D}^k = [\mathbf{D}_P^k, \mathbf{D}_N^k]$ is composed of the positive sub-matrix $\mathbf{D}_P^k$ and negative sub-matrix $\mathbf{D}_N^k$. Each target template pair, $\mathbf{d}_i^1$ and $\mathbf{d}_i^2$ ($\mathbf{d}_i^1 \in \mathbf{D}^1$ and $\mathbf{d}_i^2 \in \mathbf{D}^2$), have not only the potential similarity but also a large gap between the two elements in this pair. In Eq. (5), the projection matrices $\mathbf{P}^1$ and $\mathbf{P}^2$ are trained by emphasizing the correlation between $\mathbf{D}^1$ and $\mathbf{D}^2$. Thus we can maximize the common and useful information in $\mathbf{P}^1\mathbf{d}_i^1$ and $\mathbf{P}^2\mathbf{d}_i^1$. This advantage can guarantee the class discrimination in both $\mathbf{P}^1\mathbf{D}^1$ and $\mathbf{P}^2\mathbf{D}^2$. Specifically, when the grayscale sequence contains severe background clutter, the positive and negative templates are similar in grayscale target dictionary $\mathbf{D}^1$ but they can be discriminated in thermal target dictionary $\mathbf{D}^2$. Under this consideration, if each template projection pair, such as $\mathbf{P}^1\mathbf{d}_i^1$ and $\mathbf{P}^2\mathbf{d}_i^2$, can highlight its own common and useful information, $\mathbf{P}^1\mathbf{D}_P^1$ will be similar to $\mathbf{P}^2\mathbf{D}_P^2$ and $\mathbf{P}^1\mathbf{D}_N^1$ will be similar to $\mathbf{P}^2\mathbf{D}_N^2$. As such, the dissimilarity between $\mathbf{P}^1\mathbf{D}_P^1$ and $\mathbf{P}^1\mathbf{D}_N^1$ can be enhanced.

Combining Eqs. (4) and (5), the proposed JCDA-InvSR model is finally formulated as

$$\begin{aligned} arg \ \min_{\mathbf{U},\mathbf{W}} \Bigg\{ F(\mathbf{U},\mathbf{W}) := &\sum_{k=1}^{2}\Big\{\|(\mathbf{P}^k)^T\widetilde{\mathbf{D}}^k - \big((\mathbf{P}^k)^T\widetilde{\mathbf{Y}}^k\big)\mathbf{U}^k\|_F^2 \\ &+ \lambda\|\mathbf{U}^k\|_1 + Tr\big((\mathbf{W}^k)^T(\mathbf{S}_w^k - \mathbf{S}_b^k)\mathbf{W}^k\big)\Big\}\Bigg\}, \\ s.t. \ \{\mathbf{P}^k\}_{k=1}^2 = arg\max_{\mathbf{P}} \Big\{ &Tr\big((\mathbf{P}^1)^T\mathbf{C}^1(\mathbf{C}^2)^T\mathbf{P}^2 \\ &- \eta\sum_{k=1}^{2}((\mathbf{P}^k)^T\mathbf{C}^k(\mathbf{C}^k)^T\mathbf{P}^k - \mathbf{I})\big)\Big\} \end{aligned} \tag{6}$$

where $\widetilde{\mathbf{D}}^k = (\mathbf{W}^k)^T\mathbf{D}^k$, $\widetilde{\mathbf{Y}}^k = (\mathbf{W}^k)^T\mathbf{Y}^k$, and $\mathbf{C}^k = \widetilde{\mathbf{Y}}^k\mathbf{U}^k$. Problem (6) integrates Eqs. (4) and (5) into a joint optimization model, which can simultaneously explore the target similarity within a given image and between different types of images. Inspired by Eq. (5) that uses $\mathbf{D}^k$ to train projection matrix $\mathbf{P}^k$, problem (6) uses $\widetilde{\mathbf{Y}}^k\mathbf{U}^k$ ($\widetilde{\mathbf{D}}^k \approx \widetilde{\mathbf{Y}}^k\mathbf{U}^k$) to update matrix $\mathbf{P}^k$, which can enforce the same targets in two image domains to obtain similar inverse sparse representation results. Specifically, $(\mathbf{W}^k)^T\mathbf{D}^k$ ($k = 1, 2$) in Eq. (4) can extract the discriminative information from the grayscale and the thermal targets, respectively. If the grayscale sequence is obtained in poor light condition, $\mathbf{W}^1$ in grayscale domain is not enough to guarantee the sparsity of inverse sparse representation result $\mathbf{U}^1$ because the positive templates are very similar to the negative templates in $\mathbf{D}^1$. Maximizing $Tr\big((\mathbf{P}^1)^T\mathbf{C}^1(\mathbf{C}^2)^T\mathbf{P}^2\big)$ can make the discriminative thermal templates in $\widetilde{\mathbf{Y}}^2\mathbf{U}^2$ enforce the difference between the positive and negative templates in $\widetilde{\mathbf{Y}}^1\mathbf{U}^1$, which can simultaneously guarantee the sparsity of $\mathbf{U}^1$ and $\mathbf{U}^2$.

B. Discussion

Here we take a further look at the difference between JCDA-InvSR and related works to gain a better insight into the JCDA-InvSR model. The state-of-the-art works that are closely related to JCDA-InvSR model include [11], [12], [46], [47]. Those sparse representation models can be generally formulated as

$$\min_{\mathbf{\Theta}} \sum_{k=1}^{2}\Big\{\alpha^k\|\mathbf{Y}^k - \mathbf{D}^k\mathbf{\Theta}^k\|_F^2 + \lambda\|\mathbf{\Theta}^k\|_1\Big\} + g(\mathbf{\Theta}), \tag{7}$$

where $g(\mathbf{\Theta})$ denotes the regularizer that aims to explore the correlation between $\mathbf{\Theta}^1$ and $\mathbf{\Theta}^2$. Note that a different $g(\mathbf{\Theta})$ is used to restrict sparse representation result in [11], [46] and [47].

Similar to Eq. (7), Eq. (6) aims to minimize the sum of two inverse sparse representation errors to make $\mathbf{U}^1$ similar to $\mathbf{U}^2$. The main difference between Eqs. (7) and (6) is that Eq. (7) assumes both $\mathbf{Y}^1$ and $\mathbf{Y}^2$ contain discriminative target information, hence it directly uses $\mathbf{D}^k$ and $\mathbf{Y}^k$ to build the sparse representation model. In fact, the vectors in $\mathbf{Y}^k$ and $\mathbf{D}^k$ are extracted by rectangular box based sampling strategy. The rectangular samples may involve background clutter. What is more, if the target is immersed in darkness, the grayscale observation matrix $\mathbf{Y}^1$ contains negligible target information while the thermal observation matrix $\mathbf{Y}^2$ contains discriminative target information. In those cases, restricting the sparse codes only by using $g(\cdot)$ and l_1 norm could not guarantee the two sparse representation models yield similar sparse codes. To overcome the aforementioned limitations, Eq. (6) firstly uses $\big(\mathbf{P}^k\big)^T\big((\mathbf{W}^k)^T\mathbf{D}^k\big)$ to build the inverse sparse representation model, where $(\mathbf{W}^k)^T$ aims to extract discriminative target information from $\mathbf{Y}^k$ and $\mathbf{D}^k$, and $(\mathbf{P}^k)^T$ ($k = 1, 2$) aims to maximize the correlation between $(\mathbf{W}^1)^T\mathbf{D}^1$ and $(\mathbf{W}^2)^T\mathbf{D}^2$. The idea of introducing $\mathbf{W}^k$ in the inverse sparse representation is inspired by [48]. In the challenging scenarios, $(\mathbf{W}^k)^T\mathbf{D}^k$ may not extract useful target appearance, hence we secondly add the max optimization as a constraint in Eq. (6). This is to adopt CCA to optimize projection $\mathbf{P}^k$, making $\big(\mathbf{P}^1\big)^T\big((\mathbf{W}^1)^T\mathbf{D}^1\big)$ similar to $\big(\mathbf{P}^2\big)^T\big((\mathbf{W}^2)^T\mathbf{D}^2\big)$ through exploring the correlation between grayscale and thermal targets. In this way, target dictionary matrices in grayscale and thermal video sequences are enforced to be similar, and thus their corresponding inverse sparse codes can be guaranteed similar.

Eq. (6) is a two-model fusion based problem. Similar to Eq. (6), the two-classification model fusion based problem in [49] also considers integrating feature extraction and target candidate encoding into a unified optimization problem.

However, the unified optimization in [49] only uses low-rank regularizer to explore the similarity between different target candidates in the grayscale and thermal images, respectively. It ignores the target correlation between the grayscale and thermal images. This may not ensure two classification model to yield similar classification result.

