Incorporating Deep Background Prior Into Model-Based Method for Unsupervised Moving Vehicle Detection in Satellite Videos

Chao Xiao, Ting Liu, Xinyi Ying, Yingqian Wang, Miao Li, Li Liu, *Senior Member, IEEE*, Wei An, and Zhijie Chen

***Abstract*— Background reconstruction is a key step of moving object detection in satellite videos. Most existing model-based methods exploit low-rank prior to recover background, which has achieved good performance but suffered degradation under complex and dynamic scenes. In this article, we introduce a deep background prior into model-based methods for moving vehicle detection in satellite videos. Our deep background prior is obtained by a background reconstruction network, which can learn to reconstruct the background from consecutive frames. By applying our deep background prior into model-based methods, a closed-form solution can be obtained via the alternating direction method of multipliers (ADMM), and then, detection results can be acquired through iterative optimization. More importantly, our background reconstruction network can be trained in an unsupervised way by introducing specifically designed loss, thus relieving the dependence on large-scale labeled datasets. Extensive experimental results demonstrate the efficiency and effectiveness of the proposed method.**

***Index Terms*— Background reconstruction, iterative optimization, moving object detection (MOD), satellite videos, unsupervised learning.**

I. Introduction

With the development of remote sensing technology in recent years, video surveillance from satellites has become an effective way for many applications, such as urban monitoring [1], resource exploration [2], and traffic condition monitoring [3], [4]. For these applications, moving object detection (MOD) plays a fundamental role to locate objects of interest. However, MOD, especially moving vehicle detection in satellite videos, is extremely challenging due to the following aspects.

1) *Small Object Sizes:* Due to the low spatial resolution (e.g., the ground sampling distance (GSD) of Jilin-1 is around 1 meter), moving vehicles captured by satellite videos are with small object sizes and usually smaller than 5×5 pixels, leading to a lack of appearance and texture information.
2) *Low Contrast to the Complex and Dynamic Background:* Due to various complex and dynamic scenes, the moving vehicles are sometimes with low contrast to backgrounds, which are difficult to be distinguished from background clutters.
3) *Insufficient Labeled Data:* Although the camera on satellites can provide continuous observation of the Earth and obtain a large number of satellite videos, collecting and manually annotating large datasets for MOD require considerable time, effort, and cost, thus hindering the research process.

Due to the merit of flexibility, model-based methods [3], [5], [6], [7], [8], [9], [10] have been widely investigated for MOD in satellite videos, in which background subtraction plays an important role to segment target from its adjacent pixels. However, existing methods [6], [7], [8] generally adopt handcrafted priors (e.g., low-rank prior [6], [7], [8]) with regularization terms and hand-tuned parameters to the model background. When dealing with complex scenes, these methods can not accurately reconstruct the background, thus limiting the detection performance. Moreover, due to the introduction of various regularization terms, most of these methods are computationally expensive as they need iterative optimization.

Recently, deep neural networks have demonstrated their effectiveness in serving as implicit image priors and have achieved remarkable performance in various fields, such as image deblurring [11], single image super-resolution (SISR) [12], and color image demosaicing [13]. Inspired by these methods, in this article, we utilize deep networks to extract implicit background information (which is called deep background prior in the following text) to accurately reconstruct the background for moving vehicle detection. Specifically, we design a U-shape network to obtain deep background prior, which is incorporated into model-based methods. The whole framework can be solved by alternating direction method of multipliers (ADMM) [14] to get the closed-form solution. Based on the closed-form solution, the detection results can be obtained via iterative optimization.

Manuscript received 29 October 2022; revised 11 January 2023; accepted 4 February 2023. Date of publication 6 February 2023; date of current version 17 February 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100800; and in part by the National Natural Science Foundation of China under Grant 62001478, Grant 61872379, and Grant 62022091. *(Corresponding author: Miao Li.)*

Chao Xiao, Ting Liu, Xinyi Ying, Yingqian Wang, Miao Li, Li Liu, and Wei An are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: lm8866@nudt.edu.cn).

Zhijie Chen is with the National Airspace Technology Key Laboratory, Beijing 100085, China.

Digital Object Identifier 10.1109/TGRS.2023.3243055

It is worth noting that our background prior can be learned in an unsupervised manner by training the U-shaped network with a specifically designed loss function, thus eliminating the reliance on the large-scale annotated dataset.

The main contributions of this article are summarized as follows.

1) We incorporate deep background prior into the model-based method for moving vehicle detection in satellite videos, which combines the advantages of both model- and learning-based methods. To the best of our knowledge, we are the first to propose such a framework for MOD in satellite videos.
2) We design a background reconstruction network to recover background from multiple frames in an unsupervised way. A loss objective is designed to guide the network to reconstruct the background without ground-truth labels.
3) With the help of the incorporated deep background prior, our method achieves superior detection performance with significant acceleration compared to state-of-the-art model-based methods.

The rest of this article is organized as follows. Some related works are briefly reviewed in Section II. The notations and preliminaries are described in Section III. The proposed framework is illustrated in Section IV. Section V presents the experimental setup and results in detail, and Section VI concludes this article.

II. Related Work

MOD in satellite videos is a newly emerging field in recent years due to the availability of satellite videos provided by launched satellites, such as Jilin-1 and Skybox. In this section, we briefly review the major works on model- and learning-based methods for MOD in satellite videos. In addition, we introduce the highly related works on model-based methods with deep image priors and unsupervised MOD.

A. Model-Based Methods for MOD in Satellite Videos

Most traditional methods exploit the physical characteristics of targets to tackle MOD in satellite videos, which can be divided into frame differencing-based methods [15], [16], [17] and background subtraction-based methods [3], [6], [7], [18], [19], [20], [21].

Frame differencing-based methods [15], [16], [17] detect moving objects by computing the differences between adjacent frames and then perform segmentation to get the detection results. A variety of two- and three-frame differencing methods have been proposed. However, the detection performance would be degraded by the sudden change in the dynamic backgrounds.

Typical background subtraction-based methods [3], [18] first estimate the background by different filters (e.g., mean or median filters) and then get the detection results by subtracting the estimated background from each frame. The relatively simple estimation of background makes these methods suffer performance degradation under complex scenes, resulting in a high false alarm rate.

Another line of background subtraction-based methods exploits robust principal component analysis (RPCA) techniques to detect moving objects. These RPCA-based methods [6], [7], [19], [20], [21] assume that the image from satellite videos is a summation of background, target, and noise, and employ different regularization on each component. The detection results can be obtained by acquiring closed-form solutions and applying iterative optimization [6], [7], [21]. However, these RPCA-based methods usually employ sophisticated and handcrafted regularization terms to tackle complex scenes, which increases the computational complexity and slows down the iterative process. Moreover, when dealing with complex scenes, these methods cannot ensure the quality of the recovered background, thus degrading the detection performance.

B. Learning-Based Methods for MOD in Satellite Videos

Before the deep learning era, feature extractors and descriptors are widely used for many tasks, such as object detection [22] and image matching [23]. With powerful feature modeling capacities, deep learning has been successfully applied in object detection for natural images [24], [25], [26], [27], [28], [29], [30] and achieved promising performance. For example, Bayraktar et al. [30] proposed a framework consisting of basic image preprocessing techniques, geometrical operations, and deep neural networks to improve the performance of ornamental plant detection and counting from onboard UAV cameras. However, these object detection methods mainly rely on appearance information to detect objects. When dealing with moving objects in satellite videos with limited appearance and texture information, these methods will suffer significant performance degradation [31], [32]. For MOD in satellite videos, the spatiotemporal information is of great importance. Therefore, existing learning-based methods usually design suitable network architectures for the extraction of spatiotemporal information.

LaLonde et al. [31] proposed a two-stage network named ClusterNet to extract spatiotemporal information from consecutive airborne images to detect moving objects. Xiao et al. [32] proposed a two-stream detection network called DSFNet to incorporate the static context information and the dynamic motion cues for MOD in satellite videos. Although the learning-based methods have achieved promising performance, their performance relies heavily on large-scale labeled data. However, due to the extremely small object size and complex backgrounds, annotating moving objects in satellite videos is labor-intensive and time-consuming. In this article, we explore an unsupervised method for MOD in satellite videos to fully use a large amount of unlabeled data to relieve the labeling burden, which is more practical in real scenes.

C. Model-Based Methods With Deep Image Priors

Unlike traditional model-based methods that require explicit and handcrafted image priors, model-based methods with deep priors [33], [34], [35] can incorporate implicit deep priors from deep CNN networks for image restoration [36]. Tirer and Giryes [37] utilized the plug-and-play framework

with IRCNN [38] denoiser to tackle SISR. Li and Wu [39] introduced IRCNN denoiser into a model-based method to solve depth image inpainting. Zhang et al. [13] modulated the deep denoiser prior into traditional model-based methods to solve various image restoration problems.

Inspired by these works, we introduce an implicit deep background prior into model-based methods to generate more accurate backgrounds, which further improves the effectiveness of MOD in satellite videos. Different from the aforementioned supervised image restoration methods, we develop a background reconstruction network to obtain deep background prior in an unsupervised manner.

D. Unsupervised Moving Object Detection

Unsupervised MOD aims to perform detection without any handcrafted annotation. Recently, many unsupervised MOD methods have been proposed for natural images [40], [41], [42], [43]. Specifically, Sultana et al. [40] proposed a GAN-based moving object detector to estimate the background and employed differencing and segmentation to generate detection results. Yun et al. [42] proposed an unsupervised MOD method for pan-tilt-zoom (PTZ) cameras. They designed two background models for large and small changes, and incorporated the results from both models to get the moving objects. Bao et al. [43] modified the SlotAttention [44] framework to detect moving objects and utilized pseudo-ground truth generated by a motion segmentation method as supervision. However, the aforementioned methods are designed for general objects in natural images, where the object contains abundant appearance and texture information. They tend to suffer significant performance degradation on MOD in satellite videos due to the small sizes and low contrast to background clutters.

