Journal of Computing and Security

Research Article, 2025, Volume 12, Number 1 (pp. 65–83)

http://www.jcomsec.org

ORGRU: Online Robust Gated Recurrent Units for Real-Time Background Subtraction in Video Sequences

Arezoo Sedghi [a] Maryam Amoozegar [b] Esmat Rashedi [a,*] Fatemeh Afsari [c]

[a] *Faculty of Electrical and Computer Engineering, Graduate University of Advanced Technology, Kerman, Iran.*

[b] *Department of Computer and Information Technology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran.*

[c] *Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.*

ARTICLE INFO.

Article history:
Received: 11 December 2024
Revised: 04 July 2025
Accepted: 19 July 2025
Published Online: 18 August 2025

Keywords:
Background Subtraction, Deep Learning, Unsupervised Learning, Gated Recurrent Unit, Online Training, Robustness.

ABSTRACT

Background subtraction is a crucial task in computer vision that involves segmenting video frames into foreground and background regions. While deep learning techniques have shown promise in this field, existing approaches typically rely on supervised learning and have limited generalization capabilities for unseen video data. Moreover, many of these methods are not suitable for real-time applications due to their offline or partially online nature. This paper introduces ORGRU, an unsupervised, online, and robust deep learning-based framework for background subtraction. ORGRU uses robust Gated Recurrent Units (GRUs) to simultaneously estimate the background model (low-rank component) and detect foreground objects (sparse component) in a fully online processing framework. The model is iteratively updated in real time with an unsupervised learning algorithm utilizing only the current frame. To evaluate the effectiveness of the proposed approach, we conduct experiments on the LASIESTA dataset, which is a comprehensive, fully-labeled dataset for change detection covering various background subtraction challenges. The experimental results provide both qualitative and quantitative assessments, demonstrating the robustness and superiority of the proposed approach compared to the state-of-the-art methods.

1 Introduction

The task of background subtraction (BGS) serves as a fundamental binary classification process that plays an essential role in extracting foregrounds from backgrounds by detecting changes within a given scene [1, 2]. In computer vision, background subtraction is a crucial operation frequently employed as a preliminary step for numerous advanced tasks [3]. This task holds significant importance in various domains, including anomaly detection [4], human activity analysis [5], people re-identification [6], visual object detection and tracking [7], video content analysis [8], traffic monitoring [9], action recognition [10], the interaction between humans and machines [11], security and surveillance missions [12–14] and angiography [15].

Background subtraction algorithms encounter numerous challenges in real-world scenarios, such as illumination changes, shadows, background fluctuations, weather variations, camera jitter, occlusions, and boot-

* Corresponding author.
Email address: e.rashedi@kgut.ac.ir (E. Rashedi)
https://dx.doi.org/10.22108/jcs.2025.143648.1158
ISSN: 2322-4460

strapping [16, 17]. To overcome these challenges, it is vital to design a model that can accurately maintain a correct representation of the background while adapting to changes, as well as efficiently separate the foreground from the background [18].

Subspace learning methods have gained significant attention due to their robust theoretical foundation and favorable practical performance in unsupervised scenarios [11, 14, 19–21]. However, these methods often face limitations when dealing with complex high-dimensional data, as they may not efficiently capture the nonlinearities present in data [22, 23].

Recently, deep neural network-based approaches have gained increasing popularity in the field of background subtraction [24] due to their ability to effectively analyze large amounts of data and capture nonlinear relationships in the data [14]. These approaches can be categorized into unsupervised learning and supervised learning [13]. Additionally, there has been a growing interest in self-supervised learning methods [25], weakly supervised learning methods [26] and methods that combine different types of learning strategies to reduce the training costs. For instance, recent work proposes BgSubNet, which leverages a virtually limitless freeform dataset generation strategy and introduces the first self-supervised pre-training strategy for deep background subtraction (BgS). This framework extends to a semi-supervised approach, enabling inference on unseen scenarios through self-supervision [27].