C. Reconstruction Algorithm

Equation (6) is a joint optimization problem, in which the design of $\mathbf{P}$, $\mathbf{W}$ and $\mathbf{U}$ are implicitly related to each other. Inspired by [18], we propose an alternative method to solve problem (6), for which we firstly define the function $\mathcal{L} = \{\mathcal{L}_1, \mathcal{L}_2\}$ as

$$\mathcal{L}_k(\mathbf{U}^k, \mathbf{W}^k) = \|(\mathbf{P}^k)^T\widetilde{\mathbf{D}}^k - \big((\mathbf{P}^k)^T\widetilde{\mathbf{Y}}^k\big)\mathbf{U}^k\|_F^2 + \lambda\|\mathbf{U}^k\|_1 + Tr\big((\mathbf{W}^k)^T(\mathbf{S}_w^k - \mathbf{S}_b^k)\mathbf{W}^k\big), \quad k = 1, 2 \tag{8}$$

Based on Eq. (7), the objective function of problem (6) can then be rewritten as

$$arg\min_{\mathbf{U},\mathbf{W}}\left\{F(\mathbf{U}, \mathbf{W}) := \sum_{k=1}^{2}\mathcal{L}_k(\mathbf{U}^k, \mathbf{W}^k)\right\}, \tag{9}$$

The constraint in Eq. (6) is reformulated as

$$\widetilde{\mathbf{P}} = arg\max_{\widetilde{\mathbf{P}}} Tr\big((\widetilde{\mathbf{P}})^T(\mathbf{V}_1 - \mathbf{V}_2)\widetilde{\mathbf{P}}\big), \tag{10}$$

where

$$\widetilde{\mathbf{P}} = [(\mathbf{P}^1)^T, (\mathbf{P}^2)^T]^T, \quad \mathbf{V}_1 = \begin{pmatrix}\mathbf{0} & \mathbf{C}^1(\mathbf{C}^2)^T \\ \mathbf{0} & \mathbf{0}\end{pmatrix},$$
$$\mathbf{V}_2 = \begin{pmatrix}\mathbf{C}^1(\mathbf{C}^1)^T - \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^2(\mathbf{C}^2)^T - \mathbf{I}\end{pmatrix}$$

The objective function in Eq. (8) leads to a nonlinear and non-convex problem. Here we adopt the stochastic gradient decent [50] to solve the non-convex problem. We propose to alternately update the gradient with respect to $\mathbf{U}^k$ and $\mathbf{W}^k$ for solving problem (8). However, since Eq. (8) involves the l_1 norm of $\mathbf{U}^k$, we cannot directly obtain the gradient with respect to $\mathbf{U}^k$. Thus, we resort to [51] to obtain $lim_{\mu\to 0} b_\mu(\mathbf{U}^k) = \|\mathbf{U}\|_1$, where $b_\mu(\mathbf{U}^k)$ is formulated as

$$b_\mu(\mathbf{U}^k) = Tr\big((\mathbf{U}^k)^T\mathbf{V}^\star\big) - \frac{\mu}{2}\|\mathbf{V}^\star\|_F^2, \quad with\ \mathbf{V}^\star = S(\mu^{-1}\mathbf{U}^k). \tag{11}$$

Here parameter μ controls the approximation accuracy, and $S(\cdot)$ is a scaling operator, defined as $S(a) = \min(1, \max(-1, a))$. With this definition, $S(\mu^{-1}\mathbf{U}^k)$ outputs a matrix as the operator applies to each element of the matrix involved. Eq. (10) can approximate $\|\mathbf{U}^k\|_1$ as a differentiable convex function, which facilitates our calculation of the gradient with respect to $\mathbf{U}^k$.

Substituting Eq. (10) into Eq. (7), the gradient with respect to $\mathbf{U}^k$ is finally obtained as

$$\nabla_{\mathbf{U}^k}\mathcal{L}_k(\mathbf{U}^k, \mathbf{W}^k) = -\big((\mathbf{P}^k)^T\widetilde{\mathbf{Y}}^k\big)^T\big((\mathbf{P}^k)^T\widetilde{\mathbf{D}}^k - ((\mathbf{P}^k)^T\widetilde{\mathbf{Y}}^k)\mathbf{U}^k\big) + \lambda\nabla_{\mathbf{U}^k}b_\mu(\mathbf{U}^k), \tag{12}$$

where $\nabla_{\mathbf{U}^k}b_\mu(\mathbf{U}^k) = \mathbf{U}^k - \mu(\mathbf{V}^\star)^T S(\mu^{-1}\mathbf{U}^k)$.

Algorithm 1 The Alternative Reconstruction Method

Input: Target dictionary $\{\mathbf{D}^k\}_{k=1}^2$, parameter λ and η.
Output: $\{\mathbf{W}^k\}_{k=1}^2$, $\{\mathbf{P}^k\}_{k=1}^2$ and $\{\mathbf{U}^k\}_{k=1}^2$.
1: **Initialization:** $\{\mathbf{W}^k\}_{k=1}^2$ with $\mathbf{W}^k = \mathbf{0}$, $\{\mathbf{P}^k\}_{k=1}^2$ with $\mathbf{P}^k = \mathbf{0}$, $\{\mathbf{U}^k\}_{k=1}^2$ with $\mathbf{U}^k = \mathbf{0}$.
2: Generating observation matrices $\mathbf{Y}^1$ and $\mathbf{Y}^2$;
3: **while** $\|(\mathbf{U}^k)^{t+1} - (\mathbf{U}^k)^t\|_F^2 > 10^{-6}$ **do**
4: **for** $t = 1 : N$ iterations **do**
5: Solving $\max_{\widetilde{\mathbf{P}}} Tr\big((\widetilde{\mathbf{P}}^t)^T(\mathbf{V}_1^t - \mathbf{V}_2^t)\widetilde{\mathbf{P}}^t\big)$ for simultaneously updating $\{\mathbf{P}^1, \mathbf{P}^2\}$;
6: **for** $k = 1 : 2$ **do**
7: $(\mathbf{Z}^k)^{t+1} = (\mathbf{Z}^k)^t - \eta\nabla_{\mathbf{U}^k}\mathcal{L}_k((\mathbf{U}^k)^t, (\mathbf{W}^k)^t)$;
8: $(\mathbf{U}^k)^{t+1} = (1 - \alpha^t)(\mathbf{Z}^k)^{t+1} + \alpha^t(\mathbf{Z}^k)^{t+1}$;
9: $\alpha^t = \frac{2}{t+1}$;
10: $(\mathbf{W}^k)^{t+1} = (\mathbf{W}^k)^t - \eta\nabla_{\mathbf{W}^k}\mathcal{L}_k((\mathbf{U}^k)^t, (\mathbf{W}^k)^t)$
11: **end for**
12: **end for**
13: **end while**

Using the relationship between the Frobenius norm and the trace, the gradient with respect to $\mathbf{W}^k$ can be obtained as

$$\nabla_{\mathbf{W}^k}\mathcal{L}_k(\mathbf{U}^k, \mathbf{W}^k) = \big(\mathbf{Z}^k + (\mathbf{S}_w^k - \mathbf{S}_b^k)\big)\mathbf{W}^k, \tag{13}$$

where $\mathbf{Z}^k = 2\cdot\big(\mathbf{D}^k(\mathbf{D}^k)^T - \mathbf{Y}^k\mathbf{U}^k(\mathbf{D}^k)^T - \mathbf{D}^k(\mathbf{U}^k)^T(\mathbf{Y}^k)^T + \mathbf{Y}^k\mathbf{U}^k(\mathbf{U}^k)^T(\mathbf{Y}^k)^T\big)$.

The detailed reconstruction method is shown in Algorithm 1. It is worth noting that problem (6) can be considered as a special multiobjective programming with three variables and two objective functions. The traditional way to solve problem (6) normally contains two steps: i) change Eq. (6) into a single objective programming based problem, and ii) use Augmented Lagrangian method to solve it. However, there are two limitations on the traditional methods: 1) It is difficult to choose the optimal function weights. 2) It often involves lots of Lagrangian parameters, causing high computational complexity. Compared with the traditional method, Algorithm 1 is derived from the metaheuristic method, which has only a few parameters to be tuned. It has been proven in [52] that the metaheuristic method for nonlinear multi-objective problems is weakly Pareto optimal, and the optimization performance depends on the updating strategy of the variables. Based on this observation, our proposed method adopts accelerated proximal gradient to alternately update $(\mathbf{U}^k)^t$ and $(\mathbf{P}^k)^t$, which can guarantee the convergence of Algorithm 1. The main computational complexity of Algorithm 1 lies in steps 7 and 10. Since the projections $\mathbf{W}^k$ and $\mathbf{P}^k$ have reduced the dimension of target observation, the computational complexity of steps 7 and 10 is largely reduced.