To alleviate the annotation burden of moving vehicles in satellite videos, Zhang et al. [45] proposed a weakly supervised method to detect moving objects in satellite videos. They first generated pixelwise pseudolabels from the traditional RPCA-based method E-LSD [6] and then utilized the pseudolabels to train an encoder–decoder network to segment moving objects. Due to the inaccuracy of the generated pseudolabels, the method in [45] achieved an inferior performance than E-LSD [6].

In the field of MOD in satellite videos, unsupervised learning has not been discussed yet. In this article, we propose the first unsupervised learning method for moving vehicle detection in satellite videos.

III. Notations and Preliminaries

A. Formulation of MOD in Satellite Videos

Generally, the problem of MOD in satellite videos can be formulated as follows:

$$f_D = f_B + f_T + f_N \tag{1}$$

where f_D, f_B, f_T, and f_N represent the original image, the background image, the target image, and the noise image, respectively. Compared with matrix-based methods [6], [7], the low-rank tensor decomposition method [8] can obtain good detection performance due to the preservation of the spatiotemporal structure. Therefore, this article uses the low-rank tensor decomposition method as the basic framework. Consequently, the model in (1) can be rewritten into the tensor form as follows:

$$\mathcal{D} = \mathcal{B} + \mathcal{T} + \mathcal{N} \tag{2}$$

where $\mathcal{D}, \mathcal{B}, \mathcal{T}, \mathcal{N} \in \mathcal{R}^{n_L \times H \times W}$ represent the original patch tensor, the background patch tensor, the target patch tensor, and the noise patch tensor, respectively. The detection results (i.e., target image) can be obtained by fetching out the slices of $\mathcal{T}$.

B. Low-Rank and Sparse Component Decomposition Model

The background regions are generally assumed to change slowly over a period of time, and there are a lot of overlapped regions among different frames. Therefore, background patch tensor $\mathcal{B}$ conforms to the low-rank property [6] with suitable video length, which can be described as

$$\text{rank}(\mathcal{B}) \leq r \tag{3}$$

where r is a constant and rank$(\cdot)$ represents the rank of a tensor.

The target patch tensor $\mathcal{T}$ conforms to the sparsity prior, which can be depicted as

$$\|\mathcal{T}\|_0 \leq d \tag{4}$$

where d is an integer that is related to the target characteristic and satisfies $d \ll W \times H$.

The noise $\mathcal{N}$ is usually modeled as additive white Gaussian noise, and it satisfies the following:

$$\|\mathcal{N}\|_F \leq \sigma \tag{5}$$

where $\|\cdot\|_F$ represents the Frobenius norm of a tensor and $\sigma > 0$ denotes the Gaussian noise level.

Generally, the low-rank tensor-based framework for MOD in satellite videos can be obtained by replacing $\|\mathcal{T}\|_0$ with $\|\mathcal{T}\|_1$ [46]. Therefore, the low-rank and sparse component decomposition model can be formulated as

$$\begin{aligned} &\min_{\mathcal{B},\mathcal{T}} \ \|\mathcal{B}\|_* + \lambda\|\mathcal{T}\|_1 + \beta\|\mathcal{N}\|_F \\ &\text{s.t. } \mathcal{D} = \mathcal{B} + \mathcal{T} + \mathcal{N} \end{aligned} \tag{6}$$

where λ and β denote the weight for target and noise components, respectively. $\|\cdot\|_*$ represents the nuclear norm, which is a nonconvex approximation of rank$(\mathcal{B})$.

IV. Proposed Framework

Previous model-based methods [6], [7], [8] usually employ explicit image prior (e.g., low rank prior) as regularization terms (e.g., the nuclear norm) to accurately recover the background. Despite achieving promising performance, these methods cannot handle complex scenes well due to the quality of the reconstructed background. To address this issue, we introduce the implicit deep background prior into the model-based

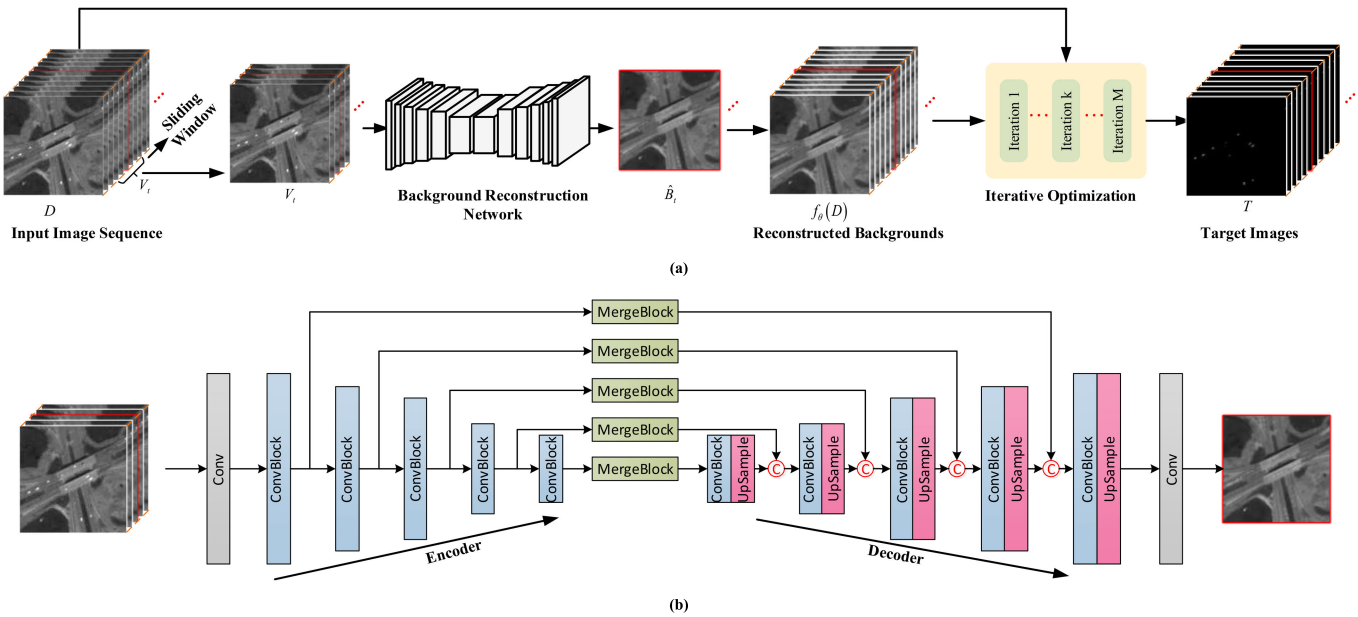

Fig. 1. Illustration of the proposed method. (a) Overall framework, which consists of two parts, including a background reconstruction network and iterative optimization. (b) Background reconstruction network. The proposed network can reconstruct the background as deep background prior, which can be incorporated into the model-based method. Then, the detection results can be obtained by solving the closed-form solution and performing iterative optimization.

method, which can be obtained by a deep background reconstruction network. In this section, we introduce the proposed framework with deep background prior, which is shown in Fig. 1(a). In the following, we first present a model-based method with deep background prior in Section IV-A. Then, the solving process of the proposed framework is illustrated in Section IV-B. Finally, the unsupervised background reconstruction network is introduced in Section IV-C.

A. Model-Based Method With Deep Background Prior

The core idea of our proposed framework is to incorporate deep background prior into the model-based method. Therefore, we introduce the deep background prior into (6) and remove the handcrafted nuclear norm on the background. The formulated model is given as follows:

$$\begin{aligned} &\min_{\mathcal{B},\mathcal{T},\mathcal{N}} \ \lambda\|\mathcal{T}\|_1+\beta\|\mathcal{N}\|_F^2 \\ &\text{s.t. } \mathcal{D}=\mathcal{B}+\mathcal{T}+\mathcal{N}, \quad \mathcal{B}=f_\theta(\mathcal{D}) \end{aligned} \tag{7}$$

where $f_\theta(\mathcal{D})$ denotes the deep background prior recovered from the input image by the background reconstruction network $f_\theta(\cdot)$. λ and β denote the positive regularization parameters. Note that one can replace $f_\theta(\cdot)$ with any designed background reconstruction network. Therefore, our proposed framework can not only retain the flexibility of the model-based method but also leverage the powerful modeling ability of deep neural networks.

B. Solving the Proposed Method

The problem in (7) can be rewritten by the inexact augmented Lagrangian multiplier (IALM) [47] approach as follows:

$$\begin{aligned} &L(\mathcal{B},\mathcal{T},\mathcal{N},y_1,y_2) \\ &\quad=\lambda\|\mathcal{T}\|_1+\beta\|\mathcal{N}\|_F^2+\frac{\mu}{2}\|\mathcal{D}-\mathcal{B}-\mathcal{T}-\mathcal{N}\|_F^2 \\ &\qquad+\frac{\mu}{2}\|\mathcal{B}-f_\theta(\mathcal{D})\|_F^2+\langle y_1,\mathcal{D}-\mathcal{B}-\mathcal{T}-\mathcal{N}\rangle \\ &\qquad+\langle y_2,\mathcal{B}-f_\theta(\mathcal{D})\rangle \end{aligned} \tag{8}$$

where y_1 and y_2 represent Lagrangian multipliers and μ is a positive penalty scalar. Since it is hard to optimize all these variables concurrently, we approximately solve this optimization problem by alternately solving one variable with the others being fixed. Thus, we apply ADMM [14] approach to decompose (8) into three optimization subproblems about $\mathcal{B}$, $\mathcal{T}$, and $\mathcal{N}$, and then alternately solve these variables. The details are given as follows.