However, it is worth noting that the majority of the existing approaches in the literature are supervised [1, 14]. The supervised methods heavily rely on a significant amount of labeled ground truth data, which is manually created by human experts [14]. Despite their effectiveness, these approaches face challenges such as high training costs [2], overfitting [28, 29] and limited generalization capability when dealing with completely unseen videos [13].

From the second perspective, it is important to highlight that most of the presented background subtraction methods operate offline, processing data in multiple epochs [28–34], which hinders their suitability for real-time applications [2]. Contrarily, in online background subtraction algorithms, processing input data, and extracting the sparse component incrementally would be more practical rather than in batches. Additionally, on-the-fly processing without a priori knowledge of future data is more efficient. This attribute makes online learning highly suitable for real-time applications, as it enables immediate adaptation to background changes and effectively handles challenging videos [18]. By continuously processing frame as it arrives, online learning algorithms offer efficient and dynamic background subtraction capabilities [35].

In the context of online object detection, some traditional methods[36, 37] initially use batches of data for setting up and later transition to online training mode[36]. However, it is important to note that most deep learning-based methods typically require access to either the entire or a partial set of training data for the learning process[38, 39]. For example, an online framework for illumination invariant moving object detection, as described in[40], incorporates batch initialization. Similarly, online anomaly detection in time series is addressed in[41], where the input data is processed within a specific time window. Additionally, motion detection methods that utilize a pair of auto-encoder networks[42] and dynamic background learning approaches[43] have been proposed, all of which employ a partially online approach and do not process frames individually in real-time.

From the third perspective, addressing the challenges in background subtraction requires a robust strategy for maintaining the background model that can effectively handle background changes and outliers. In[44], a robust deep auto-encoder approach is proposed, drawing inspiration from subspace learning-based methods. In this method, the foreground is treated as a sparse section or outlier, which is separated from the low-rank or background component constructed by the network. However, it is important to note that this method is also offline and involves multiple epochs of processing.

In summary, this paper addresses several key gaps in the background subtraction literature. Unlike many existing methods that rely on labeled data or offline batch processing, the proposed ORGRU framework operates fully online and in an unsupervised manner. It effectively captures temporal dependencies in video sequences while enabling adaptive background modeling. The proposed ORGRU framework updates the background model frame-by-frame without relying on future frames, making it highly suitable for real-time applications. By explicitly separating low-rank and sparse components during learning, ORGRU enhances robustness against outliers and sudden scene changes. Despite its lightweight architecture, ORGRU achieves significant performance improvements over strong baselines while requiring minimal hyperparameter tuning—making it a practical and efficient solution for real-world deployment. This online processing capability ensures efficient memory and computation usage and enables more adaptation to background changes. Additionally, the algorithm continuously tracks and updates the background model, leading to improved accuracy in foreground separation.

The structure of the paper is summarized as follows.

First, Section 2 briefly reviews related works. In Section 3, the proposed method is illustrated. In Section 4, the database, evaluation metrics, and experimental results are described and the proposed model is compared with state-of-the-art methods. Finally, the main conclusions are summarized, and future research is recommended in Section 5.

2 Related Works

Existing approaches for background subtraction and object detection can be categorized into traditional and deep learning-based methods [24]. The focus of the discussion is on deep learning-based algorithms, including both supervised and unsupervised learning techniques.

2.1 Traditional Approaches

To solve the background subtraction problem, several traditional methods have been proposed, including parametric and non-parametric methods and matrix decomposition-based methods. A parametric method based on the mixture of Gaussians models (MOG) is presented in [45, 46] and [47] in which MOG is enhanced by selecting variable parameters, mixing Gaussians spatially, and implementing fast initialization. The non-parametric methods are primarily inspired by kernel density estimation [48, 49] and consensus-based estimation [50, 51]. In 2011, a universal background subtraction algorithm was presented [52] that contains three significant background model maintenance policies: a background sample replacement policy to represent short and long-term history, a memoryless update policy, and spatial diffusion through background sample propagation. Other well-known methods are a variety of robust principal component analysis (RPCA) based models presented to reconstruct the background as a low-rank component and foreground as a sparse matrix [21, 53, 54].