IV. Inverse Sparse Representation Based Tracking

JCDA-InvSR can simultaneously encode the target candidates in the grayscale and thermal video sequences. When using JCDA-InvSR to achieve the SVM based grayscale-thermal tracking, we only have limited positive and

negative samples that are accurately labeled from first few frames. To overcome this limitation, we propose to adopt particle filter to randomly yield pseudo-labeled samples to enrich the training data set. Specifically, in the particle filter method, the state vector of a moving target at time t is denoted as $\mathbf{x}^t \in R^h$, and the observations of the state vector from time 1 to t are denoted as $\mathcal{Y}^t = \{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^t\}$. Using the Bayes rule, the posterior probability $p(\mathbf{x}^t|\mathcal{Y}^t)$ is calculated as $p(\mathbf{x}^t|\mathcal{Y}^t) \propto p(\mathbf{y}^t|\mathbf{x}^t)\int[p(\mathbf{x}^t|\mathbf{x}^{t-1})p(\mathbf{x}^{t-1}|\mathcal{Y}^{t-1})]d\mathbf{x}^{t-1}$, where $p(\mathbf{y}^t|\mathbf{x}^t)$ is the observation likelihood and $p(\mathbf{x}^t|\mathbf{x}^{t-1})$ denotes the motion model. As it is very difficult to calculate $p(\mathbf{x}^t|\mathcal{Y}^t)$ directly using the aforementioned equation, the posterior probability is instead approximated by $p(\mathbf{x}^t|\mathcal{Y}^t) = \sum_{j=1}^{n} \omega_j^t \delta(\mathbf{x}^t - \mathbf{x}_j^t)$, where δ is the Dirac measure, $\mathbf{x}_j^t$ is the j-th sampled particle at time t, and ω_j^t is the particle importance weight, which is updated by $\omega_j^t = \omega_j^{t-1} p(\mathbf{y}^t|\mathbf{x}_j^t)$. Based on the particle filter method, we can adopt the property of JCDA-InvSR model to design the pseudo-labeled training sets $\{\mathbf{y}_j^1\}_{j=1}^n$ and $\{\mathbf{y}_j^2\}_{j=1}^n$. Specifically, suppose we have calculated the inverse sparse representation result $\mathbf{u}_j^k$ of the particle observation $\mathbf{y}_j^k$, then we can obtain

$$p(\mathbf{y}_j^k|\mathbf{x}_j^k) \propto exp\big(-H(\mathbf{y}_j^k, \mathbf{u}_j^k)\big), \tag{14}$$

with

$$H(\mathbf{y}_j^k, \mathbf{u}_j^k) = \|\mathbf{y}_j^k - \mathbf{D}_P^k \mathbf{u}_j^k\|_2 - \|\mathbf{y}_j^k - \mathbf{D}_N^k \mathbf{u}_j^k\|_2, \tag{15}$$

where $\mathbf{u}_j^k$ is the j-th vector in matrix $\mathbf{U}^k$. In Eq. (13), $p(\mathbf{y}_j^k|\mathbf{x}_j^k)$ indicates the likelihood of observation $\mathbf{y}_j^k$, which means the probability of the random sample $\mathbf{y}_j^k$ belonging to the positive sample.

Based on the aforementioned analysis, we can obtain two data sets for training SVM, the first one is the labeled training set $\mathcal{S}$, which contains the inverse sparse codes of pre-labeled samples. The second one is the pseudo-labeled training set $\mathcal{B}$, which contains the inverse sparse codes of random particle samples. Using $\mathcal{S}$ and $\mathcal{B}$ for machine learning involves two challenges: 1) The labeled training samples are limited. 2) The particle filter is a dense sampling method, which means that most random samples have a high probability of belonging to positive samples, causing unbalanced samples. Note that traditional SVM cannot overcome those challenges. Inspired by [53], we revise the traditional SVM based optimization problem as

$$\begin{aligned} &\min_{\boldsymbol{\omega}} \|\boldsymbol{\omega}\|_2^2 + \mathcal{C}_{\mathcal{S}} \sum_{i=1}^{|\mathcal{S}|} \xi_{\mathcal{S},i}^k + \sum_{j=1}^{|\mathcal{B}|} \mathcal{C}_{\mathcal{B},j}^k \xi_{\mathcal{B},j}^k, \\ &subject\ to\ y_{\mathcal{S},i}^k(<\boldsymbol{\omega}, u_{\mathcal{S},i}^k> + b) \geq 1 - \xi_{\mathcal{S},i}^k \\ &\qquad i = 1, 2, \ldots, |\mathcal{S}| \\ &\qquad y_{\mathcal{B},j}^k(<\boldsymbol{\omega}, u_{\mathcal{B},j}^k> + b) \geq 1 - \xi_{\mathcal{B},j}^k \\ &\qquad j = 1, 2, \ldots, |\mathcal{B}| \end{aligned} \tag{16}$$

where $|\mathcal{S}|$ denotes the size of $\mathcal{S}$, $\mathcal{C}_{\mathcal{S}}^k$ and $\mathcal{C}_{\mathcal{B},j}^k$, $j = 1, 2, \ldots, |\mathcal{B}|$ are the parameters that control the trade-off between the function complexity and the training error, and moreover $\mathcal{C}_{\mathcal{B}}^k$ varies with the confidence of the pseudo-label $u_{\mathcal{B},j}^k$. Finally, $\xi_{\mathcal{S},i}^k$ and $\xi_{\mathcal{B},j}^k$ are the slack variables in $\mathcal{S}$ and $\mathcal{B}$, respectively.

Algorithm 2 The Detailed Tracking Process

```
Training SVM through using Eq. (15) in first 5 frames;
for Frame t = 6 : end do
    Randomly sampling target candidates around the (t−1)-th tracking result to generate observation matrices Y^1 and Y^2;
    Using Eq. (6) to calculate target codes U^1 and U^2;
    Putting target candidate codes into trained SVM to perform classification;
    for k = 1 : 2 do
        for each positive code (u_i^k)^+ do
            Calculating the likelihood of each positive candidate in the manner similar to that in [10];
        end for
    end for
    Choosing the best candidate with the highest likelihood as the tracking result in t-th frame;
    If t==50
        Updating SVM using Eq. (15);
    end
end for
```

Different from the traditional SVM, Eq. (15) considers the prior knowledge of those pseudo-labeled target candidates as the regularizer $\sum_{j=1}^{|\mathcal{B}|} \mathcal{C}_{\mathcal{B},j}^k \xi_{\mathcal{B},j}^k$, which can enhance the classification performance in the case of unbalanced samples.

Based on the new training process of SVM, the detailed tracking process is described in Algorithm 2. In fact, the inverse sparse representation cannot only be used as the target encoder but also considered as a classifier. The traditional inverse sparse representation based tracking frameworks [21]–[23] often consider inverse sparse representation as a classifier, in which the particle observation likelihood is directly used to estimate the tracking result. Those frameworks cannot effectively discriminate the target with severe background clutter and illumination changes. Different from traditional tracking frameworks, our framework does not directly use the particle observation likelihood to achieve online visual tacking. Instead, we use inverse sparse representation as the target encoder, and the particle observation likelihood as the prior knowledge to yield the pseudo-labeled samples for enriching the SVM training set (the first step in Algorithm 2). This can effectively enhance the tracking accuracy. It should be mentioned that although the authors of [13] also considered the inverse sparse representation as target encoder in visible spectrum camera based visual tracking, yet their method is not directly applicable to grayscale-thermal tracking because it cannot exploit the correlation between the grayscale and the thermal video sequences. Moreover, their method cannot overcome the overfitting problem caused by inaccurate and limited training samples.

V. Experiment Results

There are two public datasets for testing the grayscale-thermal tracking performance. One is called the GTOT benchmark [12], which contains 50 grayscale-thermal

video pairs under different scenarios with very challenging factors such as poor illumination, small target, *etc*. The other is RGBT234 [54], which contains 234 video pairs with 12 attributes. Here we use these two datasets to test the tracking performance of our method. Referring to [55], we use four objective measures (the position error, the overlap rate, the precision plot and the success plot) to evaluate the tracking performance. The position error is defined as the Euclidean distance between the central location of the tracked bounding box and the manually labeled ground truth. The overlap rate is defined as $\frac{area(B_T \cap B_G)}{area(B_T \cup B_G)}$, where B_T and B_G are the tracked bounding box for each video frame and the corresponding ground truth, respectively. The precision plot indicates accumulated position errors under different location error thresholds. The success plot reflects the accumulated success rates versus different overlap thresholds, where the success rate counts the number of video frames where the overlap rate is larger than 0.5.