1) Updating $\mathcal{B}$ with other variables being fixed

$$\begin{aligned} \mathcal{B}^{k+1}=\arg\min_{\mathcal{B}} \ &\frac{\mu^k}{2}\left\|\mathcal{B}-f_\theta(\mathcal{D})+\frac{y_2^k}{\mu^k}\right\|_F^2 \\ &+\frac{\mu^k}{2}\left\|\mathcal{D}-\mathcal{B}-\mathcal{T}^k-\mathcal{N}^k+\frac{y_1^k}{\mu^k}\right\|_F^2. \end{aligned} \tag{9}$$

The solution of (9) can be obtained by

$$\mathcal{B}^{k+1}=\frac{1}{2}\times\left(f_\theta(\mathcal{D})-\frac{y_2^k}{\mu^k}+\mathcal{D}-\mathcal{T}^k-\mathcal{N}^k+\frac{y_1^k}{\mu^k}\right). \tag{10}$$

2) Updating $\mathcal{T}$ with other variables being fixed

$$\begin{aligned} \mathcal{T}^{k+1}=\arg\min_{\mathcal{T}} \ &\lambda\|\mathcal{T}\|_1 \\ &+\frac{\mu}{2}\left\|\mathcal{D}-\mathcal{B}^{k+1}-\mathcal{T}-\mathcal{N}^k+\frac{y_1^k}{\mu^k}\right\|_F^2. \end{aligned} \tag{11}$$

Algorithm 1 Proposed Algorithm

Input: image sequence, parameters $\lambda, \beta, \mu > 0$
Initialize: Transform the image sequence with length n_L into the original tensor $\mathcal{D}$, $\mathcal{B}^0 = \mathcal{T}^0 = \mathcal{N}^0 = 0$, $y_1^0 = y_2^0 = 0$, $\mu_0 = 5e-4$, $\mu_{\max} = 1e7$, $\rho = 1.5$, $\beta = 100$, $\lambda = 1/\sqrt{max(H, W) \times n_L}$, $\zeta = 1e-7$, $maxIter = 500$, $k = 0$.
While not converged **do**
1 : Update $\mathcal{B}^{k+1}$ by Eq. (10)
2 : Update $\mathcal{T}^{k+1}$ by Eq. (12)
3 : Update $\mathcal{N}^{k+1}$ by Eq. (14)
4 : Update the Lagrangian multipliers by Eq. (15)
5 : Update μ^{k+1} by Eq. (16)
6 : Check the convergence conditions $\frac{\|\mathcal{D}-\mathcal{B}^{k+1}-\mathcal{T}^{k+1}-\mathcal{N}^{k+1}\|_F^2}{\|\mathcal{D}\|_F^2} \le \zeta$ or $k = maxIter$
7 : Update $k = k + 1$
end While
Output : $\mathcal{B}^{k+1}, \mathcal{T}^{k+1}, \mathcal{N}^{k+1}$

Equation (11) can be solved by performing elementwise shrinkage operation [48]

$$\mathcal{T}^{k+1} = Th_{\lambda(\mu^k)}{}^{-1}\left(\mathcal{D} - \mathcal{B}^{k+1} - \mathcal{N}^k + \frac{y_1^k}{\mu^k}\right) \quad (12)$$

where $\mathrm{Th}(\cdot)$ denotes the elementwise shrinkage operator and μ^k is the positive penalty scalar for the kth iteration.

3) Updating $\mathcal{N}^{k+1}$ with other variables being fixed

$$\mathcal{N}^{k+1} = \arg\min_{\mathcal{N}} \ \beta\|\mathcal{N}\|_F^2 + \frac{\mu^k}{2}\left\|\mathcal{D} - \mathcal{B}^{k+1} - \mathcal{T}^{k+1} - \mathcal{N} + \frac{y_1^k}{\mu^k}\right\|_F^2. \quad (13)$$

The solution of (13) can be obtained by

$$\mathcal{N}^{k+1} = \frac{\mu^k\left(\mathcal{D} - \mathcal{B}^{k+1} - \mathcal{T}^{k+1}\right) + y_1^k}{\mu^k + 2\beta}. \quad (14)$$

4) Updating multipliers y_1, y_2 with other variables being fixed

$$\begin{cases} y_1^{k+1} = y_1^k + \mu^k\left(\mathcal{D} - \mathcal{B}^{k+1} - \mathcal{T}^{k+1} - \mathcal{N}^{k+1}\right) \\ y_2^{k+1} = y_2^k + \mu^k\left(\mathcal{B}^{k+1} - f_\theta(\mathcal{D})\right). \end{cases} \quad (15)$$

5) Updating the positive penalty scalar μ^{k+1}

$$\mu^{k+1} = \min\left(\rho\mu^k, \mu_{\max}\right). \quad (16)$$

Finally, the proposed method is summarized in Algorithm 1.

C. Unsupervised Background Reconstruction Network

We build a background reconstruction network to recover background, which can serve as implicit deep background prior. Due to the lack of ground-truth background and the difficulty in acquiring such labels in real scenes, we propose to train the background reconstruction network in an unsupervised manner. In the following parts, we introduce the architecture of the proposed background reconstruction network in Section IV-C1, the merge block in Section IV-C2, and the specifically designed loss in Section IV-C3.

1) Network Architecture: We design a U-shape network [49] to recover the background from consecutive frames in satellite videos, which consists of an encoder for feature extraction, a decoder for feature reconstruction, and skip connections for feature propagation. The encoder and the decoder are composed of several convolution blocks (each convolution block consists of two Conv-BN-ReLU layers) and downsampling or upsampling operations. A merge block is added to the skip connection to aggregate the temporal information, which can propagate the spatiotemporal information from the encoder side to the decoder side. The network architecture is shown in Fig. 1(b). Specifically, a video clip V_t with n frames $I_{t+\tau}(\tau = [-r, r], r = \lfloor n/2 \rfloor)$ is first fed into a 2-D convolutional layer to generate the initial feature map $F_0^t \in \mathcal{R}^{bn \times c_0 \times H \times W}$, where b denotes the batch size. After that, the initial feature map is processed by the encoder to generate multilevel feature maps, resulting in $F_i^t \in \mathcal{R}^{bn \times c_i \times (H/2^{i-1}) \times (W/2^{i-1})}$ for the ith convolution block. Next, the generated multilevel feature maps are processed by merge blocks, which can fuse the spatial and temporal information, resulting in $G_i^t \in R^{b \times c_i \times (H/2^{i-1}) \times (W/2^{i-1})}$. Then, the fused multilevel feature maps are sent to the decoder, which can recover the resolution of the feature map. Finally, the resulting feature map from the decoder is processed by a 2-D convolution to get the reconstructed background $\hat{B}^t$.

2) Merge Block: For the skip connection, to aggregate the spatiotemporal information of the feature map generated from the video clip, we build a merge block into the skip connection to propagate spatiotemporal information from the encoder to the decoder. Since the background region overlaps among adjacent frames, merging the spatiotemporal information from multiframes and reducing the temporal dimension can help to obtain deep background prior. The merge block consists of a 3-D convolution block and a temporal average-pooling operation. To reduce the computational cost of the 3-D convolution block, we decomposed the 3-D convolution with a kernel size of $k \times k \times k$ into a spatial convolution with a kernel size of $1 \times k \times k$ and a temporal convolution with a kernel size of $k \times 1 \times 1$. Each convolution in the decomposed 3-D convolution block is followed by a batch normalization and a ReLU. Following the decomposed 3-D convolution is the temporal average-pooling operation. The temporal average-pooling operation can reduce the temporal dimension and fuse background information from multiple frames. Through the merge block, multiframe background information can be extracted and fused for background reconstruction.

3) Objective Loss Function: It is a straightforward way to utilize clean backgrounds as supervision to train the background reconstruction network. However, in practice, it is difficult to generate backgrounds as supervision from natural images. Therefore, in this article, we design a loss function to guide the network to reconstruct the background in an unsupervised manner.

Since the image can be intuitively separated into background and target regions, we can use different strategies to deal with these regions when computing loss. For the background region, it is better to make the reconstructed results approximate the

TABLE I

DETECTION PERFORMANCE ACHIEVED BY DIFFERENT METHODS. RECALL (RE) (%), PRECISION (PR) (%), AND F1 SCORE (F1) (%) ACHIEVED BY DIFFERENT METHODS ON SEVEN SATELLITE VIDEOS. TIME COST (S) FOR A SINGLE IMAGE (1024 × 1024) OF DIFFERENT METHODS IS ALSO LISTED IN THE TABLE. THE BEST RESULTS ARE SHOWN IN **BOLDFACE**, AND THE SECOND BEST RESULTS ARE SHOWN IN UNDERLINE

