Image and Vision Computing 121 (2022) 104425

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Learning dynamic background for weakly supervised moving object detection

Zhijun Zhang [a], Yi Chang [a,b], Sheng Zhong [a], Luxin Yan [a], Xu Zou [a,*]

[a] Huazhong University of Science and Technology, Wuhan, China
[b] Pengcheng Laboratory, Shenzhen, China

ARTICLE INFO

Article history:
Received 17 December 2021
Received in revised form 26 February 2022
Accepted 28 February 2022
Available online 08 March 2022

Keywords:
Moving object detection
Dynamic background
Data-driven discriminative prior
Low-rank framework

ABSTRACT

Moving Object Detection (MOD) aims at extracting foreground moving objects in videos from static cameras. While low-rank based approaches have achieved impressive success in the MOD task, their performance remains limited on dynamics background scenes. The main reason is that dynamic clutters, *e.g.*, *swaying leaves and rippers*, are easy to mix up with moving objects in the decomposition model which simply classify the sparse noise as foregrounds. In order to improve the generalization ability of low-rank based moving object detectors, we suggest adding an explicit dynamic clutter component in the decomposition framework with realistic dynamic background modeling. Then the dynamic clutter can be learned through object-free video data due to their self-similarity across time and space. Thus, the moving objects can be naturally separated by a tensor-based decomposition model which formulates the static background by a unidirectional low-rank tensor, learns the dynamic clutter by a two-stream neural network, and constrains moving objects with spatiotemporal continuity. To further provide a more accurate object detection result, an objectness prior is embedded into our model in an attention manner. Extensive experimental results on the challenging datasets of dynamic background clearly demonstrate the superior performance of our model over the state-of-the-art in terms of quantitative metrics and visual quality.

1. Introduction

We explore a problem of weakly-supervised moving object detection, where the task is to extract single or multiple moving objects in videos from static cameras with only providing a few object-free frames in the same sequence. Such a formulation of moving object detection [1–4] is attractive because it has been discarded heavy pixel-wise labeling and the object-free frames can be readily obtained in practical applications. This task plays an important role in computer vision such as video surveillance [5], traffic monitoring [6], *etc.* Due to the temporal correlation of video data, low-rank based approaches have achieved impressive success in the literature [1,7,8]. The problem is usually formulated as follow:

$$\mathcal{D} = \mathcal{B} + \mathcal{F} + \mathcal{N}, \tag{1}$$

where $\mathcal{D} \in \mathbb{R}^{H \times W \times T}$ is the input video, $\mathcal{B}$ is the static background, $\mathcal{F}$ is the moving foreground, and $\mathcal{N}$ is the random noise. These methods [9,10] usually assume the background is low-rank due to the similarity of temporal space, and consider the foreground as sparse noise. However, these solutions often fail on the dynamic background scenes, *e.g.*, *swaying leaves or rippers*, since these dynamic variations caused by background have similar sparse and changeable properties with moving objects, thus resulting in ambiguities on sparse noise as well as poor detection results. An example is illustrated in Fig. 1(a).

To handle this issue, existing methods mainly start from three different perspectives respectively: foregrounds $\mathcal{F}$, noise $\mathcal{N}$, and background $\mathcal{B}$. An intuitive idea to differentiate foregrounds and dynamic clutter is that moving objects $\mathcal{F}$ often have spatial continuity and smooth movements in temporal space in contrast to the randomness of dynamic background variations, thus a lot of works [11–14] focus on modeling these properties by constraining foregrounds with total variation [12,15], MRF [14], segmentation [16], structural sparse [11,17], or block sparsity [18,19] regularization. However, this assumption is violated when the object size is small or the dynamic clutter (*e.g.*, *fountain*) exhibits structural and smooth properties in the spatiotemporal domain.

Another research line [20–22] regards the dynamic clutter as structural noise, and thus model it into the noise $\mathcal{N}$, or mixed noise both $\mathcal{N}$ and sparse error noise $\mathcal{F}$. They introduce the Mixture of Gaussian (MoG) to model the multiple modalities of the complex 'noises'.

[*] Corresponding authors.
E-mail address: zx@zoux.me (X. Zou).

https://doi.org/10.1016/j.imavis.2022.104425
0262-8856/© 2022 Elsevier B.V. All rights reserved.

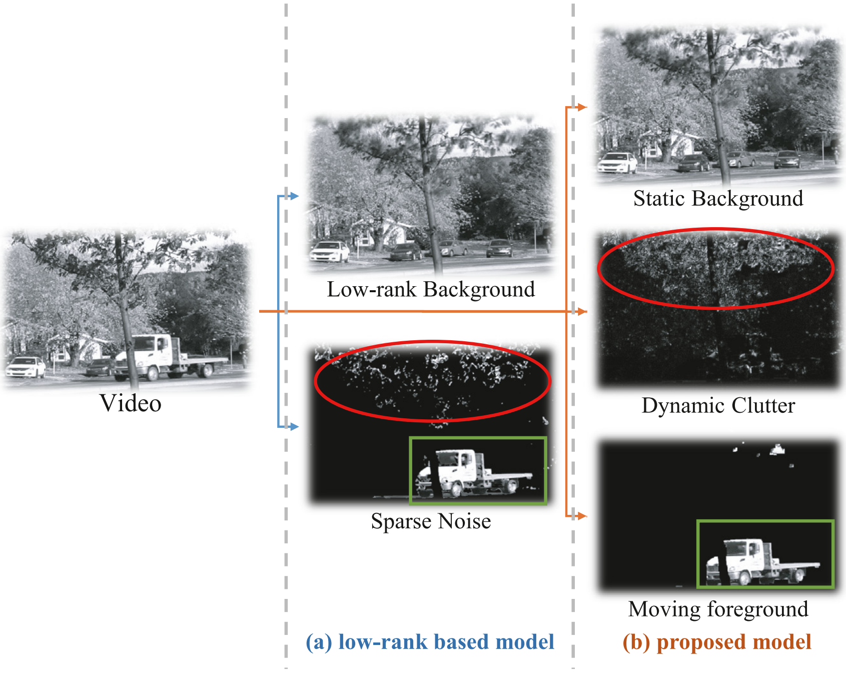

Fig. 1. An example of video decomposition models of dynamic background videos. From left to right: video data, low-rank based decomposition model, proposed model. (a), the sparse characteristic of dynamic clutter makes it differentiating from foregrounds much difficult in the low-rank based model. (b), to address this issue, we present a novel decomposition model, where the dynamic clutter is modeled explicitly. The additional dynamic clutter modeling facilitates us to decouple the moving objects from dynamic backgrounds.

However, the distributions of the moving object foreground and the dynamic clutter vary with different scenes, thus it is very hard for the Gaussian models to accommodate such complex mixed noise. A few methods [7,23] take the dynamic clutter background into the static background $\mathcal{B}$. They resort to the higher-order low-rank technique [23,24] to directly describe a background scene with dynamic textures, or utilize tensor nuclear norm representation [7] to accommodate the background variation. Unfortunately, this would inevitably result in spatial distortion.

To avoid the confusion between dynamic clutter and foregrounds, we start from the idea of keeping each component simple, and propose a novel tensor-based video decomposition framework to Learn Dynamic Background (LDB), which additionally model the dynamic clutter explicitly, as shown in Fig. 1(b). The separated dynamic clutter makes the representation of both the static background and moving object foreground easier and more precise, as well as the dynamic background clutter itself. On the one hand, the dynamic clutter can be regarded as a specific structural noise, and learned from a single video by considering one object-free video sequence itself as a dataset. On the other hand, the foreground can be extracted based on the MAP framework with the integration of data-driven discriminative prior for explicit dynamic clutter modeling. The overview of our approach is shown in Fig. 2. To complete the framework, we introduce the low-rank tensor approximation to exploit spatiotemporal correlations of the background, as well as a 3D MRF regularization for constraining the continuity of the foregrounds in spatial and temporal space. Furthermore, an objectness prior mask is served as adaptive guidance for the dynamic clutter component to improve the detection results. The contributions of our approach are four-folds:

- We formulate moving object detection as a novel tensor-based video decomposition framework, consisting of the static background, moving foregrounds and dynamic clutter. To the best of our knowledge, we are the initial work to model the dynamic clutter explicitly.

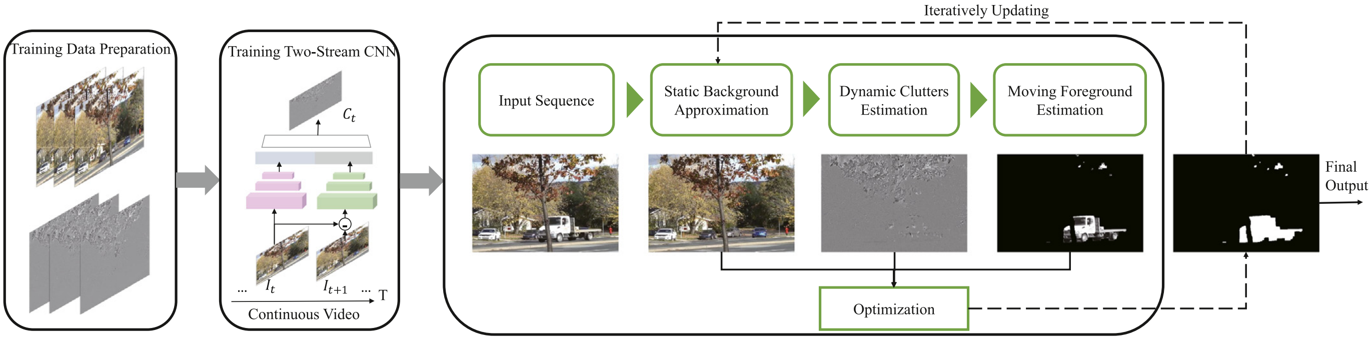

Fig. 2. The overview of the proposed approach. Our approach is mainly based on the video decomposition framework consisting of three components: Static Background, Dynamic Clutter and Moving Foregrounds. The dynamic clutter is explicitly represented by a learned context-aware two-stream CNN in a weakly supervised manner. Incorporated with this data-driven discriminative prior, the unified MAP framework can be solved by the alternating minimization algorithm iteratively, with the regularization on the static background (unidirectional low-rank) and moving foregrounds (3D MRF).