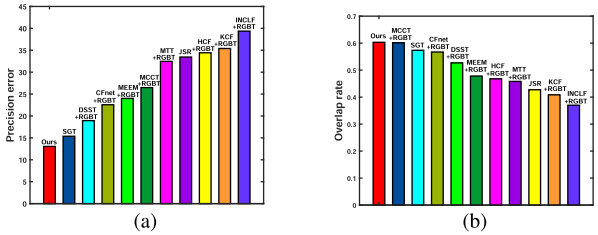
Experiment Setting: In the target coding of our tracking framework, the target dictionaries are set as $\mathbf{D}^k \in R^{256\times 300}$ $(k = 1, 2)$, in which the number of target templates is 300 (200 for foreground templates and 100 for background templates). The target dictionaries are updated in every 10 frames, and the updating strategy is similar to that in [21]. The number of randomly sampled target candidates in each frame is 600. The two parameters in Algorithm 1 are empirically set as $\lambda = 0.1$ and $\eta = 0.01$. In the target classification process, SVM is pre-trained by the target codes obtained through JCDA-InvSR from the first 5 frames, and it is online updated every 50 frames.

A. Quantitative Tracking Experiments

Existing visual tracking methods mainly focus on visible spectrum camera based tracking methods, and those methods only use grayscale video sequence to carry out visual tracking. By comparison, our tracking method makes use of grayscale-thermal video pairs and its advantage is that robust tracking in both thermal and grayscale sequences is guaranteed through making the thermal and the grayscale information complement with each other. In the following texts, we conduct the experiments to show that our method can effectively utilize the grayscale-thermal video pairs to enhance the grayscale tracking performance in the challenging video sequences with the help of thermal information. The selected tracking methods for comparison include: DSST [56], MTT [15], MEEM [57], KCF [37], INLCF [13], HCF [38], SOWP [58], CFnet [59], MCCT [60], JSR [10], CSR [12], SGT [61], GLT [54]. The detailed information of these methods are shown in Table I, where DSST, MTT, MEEM, KCF, MCCT *et.al* are spectrum camera based trackers. To make a fair comparison, we extend the spectrum camera based trackers to grayscale-thermal version. Specifically, we concatenate grayscale and thermal features into a single vector for MTT, MEEM and INLCF. For correlation filter and deep learning based trackers such as KCF, MCCT, HCF and CFnet, we consider the thermal video sequence as an extra channel. We select the state-of-the-art grayscale-thermal trackers for comparison. Since inverse sparse representation based collaborative encoding is the core of our tracking framework, we select well-known sparse representation and inverse sparse representation methods to illustrate the efficiency of collaborative encoding. In addition to those tracking methods that are similar to ours, we also select representative correlation and deep learning based methods KCF, HCF, CFnet and MCCT for comparison.

TABLE I
THE DETAIL INFORMATION ABOUT THE COMPARISON METHODS

Tracker	Type	Technique	Publication	Year
JSR	grayscale&thermal	Sparse representation	IS	2012
MTT	grayscale	Sparse representation	IJCV	2013
DSST	grayscale	Inverse sparse representation	BMVC	2014
MEEM	grayscale	SVM	ECCV	2014
KCF	grayscale	Correlation filter	PAMI	2015
HCF	grayscale	Deep learning	ICCV	2015
SOWP	grayscale	Structured SVM	ICCV	2015
CSR	grayscale&thermal	Sparse representation	TIP	2016
CFnet	grayscale	Deep learning	CVPR	2017
SGT	grayscale&thermal	Graph learning	ACM MM	2017
INLCF	grayscale	Inverse sparse representation	TCSVT	2018
MCCT	grayscale	Correlation filter	CVPR	2018
GLT	grayscale&thermal	Graph learning	PR	2019



(a) (b)

Fig. 3. The average tracking performance in GTOT dataset: (a) the mean value of position error (b) the mean value of overlap rate.

1) GTOT Dataset:

a) Overall performance: The 50 video pairs in GTOT dataset are obtained from sixteen scenarios. Here we first test the overall tracking performance with the GTOT dataset. Specifically, the average position error and the average overlap rate for one video frame are denoted as ave_p and ave_0, respectively. Based on this definition, the mean values of ave_p and ave_0 over 50 video sequences are shown in Fig. 3. The reference method that has "+**RGBT**" denotes that this visible spectrum camera based tracking method has been extended to grayscale-thermal version. From this test we can clearly see that the mean value of position error of our method is lower than the state-of-the art grayscale-thermal tracker SGT in Fig. 3(a), while our mean value of overlap rate is higher than SGT over 3% in Fig. 3(b). Although the average overlap rate of our method is similar to MCCT + RGBT, it still can illustrate the effectiveness of our collaborative encoding. The reasons are: 1) The average position error of MCCT + RGBT is obviously higher than our method; 2) MCCT + RGBT uses deep feature to represent the target appearance, while our tracking method directly uses two rough handcraft features (grayscale and thermal pixels) to achieve grayscale-thermal tracking. Effectively exploring the correlation between the grayscale and thermal targets during target encoding can make

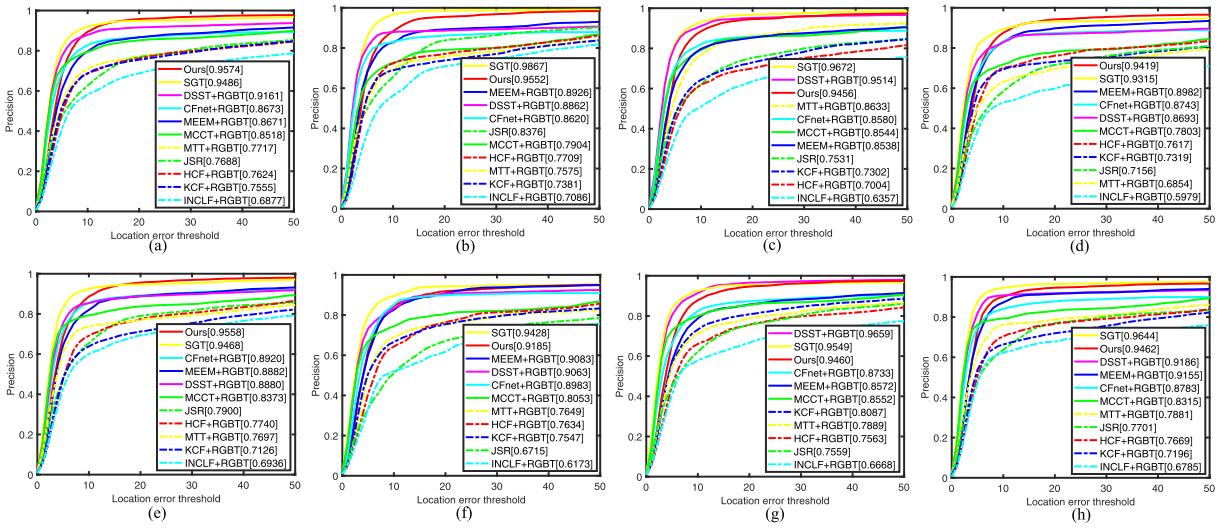


Fig. 4. The precision plots with entire dataset and different adverse factors against visual tracking in GTOT dataset: (a) Entire dataset (b) DEF subset (c) LI subset (d) OCC subset (e) TC subset (f) FM subset (g) LSV subset (h) SO subset. We show the distance precision score in the legend of precision plot, which can indicate the precision performance of different curves.

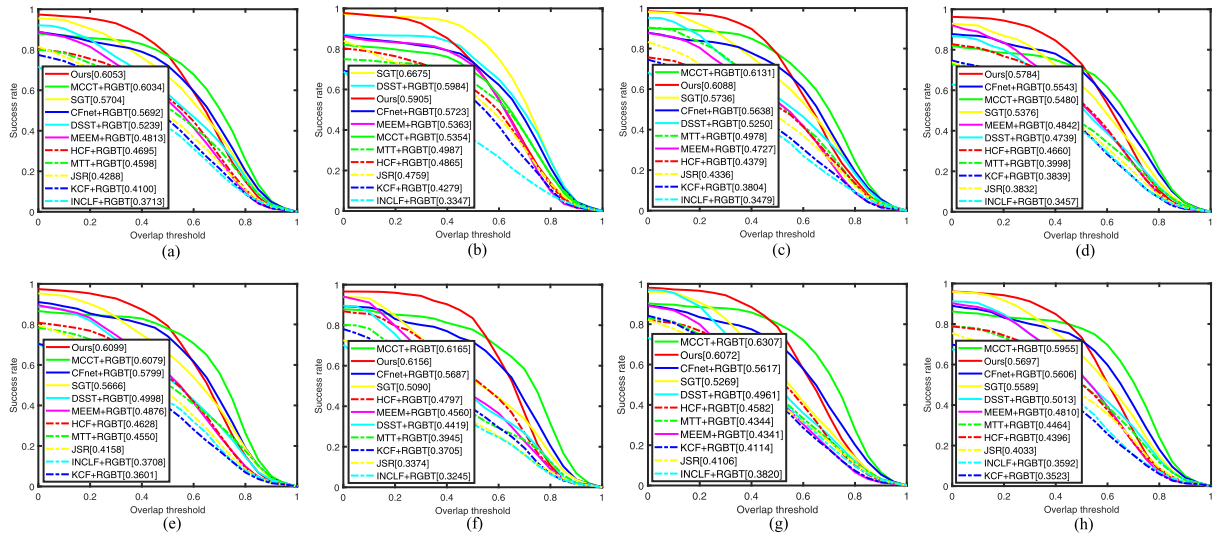
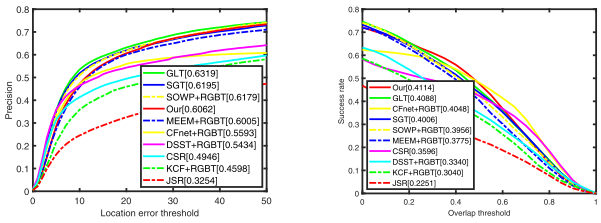