Method	Video1			Video2			Video3			Video4			Video5			Video6			Video7			AVERAGE			Time Cost
	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	Re	Pr	F1	
GoDec [19]	84.2	77.5	80.7	77.9	78.3	78.1	77.1	82.1	79.5	63.6	78.6	70.3	78.0	71.2	74.4	70.5	68.3	69.4	35.8	49.8	41.7	69.6	72.2	70.6	5.10
DECOLOR [5]	35.9	**99.3**	52.8	79.4	80.2	79.8	88.6	76.6	82.2	43.4	**99.0**	60.4	78.4	74.0	76.1	77.7	63.8	70.1	33.3	70.0	45.2	62.4	80.4	66.6	8.20
E-LSD [6]	70.6	75.8	73.1	63.8	38.5	48.0	74.6	87.0	80.3	57.2	87.7	69.2	61.1	82.1	70.0	53.7	75.9	62.9	58.4	57.4	57.9	62.7	72.0	65.9	34.20
B-MCMD [7]	75.1	92.1	82.7	72.1	80.6	76.1	82.0	77.1	79.5	69.0	69.7	69.3	54.3	77.1	63.7	66.8	68.3	67.6	82.1	43.1	56.6	71.6	72.6	70.8	45.70
D&T [9]	70.7	90.6	79.4	67.1	82.3	73.9	81.7	81.1	81.4	72.5	80.0	76.1	61.3	78.3	68.7	62.8	72.9	67.5	83.4	38.7	52.9	71.4	74.8	71.4	0.18
MMB [10]	79.9	94.0	86.4	71.9	**88.6**	79.4	85.3	**90.1**	87.6	75.8	82.5	79.0	68.1	83.9	75.2	66.1	84.0	74.0	**85.7**	66.0	74.6	76.1	84.2	79.5	0.50
SAHI [29]	37.2	65.7	47.5	61.0	73.4	66.6	59.1	59.3	59.2	40.1	58.1	47.4	52.8	71.9	60.9	49.0	57.9	53.1	64.3	41.1	50.1	51.9	61.1	55.0	**0.07**
ClusterNet [31]	74.3	63.7	68.6	62.5	76.8	68.9	85.3	68.4	75.9	49.2	70.0	57.8	72.9	78.2	75.5	74.5	68.9	71.6	85.3	65.3	74.0	72.0	70.2	70.3	0.40
DSFNet [32]	**91.6**	92.1	**91.8**	**87.3**	84.7	**86.0**	**94.6**	81.0	87.3	**82.6**	91.1	**86.6**	**94.8**	66.9	78.5	**81.7**	85.8	**83.7**	82.4	**86.6**	**84.4**	**87.8**	84.0	**85.5**	0.29
Ours	83.3	89.8	86.4	76.0	87.5	81.3	89.7	87.7	**88.7**	64.5	82.8	72.5	72.2	**89.3**	**79.8**	73.1	**85.9**	79.0	83.0	74.0	78.2	77.4	**85.3**	80.9	0.48

original images. In contrast, for the target areas, it is better to make the reconstructed results approximate the adjacent background area instead of the original target pixels. Based on these motivations, we separate the reconstructed background into two disjoint subsets (i.e., target region and background region) and employ different supervisions to compute the loss of these two regions. For the background region, we use the original input image as supervision. For the target region, to alleviate the influence of target pixels, we utilize the temporal median filtered image as supervision since targets are moving, and temporal median filtering can filter out most target pixels to reduce their influence in the target region. Therefore, the loss objective consists of two parts, including background region-related loss L_{back} and the target region-related loss L_{tar}, which are defined as follows:

$$L_{\text{back}} = \frac{1}{HW}\left\|\hat{B}_t \odot (1 - M_t) - I_t \odot (1 - M_t)\right\|_F \quad (17)$$

and

$$L_{\text{tar}} = \frac{1}{HW}\left\|\hat{B}_t \odot M_t - I_m \odot M_t\right\|_F \quad (18)$$

where $\odot$ represents element multiplication and M_t represents the generated binary mask of the target areas with 1 denoting the target region and 0 indicating the background region. I_m represents the temporal median filtered image of the input video clip. To obtain the target region mask, we first feed the reconstructed background and the input images to the iterative optimization to generate detection results and then apply segmentation to get the target mask.

The background region-related loss L_{back} and the target region-related loss L_{tar} work jointly to guide the network to reconstruct the background. The total loss objective is defined as

$$L = L_{\text{back}} + L_{\text{tar}}. \quad (19)$$

Since the designed loss objective is label-independent, the proposed method can alleviate the dependence on large-scale labeled data.

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we conduct extensive experiments to evaluate the detection performance of the proposed framework on the dataset collected from Jilin-1 satellite [10].

A. Dataset Description and Experimental Details

The detection performance of the proposed method is evaluated on satellite videos from the Jilin-1 satellite. The GSD of the dataset is around 1 m, and the frame rate is 10 frames per second. The moving vehicles in the dataset are labeled by bounding boxes as the ground truth. The videos in the dataset contain complex and dynamic backgrounds, which are challenging for MOD.

For the background reconstruction network, we used seven consecutive frames with a frame interval of 3 as input to the network. The batch size was set to 10 with a random crop image patch size of 256 × 256. We trained our network using the Adam optimizer [50] for 100 epochs with a learning rate of 1×10^{-4}. All the models were implemented with Pytorch on one Nvidia RTX 3090Ti GPU.

For the iterative optimization, we set $\rho = 1.5$, $\mu = 0.0005$, $n_L = 16$, $\beta = 100$, and $\lambda = 1/(\max(H, W) \times n_L)^{1/2}$, where H, W, and n_L represent the height, width, and video length of the input video, respectively.

B. Evaluation Criteria

In order to make a fair comparison with other compared methods, we follow [7], [10], and [31] to use precision, recall, and F1 score as the evaluation metrics, which are defined as follows.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (20)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (21)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (22)$$

where TP, FN, and FP represent the number of true positives (correct detections), false negatives (missed targets), and false positives (false alarms), respectively. Specifically, the precision

metric measures the fraction of the detections of TPs, and the recall metric indicates the fraction of positives that are correctly identified. The F1 measure is a combination of precision and recall, and is a more reliable and comprehensive evaluation metric.

It is worth noting that, although IoU is widely used for the performance evaluation of generic object detection [24], [25], [26], [27], [28], [29], it is not suitable for the evaluation of extremely small objects in satellite videos. Due to the small size of moving targets in satellite videos, tiny shifts of the predicted bounding box will cause a large fluctuation in the IoU value. Therefore, in this article, we follow [31] to consider a predicted bounding box as a TP if the distance between the center of this bounding box and the ground-truth one is smaller than a predefined threshold. In this article, we set the distance threshold to 5 pixels, which represents around 5 m considering the GSD of the Jilin-1 satellite.

C. Comparison to the State of the Arts

In this section, we present the detection results and analyses of MOD in satellite videos. We compare the proposed method with nine state-of-the-art methods, including two frame differencing-based methods (i.e., D&T [9] and MMB [10]), four RPCA-based methods (i.e., GoDec [19], DECOLOR [5], E-LSD [6], and B-MCMD [7]), and three deep learning-based methods (i.e., SAHI [29], ClusterNet [31], and DSFNet [32]).

1) Quantitative Results: The quantitative results are shown in Table I. It can be observed that, compared with the model-based methods, our framework achieves higher average recall, precision, and F1 score, outperforming the second best model-based method MMB [10] by 1.4 in terms of F1 score. That is because our method introduces deep background prior into model-based iterative optimization, which can recover background more accurately, thus achieving superior performance. Compared with the deep learning-based method SAHI [29], our method achieves superior performance. That is because SAHI [29] is designed for generic small object detection in a single image and would suffer significant performance degradation when applied to an extremely small moving object in satellite videos. Note that our method achieves the best average precision, outperforming the second best method DSFNet [32] by 1.3 in terms of precision rate, which means that our method can improve the detection performance with reduced false alarms due to the accurately reconstructed background. Moreover, although our framework performs inferior to DSFNet [32] (80.9 versus 85.5 in terms of F1 score), our method can detect moving objects in an unsupervised way, which can relieve the dependence on the large-scale dataset with labor-intensive and time-consuming annotation process.

2) Time Efficiency Analyses: To compare the efficiency of different methods, we record the average time cost (s) of different methods on an input image with a size of 1024 × 1024. The results are listed in Table I. It can be observed that, compared with LRSD-based methods (i.e., GoDec [19], DECOLOR [5], E-LSD [6], and B-MCMD [7]), our method is faster and achieves higher F1 score. Compared with the LRSD-based method GoDec [19], our method can achieve nearly 10× acceleration. That is because our method substitutes the low-rank regularization term with deep background prior, which can reduce the computational burden of the low-rank regularization term. Moreover, due to the removal of the regularization nuclear term in the background, our method can exploit CUDA acceleration techniques to improve efficiency, which can further speed up the detection process. Moreover, compared with the fastest deep learning-based method SAHI [29] and the second fastest frame differencing-based method D&T [9], our method runs relatively slowly, while the detection performance of our method is superior, which demonstrates the effectiveness of our method.

3) Qualitative Results: Qualitative results of different methods are shown in Figs. 2 and 3. It can be observed that, compared to the complex backgrounds, moving vehicles occupy only a few pixels, and there are many distractors in the surroundings. Compared with the state-of-the-art model-based methods, our method can produce more reliable detection results with fewer false alarms (as can be seen from the numbers of the TP, FP, and FN), which demonstrates the superiority of our method in tackling challenging scenes. It can also be observed that the existing model-based methods exhibit many false alarms on stationary background objects (e.g., residential area of video 7 in Fig. 3), while our method produces fewer false alarms on these objects. We attribute this to the accurately reconstructed background produced by our deep background reconstruction network.

D. Ablation Study

In this section, we conduct different ablation studies to investigate the design of our proposed framework.

1) Effectiveness of Background Reconstruction Network: To validate the effectiveness of our background reconstruction network, we replace the background reconstruction network with other reconstruction methods, including the spatial mean filter, the spatial median filter, the temporal mean filter, and the temporal median filter. The quantitative results are shown in Table II. It can be observed that our proposed method achieves the best F1 score and outperforms the second best background reconstruction method by 1.9 in terms of F1 score. The backgrounds reconstructed by different methods are shown in Fig. 4. It can be observed that our method can restore a more clean background, which can be used to obtain better detection results. In contrast, other background reconstruction methods have target residuals in the target region and, thus, have inferior detection performance.