A sparse coding-based background modeling technique was introduced by Stagliano et al. in 2015 [55]. The method uses space-variant analysis, in which each image patch is learned by a dictionary of atoms whose size is determined by the background variability. Recently, to process backgrounds containing moving textures, the RPCA framework is used [56]. These approaches generally work offline and deal with data in batches, making them unsuitable for real-time applications. Online RPCA methods have also been proposed to solve this problem [57]. However, the optimization method and high computational complexity are still challenging. Also, these methods cannot handle large data sets and long videos due to the memory limitations.

2.2 Deep Learning-Based Approaches

In the past years, deep learning-based methods have been demonstrated to be capable of extracting high-level abstractions from large datasets. Supervised and unsupervised techniques are described in the deep learning-based literature. Model parameters are tuned through the use of labeled data in supervised approaches. Unsupervised techniques, however, do not require ground truth to optimize the model.

2.2.1 Supervised Learning

Braham and Van Droogenbroeck [58] presented the first background subtraction algorithm based on convolutional neural networks (ConvNets) that is built specifically for a particular scene. For each frame in a video sequence, image patches that are centered on each pixel are extracted and then combined with the corresponding patches from the background model. Then, ConvNet performs a subtraction operation to compute the foreground probability of the pixel.

In 2018, Babaee et al. [38] proposed a deep CNN-based method for background subtraction, which consists of three processing steps, including the background model generation, CNN for feature learning, and a spatial median filter post-processing module. In their study, authors generated background models using handcrafted approaches, including SuBSENSE [59] and flux tensors [60]. More precisely, Babaee et al. proposed a combination of the segmentation mask created by the SuBSENSE algorithm and the output of the Flux Tensor algorithm, which dynamically adjusts the parameters used in the background model in response to the motion changes.

Zeng et al. [31] proposed a multiscale fully convolutional network (MFCN) architecture that takes advantage of different layer features for background subtraction and has no need to extract the background images. The input of the neural network is the RGB frame from different sequences and the output is a probability map.

In 2019, Bian et al. [61] presented a moving object detection method in complex scenes using ResNet-18 to achieve pixel-level classification. This method is based on supervised training and networks with an encoder-decoder structure.

Recently, a deep neural network employing an encoder-decoder structure based on VGG-16 and ResNet-50 was introduced to enhance hierarchical feature extraction and use a transposed convolutional network for foreground and background classification [62].

2.2.2 Unsupervised Learning

Although the previously mentioned deep learning-based algorithms perform better than the traditional ones, they are all supervised. For online video surveillance, unsupervised background subtraction methods are preferred [63].

In 2018, Bakkay et al. [64] introduced a background subtraction method based on a conditional Generative Adversarial Network (cGAN). For the foreground mask, the generator learns the mapping from the background and the current image. The discriminator then learns a loss function to train this mapping by comparing the ground truth and the predicted output by considering the input image and the background. Deep Context Prediction (DCP) was designed by Sultana et al. [65] for estimating background and segmenting foreground using a combination of GAN and image inpainting. Based on context prediction, DCP is an unsupervised hybrid GAN for visual feature learning. In a compact network architecture, Rahmon et al. [32] developed Motion U-Net, a deep convolutional neural network, into a multi-cue auto-encoder deep architecture that integrates motion and changes cues with appearance cues for robust moving object detection performance. The current frame, the frame with change clues, and the frame with flux motion clues are inputs to the network. By using unsupervised methods, motion and change clues can be learned. In 2019, Rezaei et al. [33] presented an unsupervised Deep Probabilistic Background Model (DeepPBM) based on a framework of Variational Auto-Encoders (VAEs) for moving object detection [66, 67]. In DeepPBM, a generative low-dimensional model of the background is estimated, and the task of background subtraction is performed by thresholding the differences between the output of the model and the original input frame. The background is a low-dimensional subspace represented by latent variables embedded in a non-linear mapping of video frames that follows a Gaussian distribution.