Fig. 5. The success plots with entire dataset and different adverse factors against visual tracking in GTOT dataset: (a) Entire dataset (b) DEF subset (c) LI subset (d) OCC subset (e) TC subset (f) FM subset (g) LSV subset (h) SO subset. We show the area under curve (AUC) score in the legend of success plot, which can indicate the success performance of different curves.

grayscale and thermal information complement with each other.

b) Attribute Based Performance: In the GTOT dataset, the 50 video pairs are tagged by 7 attributes, which indicate the challenging aspects in visual tracking. These 7 attributes include: Occlusion (OCC), Large Scale Variation (LSV), Fast Motion (FM), Low Illumination (LI), Thermal Crossover (TC), Small Object (SO) and Deformation (DEF). Based on different attributes, we next refer to [12] to divide 50 video pairs into 7 subsets, and give the precision and the success plots over different groups (see Fig. 4 and Fig. 5). This can give a detailed experiment on our tracking framework with different adverse factors against visual tracking.

From Fig. 4(a) and Fig. 5(a) we could see that our tracking method gives the best precision and success performance on entire dataset. The moving target often loses some important information when facing occlusion. Only using one kind of video sequence may not make up for the loss well. Different from the competitors, our tacking method can make the thermal information complement the lost information in grayscale video sequence, hence it can obviously give higher distance precision and AUC score than other 11 methods in Fig. 4(d) and Fig. 5(d). Besides OCC, the proposed tracking framework also gives the higest AUC score in TC scenario (see Fig. 5(e)). This can indicate the advantage of collaborative encoding. Low illumination is very challenging for visible spectrum camera based tracking methods. Since the JCDA-InvSR model used in our tracking framework can explore the correlation between the grayscale and the thermal video sequences, it can give higher AUC score than SGT over 9% (see Fig. 5(c)). Some video pairs in GTOT are not well aligned, this may make the grayscale and thermal



(a) (b)

Fig. 6. The overall tracking performance on RGBT234 dataset via precision and the success plots: (a) the precision plot (b) the success plot. We show the distance precision score in the legend of precision plot and the AUC score in the legend of success plot.

models capture different deformation appearance. In this case, it may bring some disturbance in correlation analysis. Due to this reason, our method could not give the highest AUC score in Fig. 5(b). Small object is very challenging for both visible spectrum camera based tracking and grayscale-thermal tracking. Small target means that the target area is less than 3%, in this case, it is hard to discriminate the target from background in both grayscale and thermal video sequences. The thermal information cannot give a strong support to the grayscale information in SO scenario. Due to this reason, our method gives a slightly lower AUC score than MCCT (see Fig. 5(h)). Although MCCT gives similar AUC score as our method in Fig 5. (c), (f), its distance precision score is obviously lower than our method.

2) RGBT234 Dataset:

a) Overall performance: RGBT234 [54] contains 234 grayscale-thermal video pairs. In this dataset, the total frame number is 210K and the maximum frames in one sequences is 8K. The video pairs captured by both static and moving cameras can be well aligned in RGBT234. Since RGBT234 contains more challenging video pairs than GTOT, it can give a comprehensive testing for our method. The overall tracking performance in RGBT234 is shown in Fig. 6, from which we clearly see that our method gives a slightly higher AUC score than GLT by 1.5%. When a tracking method loses the target in the tracking process, the output location becomes random, thus the position error does not measure the tracking accuracy correctly. Due to this reason, the precision plot in Fig. 6(a) may not fairly reflect the tracking performance. Comparing with precision plots, the success plot can fairly evaluate the overlap rate performance.

b) Attribute Based Performance: In the RGBT234 dataset, the 234 video pairs are tagged by 12 attributes. Those attributes includes: Scale Variation (SV), Fast Motion (FM), Low Illumination (LI), Thermal Crossover (TC), Deformation (DEF), None Occlusion (NO), Partial Occlusion (PO), Heavy Occlusion (HO), MB (Motion Blur), CM (Camera Moving), LR (Low Resolution) and BC (Background Clutter). Based on these attributes, we divide the whole video pairs into 12 subsets. The overlap rate score over different challenging factors is shown in Table II. Clearly, our method gives highest overlap rate score among 7 attributes. Especially in TC and LI, our overlap rate score is obviously higher than other 9 methods, which validates the effectiveness of our inverse sparse model.

B. Qualitative Tracking Experiments

In this section, we select 6 scenarios as examples to show the qualitative tracking performance (see Fig. 7). The video sequence selecting strategy is that: we randomly select 3 video sequences from each scenario. This test can give a direct impression of the tracking performance in challenging scenarios. The target is occluded by the bushes in Fig. 7(a), which would cause drift for traditional tracking methods. From Fig. 7(a) we can see that the thermal sequence can highlight the moving target. Effectively using the advantage of thermal information, our tracking method can give the best tracking performance. There is a big pool in Fig. 7(b). The reflection of the target in the water and the shade of the tree would pose great challenges for grayscale based visual tracking. With the help of thermal information, our tracking method can still follow the target. The test in Fig. 7(c) is very challenging because the moving target is stuck in the heavy rain. Since the target is very small, it is difficult to discriminate the target from the background in both grayscale and thermal video sequences. In this case, our tracking method can still follow the target with an appropriate bounding box. This test indicates the proposed JCDA-InvSR model can adaptively adjust the scale of tracking result. The video sequences in Fig. 7(d) are obtained in the dawn, which contains a lot of fog. Since our tracking framework uses JCDA-InvSR to effectively explore the correlation between the grayscale and the thermal video sequences, a better tracking performance is obtained as compared with other 10 methods. Fig 7(e) and (f) show the test in the dark scenario. In this test, the illumination is very limited. Traditional deep learning and the correlation filter based tracking methods (CFnet + RGBT and MCCT + RGBT) give inevitable drift in those two tests because they could not effectively utilize extra information to enforce the target appearance. By contrast, our tracking framework can use feature extraction and CCA to make the thermal information enforce the target discrimination in grayscale video sequence, and hence gives the best tracking performance.

C. Analysis of Collaborative Encoding Model

In this section, we will test the performance of the proposed JCDA-InvSR model which is the core in our tracking framework through following experiments.

1) Testing the Generality of Collaborative Encoding Model: The proposed JCDA-InvSR model can jointly encode the grayscale and thermal target candidates. In fact, JCDA-InvSR can not only jointly encode the grayscale and thermal target candidates, but also can be used to jointly encode multi-view observations (multi-view refers to different feature subsets used to represent particular characteristics of an object). In this test, we use TU-VDN dataset [62] as an example to test the collaborative encoding performance on multi-view observations (see Fig. 8). TU-VDN dataset is a thermal dataset, which contains four scenarios: fog, dust, low light and rain. Each scenario contains three adverse factors such as flat cluttered

TABLE II

MEAN VALUE OF OVERLAP RATE OVER DIFFERENT VIDEO SUBSETS IN RGBT234 DATASET. THE BEST TWO RESULTS ARE DENOTED AS RED AND BLUE