- We introduce a data-driven discriminative prior represented by a spatiotemporal two-stream CNN for dynamic clutter, which could better approximate the complex distribution of the clutter. The prior is embedded into the MAP framework with an efficient alternating algorithm to solve it.
- We further enhance the proposed method by incorporating an objectness mask of the static and moving objects, functioning as an attention map to adaptively differ moving objects from dynamic clutter.
- Extensive results demonstrate that the proposed algorithm outperforms the state-of-the-art methods both quantitative and qualitative in dynamic scenes. Moreover, the generalization ability of our method is validated in other challenging conditions, such as noises and rains.

2. Related work

Video Decomposition: In recent years, the low-rank matrix recovery [9,10,13,14,25,26] and tensor-based recovery methods [7,12,15,27,28] have been widely used and achieved state-of-the-art performance for moving object detection, where they split the input video into the low-rank approximation of background component and a sparse error of foreground component based on the Eq. (1). However, the classical RPCA method is vulnerable to the dynamic background variations, due to the violation of the low-rank assumption [9]. From the noise modeling viewpoint, Meng et al. [20] proposed to treat dynamic clutter as a special structural noise, and introduced the mixture of Gaussian (MoG) to accommodate the complex distribution of the mixed random $\mathcal{N}$ and clutter noise. A lot of works follow this research line [21,22]. Another research direction tends to model the dynamic clutter along with the foreground object $\mathcal{F}$, since both of them are moving and structural sparse. Cao et al. [12] introduced the conventional L_0 for both the dynamic clutter and foreground object, which implicitly assumed that they were sparsely distributed in the video. However, such an implicit modeling manner for dynamic clutter would increase the difficulty of differentiating it from the other components. In this work, we propose to explicitly model the dynamic clutter, and additionally incorporate the dynamic clutter into the degradation model Eq. (1) (Section 3.1) which would significantly reduce the difficulty of the model each term.

Spatiotemporal Representation: The success of the video decomposition depends greatly on the representation of each term. Most of the existing methods follow the conventional RPCA model and start from the 2D low-rank matrix methods [9]. Both the low-rank matrix recovery [13,18,28] and the matrix factorization [29–31] handle the video decomposition problem from the 2D perspective. However, transforming the 3D video into a matrix would unexpectedly cause damage to the spatiotemporal structural correlation. The latest works consistently indicate that the tensor-based methods substantially preserve the intrinsic structure correlation with better results, such as the background subtraction [7,32], video completion [33–35], multispectral image restoration [36]. Hu et al. [7] proposed a tensor-based low-rank and saliently fused sparse decomposition model with better performance. In this work, we propose a unidirectional low-rank tensor prior to model the spatiotemporal correlation of the static background (Section 3.3), and 3D MRF to model the spatiotemporal continuity of the moving object (Section 3.4).

CNNs for Discriminative Prior: Most previous methods make use of the hand-crafted priors [7,12] for each component. However, it is extremely hard to provide an accurate mathematical expression to fit the complex distribution of the dynamic clutter, since the distribution of it is usually non-Gaussian or non-Laplacian. Recently, CNN has received great success in both low-level and high-level vision tasks. Compared with the hand-craft priors, CNN could perfectly fit any regular components, such as the structural noise [37], snow and rain [38], guaranteed by the universal approximation theory [39]. In this work, to the best of our knowledge, this is the first time the discriminative CNN is introduced to model the dynamic clutter (Section 3.2), and has been embedded into the MAP framework for (Section 3.5) for better performance.

CNNs for Background Modeling: Several recent papers introduce end-to-end CNNs to learn background in a supervised manner, which could be divided into video-optimized [40–43] and video-agnostic [44,45] categories. The first category [41,42] trains and tests on the same set of videos, obtaining astonishing performance, but suffers from severe overfitting and inevitably fails when transferring into new scenes. The second category [44,45] tries to eliminate the overfitting by training and testing on different sets of videos, while these methods still rely on pixel-wise annotations on many sequences, and are sensitive to the training/testing split and sematic prior. In contrast to these algorithms, our method does not require pixel-wise annotations, thus would significantly reduce labeling burdens, and has a better generalization ability and stability than deep learning based methods, which would be discussed in the experiments.

3. Proposed method

3.1. The tensor decomposition model

To decouple the dynamic background and moving objects, we propose a novel video decomposition model that separates the dynamic clutter from the data. For a video sequence tensor $\mathcal{D} \in \mathbb{R}^{H \times W \times T}$, the model is formulated as follow,

$$\mathcal{D} = \mathcal{B} + \mathcal{F} + \mathcal{C} + \mathcal{N}, \tag{2}$$

where $\mathcal{B}, \mathcal{F}, \mathcal{C}, \mathcal{N} \in \mathbb{R}^{H \times W \times T}$ denote static background, moving foreground, dynamic clutter and Gaussian noise tensor respectively. Compared with Eq. (1), we additionally introduce a dynamic clutter term $\mathcal{C}$ in an explicit manner, which represents the dynamic background variations. Our goal is to decompose the static background $\mathcal{B}$, the moving foreground $\mathcal{F}$ and the dynamic clutter $\mathcal{C}$ from the input video $\mathcal{D}$. To solve this ill-posed problem, we need to analyze the priors of $\mathcal{B}$, $\mathcal{F}$ and $\mathcal{C}$, and then introduce the corresponding regularizers, which will be discussed in the next subsections.

3.2. Dynamic background: Two-stream CNN

In this section, we first illustrate the reason for choosing convolutional neural networks to model the dynamic clutter, then the architecture of proposed two-stream spatiotemporal CNN and the details of annotation generation are described in the following subsection.

3.2.1. Why CNN modeling for dynamic clutter?

Existing moving object detection methods often assume a probability distribution for dynamic background clutter, such as Gaussian noise [10] (*i.e.*, L_2-norm), sparse error [12] (*i.e.*, L_1-norm) and Mixture of Gaussian [22]. Such assumption is substantiated to be effective in some video sequences, however, still insufficient to fit complex dynamic background in real scenes.

To understand it, we show three typical complex dynamic clutters in Fig. 3. It can be observed that the dynamic clutter, like *swaying leaves* and *fountain*, has strong structural property varying from scene contexts, and their distributions are quite complex and diverse. Here, we select an effective algorithm, OMoGMF [22], to model the dynamic clutter by the Mixture of Gaussian (MOG). Even though MoG has a good approximation capability to a wide range of distribution, a gap between the estimated results and the ground truth can be observed easily in Fig. 3. The phenomenon reflects that current generative models are difficult to handle diverse complex background variations.

On the contrary to these methods, we introduce a context-aware learning prior to model the distribution of dynamic clutter. CNN, as well known for its powerful representation capability, is a natural choice

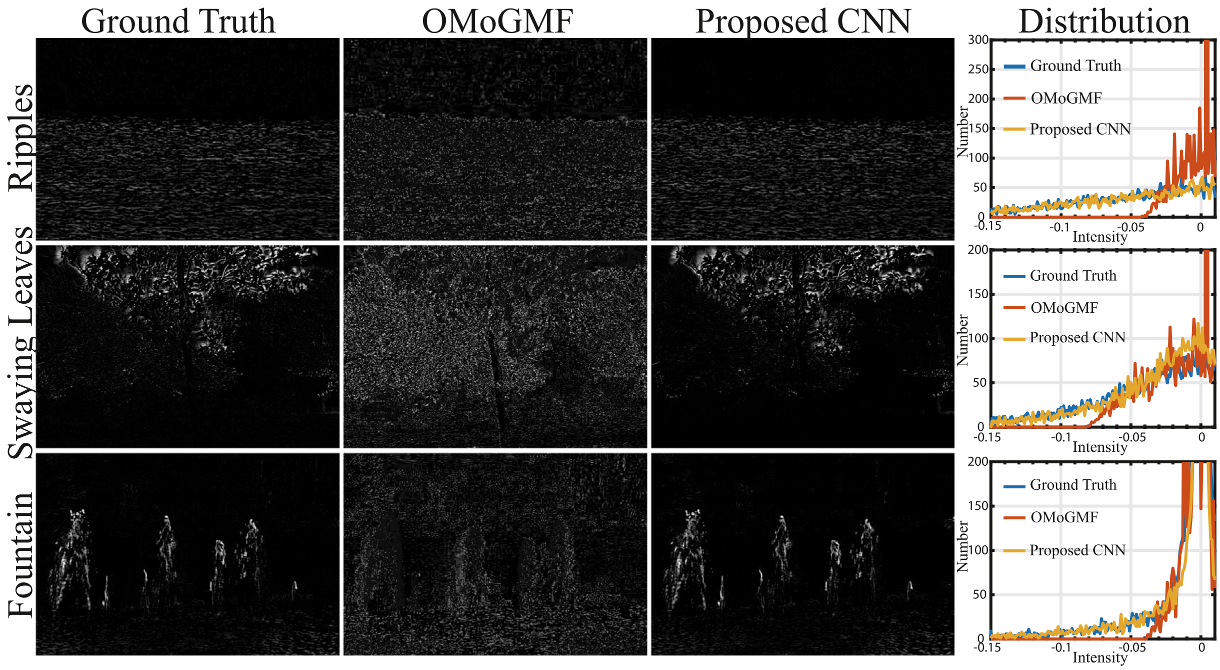

Fig. 3. A comparison of dynamic clutter estimation between OMoGMF [22] and proposed CNN on three videos. From left to right: ground truth of dynamic clutter, estimated results of OMoGMF, our results, and corresponding distributions of them. Clearly, our model has a more powerful ability to handle the complex and changeable distribution of dynamic clutter, rather than the mixture of Gaussian.

for the problem. Therefore, we design a two-stream spatiotemporal CNN to model the dynamic clutter. As shown in Fig. 3. the results of proposed CNN (the third column) are quite similar to the ground truth (the first column) in visual, and the last column shows that the distribution of proposed CNN is much closer to the ground truth, compared with OMoGMF.