3 Proposed Method

This section introduces the proposed deep GRU-based background subtraction approach, which makes GRU cells robust by utilizing a low-rank (L) and sparse (S) framework. The low-rank matrix models the background, and the sparse matrix represents the foreground. The proposed method is designed to be robust against outlier data. Also, this is an online learning method that updates the parameters of the model incrementally and instantly. In order to capture the low-rank data, the training can continue as long as the task is continued and adapt to the changes. To the best of our knowledge, the proposed model is the first unsupervised, online, robust, deep recurrent neural network that processes one frame per time step and updates the weights of the GRU-based network accordingly when a new frame arrives. The capability of processing long-term videos and not relying on the entire video sequence make the proposed model desirable for real applications and long-term videos. The following subsections provide details about the network architecture.

3.1 Network Architecture

In video processing tasks, sequential data is processed, and each frame has information similar to the previous one. The proposed online and robust GRU-based method, which we hereafter we call ORGRU, considers this property of data and shows how to process new data without removing the effect of the previous one, by using gated recurrent units. The network contains two GRU layers and one linear layer. ORGRU using recurrent layers is able to learn sequential patterns, reconstruct low-rank data, and separate a video frame into two scenes. In this approach, the least complex network architecture is used, and the advantages of recurrent neural networks are utilized for sequential data processing without the need for labeled data. In real-time applications, each incoming frame has an impact on the training and outputs of the network.

The overview of the proposed method is depicted in Figure 1. As shown in Figure 1, at each time step, a video frame with size $W \times H$ is entered into the model as a $WH \times 1$ one-dimensional signal and is separated into a low-rank matrix L and a sparse matrix S based on the current model of the background. A fully connected layer with size WH comes last. Then, the current background model is updated and trained based on the newly received frame.

3.2 Online Robust GRU Method

Consider the video sequence

$$\mathbf{X}_S = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t, \ldots, \mathbf{X}_T\}, \tag{1}$$

where $\mathbf{X}_t$ is the incoming frame at time step t and T is the length of the video. The RGB video frames are changed into the gray scale ones while maintaining the original size $W \times H$. $\mathbf{X}$ (to simplify the notation, we omit the subscript t) can be expressed as shown in Eq. (2), when referring to the reconstruction of low-rank and sparse extraction identification in large data matrices [44]:

$$\mathbf{X} = \mathbf{L} + \mathbf{S}, \tag{2}$$

The robustness of the model against sparsity is guaranteed by Eq. (2). The GRU net can estimate the low-rank component, but the sparse component