Seq. \ Meth.	Our	CFnet+RGBT	SGT	SWOP+RGBT	GLT	MEEM+GRBT	JSR	KCF+RGBT	CSR	DSST+RGBT
BC	0.29	0.24	0.27	0.32	0.31	0.30	0.15	0.21	0.23	0.25
CM	0.38	0.32	0.39	0.40	0.41	0.37	0.21	0.29	0.33	0.28
DEF	0.43	0.37	0.42	0.42	0.44	0.38	0.19	0.31	0.35	0.34
FM	0.34	0.23	0.32	0.30	0.32	0.30	0.13	0.24	0.27	0.25
HO	0.36	0.27	0.34	0.32	0.34	0.31	0.17	0.23	0.26	0.25
LI	0.20	0.16	0.15	0.15	0.15	0.14	0.10	0.11	0.13	0.12
LR	0.29	0.26	0.34	0.31	0.33	0.29	0.21	0.25	0.21	0.30
MB	0.34	0.22	0.30	0.31	0.32	0.27	0.15	0.19	0.25	0.19
NO	0.47	0.56	0.49	0.45	0.48	0.43	0.31	0.38	0.44	0.41
PO	0.46	0.45	0.41	0.43	0.43	0.41	0.24	0.34	0.40	0.38
SV	0.34	0.34	0.28	0.27	0.28	0.26	0.17	0.21	0.29	0.24
TC	0.47	0.37	0.42	0.40	0.42	0.36	0.26	0.29	0.38	0.37
Average	0.37	0.31	0.34	0.34	0.35	0.31	0.19	0.25	0.29	0.28

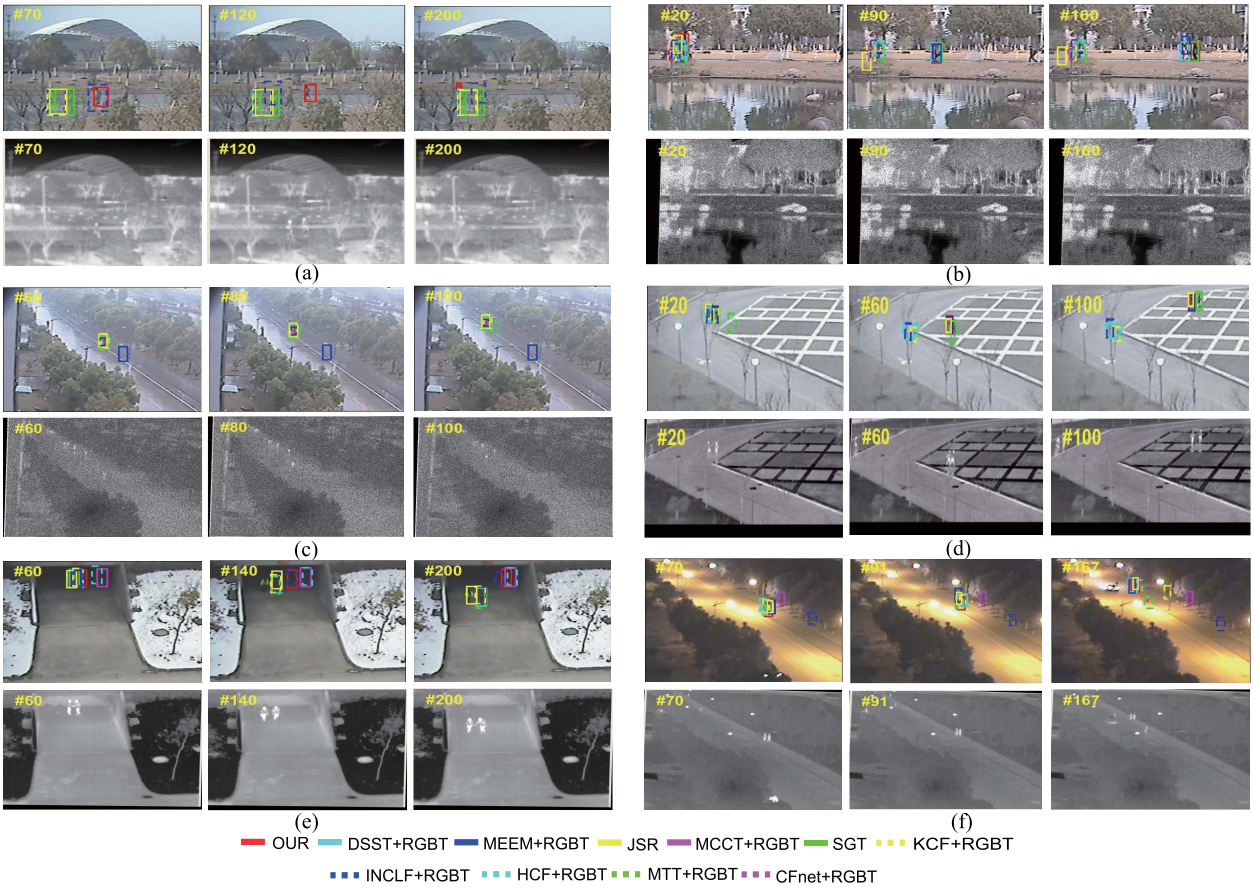


Fig. 7. The qualitative results on six video pairs. (a) FastCar2 video pair (b) Pool video pair (c) RainyMotor2 video pair (d) Running video pair (e) Tunnel video pair (f) WalkingNig video pair.

background, temperature polarity changes and background dynamics. In the test, the two view observations are obtained from the thermal pixel and the LBP method, respectively. From Fig. 8, we can clearly see that our JCDA-InvSR model only uses two rough observations but can give a similar overlap rate score as SRDCF does.

2) Ablation Experiment With Robustness Evaluation: In the grayscale-thermal tracking, the initial position is obtained from the ground truth in the first frame trackers. In fact, the tracking performance may be sensitive to the position initialization. Considering the observation, the robustness of a tracking method is often evaluated by two objective measures: TRE and SRE. TRE is to calculate the average overlap rate by using several position initializations at different time instances. SRE is defined as the average overlap rate by shifting the ground truth 12 times at the first frame. TRE is suitable for evaluating the model robustness for long term video sequences. Since the GTOT dataset contains a large number of short term sequences (less than 200 frames), TRE may not give accurate evaluation. Thus, we follow the same procedure of using SRE as in [55] to carry out the ablation experiment. Specifically, we add three competitors in the experiments:

TABLE III

ABLATION EXPERIMENT WITH ROBUSTNESS EVALUATION. THE BEST RESULTS ARE DENOTED AS RED

Challenges	grayscale performance					thermal performance				
	CSR	Our-III	Our-II	Our-I	Our	CSR	Our-III	Our-II	Our-I	Our
DEF	0.49	0.23	0.35	0.30	0.45	0.43	0.25	0.37	0.32	0.44
LI	0.47	0.30	0.39	0.37	0.51	0.49	0.21	0.34	0.21	0.50
OCC	0.44	0.31	0.45	0.42	0.52	0.40	0.27	0.42	0.35	0.52
TC	0.46	0.30	0.30	0.34	0.51	0.41	0.21	0.33	0.25	0.47
FM	0.51	0.31	0.37	0.35	0.49	0.40	0.32	0.45	0.43	0.51
LSV	0.49	0.35	0.41	0.39	0.54	0.54	0.34	0.42	0.44	0.52
SO	0.42	0.35	0.21	0.37	0.43	0.29	0.27	0.36	0.30	0.40
Average	0.46	0.30	0.35	0.36	0.49	0.42	0.26	0.38	0.32	0.48

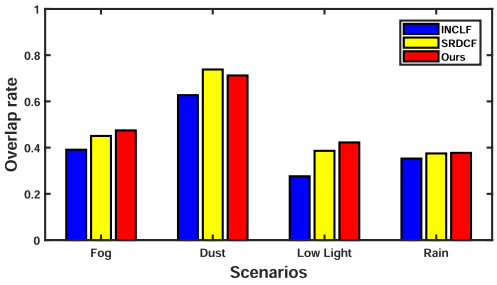


Fig. 8. Average overlap rate performance of four scenarios on TU-VDN dataset.

i) **Our-I**: removing the optimization element of $\mathbf{P}^k$, and only introducing projection $\mathbf{W}^k$ in inverse sparse representation based collaborative encoding model (see equation (4)). *ii*) **Our-II**: removing the optimization element of $\mathbf{W}^k$, and only introducing projection $\mathbf{P}^k$ in the inverse sparse representation based collaborative encoding model. *iii*) **Our-III**: removing both of $\mathbf{W}^k$ and $\mathbf{P}^k$ in the inverse sparse representation based target encoding model, which is similar to INCLF-RGBT method. Our proposed JCDA-InvSR model (see equation (6)) simultaneously introduces projection $\mathbf{W}^k$ and $\mathbf{P}^k$ in the inverse sparse representation, which is called **Our** here.