3.2.2. Two-stream spatiotemporal CNN

To model the dynamic clutter efficiently, both appearance and motion information are utilized in our model. We construct two branches to incorporate spatial and temporal information: one branch is to extract appearance features of the target frame I_t, and another branch can obtain the motion feature. Since the camera is static, the simple difference of consecutive frames is sufficient to represent motion characteristics of dynamic variations, thus we formulate $I_{t+1} - I_t$ as the input of motion branch. The architecture of each branch is based on a patch-based denoising network [46] due to self-similarity across time and space, which removes the last convolution layer for noise estimation. Fused by the features from two branches, the network could estimate the dynamic clutter of each frame, *i.e.* $f_{CNN}(I_t, I_{t+1})$, where f_{CNN} denotes the prediction of proposed two-stream spatiotemporal CNN. The detail of proposed CNN can be found in supplementary materials.

An important problem is that we have no annotations of dynamic clutter. To solve the problem, We introduce a weakly supervised strategy to generate annotations from object-free background frames, which are usually available due to the redundancy of video sequence. For each video, we select around 200 consecutive object-free background frames manually. These frames do not contain moving foregrounds, but show a strong dynamic property of background variations. We perform RPCA algorithm to extract the dynamic variation of these background frames. The obtained sparse error is regarded as the ground truth of dynamic clutter. Due to the temporal consistency of clutter, the learned network could model the distribution of whole sequences. It is worth noting that there is no overlap between training and testing frame sets.

3.3. Static background: Low-rank tensor

Since we decouple the dynamic clutter from background explicitly, the remaining static background should be unchanged. Thus, we model the static background in a low-rank tensor. Notably, the low-rank property of static background tensor is anisotropy, where the spatial dimensions have no similar low-rank property with temporal, as shown in Fig. 4. For this reason, we only apply temporal dimensional low-rank constraint on static background tensor. Mathematically, given a static background tensor $\mathcal{B} \in \mathbb{R}^{H \times W \times T}$, we define the unidirectional rank as $rank_3(\mathcal{B})$, where $rank_3(\mathcal{B}) = rank(B_3) = \sum_j |\sigma_j(B_3)|_1$ [47,48] is the sum of the singular values of the tensor unfolded along the temporal dimension, B_3 is the mode-3 unfold matrix of $\mathcal{B}$, and σ_j (B_3) means the j-th singular value of B_3.

3.4. Moving foreground: 3D MRF

For moving foreground, the hypothesis is that foreground objects are contiguous and sparse in spatial and temporal space. Thus we use Markov Random Field (MRF) [49,50] to model the probability and correlations of foreground pixels.

Define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices denoting all $H \times W \times T$ pixels in the video frames and $\mathcal{E}$ is the set of edges connecting spatial and temporal neighboring pixels respectively. The constructed graph is shown in Fig. 5. Similar to [14], we define a binary tensor $\mathcal{S} \in \{0, 1\}^{H \times W \times T}$ as the foreground support,

$$S_{ijk} = \begin{cases} 1, & \textit{if pixel ij at frame k is foreground} \\ 0, & \textit{otherwise}. \end{cases} \tag{3}$$

We set the weights of pairwise edges on spatial and temporal are constant, *i.e.* $\omega_s = \eta$, $\omega_t = \theta$. And the unary probability is β when $S_{ijk} = 1$. Thus the energy of moving foreground

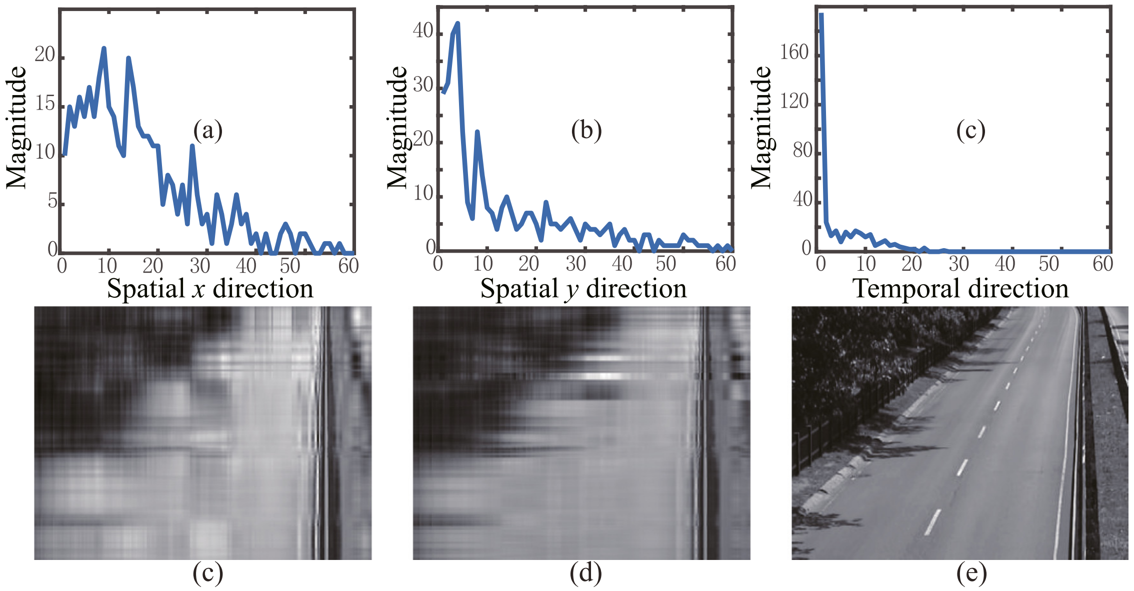

Fig. 4. Low-rank property analysis of video tensor along each dimension *via* HOSVD. From left to right is the mean profile of singular values bigger than 5 elements in the core tensor and reconstructed images across spatial *x*, spatial *y*, and temporal dimension, respectively. The video tensor exhibits unidirectional low-rank property.

ort can be written as follow,

$$E(\mathcal{S}) = \beta \|\mathcal{S}\|_1 + \eta \sum_{(ijk,mnk)\in\mathcal{E}_s} \left|S_{ijk} - S_{mnk}\right| + \theta \sum_{(ijk,ijt)\in\mathcal{E}_t} \left|S_{ijk} - S_{ijt}\right|, \tag{4}$$

where $\mathcal{E}_s$, $\mathcal{E}_t$ are edges of spatial and temporal neighborhoods. The hyper-parameters β penalizes the $S_{ijk}=1$, and η, θ reflect the importance of spatial and temporal continuity. Thus the sparsity and spatiotemporal continuity are modeled in the energy function of MRF, which can be integrated into the unified framework to solve the problem.

3.5. Formulation

Based on the modeling of three components aforementioned, our moving object detection method builds on Learning Dynamic Background (LDB), which is formulated as:

$$\min_{\mathcal{B},\mathcal{C},\mathcal{S}_{ijk}\in\{0,1\}} \frac{1}{2} \|P_{\mathcal{S}^{\perp}}(\mathcal{D}-\mathcal{B}-\mathcal{C})\|_F^2 + \alpha rank_3(\mathcal{B}) + \beta E(\mathcal{S}) \\ + \gamma\left(\mathcal{M}\circ \|\mathcal{C}-f_{CNN}(\mathcal{D})\|_F^2\right), \tag{5}$$

where $P_{\mathcal{S}^{\perp}}$ represents the complementary orthogonal projection of a tensor $\mathcal{X}$ onto the linear space of tensor supported by $\mathcal{S}$, refers to [14], and the Frobenius norm of a tensor is the square root of the sum of the squares of all its elements. The first term in Eq. (5) is the data constrain respect to the decomposition model, *i.e.* video D_{ijk} is best fitted by static background B_{ijk} plus dynamic clutter C_{ijk} when no moving foreground $S_{ijk}=0$. The second term forces that static background $\mathcal{B}$ is temporal unidirectional low-rank. The third term $E(\mathcal{S})$ is the MRF energy function of $\mathcal{S}$, which models the sparsity and spatiotemporal smoothness of moving foreground. The last term means the dynamic clutter can be well represented by a constructed two-stream spatiotemporal CNN. $\mathcal{M}$ is an indicator tensor with objectness prior to force the objects not to mingle with the background variations:

$$\mathcal{M}_{ijk} = \begin{cases} \zeta, & if\ pixel_{ijk} \in objectness\ regions \\ 1, & otherwise, \end{cases} \tag{6}$$

where $\zeta=0.5$ is a hyper-parameter. The objectness prior is utilized and validated its effectiveness for moving object detection in [7,51]. We calculate objectness regions by a category-agnostic object segmentation approach [52] trained on PASCAL VOC dataset [53]. We impose the indicator tensor on dynamic clutter estimation as an attention map, to