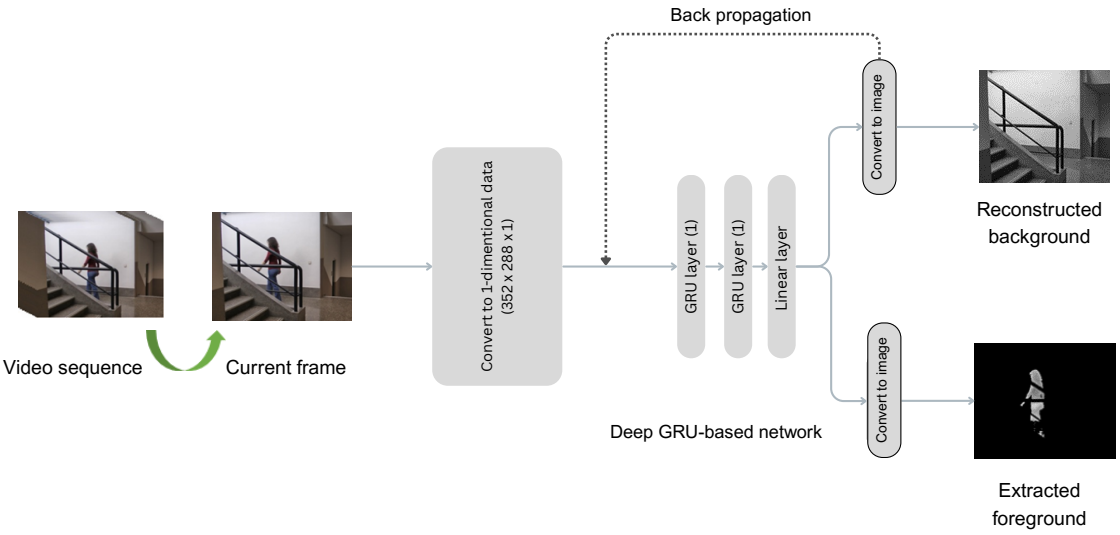


Figure 1. The Process of Entering Frames into the Network and Decomposing Them into the Background and Foreground Components.

has outliers that cannot be represented by GRU. By removing outliers from $\mathbf{X}$, GRU net can estimate and maintain the low-rank data more accurately. The main goal of a robust model is to minimize the following non-convex optimization problem [44]:

$$\begin{aligned} \min \quad & \|\mathbf{L} - \hat{\mathbf{L}}\|_2 + \lambda \|\mathbf{S}\|_1 \\ \text{s.t.} \quad & \mathbf{X} - \mathbf{L} - \mathbf{S} = 0 \end{aligned} \tag{3}$$

Where $\mathbf{L}$ is the current estimated low-rank component, $\hat{\mathbf{L}}$ is the reconstructed low-rank component, $\mathbf{S}$ represents the sparsity matrix, and $\| * ||_i$ is l_i-norm. After estimating the low-rank component by the GRU, the matrix $\mathbf{S}$ can be extracted. Therefore, by fixing all variables except $\mathbf{S}$ in Eq. (3), the above non-convex minimization problem is changed to an l_1-norm minimization. So, to calculate $\mathbf{S}$, proximal approaches [68–70] and a new formulation incorporating correlated and adaptive sparsity is recommended. Optimization problems can also be phrased as proximal problems by using proximal mapping, which is also known as shrinkage functions and is widely used in l_1-norm optimization. It is particularly similar to soft-thresholding functions for the proximal operator as Eq. (4) for optimizing $\mathbf{S}$ [69]:

$$\text{Proximal operator}\,(S_i) = \begin{cases} s_i - \lambda & s_i > \lambda \\ s_i + \lambda & s_i < -\lambda \\ 0 & s_i \in [-\lambda, \lambda] \end{cases} \tag{4}$$

Where s_i is the i-th element of the matrix $\mathbf{S}$ and λ is the balancing parameter that controls the degree of sparsity of the sparse matrix $\mathbf{S}$ [71], which plays an essential role in the ORGRU method. A small λ encourages much of the data to belong to $\mathbf{S}$ as sparsity. In particular, λ should be tuned considering the desired task and the trade-off between low-rank and sparse components. Moreover, outlier data cannot be included in $\mathbf{S}$ when a large value is applied for λ [44], and Eq. (3) cannot be solved computationally. More details are explained in Section 4.2.

As mentioned above, after removing sparsity from input data $\mathbf{X}$, GRU_net can reconstruct the low-rank data precisely. So, $\hat{\mathbf{L}}$ is obtained and the ORGRU approach can be seen as minimizing the reconstruction error using back-propagation, a typical deep neural network training algorithm, and describes what should be done for optimization when $\mathbf{S}$ is constrained. The objective function to solve the minimization problem in Eq. (3) is defined as Eq. (5):

$$\mathcal{L} = \sqrt{\frac{\sum_{i=1}^{W \times H} |l_i - \hat{l}_i|^2}{W \times H}} \tag{5}$$

Where $\mathcal{L}$ denotes the root mean square error

(RMSE) measuring the error of a model in reconstructing the data, $W \times H$ is the size of $\mathbf{X}$, l_i is the i-th element in the low-rank matrix, and $\hat{l}_i$ is its corresponding reconstruction element, respectively. The RMSE loss function is used to train the GRU_net in the ORGRU algorithm by measuring the difference between the low-rank and the reconstructed low-rank in each stage. Based on the RMSE loss function, it can be gauged how closely the real low-rank component matches its reconstructed low-rank component.