The 50 video pairs in GTOT dataset can be divided into seven subsets. In the ablation experiment, we evaluate the mean value of overlap rate over seven subsets as shown in Table III. This table shows that: **1**) The average overlap rate score of **Our** is obviously higher than **Our-I**, **Our-II** and **Our-III**. This indicates that alternately updating $\mathbf{W}^k$ and $\mathbf{P}^k$ can enhance the robustness of our JCDA-InvSR model; **2**) The robustness of our tracking method with JCDA-InvSR model is better than the CSR method; **3**) The second and the seventh rows of Table III indicate that **Our-II** and **Our-III** give different overlap score in grayscale and thermal images while **Our** still gives similar overlap score in those two rows. This implies that JCDA-InvSR model can yield similar grayscale and thermal target codes.

D. Parameter Analysis

There are two parameters η and λ required to be tuned in Algorithm 1, where η controls the convergence rate of reconstruction algorithm and λ controls the sparsity of the inverse sparse representation result. We refer to [63] to carry out the parameter analysis in Table IV. Specifically, we firstly fix $\eta = 0.01$, then use AUC and distance precision scores [55] to evaluate the tracking accuracy on GTOT dataset with different settings of λ. Similarly, we fix $\lambda = 0.1$ and evaluate the FPS (Frame number Per Second) with different η values. From Table IV we can see that the proposed JCDA-InvSR model is not sensitive to the variation of η and λ.

TABLE IV

THE TRACKING PERFORMANCE WITH DIFFERENT PARAMETER VALUE

λ	Setting	0.001	0.01	0.1	1
	AUC	0.52	0.56	0.60	0.51
	Precision	0.82	0.86	0.90	0.80
η	Setting	0.0001	0.001	0.01	0.1
	FPS	0.3	0.9	1.6	NaN

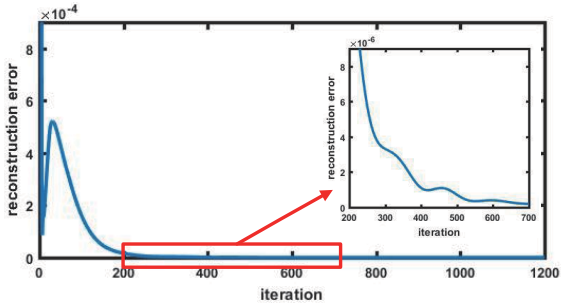


Fig. 9. The convergence of Algorithm 1.

E. Convergence and Computational Complexity

The convergence experiment for Algorithm 1 is shown in Fig. 9. It is clearly seen that the proposed reconstruction method begins to converge after 100 iterations, and after 400 iterations, the reconstruction error reaches a steady state. The main computational complexity of our tracking framework lies in the JCDA-InvSR model. Here, we use FPS (Frame num Per Second) to evaluate the computational complexity of our method. The average FPS is carried out on a desktop with Inter (R) Core (TM) i3-2310M CPU @ 2.10Hz (2GB RAM) (see Table III), where different methods are all implemented on GTOT dataset.

Table V gives the computational complexity comparison of our method and traditional visible spectrum camera based

TABLE V
FPS PERFORMANCE ON DIFFERENT SPECTRUM CAMERA BASED TRACKERS

Tracker	Our	DSST	CFnet	KCF	MEEM	MTT	INCLF	MCCT	SWOP	HCF	MDnet	MCPF	SimaFC
Compiler	matlab	matlab	python	matlab	matlab & C++	matlab	matlab	matconvnet	matlab	python	python	matconvnet	python
FPS	1.6	10.6	29.5	40.2	33.2	4.3	2.5	2.9	4.9	7.8	0.8	0.2	31.5

TABLE VI
FPS PERFORMANCE ON DIFFERENT GRAYSCALE-THERMAL TRACKERS

Tracker	Our	L1-PF	CSR	SGT	LGMG [64]	MCSR [48]
Compiler	matlab	matlab&C++	matlab	matlab	matlab	matlab
FPS	1.6	7.0	0.9	2.3	1.3	0.3

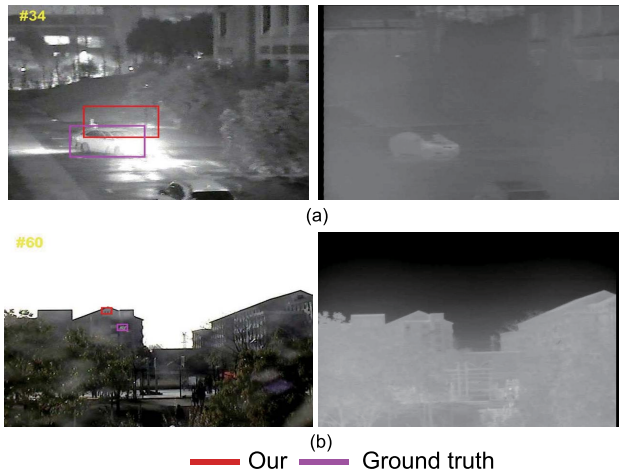


Fig. 10. Two examples of failure case. (a) carlight video pair (b) kite2 video pair.

tracking methods, where HFC, CFnet, MDnet and SimaFC are deep learning based trackers whose FPS counts only the online tracking speed without including off-line training. Moreover, we give the computational complexity of our method and the state-of-the-art grayscale-thermal tracking methods in Table VI. Clearly, our tracking speed is faster than the sparse representation based grayscale-thermal tracking method CSR. From Tables IV and V we can conclude that although our tracking speed is slower than fast tracking methods such as CFnet, KCF and SimaFC, its speed is comparable to some correlation filter and deep learning trackers such as MCCT, MCPF and HCF. This means that our joint optimization has moderate computational complexity only.

F. The Experiment Discussion

The aforementioned experiments have validated the effectiveness of our JCDA-InvSR model. However, it should be mentioned that our model does not work well in some special cases (see Fig. 10). The video pair in Fig. 10(a) is very challenging because the grayscale video sequence is captured by moving camera, and the appearance of the car is seriously disturbed by the motion blur and headlamps. In addition to above challenges, the outline of the car is not clear in thermal sequence. In this scenario, the handcraft feature obtained from thermal image could not give strong support to the severe disturbance of the grayscale image. This would incur side effect when using CCA to enforce the commonality between grayscale and thermal target. The case in Fig. 10(b) is more challenging than Fig. 10(a) because it is hard to discriminate the flying kite from the background in both grayscale and thermal images. In this case, the feature selection matrices used in our JCDA-InvSR model could not extract useful information to make two handcraft features complement with each other. Introducing deep features in our JCDA-InvSR model may enhance the tracking accuracy in the aforementioned two cases, and thus will be considered in our future work.

VI. CONCLUSION

In this paper, we have proposed an inverse sparse representation based tracking framework by using both grayscale and thermal video sequences. Our tracking framework has benefited from the proposed JCDA-InvSR model that can adopt multi-objective programming to integrate the feature selection and the multi-view correlation analysis into a unified optimization. This can simultaneously highlight the special characters of the grayscale and thermal targets through alternately optimizing two aspects: the target discrimination within a given image and the target correlation in different images. Extensive experiments on GTOT and RGBT234 datasets indicate that our tracking framework can give a superior performance as compared to many other state-of-the-art techniques.

REFERENCES

[1] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3249–3260, Jul. 2016.
[2] S. Y. Huang *et al.*, "Deep learning driven visual path prediction from a single image," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5892–5904, Dec. 2016.
[3] R. Ji, H. Liu, L. Cao, D. Liu, Y. Wu, and F. Huang, "Toward optimal manifold hashing via discrete locally linear embedding," *IEEE Trans. Image Process*, vol. 26, no. 11, pp. 5411–5420, Nov. 2017.
[4] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, 2014.
[5] N. Cvejic *et al.*, "The effect of pixel-level fusion on object tracking in multi-sensor surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.
[6] A. Leykin and R. I. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Mach. Vis. Appl.*, vol. 21, no. 4, pp. 587–595, Jun. 2010.
[7] M. Talha and R. Stolkin, "Particle filter tracking of camouflaged targets by adaptive fusion of thermal and visible spectra camera data," *IEEE Sensors J.*, vol. 14, no. 1, pp. 159–166, Jan. 2014.
[8] X. Zhang, D.-S. Pham, S. Venkatesh, W. Liu, and D. Phung, "Mixed-norm sparse representation for multi view face recognition," *Pattern Recognit.*, vol. 48, no. 9, pp. 2935–2946, Sep. 2015.
[9] Y. Wu, E. Blasch, G. Chen, L. Bai, and H. Ling, "Multiple source data fusion via sparse representation for robust visual tracking," in *Proc. Int. Conf. Inf. Fusion*, 2011, pp. 1–8.
[10] H. Liu and F. Sun, "Fusion tracking in color and infrared images using joint sparse representation," *Inf. Sci.*, vol. 55, no. 3, pp. 590–599, 2012.

[11] C. Li, X. Sun, X. Wang, L. Zhang, and J. Tang, "Grayscale-thermal object tracking via multitask Laplacian sparse representation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 673–681, Apr. 2017.