To accurately evaluate the background reconstruction capability of different methods, we add synthetic moving targets on the clean backgrounds and then apply different methods for background reconstruction. The reconstructed background is compared with the ground-truth clean background. Following [51], we use PSNR calculated between the reconstructed background and the ground-truth one as quantitative metrics for reconstruction performance evaluation. We compare our proposed method with three RPCA-based methods, including

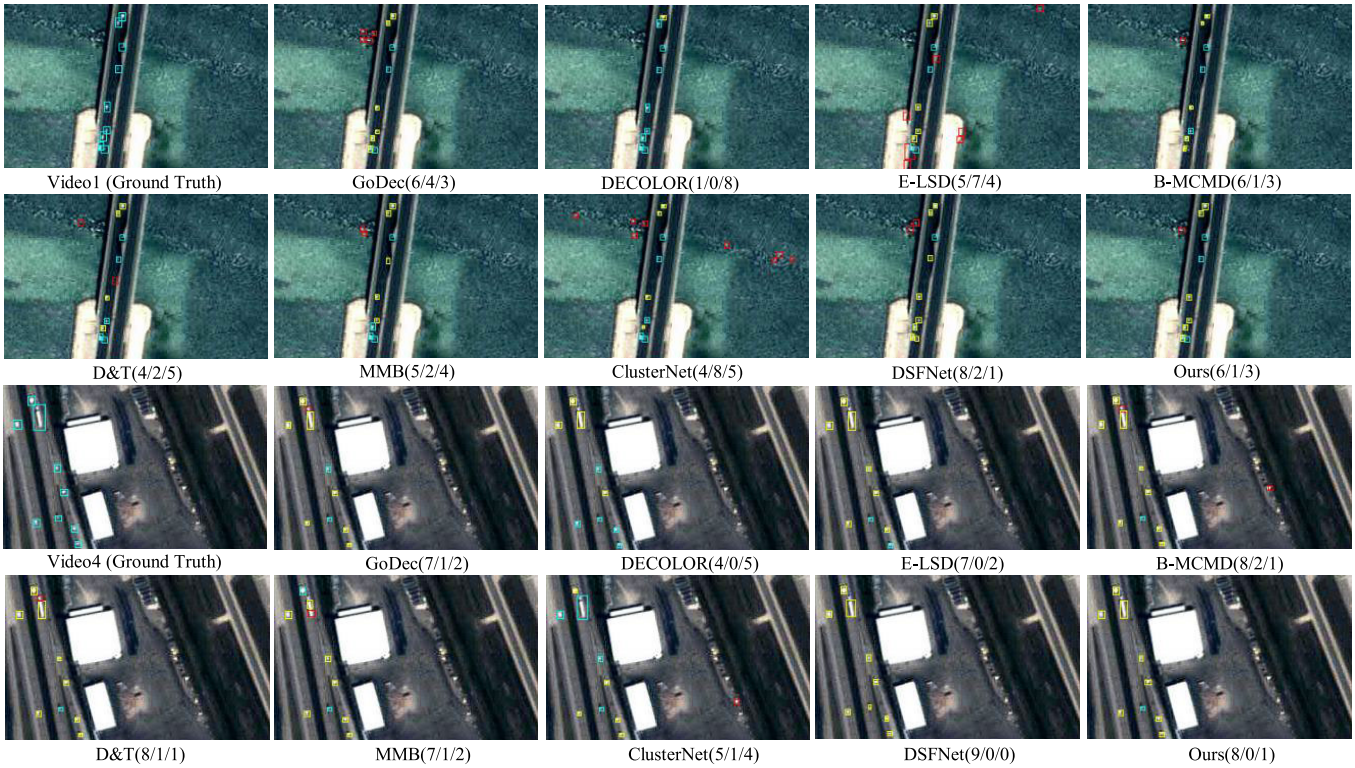

Fig. 2. Visual comparison of detection results on Video 1 (bridge area) and Video 4 (airport area). The light green, yellow, and red boxes denote the annotated ground truth, correct detections, and false alarms, respectively. The TP/FP/FN numbers achieved by different methods on the presented scenes are reported below the image regions.

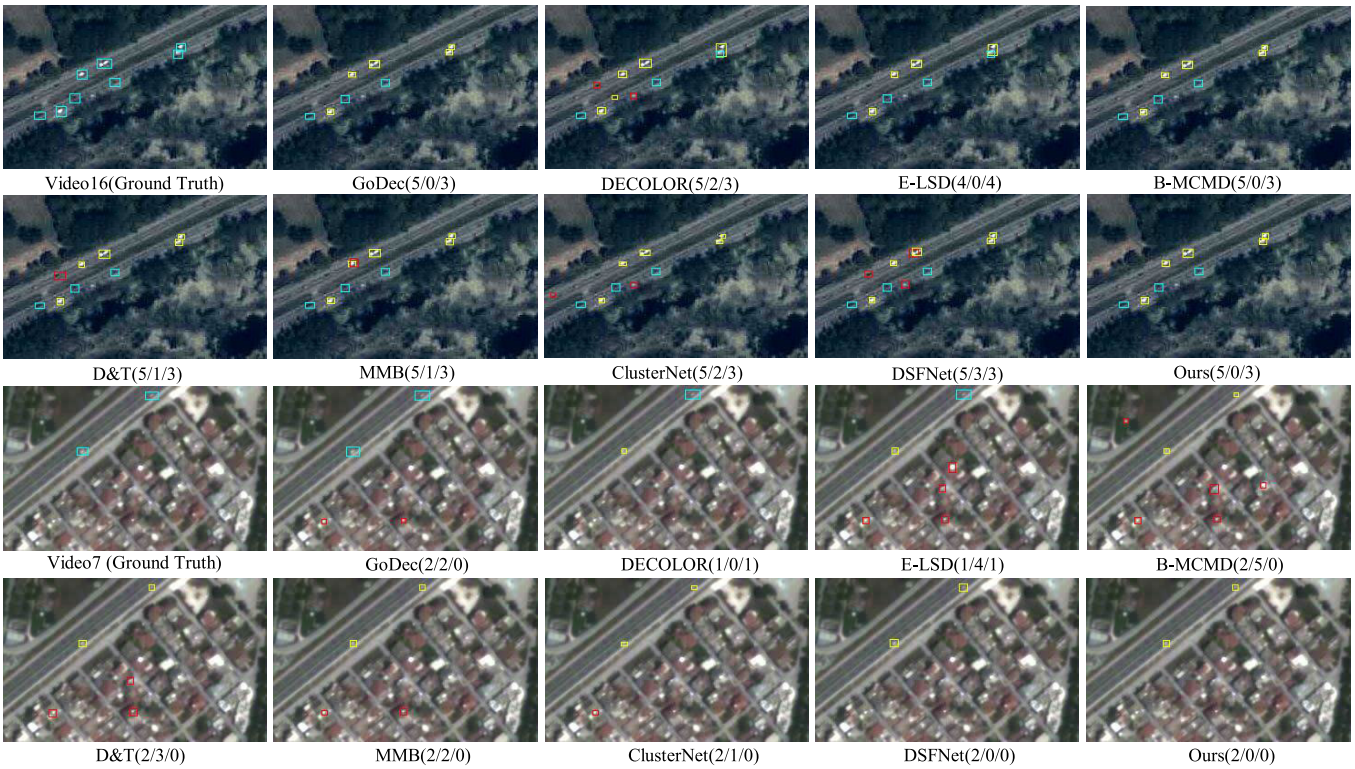

Fig. 3. Visual comparison of detection results on Video 6 (wild road) and Video 7 (residential area). The light green, yellow, and red boxes denote the annotated ground truth, correct detections, and false alarms, respectively. The TP/FP/FN numbers achieved by different methods on the presented scenes are reported below the image regions.

DECOLOR [5], E-LSD [6], and B-MCMD [7]. The quantitative results are shown in Table III. It can be observed that our method achieves the best PSNR and F1 score, which demonstrates the effectiveness of our deep background prior. The qualitative results are shown in Fig. 5. It can be observed that our method can reconstruct a more accurate

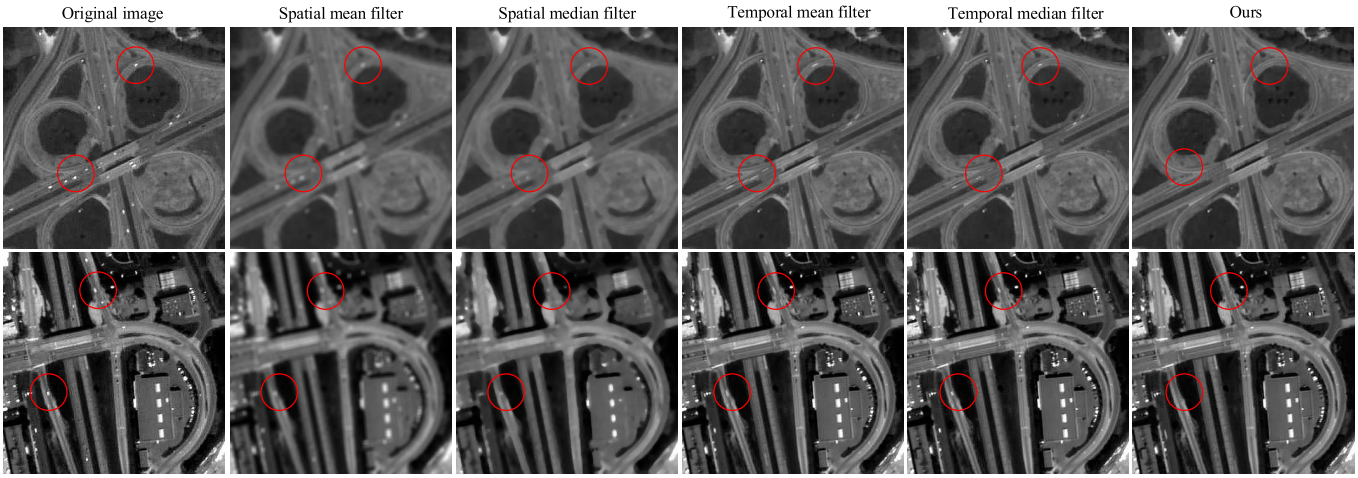

Fig. 4. Background reconstruction results generated by different methods. The red circles indicate the target region to show the target residuals. Fewer target residuals indicate better background reconstruction quality.