In order to achieve a real-time background subtraction model, each turn involves updating the weights of the ORGRU based on the incoming frame and the reconstructed background scene, and extracting the foreground scene. The proposed algorithm based on the mentioned explanation is presented in the following algorithm.

Algorithm 1 Online Robust GRU (ORGRU).

Input: Video frame:
$\mathbf{X_t} \in \mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_t}, \ldots, \mathbf{X_T}\}$
Output: Background and foreground of $\mathbf{X_t}$
Initialization:
Low-rank matrix is set to zero ($\mathbf{L} = 0$).
Initialize the GRU_net parameters randomly.
Online Robust GRU (ORGRU) Execution:
Performing real-time analysis on video frames
For each frame:

(1) Estimating **L:** Estimate the low-rank matrix $\mathbf{L}$ using GRUs:

$$\mathbf{L} = \text{GRU\_net}(\mathbf{X_t}, \text{train} = \text{False})$$

(2) **Calculating S by the proximal operator:** Calculate the sparse matrix $\mathbf{S}$ by subtracting $\mathbf{L}$ from $\mathbf{X_t}$ and apply Eq. (3) to enforce sparsity:

$$\mathbf{S} = \mathbf{X_t} - \mathbf{L}$$

$$\mathbf{S} = \text{Proximal operator}(\mathbf{S})$$

(3) **Training:** Update the low-rank matrix $\mathbf{L}$ using the difference between $\mathbf{X_t}$ and $\mathbf{S}$ and train the GRUs with the updated $\mathbf{L}$ by minimizing Eq. (5):

$$\mathbf{L} = \mathbf{X_t} - \mathbf{S}$$

$$\hat{\mathbf{L}} = \text{GRU\_net}(\mathbf{L}, \text{train} = \text{True})$$

End for

4 Experimental Results

The proposed algorithm has been evaluated on a comprehensive and challenging dataset to demonstrate its generalizability and the results are compared with DeepPBM, a state-of-the-art method. The code shared in [72] is used to implement the DeepPBM model. Furthermore, the performance of the non-robust mode of the model (OGRU) [73] is evaluated to analyze the robustness effect on the proposed method in various challenges. The quantitative and qualitative results of all three methods are presented in the following subsections.

The proposed deep GRU-based model is implemented using Python and the Pytorch framework with two gated recurrent unit layers, and a hidden_size = 40 and a linear layer with 101376 neurons. For minimizing the loss function through backpropagation, the optimization is performed using the R-prop [74] optimizer with a learning rate of 0.02.

4.1 Dataset Description

In [16], the LASIESTA (Lagged and Annotated Sequences for Integrated Evaluation of Segmentation Algorithms) dataset is presented that contains 18425 video frames of size 352×288 in 48 videos and 14 different categories, which are free and openly available. In LASIESTA, the creation of a comprehensive, compact, and well-structured database for analyzing moving object detection strategies is proposed. There are many real indoor and outdoor sequences in the database, organized into various categories. Each category covers a specific challenge (e.g. occlusions, bootstrapping, illumination changes, camouflage, modified background, etc.). This database is fully annotated at the pixel and object levels.