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

[13] F. Liu, T. Zhou, C. Gong, K. Fu, L. Bai, and J. Yang, "Inverse nonnegative local coordinate factorization for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1752–1764, Aug. 2017.

[14] W. Ding, B. Kang, Q. Zhou, M. Lin, and S. Zhang, "Grayscale-thermal tracking via canonical correlation analysis based inverse sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 3985–3989.

[15] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 367–383, Jan. 2013.

[16] B. Ma, J. Shen, Y. Liu, H. Hu, L. Shao, and X. Li, "Visual tracking using strong classifier and structural local sparse descriptors," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1818–1828, Oct. 2015.

[17] T. Zhang, A. Bibi, and B. Ghanem, "In defense of sparse tracking: Circulant sparse tracker," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3880–3888.

[18] H. Fan and J. Xiang, "Robust visual tracking with multitask joint dictionary learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 1018–1030, May 2017.

[19] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M.-H. Yang, "Structure-aware local sparse coding for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3857–3869, Aug. 2018.

[20] F. Liu, C. Gong, T. Zhou, K. Fu, X. He, and J. Yang, "Visual tracking via nonnegative multiple coding," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2680–2691, Dec. 2017.

[21] B. Zhuang, H. Lu, Z. Xiao, and D. Wang, "Visual tracking via discriminative sparse similarity map," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1872–1881, Apr. 2014.

[22] D. Wang, H. Lu, Z. Xiao, and M.-H. Yang, "Inverse sparse tracker with a locally weighted distance metric," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2646–2657, Sep. 2015.

[23] Y. Zhou, J. Han, X. Yuan, Z. Wei, and R. Hong, "Inverse sparse group lasso model for robust object tracking," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1798–1810, Aug. 2017.

[24] Y. Yang, W. Hu, W. Zhang, T. Zhang, and Y. Xie, "Discriminative reverse sparse tracking via weighted multitask learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 1031–1042, May 2017.

[25] Y. Yang, W. Hu, Y. Xie, W. Zhang, and T. Zhang, "Temporal restricted visual tracking via reverse-low-rank sparse learning," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 485–498, Feb. 2017.

[26] Z. Chi, H. Li, H. Lu, and M.-H. Yang, "Dual deep network for visual tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2005–2015, May 2017.

[27] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4293–4302.

[28] Y. Song, M. Chao, L. Gong, J. Zhang, and M. H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 2555–2564.

[29] Y. Song *et al.*, "VITAL: Visual tracking via adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8990–8999.

[30] H. Fan and H. Ling, "SANet: Structure-aware network for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Jun. 2017, pp. 42–49.

[31] T. Yang and A. B. Chan, "Recurrent filter learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2017, pp. 2010–2019.

[32] Q. Wang, Z. Teng, J. Xing, J. Gao, and S. Maybank, "Learning attentions: Residual attentional Siamese network for high performance online visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

[33] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[34] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[35] Z. Zhu, Q. Wang, L. Bo, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[36] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. Liu, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[37] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[38] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 3074–3082.

[39] Y. Qi *et al.*, "Hedged deep tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4303–4311.

[40] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4335–4343.

[41] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1387–1395.

[42] C. Sun, D. Wang, H. Lu, and M.-H. Yang, "Learning spatial-aware regressions for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8962–8970.

[43] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4904–4913.

[44] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.

[45] S. Sun, X. Xie, and M. Yang, "Multiview uncorrelated discriminant analysis," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3272–3284, Dec. 2016.

[46] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for rgb-infrared object tracking," *Pattern Recognit. Lett.*, to be published.

[47] X. Lan, M. Ye, R. Shao, B. Zhong, P. C. Yuen, and H. Zhou, "Learning modality-consistency feature templates: A robust RGB-infrared tracking system," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9887–9897, Dec. 2019.

[48] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, and S. Ji, "Discriminant sparse neighborhood preserving embedding for face recognition," *Pattern Recognit.*, vol. 45, no. 8, pp. 2884–2893, Aug. 2012.

[49] X. Lan, M. Ye, S. Zhang, and P. C. Yuen, "Robust collaborative discriminative learning for RGB-infrared tracking," in *Proc. AAAI Conf. Artif. Intell.*, 2018.

[50] H. Hu, K. Wang, C. Lv, J. Wu, and Z. Yang, "Semi-supervised metric learning-based anchor graph hashing for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 739–754, Feb. 2019.

[51] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, and T.-S. Chua, "Multi-label visual classification with label exclusive context," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 834–841.

[52] X. Gandibleux, *Multiple Criteria Optimization: State of the Art Annotated Bibliographic Surveys*. Dordrecht, The Netherlands: Springer, 2006.

[53] F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector machines for classification: A review," *Neurocomputing*, vol. 71, no. 7, pp. 1578–1594, 2008.

[54] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, 2019, doi: 10.1016/j.patcog.2019.106977.

[55] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[56] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[57] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Eur. Conf. Comput. Vis.*, 2014.

[58] H.-U. Kim, D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "SOWP: Spatially ordered and weighted patch descriptor for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2016, pp. 3011–3019.

[59] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5000–5008.

[60] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4844–4853.
[61] C. Li, N. Zhao, Y. Lu, C. Zhu, and J. Tang, "Weighted sparse representation regularized graph learning for RGB-T object tracking," in *Proc. ACM Int. Conf. Multimedia*, 2017.
[62] A. Singha and M. K. Bhowmik, "TU-VDN: Tripura University video dataset at night time in degraded atmospheric outdoor conditions for moving object detection," in *Proc. IEEE Conf. Image Process.*, Sep. 2019, pp. 2936–2940.
[63] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, "Robust visual tracking via consistent low-rank sparse learning," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 171–190, 2015.
[64] C. Li, C. Zhu, J. Zhang, B. Luo, X. Wu, and J. Tang, "Learning local-global multi-graph descriptors for RGB-T object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2913–2926, Oct. 2019.

Bin Kang received the M.S. degree in circuits and systems from Lanzhou University in 2011, and the Ph.D. degree in electrical engineering from the Nanjing University of Posts and Telecommunications in 2016. He is currently working with the College of Internet of Things, Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition.

Dong Liang received the B.S. degree in telecommunication engineering and the M.S. degree in circuits and systems from Lanzhou University, China, in 2008 and 2011, respectively, and the Ph.D. degree from the Graduate School of IST, Hokkaido University, Japan, in 2015. He is currently an Assistant Professor with Pattern Recognition and Neural Computing Laboratory, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA). His research interests include computer vision and pattern recognition.

Wan Ding received the M.S. degree from the Department of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, China. Her research interests include visual tracking.

Huiyu Zhou received the B.Eng. degree in radio technology from the Huazhong University of Science and Technology, China, the M.Sc. degree in biomedical engineering from the University of Dundee, U.K., and the Dr. Phil. degree in computer vision from Heriot-Watt University, Edinburgh, U.K. He is currently a Reader with the Department of Informatics, University of Leicester, U.K. He has authored over 250 peer-reviewed articles in the field. His research work has been or is being supported by U.K. EPSRC, MRC, EU, Royal Society, Leverhulme Trust, Puffin Trust, Invest NI, and industry.

Wei-Ping Zhu (SM'97) received the B.E. and M.E. degrees from the Nanjing University of Posts and Telecommunications,, Nanjing, China, and the Ph.D. degree from Southeast University, Nanjing, in 1982, 1985, and 1991, respectively, all in electrical engineering.

He was a Postdoctoral Fellow from 1991 to 1992, and a Research Associate at the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada, from 1996 to 1998. From 1993 to 1996, he was an Associate Professor with the Department of Information Engineering, Nanjing University of Posts and Telecommunications. From 1998 to 2001, he worked with hi-tech companies in Ottawa, Canada, including Nortel Networks and SR Telecom Inc. Since July 2001, he has been with Concordia's Electrical and Computer Engineering Department as a full-time Faculty Member, where he is currently a Full Professor. Since 2008, he has been an Adjunct Professor with the Nanjing University of Posts and Telecommunications. His research interests include digital signal processing fundamentals, speech, and audio processing. He was the Chair-Elect of the Digital Signal Processing Technical Committee (DSPTC) of the IEEE Circuits and System Society from 2012 to 2014, and is currently the Chair of the DSPTC. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART I: FUNDAMENTAL THEORY AND APPLICATIONS from 2001 to 2003, and an Associate Editor of the *Circuits, Systems and Signal Processing* from 2006 to 2009. He was also a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special issues of: Broadband Wireless Communications for High Speed Vehicles, and Virtual MIMO from 2011 to 2013. Since 2011, he has served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS PART II: EXPRESS BRIEFS. He currently serves as an Associate Editor for the *Journal of The Franklin Institute*.