TABLE II

PERFORMANCE OF DIFFERENT BACKGROUND RECONSTRUCTION METHODS. AVERAGE RECALL (AVG RE) (%), AVERAGE PRECISION (AVG PR) (%), AND AVERAGE F1 SCORE (AVG F1) (%) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Methods	Avg Re	Avg Pr	Avg F1
Spatial mean filter	31.7	19.1	22.6
Spatial median filter	45.5	18.5	25.3
Temporal mean filter	73.5	85.5	78.7
Temporal median filter	72.6	**87.8**	79.0
Ours	**77.4**	85.3	**80.9**

TABLE III

PERFORMANCE OF DIFFERENT METHODS ON THE SYNTHETIC DATA. PSNR AND F1 SCORE (%) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Methods	DECOLOR [5]	E-LSD [6]	B-MCMD [7]	Ours
PSNR	26.42	26.37	26.52	**29.42**
F1	65.7	73.8	79.0	**86.0**

TABLE IV

ABLATION STUDY ON MERGE BLOCK. AVERAGE RECALL (AVG RE) (%), AVERAGE PRECISION (AVG PR) (%), AVERAGE F1 SCORE (AVG F1) (%), AND TIME COST (S) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Block	Avg Re	Avg Pr	Avg F1	Time Cost (s)
Block2D	75.8	81.3	78.0	**0.38**
Block3D	76.4	**85.7**	80.5	0.72
Ours	**77.4**	85.3	**80.9**	0.48

TABLE V

PERFORMANCE OF DIFFERENT DETECTION METHODS. AVERAGE RECALL (AVG RE) (%), AVERAGE PRECISION (AVG PR) (%), AND AVERAGE F1 SCORE (AVG F1) (%) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Methods	Avg Re	Avg Pr	Avg F1
frame differencing	71.8	**85.7**	78.0
Ours	**77.4**	85.3	**80.9**

background (smaller errors between the generated background and the ground-truth one) and, thus, achieves better detection performance.

2) Effectiveness of Merge Block: As a component of our background reconstruction network, the merge block can integrate the spatial and temporal information, and propagate the fused spatiotemporal information from the encoder to the decoder. Here, we investigate the merge block by introducing two variants, i.e., Block2D and Block3D. Block2D merges spatial and temporal information by first concatenating multi-frame features along channel dimension and then performing a 2-D convolution with a kernel size of 3×3 (with BN and ReLU layers). Block3D integrates the spatiotemporal information explicitly by a 3-D convolution with BN, ReLU, and a temporal average-pooling layer.

The detection performance of different variants is shown in Table IV. It can be observed that our method achieves the best F1 score and outperforms Block2D by 2.9 in terms of F1 score. That is because Block2D utilizes 2-D convolution and, thus, cannot fully extract and fuse the spatial and temporal information. Moreover, compared with Block3D, our method reduces the processing time of a single image by 33% (0.48 s versus 0.72 s) and achieves a better F1 score. That is because the decomposed 3-D convolution in the merge block can not only reduce the computational cost but also introduce extra nonlinear operations to enhance the modeling ability of the network. In conclusion, our designed merge block can achieve higher accuracy and efficiency.

3) Effectiveness of the Iterative Optimization: To validate the effectiveness of the iterative optimization, we directly use frame differencing operation between the input image and the reconstruction background and segment detection results from the residual images. The quantitative results are shown in Table V. It can be observed that the iterative optimization achieves the best average F1 score and outperforms the frame differencing method by 2.9 in terms of F1 score. That is

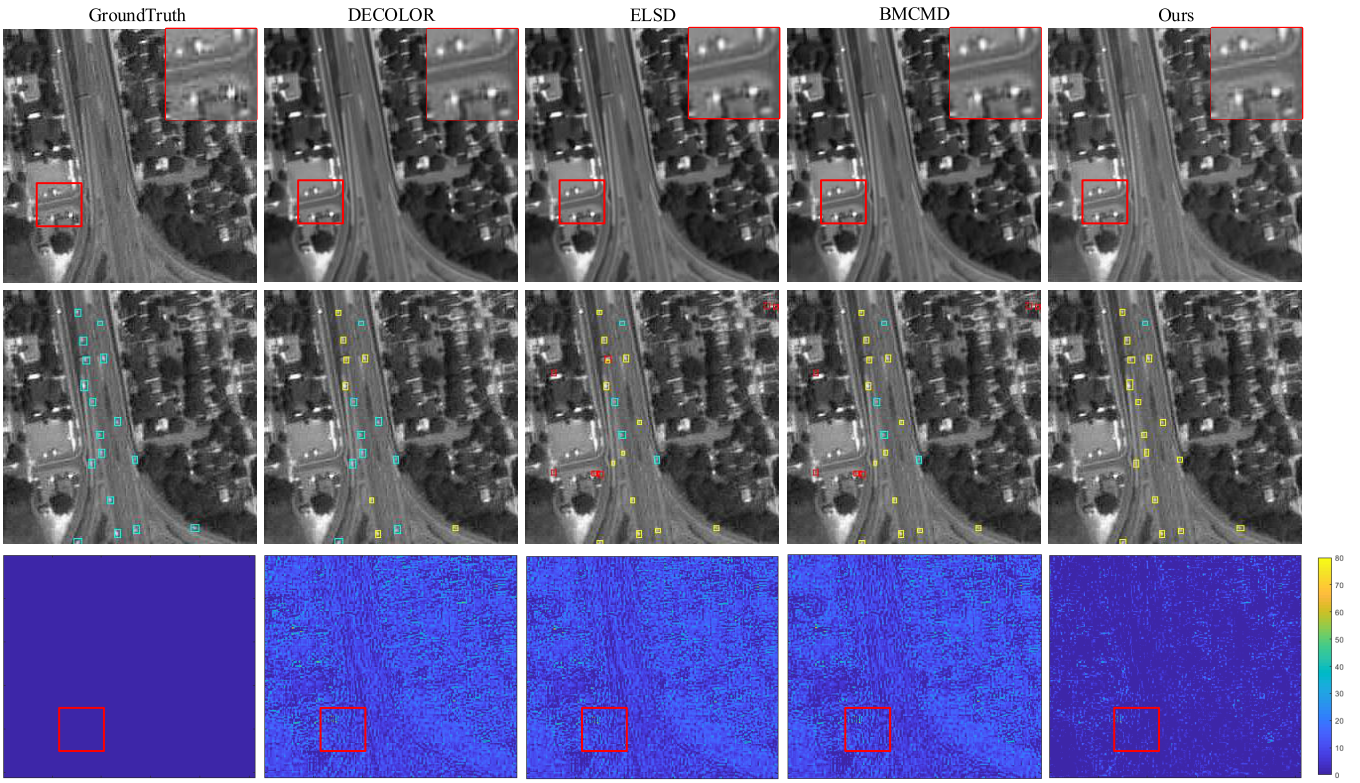

Fig. 5. Experimental results on the synthetic sequence. The first row illustrates the background reconstruction results obtained by different methods, and the zoomed-in area is utilized for a better illustration of details. The second row draws the detection results, and the light green, yellow, and red rectangles indicate the ground truth, correct detections, and false alarms, respectively. The third row shows the differencing heatmaps between the generated background and the ground-truth one, and lower errors indicate better reconstructed background quality.

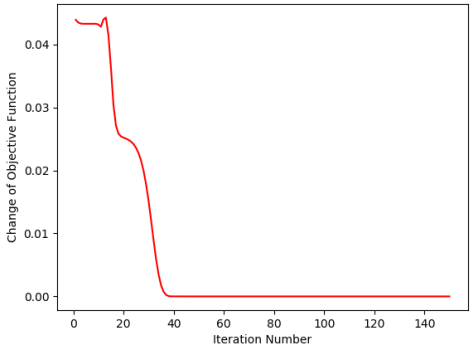

Fig. 6. Convergence curve of the iterative optimization process.

TABLE VI

IMPACT OF NETWORK DEPTH ON DETECTION PERFORMANCE. AVERAGE RECALL (AVG RE) (%), AVERAGE PRECISION (AVG PR) (%), AVERAGE F1 SCORE (AVG F1) (%), AND TIME COST (S) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Network Depth	Avg Re	Avg Pr	Avg F1	Time Cost
3	77.1	83.8	80.0	**0.40**
4	**78.1**	83.4	80.2	0.44
5	77.4	**85.3**	**80.9**	0.48
6	77.5	84.6	80.5	0.51

because the iterative optimization process can optimize the detection results to achieve optimal performance.

Moreover, we investigate the convergence of iterative optimization. Here, we study numerical convergence instead of analytical convergence since our method is a combination of deep learning and model-based approaches. Following [52], we use $((\|\mathcal{D}-\mathcal{B}^{k+1}-\mathcal{T}^{k+1}-\mathcal{N}^{k+1}\|_F^2)/(\|\mathcal{D}\|_F^2)) \leq \zeta$ as criterion to measure the convergence. Taking video 1 as an example, the convergence curve is shown in Fig. 6. It can be observed that the proposed method converges to an optimal objective value after about 40 iterations and maintains stable.