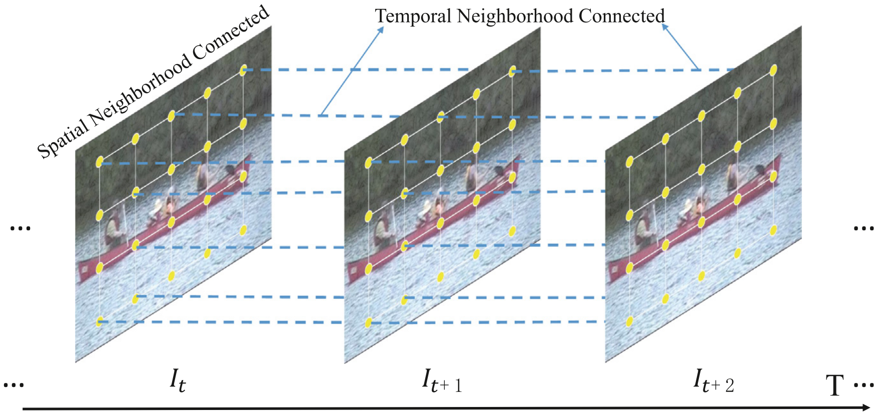
Spatial Neighborhood Connected
Temporal Neighborhood Connected
...
...
...
I_t
I_{t+1}
I_{t+2}
T ...

Fig. 5. Graph construction for 3D MRF. We connect each pixel with its 6-neighborhood in spatial and temporal space to constrain foregrounds with spatiotemporal continuity.

adaptively reduce the negative impacts from inaccurate clutter estimation. The stable improvement is illustrated in Fig. 6.

The basic idea of the model is that the dynamic clutter can be satisfactorily represented by a two-stream CNN, and the sparse and smooth moving objects modeled by 3D MRF, meanwhile unidirectional low-rank regularizes the temporal similarity of static background. The unified framework could benefit from the data-driven discriminative prior of dynamic clutters, yielding separation of the moving foreground with the dynamic background. Note the proposed method is a general framework that can handle different kinds of clutter, *e.g.*, dynamic variations, bad weathers, noises.

3.6. Optimization

Since the objective function defined in (5) is nonconvex, joint optimization over $\mathcal{B}, \mathcal{C}, \mathcal{S}$ is extremely difficult. Hence, we adopt an alternating algorithm that separates the energy minimization over $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{S}$ into three steps.

1) **Update for $\mathcal{B}$: Static background approximation.** Fixed moving foreground support $\mathcal{S}$ and dynamic clutter $\mathcal{S}$, the minimization in (5) over $\mathcal{B}$ by its tensor unfolding formation can be formulated,

$$\min_{B_3} \frac{1}{2} \|P_{S_3^\perp}\left(\mathrm{D}_3 - \mathrm{B}_3 - \mathrm{C}_3\right)\|_\mathrm{F}^2 + \alpha \|\mathrm{B}_3\|_*. \tag{7}$$

The Eq. (7) is a low-rank matrix approximation problem which can be solved by the SOFT-IMPUTE algorithm [54] iteratively, *i.e.*,

$$B_3 \leftarrow \Theta_\alpha \left(P_{S_3^\perp}(D_3 - C_3) + (P_{S_3} B) \right), \tag{8}$$

where Θ_α the singular value thresholding [55]. After B is obtained, the tensor folding is performed to transform it into 3-order tensor $\mathcal{B}$.

2) **Update for $\mathcal{C}$: Dynamic clutter estimation.** Dropping out the irrelevant variable in Eq. (5), we can get the following subproblem,

$$\min_{\mathcal{C}} \frac{1}{2} \|P_{\mathcal{S}^\perp}(\mathcal{D} - \mathcal{B} - \mathcal{C})\|_F^2 + \gamma \left(\mathcal{M} \circ \|\mathcal{C} - f_{CNN}(\mathcal{D})\|_F^2 \right), \tag{9}$$

the clutter can be computed iteratively by [54],

$$\mathcal{C} \leftarrow \mathcal{M} \circ \frac{P_{\mathcal{S}^\perp}(\mathcal{D} - \mathcal{B}) + P_\mathcal{S}(\mathcal{C}) + 2\gamma f_{CNN}(\mathcal{D})}{(2\gamma + 1)}. \tag{10}$$

3) **Update for $\mathcal{S}$: Moving foreground estimation.** Given the current static background $\mathcal{B}$ and dynamic clutter $\mathcal{C}$, define $\mathcal{Z} = \mathcal{D} - \mathcal{B} - \mathcal{C}$, $E_p(\mathcal{S})$ as the pairwise energy of Eq. (4), the formulation (5) can be rewritten as follow,

$$\begin{aligned} &\frac{1}{2} \|P_{\mathcal{S}^\perp}(\mathcal{Z})\|_F^2 + \beta \|\mathcal{S}\|_1 + E_p(\mathcal{S}) \\ &= \frac{1}{2} \sum_{ijk} Z_{ijk}^2 (1 - S_{ijk}) + \beta \|\mathcal{S}\|_1 + E_p(\mathcal{S}) \\ &= \sum_{ijk} \left(\beta - \frac{1}{2} (Z_{ijk})^2 \right) (S_{ijk}) + E_p(\mathcal{S}) + \mathcal{O}, \end{aligned} \tag{11}$$

where $\mathcal{O} = \frac{1}{2} \sum_{ijk} \left(\mathcal{Z}_{ijk} \right)^2$ is a constant when $\mathcal{B}, \mathcal{C}$ is fixed. The above energy is the standard form of MRFs with binary labels, which can be solved exactly by graph cuts [56].

4. Experimental results and discussion

Extensive experimental validation is built on both synthetic (SABS [57]) and real (I2R [58] and CDnet [59]) datasets. In order to verify the effectiveness of proposed method in dynamic scenes, we select three dynamic outdoor scenes in I2R [58] database and one typical sequence in SABS [57] dataset. For CDnet [59], the challenging *dynamicBackground* and *badWeather* categories are employed in our experiments. We compare LDB with six state-of-the-art algorithms, including RPCA [9], DECOLOR [14], GOSUS [30], OMoGMF [22], LSD [11], TVRPCA [12].

4.1. Implementation details

Parameter setting. In our model, the low-rank parameter α, the sparsity parameter β and the clutter parameter γ control the balance of the temporal low-rank property of static background, the sparsity and continuity of moving foreground and similarity of estimated dynamic clutter respectively. For α, we start it with a relatively large value, (*i.e.*, the second largest singular value of D_3), and decrease it after each iteration by a factor $\eta_1 = 1/\sqrt{2}$ until the temporal-dimension rank less than $K = 5$. For β, we similarly set it with a relatively large value, and decrease with each iteration by a factor $\eta_2 = 0.5$ until β reaches $4.5\hat{\sigma}^2$, where $\hat{\sigma}$ is estimated online by the variance of residual $\mathcal{D} - \mathcal{B} - \mathcal{C}$. Empirically, the γ is set to be $\eta_3\sqrt{\beta}$, where $\eta_3 = 30$ for most sequences. Specifically, the parameter η, and θ in MRF constraint the spatial and temporal continuity of moving foreground, which are set 5β, 0.5β respectively.

Training details. For each video, we select around 200 continuous pure background frames to generate annotations. These training frames are performed by RPCA method to obtain static background and sparse outliers which can be regarded as the ground truth of dynamic clutters. We use the target frame I_t and its neighbor frame I_{t+1} as the input of two-stream spatiotemporal CNN, and the dynamic clutter C_t of the

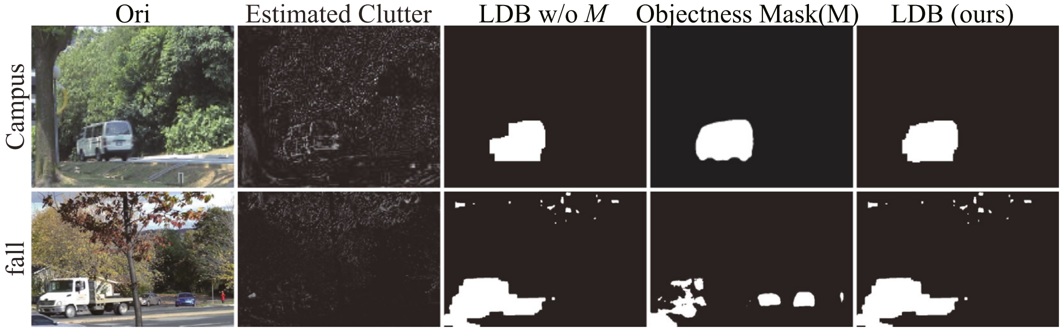

Fig. 6. Illustrations of two example results on estimated clutter, objectness mask, our results without and with objectness prior. It is shown that our approach obtains a better result by utilization of objectness prior (the first row), and avoid the negative impact of some unsatisfactory objectness predictions (the second row).

target frame as the output of the network. For the appearance branch, the input is the target frame I_t to obtain the appearance information. For the motion branch, we use an informative difference image of the target frame and its neighbor frame $I_{t+1} - I_t$ to capture the motion patterns. All frames are cropped into 40×40 image patches for training. Based on Pytorch [60], we use Adam [61] with 32 batch size for 50 epochs. The learning rate is set to 10^{-3}, and is gradually decreased. The details of the network can be found in the supplementary material.

4.2. Experimental results

CDnet dataset. CDnet dataset is considered as one of the most difficult tracking benchmarks, which consists of 31 real-world videos over 80,000 frames and spanning six categories including diverse motion and change detection challenges. In this dataset, we select the challenging *dynamicBackground* and *badWeather* categories to verify the effectiveness of proposed methods in complex dynamic conditions. The results are shown in Fig. 7 and Table 1.

The *dynamicBackground* category contains six video sequences with dynamic background motions and shelters, and some dynamic clutters have strong structural properties, *e.g.*, *swaying leaves*, *fountain*. Due to the complexity of dynamic background, many methods do not achieve satisfactory results, the best result is 0.56 on average of six videos by TVRPCA [12]. However, our approach (LDB) obtains a great improvement of 39%, and the F-measure score is 0.78. It is worth noting that for the most challenging sequences *fountain01* and *fountain02*, our approach improves the performance by 75% and 43% respectively.

The *badWeather* category has slight background variations due to rain or snow. These variations are quite small and sparse, which could be eliminated well by some powerful approaches, and our approach still achieves a comparable performance of 0.91 F-measure.

I2R dataset. As a widely used benchmark in the tasks of foreground and background separation, the I2R dataset contains nine real videos including static background, dynamic background, *etc. Campus, Fountain, WaterSurface*, which are three typical dynamic outdoor scenes, are selected. Table 3 lists the average F-measure over all foreground-annotated frames calculated from all competing methods on the three sequences. It can be observed that on average LDB has an evidently better performance of 0.88 than other competing methods, while the second result is calculated by DECOLOR [14] of 0.85. This superiority of the LDB method can also be observed from Fig. 7. It is easy to see that the detected foregrounds by LDB are closer to the groundtruth ones, which conducts its larger F-measures in experiments.

SABS dataset. The SABS dataset is a synthetic dataset for pixel-wise evaluating the performance of background subtraction. We employ only *NoCamouflage* sequence, which exists shelters (trees) and dynamic textures (periodically swaying leaves). The quantitative result is shown in Table 3. Our approach achieves the best F-measure score of 0.71, significantly better than other competing methods, where the best one is 0.38 of TVRPCA [12]. Fig. 7 (last row) also validates the superiority of proposed methods. Results evidently eliminate the interference of dynamic textures of trees.

4.3. Ablation studies

Low-rank matrix *vs.* low-rank tensor. We begin our ablation study by exploring the effectiveness of low-rank tensor modeling on static background in our LDB. We implement fairly by removing other component design, *i.e.*, 3D MRF and Clutter in our model and utilize the L1 sparse norm regularization on foreground modeling. Table 2a shows the results of low-rank tensor(LT) modeling and low-rank matrix(LM) modeling. It can be seen that unidirectional low-rank tensor greatly improves the results by 0.24 in *I2R*, 0.15 in *dB*, and 0.08 in *bW*, while slightly decrease the synthetic *NoCamouflage* by 0.03. We analyze that the low-rank tensor modeling would capture spatial–temporal correlations in videos, inducing better performance, except for the synthetic sequence which has strict alignment in temporal dimension.

Study on the effect of each component. In order to verify the effectiveness of each component presented, we experiment on three datasets shown in Table 2b. From the results, we can see that 3D MRF modeling significantly improves the results of I2R (0.28) and SABS (0.28) due to the spatial–temporal continuity constraints on foregrounds, additional clutter component increases the performance by a large margin on typical dynamic scences, including dB (0.26), and SABS (0.35), and objectness modeling slightly improves the final results by means of the semantic prior of objects.

Comparison with objectness detection. We conduct comparative study to investigate the effectiveness of our model with pure objectness detection, reported in Table 2c. The superior performance of LDB might be ascribed to the fact that low-rank decomposition model is crucial for moving object detection. The boost performance might come from two aspects: (a) our foreground modeling could represent arbitrary moving objects instead of pre-defined object categories in PASCAL VOC, and (b) the low-rank constraint on static background would suppress the ambiguity of static objects detected in objectness detection.

4.4. Generalization analysis

To further explore the generalization of the proposed method, we extend to apply LDB to other challenging conditions, like Gaussian noise and rain. Here we select one simple video sequence without dynamic background *highway* while adding three levels of Gaussian noises as well as rains to simulate various challenging dynamic conditions. Table 4 shows the comparison results with state-of-the-art algorithms. The superiority of LDB can be easily observed: It performs the best/the second best in 4/1 out of 5 conditions. Fig. 8 shows that LDB is robust to these challenging conditions, which naturally confirms the generalization ability of our proposed method.

Despite our method being designed for the dynamic background scenes, we also conduct the generalization experiments on other regular and challenging scenes in the CDnet dataset, demonstrated in Table 5. We compare our LDB with other state-of-the-art methods on the other challenging categories of CDnet, including *baseline*, *cameraJitter*, *intermittentObjectMotion*, *shadow*, *thermal*, *turbulence*, and *nightVideos*. We observe that for these categories, our LDB still achieves comparable results, which indicate that our low-rank tensor constraints on static background and the sptaiotemporal continutity modeling on foregrounds can adapt flexibly on various challenging videos.

4.5. Compared with deep learning methods

We also compare our approach with end-to-end deep learning methods from three aspects: annotation burdens, generalization ability, and stability. We select two state-of-the-art algorithms FgSegNet v2 [42] and BSUV-Net [44], and implement their source codes to perform evaluations on the CDnet dataset, shown in Table 6. (1) Annotation burden: FgSegNet v2 and BSUV-Net both require pixel-wise annotations of moving objects, causing higher labeling burdens than our approach, which only utilizes a few object-free frames to support weakly supervision. (2) Generalization ability: we evaluate it by splitting the training and testing sets on different videos, and report the results on unseen videos. As shown in Table 6, although FgSegNet v2 achieves the best result on seen videos (row 1), it fails catastrophically on unseen sequences (row 2), while our method has a stable result on unseen videos (row 5) since the majority of the background is modeled by the low-rank component, not discriminative prior of sparse dynamic clutter. The BSUV-Net only provides one trained model without knowing its training/testing split configuration, thus the result on row 3 is on both seen and unseen sequences. (3) We also compare the stability of BSUV-Net and our approach on additional segmentation masks, which are used in both BSUV-Net and LDB to improve the results. By setting the value

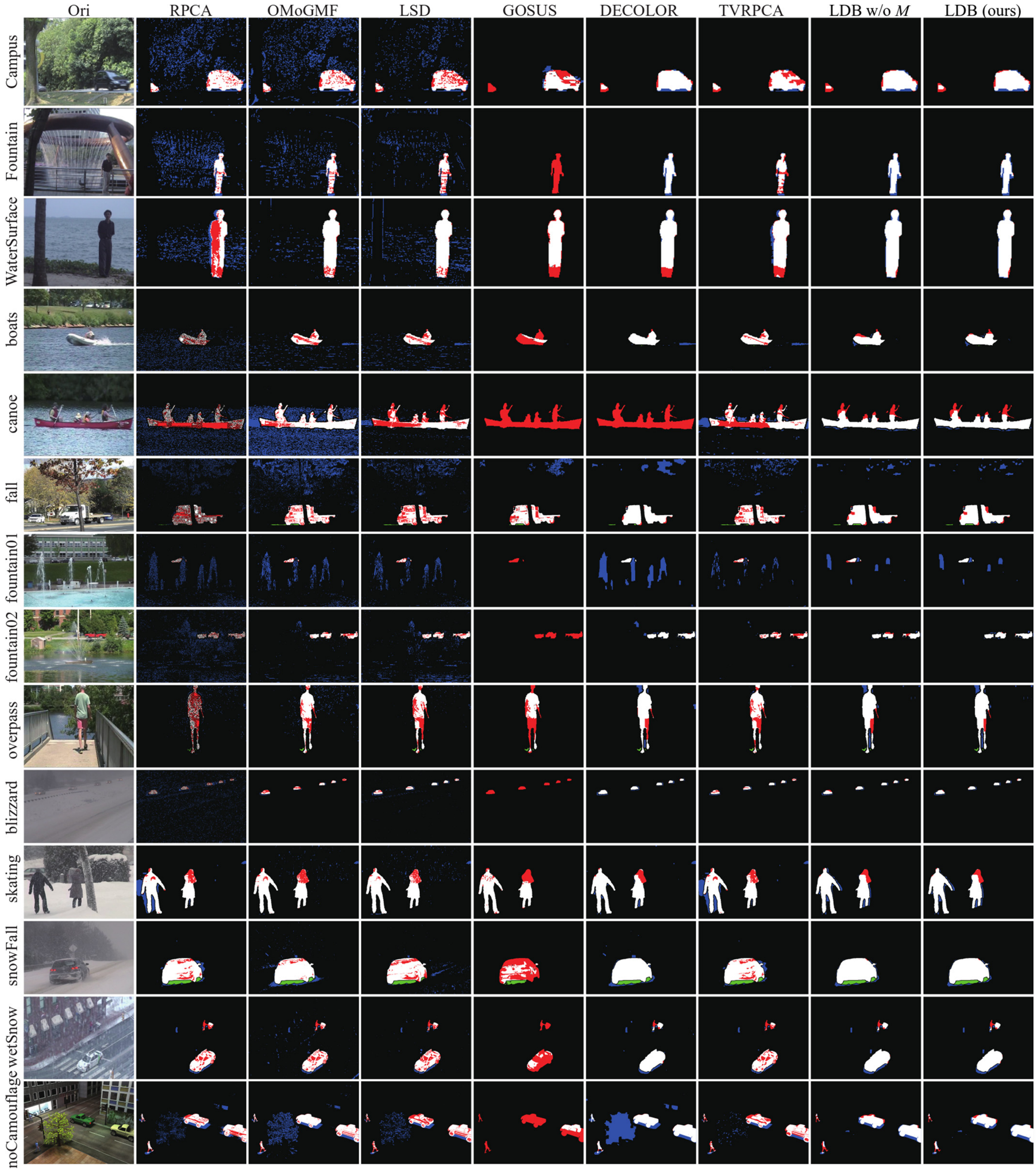

Fig. 7. Qualitative result comparisons on I2R [58] (from first to third rows), CDnet [59] (from fourth to thirteenth rows) and SABS [57] (the last row) datasets.

of the segmentation mask as zero, we obtain the final results (row 4 and 5) to compare the stability. It is seen that the BSUV-Net highly relies on the segmentation mask results, while our approach shows the best results on unseen videos without semantic segmentation prior, as shown on gray background in the Table 6.