4) Impact of Network Depth: We investigate the impact of network depth on detection performance. We set the number of convolution blocks in the encoder and decoder to 3, 4, 5, and 6, respectively, and investigate the accuracy and efficiency of different variants. The quantitative results are shown in Table VI. It can be observed that, when the network depth increases from 3 to 5, the detection performance is improved with the increase in the network depth but at the cost of a higher computational burden with more processing time. When network depth increases from 5 to 6, the average F1 score slightly drops. That is because, when the depth goes deeper, it tends to overfit the limited training data and, thus, damages the performance. Therefore, we choose a five-layer U-shape network as our reconstruction network.

5) Impact of Frame Number: Our background reconstruction network reconstructs the background from n consecutive frames. We evaluate the background reconstruction network with different frame numbers, i.e., $n = 3, 5, 7, 9$. The results are shown in Table VII. It can be observed that, when n increases from 3 to 7, the detection performance is improved

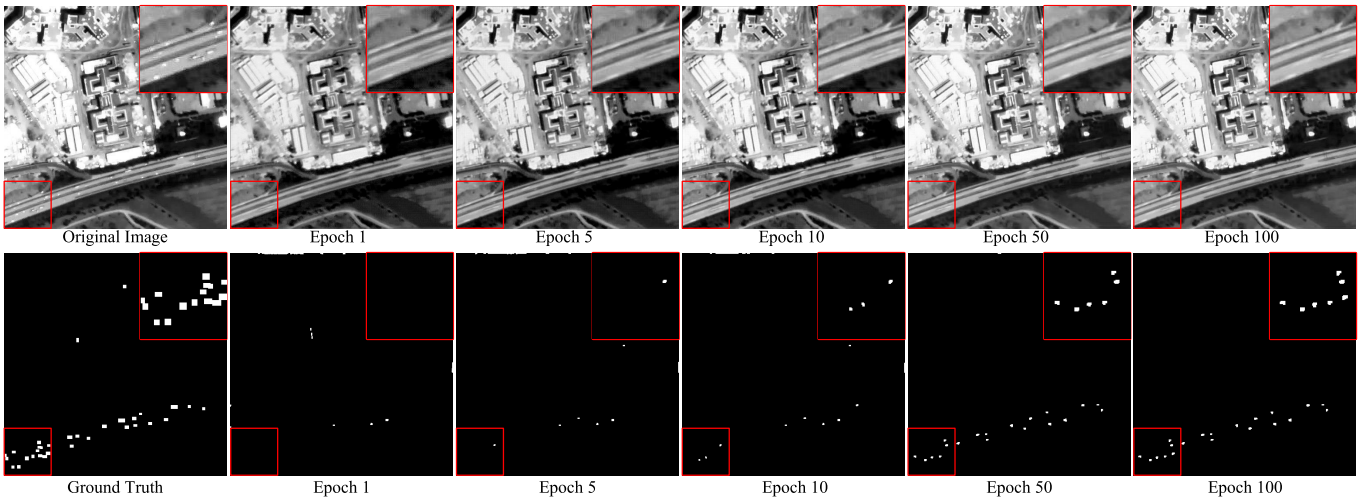

Fig. 7. Background reconstruction results and generated masks in loss objective during training. The first row presents the original image and the reconstructed backgrounds. The second row illustrates the generated target masks. The red rectangles indicate the background details and target mask regions.

TABLE VII

IMPACT OF INPUT FRAME NUMBER ON DETECTION PERFORMANCE. AVERAGE RECALL (AVG RE) (%), AVERAGE PRECISION (AVG PR) (%), AVERAGE F1 SCORE (AVG F1) (%), AND TIME COST (S) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Frame Num	Avg Re	Avg Pr	Avg F1	Time Cost
3	73.0	85.2	78.3	**0.25**
5	77.4	84.1	80.4	0.36
7	77.4	**85.3**	**80.9**	0.48
9	**77.8**	85.2	**80.9**	0.63

as the frame number is increased. That is because additional frames can provide more information about the background, which is beneficial to background reconstruction. It is also notable that the detection performance tends to be saturated when the frame number is increased from 7 to 9 (the average F1 score remains unchanged). That is because the information provided by the seven frames is already sufficient for background reconstruction. Since the spatial and temporal information has been fully exploited for seven input frames, a further increase in frames cannot provide performance improvement but bring extra computational burdens. Therefore, we utilize seven frames as input to the proposed network.

E. Analyses of Loss Objective

To reconstruct the background in an unsupervised manner, we design a loss objective and adapt different strategies for different image regions. To verify the effectiveness of the proposed loss objective, we train our background reconstruction network under L_{back}, L_{tar}, and the combination of both losses, respectively. The quantitative results are shown in Table VIII. It can be observed that, with only L_{back}, the trained model only suffers a minor performance degradation (80.1 versus 80.9 in terms of F1 score). That is because, due to the ignoring of target regions, the network cannot learn to reconstruct a fine-grained background. It can also be observed that, with only L_{tar}, the F1 score drops nearly half compared to our proposed method. That is because the limited background information is insufficient to reconstruct the background. Thanks to the discriminative treatment of target and background areas, our method can learn to reconstruct a fine-grained background and, thus, achieves higher performance.

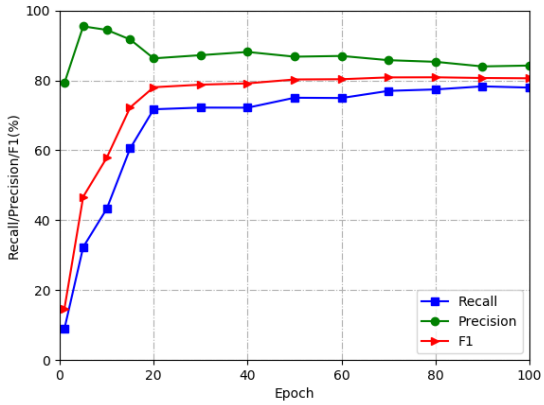

Fig. 8. Evaluated detection performance during training.

To further investigate the effectiveness of our proposed loss objective, we visualize the reconstructed background and generated masks during training in Fig. 7. It can be observed that, with the increase in training epochs, the generated masks can cover more target regions, and the quality of the reconstructed background can be improved gradually. Since the quality of the reconstructed background is gradually improved, the detection performance increases with epochs and reaches saturation at around 100 epochs, as shown in Fig. 8.

F. Parameter Sensitivity Analyses

In this section, we conduct experiments to investigate the impact of two important parameters λ and β in the iterative optimization on the MOD performance.

1) Impact of λ: To make it concise, while keeping β fixed to 100, we use various values of λ_0 to control the values of λ ($\lambda = \lambda_0/(\max(H, W) \times n_L)^{1/2}$). The results are shown in

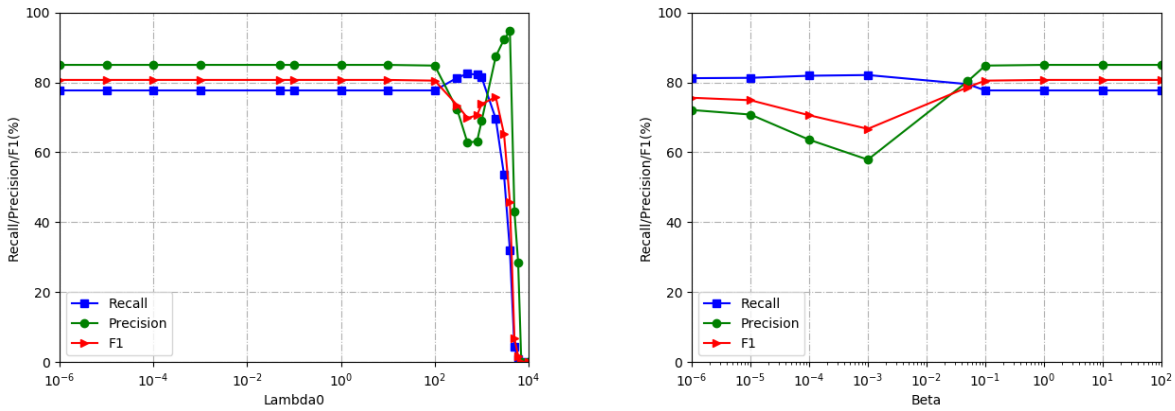

Fig. 9. Average performance evaluation with different parameters. (a) Average performance evaluation with different λ_0's and fixed $\beta = 100$. (b) Average performance evaluation with different β's and fixed $\lambda_0 = 1$.

TABLE VIII

RESULTS OF DIFFERENT LOSS OBJECTIVES. AVERAGE RECALL (AVG RE) (%), AVERAGE PRECISION (AVG PR) (%), AND AVERAGE F1 SCORE (AVG F1) (%) ARE LISTED IN THE TABLE FOR PERFORMANCE COMPARISON. THE BEST RESULTS ARE SHOWN IN BOLDFACE

Model	L_{back}	L_{tar}	Avg Re	Avg Pr	Avg F1
Model1	✗	✓	56.6	42.9	45.5
Model2	✓	✗	77.2	83.8	80.1
Ours	✓	✓	**77.4**	**85.3**	**80.9**

Fig. 9(a). It can be observed that, when λ_0 increases from 10^{-6} to 10^2, the detection performance remains unchanged. However, when λ_0 becomes too large, the sparsity of the target would be overemphasized, leading to overshrinkage of the target and a dramatic drop in detection performance. Theoretically, when λ_0 approximates 0, the sparsity term will be ignored, which will damage the detection performance. It can be observed that our proposed method can still achieve good performance when λ_0 approximates 0. We attribute this to the introduction of deep background prior, which would prevent the performance from dropping to 0 when λ_0 is too small.