We also compare our approach with unsupervised deep learning based methods, where UMOD [62] is a typical work. They utilize images and the associated optical flow as data input, and train two adversarial networks to separate the foreground and background. Since this paper does not provide their results on the CDNet dataset, we try to implement it on the CDNet with their released codes for fair comparison. We use all annotated CDNet data to train 40 epoches for the generator and the inpainter on a single GPU Nvidia 2080Ti, with PWC [63] to generate optical flow following the paper setting. The results are shown as

Table 1
Performance evaluation on CDnet [59] dataset using F-measure. The average F-measures of the *dynamicBackground* and *badWeather* categories are shown in the eighth (Avg) and the last column (Avg). Red: best, blue: the second best. Our approach exhibits a great improvement on *dynamicBackground* category, and achieves state-of-the-art performance on *badWeather* category.

	dynamicBackground							badWeather				
Video	boats	canoe	fall	foun.01	foun.02	over.	**Avg**	blizz.	skat.	snowF.	wetS.	**Avg**
RPCA[55]	0.13	0.14	0.36	0.02	0.02	0.52	0.20	0.10	0.59	0.58	0.69	0.49
OMoGMF[22]	0.37	0.26	0.23	0.04	0.07	0.78	0.29	0.63	0.86	0.81	0.72	0.76
LSD[11]	0.35	0.47	0.35	0.04	0.09	0.72	0.34	0.70	0.85	0.80	0.74	0.77
GOSUS[30]	0.34	0.74	0.60	0.00	0.08	0.79	0.43	0.00	0.80	0.20	0.44	0.36
DECOLOR[14]	0.92	0.00	0.69	0.03	0.56	0.94	0.53	0.90	0.93	0.90	0.89	**0.91**
TVRPCA[12]	0.79	0.52	0.46	0.08	0.65	0.83	**0.56**	0.84	0.84	0.83	0.78	0.83
LDB (ours)	0.92	0.92	0.79	0.14	0.93	0.95	**0.78**	0.90	0.90	0.93	0.89	**0.91**

Table 2
Ablations on I2R, dynamicBackground(dB), badWeather(bW) of CDnet, and SABS dataset with average F-measure. Red: best, blue: the second best.

LM+L1	LT+L1	I2R	dB	bW	SABS
✓		0.33	0.49	0.20	**0.07**
	✓	**0.56**	**0.64**	**0.28**	0.04

(a) The effectiveness of low-rank tensor (LT) modeling compared with low-rank matrix (LM). The L1 sparse norm is used to constrain foregrounds for fair comparison.

B	F	C	O	I2R	dB	bW	SABS
✓				0.56	0.64	0.28	0.04
✓	✓			0.84	0.52	0.91	0.36
✓	✓	✓		0.83	0.76	0.90	0.71
✓	✓	✓	✓	**0.88**	**0.78**	**0.91**	**0.71**

(b) Ablation study of each component, *i.e.*, Background(B), Foreground(F), Clutter(C), and Objectness(O) in our model.

	I2R	dB	bW	SABS
Objectness	0.88	0.68	0.79	0.66
LDB(ours)	**0.88**	**0.78**	**0.91**	**0.71**

(c) Comparison of our model with pure semantic objectness modeling. The results imply that our LDB is better than pure objectness detection in all conditions.

Table 7 and Fig. 9. We observe that our approach achieves better results on all sequences in the dynamicBackground and badWeather categories. From Fig. 9, it is seen that UMOD fails on many sequences, especially on low illumination (*e.g.*, blizzard), and dynamic scenes (*e.g.*, canoe). The reasons may be that: (1) Optical flow cannot work well for these sequences due to the data distribution gap between these scenes and pre-trained datasets, thus the severe optical flow results limit the performance of UMOD. (2) The dynamic clutter in the background breaks the assumption of UMOD that motion is only highly relevant with the moving objects, so that UMOD cannot handle dynamic background, which is the typical issue we address. In contrast, our method is built on temporal low-rank property which exists on arbitrary sequences, and learns dynamic background only based on the video sequence itself. That means that our approach does not rely on any pre-trained datasets, causing better generalization ability for various scenes in the real world.

5. Discussions and limitations

Our work follows the research line of low-rank based methods, which have a long history and remain reasonability on the moving object detection task. With the low-rank assumption, these methods usually achieve satisfactory results and have a good generalization ability. However, they still suffer from several shortages compared with other categories of approaches, *e.g.*, deep learning based algorithms. (1) They model the similarity of adjacent frames in sequences, which is not suitable in a few cases, like abandon illumination changes by turning on/off the light. While deep learning methods can address them by data augmentation strategies. (2) These methods usually consider multiple frames simultaneously to model the low-rank property, resulting in more computational costs than single image processing in the deep learning based algorithm. Despite the disadvantages, they still contribute to the computer vision community for two reasons. (1) There is a comprehensive theory of low-rank tensor framework, and moving object detection task is a typical application for this theory, obtaining

Table 3
Performance evaluation on I2R [58] and SABS [57] dataset using F-measure. From left to right: *Campus*, *Fountain*, *WaterSurface*, the average of I2R, and *NoCamouflage* sequence in SABS. Red: best, blue: the second best. LDB outperforms state-of-the-art methods on I2R, and significantly boosts on the SABS dataset.

	Camp.	Foun.	Wate.	**I2R-A.**	**NoCamou.**
RPCA[9]	0.21	0.39	0.38	0.33	0.07
OMoGMF[22]	0.20	0.47	0.79	0.49	0.08
LSD[11]	0.32	0.54	0.73	0.53	0.26
GOSUS[30]	0.15	0.00	0.66	0.27	0.01
DECOLOR[14]	0.81	0.86	0.88	**0.85**	0.19
TVRPCA[12]	0.71	0.78	0.86	0.78	**0.38**
LDB (ours)	0.85	0.85	0.94	**0.88**	**0.71**

Table 4
Generalization analysis of proposed method on noisy and rainy sequences, where σ is the noise level of added Gaussian noise. Red: best, blue: the second best.

	RPCA	OMoGMF	LSD	DECOLOR	TVRPCA	LDB (ours)
Ori	0.70	0.82	0.77	0.93	0.90	0.98
σ=15	0.15	0.83	0.65	0.98	0.90	0.98
σ=25	0.08	0.45	0.46	0.88	0.89	0.98
σ=50	0.05	0.05	0.29	0.96	0.86	0.92
Rain	0.36	0.83	0.74	0.98	0.89	0.98

Table 5
Evaluation experiments on other categories of CDnet. The metric is F-measure. Despite that our approach is specific designed for the dynamic clutter, we still have a good ability for handling various challenging conditions due to flexible background and foreground modeling, thus resulting in advanced performance. Red: best, blue: the second best.

	base.	came.	inte.	shad.	ther.	turb.	nigh.
RPCA	0.36	0.48	0.27	0.35	0.24	0.46	0.34
OMoGMF	0.58	0.36	0.36	0.20	0.20	0.49	0.26
LSD	0.52	0.42	0.29	**0.46**	0.31	0.50	0.36
GOSUS	0.07	0.07	0.10	0.02	0.03	0.06	0.06
DECOLOR	0.65	0.61	0.43	0.35	0.57	0.54	0.38
TVRPCA	**0.81**	0.77	0.27	0.36	0.52	0.22	0.36
LDB(ours)	0.79	**0.77**	**0.45**	0.38	**0.57**	**0.59**	**0.38**

Table 6
Comparison experiments with deep learning methods on CDnet dataset.

	FS	WS	Seen	Unseen	w/ Seg.	dyna.	badw.	base.	came.	inte.	shad.	ther.	turb.	nigh.
FgSegNet v2 [42]	✓		✓			0.99	0.99	0.99	0.99	0.83	0.82	0.41	0.74	0.81
FgSegNet v2 [42]	✓			✓		0.28	0.16	0.41	0.28	0.37	0.06	0.00	0.00	0.03
BSUV-Net [44]	✓		✓	✓	✓	0.87	0.79	0.96	0.85	0.69	0.66	0.48	0.57	0.44
BSUV-Net [44]	✓		✓	✓		0.34	0.18	0.34	0.59	0.14	0.27	0.04	0.49	0.38
LDB(ours)		✓		✓		0.76	0.90	0.79	0.77	0.45	0.38	0.57	0.59	0.38

FS means Fully supervised, WS indicates Weakly supervised. FgSegNet v2 achieves the best results on seen videos (row 1), but has a poor generalization ability on unseen videos (row 2). BSUV-Net highly relies on segmentation results as inputs (row 3 v.s. row 4). Our approach achieves the best results on unseen videos without segmentation prior (as shown on gray background), meanwhile avoids high annotation burdens due to the weakly supervised manner.