2) Impact of β: While keeping λ_0 fixed to 1, we conduct experiments to verify the influence of β. The results are shown in Fig. 9(b). It can be observed that, when β exceeds 10, the detection performance tends to be fixed. That is because, when β is sufficiently large, the noise term $\mathcal{N}$ tends to zero, which will negligibly influence the detection performance. Theoretically, when β becomes too small, the noise term $\mathcal{N}$ will be less emphasized, leading to the increase in the residual in noise term $\mathcal{N}$ and significant performance degradation. However, in our method, when β turns very small, the detection performance remains at a certain level. We attribute this to the introduction of recovered background, which tends to prevent $\mathcal{N}$ from including too many residuals into the noise term.

VI. CONCLUSION

In this article, we have introduced deep background prior into the model-based method for MOD in satellite videos. The deep background prior is obtained by a background reconstruction network, which is trained in an unsupervised manner with the help of a specifically designed loss. Combining the learned deep background prior with the model-based iterative optimization, the proposed framework benefits from both worlds. Extensive experiments have demonstrated the effectiveness and efficiency of the proposed framework.

It is worth noting that there remains room for further improvements. On the one hand, our deep background prior can be generated by any background reconstruction network, and the quality of the reconstructed background has a great impact on the detection performance. One possible direction would be how to design a more powerful background reconstruction network for effective background reconstruction. On the other hand, the background reconstruction and the iterative optimization are divided into two separate steps, and the parameters of iterative optimization need to be tuned by manual efforts. One can explore how to make the parameters in iterative optimization learnable and how to combine the deep background prior and iterative optimization into an end-to-end network.

REFERENCES

[1] Z. Shao, H. Fu, D. Li, O. Altan, and T. Cheng, "Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation," *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111338.

[2] H. Shirmard, E. Farahbakhsh, R. D. Muler, and R. Chandra, "A review of machine learning in processing remote sensing data for mineral exploration," *Remote Sens. Environ.*, vol. 268, Jan. 2022, Art. no. 112750.

[3] S. Ahmadi, A. Ghorbanian, and A. Mohammadzadeh, "Moving vehicle detection, tracking and traffic parameter estimation from a satellite video: A perspective on a smarter city," *Int. J. Remote Sens.*, vol. 40, no. 2, pp. 8379–8394, May 2019.

[4] Q. He, X. Sun, Z. Yan, B. Li, and K. Fu, "Multi-object tracking in satellite videos with graph-based multitask modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619513.

[5] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[6] J. Zhang, X. Jia, and J. Hu, "Error bounded foreground and background modeling for moving object detection in satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2659–2669, Apr. 2020.

[7] J. Zhang, X. Jia, J. Hu, and K. Tan, "Moving vehicle detection for remote sensing video surveillance with nonstationary satellite platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5185–5198, Sep. 2022.

[8] Q. Yin, T. Liu, Z. Lin, W. An, and Y. Guo, "Moving object detection in satellite videos via spatial–temporal tensor model and weighted Schatten p-norm minimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[9] W. Ao, Y. Fu, X. Hou, and F. Xu, "Needles in a Haystack: Tracking city-scale moving vehicles from continuously moving satellite," *IEEE Trans. Image Process.*, vol. 29, no. 7, pp. 1944–1957, Oct. 2019.

[10] Q. Yin et al., "Detecting and tracking small and dense moving objects in satellite videos: A benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612518.

[11] T. Tirer and R. Giryes, "Image restoration by iterative denoising and backward projections," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1220–1234, Mar. 2019.

[12] S. Gu, R. Timofte, and L. Van Gool, "Integrating local and non-local denoiser priors for image restoration," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2923–2928.

[13] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. V. Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6360–6376, Oct. 2022.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Nov. 2010.

[15] I. Saleemi and M. Shah, "Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 198–219, 2013.

[16] M. Keck, L. Galup, and C. Stauffer, "Real-time tracking of low-resolution vehicles for wide-area persistent surveillance," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 441–448.

[17] G. Saur, W. Krüger, and A. Schumann, "Extended image differencing for change detection in UAV video mosaics," in *Proc. SPIE*, vol. 9026, Mar. 2014, Art. no. 90260L.

[18] 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.

[19] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1–16.

[20] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.

[21] J. Zhang, X. Jia, J. Hu, and J. Chanussot, "Online structured sparsity-based moving-object detection from satellite videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6420–6433, Sep. 2020.

[22] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[23] E. Bayraktar and P. Boyraz, "Analysis of feature detector and descriptor combinations with a localization experiment for various performance metrics," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 25, no. 3, pp. 2444–2454, 2017.

[24] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Jan. 2020.

[25] L. Jiao et al., "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3195–3215, Aug. 2022.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[27] X. Zhou, D. Wang, and P. Krähenbuhl, "Objects as points," 2019, *arXiv:1904.07850*.

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[29] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 966–970.

[30] E. Bayraktar, M. E. Basarkan, and N. Celebi, "A low-cost UAV framework towards ornamental plant detection and counting in the wild," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 1–11, Sep. 2020.

[31] R. Lalonde, D. Zhang, and M. Shah, "ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4003–4012.

[32] C. Xiao et al., "DSFNet: Dynamic and static fusion network for moving object detection in satellite videos," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[33] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2013, pp. 945–948.

[34] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 479–486.

[35] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5546–5557.

[36] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1671–1681.

[37] T. Tirer and R. Giryes, "Super-resolution via image-adapted denoising CNNs: Incorporating external and internal learning," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1080–1084, Jul. 2019.

[38] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3929–3938.

[39] Z. Li and J. Wu, "Learning deep CNN denoiser priors for depth image inpainting," *Appl. Sci.*, vol. 9, no. 6, p. 1103, Mar. 2019.

[40] M. Sultana, A. Mahmood, and S. K. Jung, "Unsupervised moving object detection in complex scenes using adversarial regularizations," *IEEE Trans. Multimedia*, vol. 23, pp. 2005–2018, 2021.

[41] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Unsupervised online video object segmentation with motion property understanding," *IEEE Trans. Image Process.*, vol. 29, pp. 237–249, 2020.

[42] K. Yun, H. Kim, K. Bae, and J. Park, "Unsupervised moving object detection through background models for PTZ camera," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3201–3208.

[43] Z. Bao, P. Tokmakov, A. Jabri, Y.-X. Wang, A. Gaidon, and M. Hebert, "Discovering objects that can move," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11789–11798.

[44] F. Locatello et al., "Object-centric learning with slot attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 11525–11538.

[45] J. Zhang, J. Zhang, and X. Jia, "Learning via watching: A weakly supervised moving object detector for satellite videos," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2333–2336.

[46] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 35, no. 1, pp. 225–253, 2014.

[47] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," 2010, *arXiv:1009.5055*.

[48] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[51] Y. Wang et al., "Disentangling light fields for super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425–443, Jan. 2023.

[52] Y. Xie, S. Gu, Y. Liu, W. Zuo, W. Zhang, and L. Zhang, "Weighted schatten p-norm minimization for image denoising and background subtraction," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4842–4857, Oct. 2016.

Chao Xiao received the B.E. degree in communication engineering and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the College of Electronic Science.

His research interests include deep learning, small object detection, and object tracking.

Ting Liu received the B.E. degree in electrical engineering and automation from the Hunan Institute of Engineering, Xiangtan, China, in 2017, and the M.E. degree in control engineering from Xiangtan University (XTU), Xiangtan, in 2020. She is currently pursuing the Ph.D. degree with the College of Electronic Science, National University of Defense Technology (NUDT), Changsha, China.

Her research interests focus on signal processing, target detection, and image processing.

Xinyi Ying received the M.E. degree in information and communication engineering from the National University of Defense Technology (NUTD), Changsha, China, in 2020, where she is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology.

Her research interests focus on object detection and image super-resolution.

Yingqian Wang received the B.E. degree in electrical engineering from Shandong University (SDU), Jinan, China, in 2016, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018, where he is currently pursuing the Ph.D. degree with the College of Electronic Science and Technology.

His research interests focus on low-level vision, particularly on light field imaging and image super-resolution.

Miao Li received the M.E. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2012 and 2017, respectively.

He is currently an Associate Professor with the College of Electronic Science and Technology, NUDT. His research interests include dim and small target detection, and deep learning.

Li Liu (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2012.

During her Ph.D. study, she spent more than two years as a Visiting Student at the University of Waterloo, Waterloo, ON, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong. From December 2016 to November 2018, she worked as a Senior Researcher at the Machine Vision Group, University of Oulu, Oulu, Finland. She is currently a Full Professor with NUDT. Her papers have currently over 7700 citations in Google Scholar. Her research interests include computer vision, machine learning, artificial intelligence, trustworthy AI, and synthetic aperture radar.

Wei An received the Ph.D. degree from the National University of Defense Technology (NUDT), Changsha, China, in 1999.

She was a Senior Visiting Scholar with the University of Southampton, Southampton, U.K., in 2016. She is currently a Professor with the College of Electronic Science and Technology, NUDT. She has authored or coauthored over 100 journal and conference publications. Her research interests include signal processing and image processing.

Zhijie Chen received the M.S. and Ph.D. degrees from the Nanjing University of Science and Technology, Nanjing, China, in 1989 and 2006, respectively.

He is currently a Professor with the National Airspace Technology Key Laboratory. His research interest is air traffic control.

Dr. Chen is a member of the Chinese Academy of Engineering.