Table 7
Comparison experiments between unsupervised deep learning based method UMOD [62] and our approach on dynamicBackground and badWeather categories in CDNet dataset using F1 Measure. It is seen that our LDB achieves better results on all sequences.

Video	dynamicBackground							badWeather				
	boats	canoe	fall	foun.01	foun.02	over.	**Avg**	blizz.	skat.	snowF.	wetS.	**Avg**
UMOD [62]	0.26	0.23	0.48	0.03	0.24	0.69	0.32	0.06	0.57	0.28	0.47	0.35
LDB (ours)	0.92	0.92	0.79	0.14	0.93	0.95	0.78	0.90	0.90	0.93	0.89	0.91

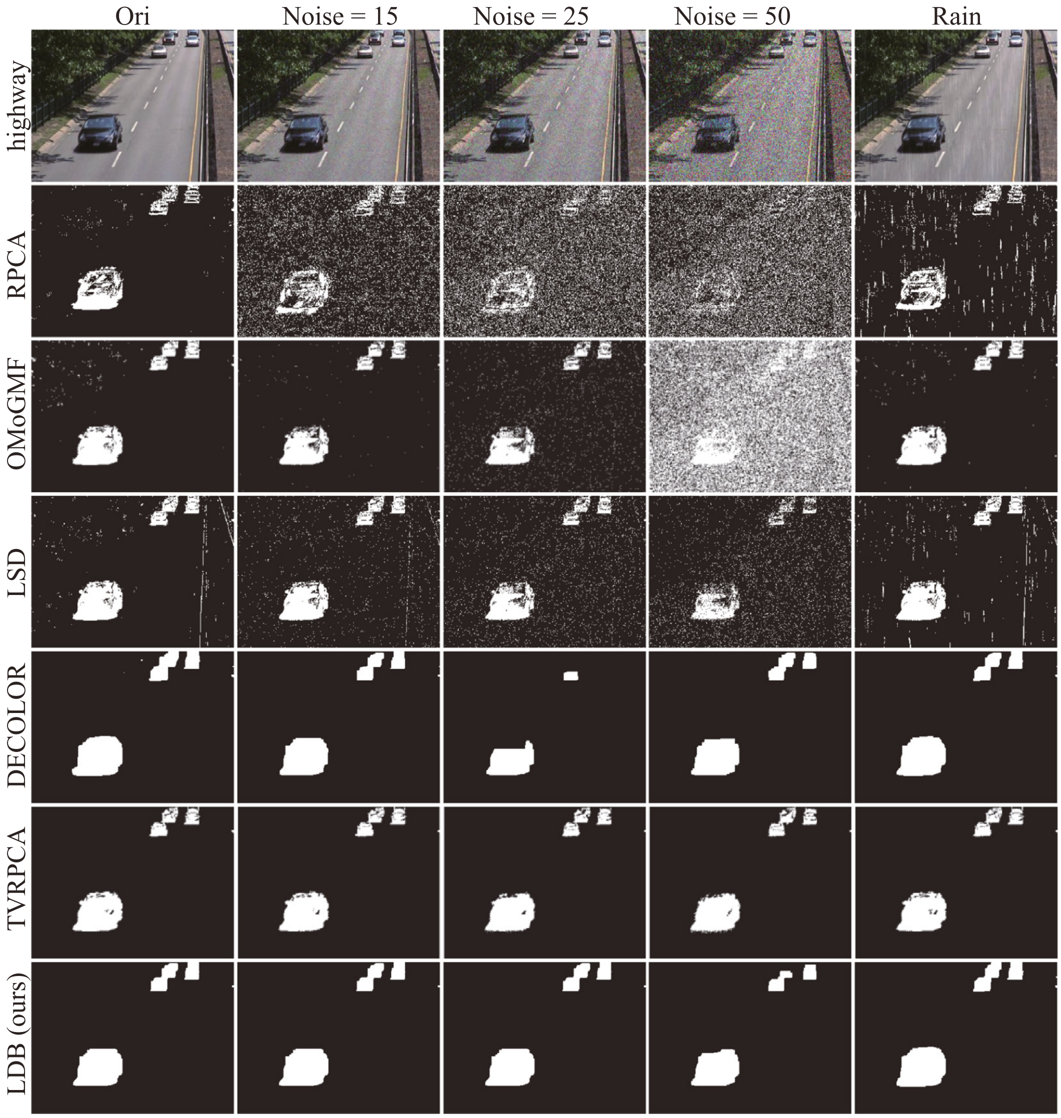

Fig. 8. Generalization experimental results of LDB and other state-of-the-art methods on noisy and rainy sequences, where the noise levels are (15, 25, 50).

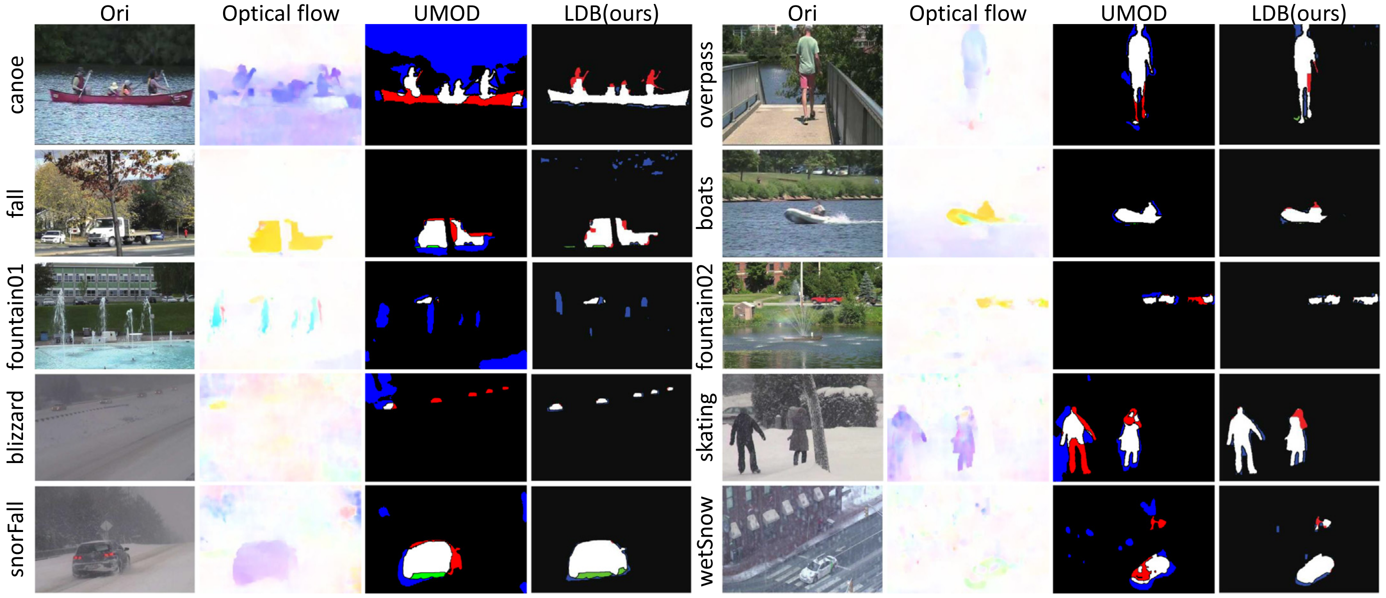

Fig. 9. Qualitative comparable results between unsupervised deep learning based method UMOD [62] and our approach on CDnet dataset. From left to right: original images, optical flow computed by PWC [63], mask prediction by UMOD, and our results. Notice that the optical flow prediction fails on dynamic background (*e.g.*, canoe, fountain01) and bad weather (*e.g.*, blizzard, wetSnow), resulting in worse moving object detection by UMOD, in contrast to that our approach (LDB) shows more stable results on these sequences.

satisfactory results. (2) As generative methods, they avoid the inductive bias in learning based methods, and could apply on any scenes without additional training process, achieving better generalization ability. Our work improves the low-rank based methods on dynamic scenes, thus further extending their flexibility and availability on various scenes.

6. Conclusion

In this paper, we propose a novel and unified model to address moving object detection from the dynamic background. In this model, a video is decomposed into a unidirectional low-rank static background, a sparse and smooth moving foreground, as well as a CNN represented dynamic clutter. The dynamic clutter is thoroughly represented by a two-branch patch-based neural network in a weakly supervised manner. Embedded with the data-driven discriminative prior of dynamic clutter, the unified model can properly decouple the variations of background with moving foreground. Experiment results show that the proposed method outperforms the state-of-the-art methods both qualitatively and quantitatively in the challenging datasets.

CRediT authorship contribution statement

Zhijun Zhang: Methodology, Validation, Writing – original draft. **Yi Chang:** Conceptualization, Writing – review & editing. **Sheng Zhong:** Supervision. **Luxin Yan:** Supervision. **Xu Zou:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) grant 62176100, and the Central Guidance on Local Science and Technology Development Fund of Hubei Province grant 2021BEE056.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imavis.2022.104425.

References

[1] T. Bouwmans, E.H. Zahzah, Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance, Comput. Vis. Image Underst. 122 (2014) 22–34.

[2] A. Sobral, A. Vacavant, A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos, Comput. Vis. Image Underst. 122 (2014) 4–21.

[3] T. Bouwmans, S. Javed, M. Sultana, S.K. Jung, Deep neural network concepts for background subtraction:a systematic review and comparative evaluation, Neural Netw. 117 (2019) 8–66.

[4] A.M. Maryam Sultana, S.K. Jung, Unsupervised moving object detection in complex scenes using adversarial regularizations, IEEE Trans. Multimed. 23 (2020) (1–1).

[5] Y. Tian, A. Senior, M. Lu, Robust and efficient foreground analysis in complex surveillance videos, Mach. Vis. Appl. 23 (5) (2012) 967–983.

[6] B.-H. Chen, S.-C. Huang, An advanced moving object detection algorithm for automatic traffic monitoring in real-world limited bandwidth networks, IEEE Trans. Multimed. 16 (3) (2014) 837–847.

[7] W. Hu, Y. Yang, W. Zhang, Y. Xie, Moving object detection using tensor-based low-rank and saliently fused-sparse decomposition, IEEE Trans. Image Process. 26 (2) (2017) 724–737.

[8] S. Javed, A. Mahmood, T. Bouwmans, S.K. Jung, Background–foreground modeling based on spatiotemporal sparse subspace clustering, IEEE Trans. Image Process. 26 (12) (2017) 5840–5854.

[9] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 1–37.

[10] T. Zhou, D. Tao, Godec: Randomized Low-Rank & Sparse Matrix Decomposition in Noisy Case, ICML, 2011 33–40.

[11] X. Liu, G. Zhao, J. Yao, C. Qi, Background subtraction based on low-rank and structured sparse decomposition, IEEE Trans. Image Process. 24 (8) (2015) 2502–2514.

[12] X. Cao, L. Yang, X. Guo, Total variation regularized rpca for irregularly moving object detection under dynamic background, IEEE Trans. Cybern. 46 (4) (2016) 1014–1027.

[13] B. Xin, Y. Tian, Y. Wang, W. Gao, Background subtraction via generalized fused lasso foreground modeling, CVPR (2015) 4676–4684.

[14] X. Zhou, C. Yang, W. Yu, Moving object detection by detecting contiguous outliers in the low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 597–610.

[15] A.J. Tom, S.N. George, Simultaneous reconstruction and moving object detection from compressive sampled surveillance videos, IEEE Trans. Image Process. 29 (2020) 7590–7602.

[16] Z. Hu, Y. Wang, R. Su, H.W. Xinxin Bian, G. He, Moving object detection based on non-convex rpca with segmentation constraint, IEEE Access 8 (2020) 41026–41036.

[17] S.E. Ebadi, E. Izquierdo, Foreground segmentation with tree-structured sparse rpca, IEEE Trans. Pattern Anal. Mach. Intell. 40 (9) (2018) 2273–2280.

[18] Z. Gao, L.-F. Cheong, Y. Wang Xiang, Block-sparse rpca for salient motion detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (10) (2014) 1975–1987.

[19] C. Guyon, T. Bouwmans, E. Zahzah Hadi, Foreground detection based on low-rank and block-sparse matrix decomposition, ICIP (2012) 1225–1228.

[20] D. Meng, F. De La Torre, Robust Matrix Factorization with Unknown Noise, ICCV, 2013 1337–1344.

[21] Q. Zhao, D. Meng, Z. Xu, W. Zuo, L. Zhang, Robust Principal Component Analysis with Complex Noise, ICML, 2014 55–63.

[22] H. Yong, D. Meng, W. Zuo, L. Zhang, Robust online matrix factorization for dynamic background subtraction, IEEE Trans. Pattern Anal. Mach. Intell. 40 (7) (2018) 1726–1740.

[23] M.E. Kilmer, K. Braman, N. Hao, R.C. Hoover, Third-order tensors as operators on matrices: a theoretical and computational framework with applications in imaging, SIAM J. Matrix Anal. Appl. 34 (1) (2013) 148–172.

[24] Z. Zhang, G. Ely, S. Aeron, N. Hao, M. Kilmer, Novel methods for multilinear data completion and de-noising based on tensor-svd, CVPR (2014) 3842–3849.

[25] X. Cui, J. Huang, S. Zhang, D.N. Metaxas, Background subtraction using low rank and group sparsity constraints, ECCV (2012) 612–625.

[26] J. Yang, W. Shi, H. Yue, K. Li, J. Ma, C. Hou, Spatiotemporally scalable matrix recovery for background modeling and moving object detection, Signal Process. 168 (2020), 107362.

[27] S. Javed, T. Bouwmans, S.K. Jung, Stochastic decomposition into low rank and sparse tensor for robust background subtraction, International Conference on Imaging for Crime Prevention and Detection 2015, pp. 1–6.

[28] M. Shakeri, H. Zhang, Moving object detection in time-lapse or motion trigger image sequences using low-rank and invariant sparse decomposition, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[29] M. Shakeri, H. Zhang, Corola: a sequential solution to moving object detection using low-rank approximation, Comput. Vis. Image Underst. 146 (2016) 27–39.

[30] J. Xu, V.K. Ithapu, L. Mukherjee, J.M. Rehg, V. Singh, Gosus: Grassmannian online subspace updates with structured-sparsity, ICCV (2013) 3376–3383.

[31] N. Wang, D.-Y. Yeung, Bayesian robust matrix factorization for image and video processing, ICCV (2013) (pp. 1798–1792).

[32] A. Sobral, S. Javed, S. Ki Jung, T. Bouwmans, E.-H. Zahzah, Online stochastic tensor decomposition for background subtraction in multispectral video sequences, ICCVW (2015) 946–953.

[33] Z. Zhang, S. Aeron, Exact tensor completion using t-svd, IEEE Trans. Signal Process. 65 (6) (2017) 1511–1526.

[34] Y. Mu, P. Wang, L. Lu, X. Zhang, L. Qi, Weighted tensor nuclear norm minimization for tensor completion using tensor-svd, Pattern Recogn. Lett. 130 (2020) 4–11.

[35] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, S. Yan, Tensor robust principal component analysis with a new tensor nuclear norm, IEEE Trans. Pattern Anal. Mach. Intell. 42 (4) (2020) 925–938.

[36] Y. Chang, L. Yan, S. Zhong, Hyper-laplacian regularized unidirectional low-rank tensor recovery for multispectral image denoising, CVPR (2017) 5901–5909.

[37] S. Lefkimmiatis, Universal denoising networks: a novel cnn architecture for image denoising, IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3204–3213.

[38] W. Yang, R.T. Tan, J. Feng, J. Liu, Z. Guo, S. Yan, Deep Joint Rain Detection and Removal from a Single Image, CVPR, 2017 1685–1694.

[39] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (5) (1989) 359–366.

[40] D. Zeng, M. Zhu, Background subtraction using multiscale fully convolutional network, IEEE Access 6 (2018) 16010–16021, https://doi.org/10.1109/ACCESS.2018.2817129.

[41] L.A. Lim, H.Y. Keles, Foreground segmentation using a convolutional neural network for multiscale feature encoding, Pattern Recogn. Lett. 112 (2018).

[42] L.A. Lim, H.Y. Keles, Learning multi-scale features for foreground segmentation, Pattern. Anal. Applic. 23 (2020).

[43] M. Bakkay, H. Rashwan, H. Salmane, L. Khoudour, D. Puig, Y. Ruichek, Bscgan: Deep background subtraction with conditional generative adversarial networks, 2018 25th IEEE International Conference on Image Processing (ICIP) 2018, pp. 4018–4022, https://doi.org/10.1109/ICIP.2018.8451603.

[44] P.I.M.O. Tezcan, J. Konrad, Bsuv-net:afully-convolutional neural network for background subtraction of unseen videos, WACV (2020) 2774–2783.

[45] M.O. Tezcan, P. Ishwar, J. Konrad, Bsuv-net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction, IEEE Access 9 (2021) 53849–53860, https://doi.org/10.1109/ACCESS.2021.3071163.

[46] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: residual learning of deep cnn for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.

[47] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.

[48] M. Fazel, Matrix Rank Minimization with Applications, Ph.D. thesis, PhD thesis Stanford University, 2002.

[49] S. Geman, D. Geman, Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, Readings in Computer Vision, Elsevier 1987, pp. 564–584.

[50] S.Z. Li, Markov Random Field Modeling in Image Analysis, Springer Science & Business Media, 2009.

[51] D. Zeng, X. Chen, M. Zhu, M. Goesele, A. Kuijper, Background subtraction with realtime semantic segmentation, IEEE Access 7 (2019) 153869–153884.

[52] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, ECCV (2018) 833–851.

[53] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html 2012.

[54] R. Mazumder, T. Hastie, R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, J. Mach. Learn. Res. 11 (8) (2010) 2287–2322.

[55] J.-F. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.

[56] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE International Conference on Computer Vision, 1999.

[57] S. Brutzer, B. Höferlin, G. Heidemann, Evaluation of background subtraction techniques for video surveillance, CVPR (2011) 1937–1944.

[58] L. Li, W. Huang, I.Y.-H. Gu, Q. Tian, Statistical modeling of complex backgrounds for foreground object detection, IEEE Trans. Image Process. 13 (11) (2004) 1459–1472.

[59] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, P. Ishwar, Cdnet 2014: An expanded change detection benchmark dataset, CVPRW (2014) 393–400.

[60] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic Differentiation in Pytorch, NeurIPS, 2017.

[61] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, CoRR, 2014 (arXiv preprint arXiv:1412.6980).

[62] Y. Yang, A. Loquercio, D. Scaramuzza, S. Soatto, Unsupervised moving object detection via contextual information separation, CVPR (2019) 879–888.

[63] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, CVPR (2018) 8934–8943.