4.2 Parameter Setting

As discussed in section 3.2, λ plays a critical role in the proximal operator, as it controls the degree of sparsity in the extracted foreground component. In order to investigate the impact of different λ values on performance, experiments on different parameter values are conducted. We performed a visual analysis by varying λ across a range of values and selecting the one that gave the best separation between background and foreground. Figure 2 depicts the effect of changing λ on a frame. Even though λ impacts the sparsity level, a single fixed value of $\lambda = 55$ works robustly and consistently across all categories in the LASIESTA dataset. This means that once λ is tuned initially, it can be applied directly to different types of videos without any need for re-tuning for each scenario. This characteristic is an important advantage of ORGRU, especially compared to other methods such as DeepPBM, which require tuning parameters like the latent space dimension (d) individually for each new sequence. We will highlight this point to emphasize that ORGRU requires only minimal hyperparameter tuning, making it more practical and efficient for real-time use. In the TP, TN, FP, and

FN denote the true positive, true negative, proposed model, a single value of the regularizing parameter λ in Eq. (4) can be able to detect the foreground for all kinds of challenges (captured by static camera). It is robust to a broad variety of videos since it does not require changing model hyperparameter values. The constant value of 55 is applied for all challenges, unlike other methods [44, 75] that demand individual re-tuning for each challenge.

4.3 Generating Foreground Masks

While the proposed model is supposed to perform background subtraction based on change detection, its outputs include both background and foreground frames simultaneously. Therefore, it can be used for both background subtraction and foreground detection tasks. On the other hand, in order to evaluate quantitatively and compare the model with other background subtraction, change detection, and foreground detection methods, obtained foregrounds are mapped into binary masks. It is performed by a post-processing binarization module and then compared with the ground truth. As shown in Figure 3, post-processing is applied only to the obtained foreground scenes and does not interfere with backgrounds and network training.

4.4 Evaluation Metrics

To experiment with the method on the mentioned data set, a confusion matrix is constructed by comparing the binary masks with the ground truth for each foreground frame. Where false positive, and false negative pixel values, respectively, are defined in Table 1.

The evaluation process was performed using the framework presented for quantitative evaluation in [76], which measures F-Measure, recall, and precision as defined by Eqs. (6), (7), and (8), respectively, for the i-th frame. An effective model should achieve high recall while maintaining precision [77].

$$\text{F-Measure} = \frac{1}{n}\sum_{i=1}^{n}\frac{2\cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{6}$$

$$\text{Recall}_i = \frac{1}{2}\left(\frac{TP_i}{TP_i + FN_i} + \frac{TN_i}{TN_i + FP_i}\right) \tag{7}$$

$$\text{Precision}_i = \frac{1}{2}\left(\frac{TP_i}{TP_i + FP_i} + \frac{TN_i}{TN_i + FN_i}\right) \tag{8}$$

For further evaluations, IoU (Intersection over Union), DR (Detection Rate), and MCC (Matthews Correlation Coefficient) are also calculated based on their formal definitions as Eqs. (9) to (11), respectively. IoU quantifies the percentage of overlap between the ground truth and the predicted foreground, the detection rate gives the percentage of the corrected pixels to the total number of foreground pixels in the ground truth, and the purpose of the MCC metric is to assess the performance of binary classifiers when faced with unbalanced problems [59]. In Figure 4, the overall values of the mentioned metrics are shown for each method.

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{9}$$

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

$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)\cdot(TP + FN)\cdot(TN + FP)\cdot(TN + FN)}} \tag{11}$$

4.5 Quantitative Results

Despite using unsupervised learning, the evaluation process has started from the first frame and continued through the end of the video. In Table 2, average experimental results are shown for each mentioned method in videos presented by LASIESTA using a static camera. In each category, there are two videos, and the average metric value of them is presented as the performance of the model against the challenge. The overall value of metrics represents the averag of the metrics of all ten challenges. This process is done for all three methods. In DeepPBM, the dimension of the latent variables, d, is tuned, and the best performance is considered. Also, all input video frames are divided into batches of 140 frames for 200 epochs. As discussed above, the proposed method has just one parameter and does not need to be tuned for every sequence separately.

Table 2 shows well-known metrics for evaluating background subtraction methods. As a most common criterion, average F-Measures are typically used in related works. F-Measure is calculated by combining two other evaluation metrics, recall, and precision. Thus, the overall performance of a background subtraction algorithm is highly dependent on its F-Measure value and it can be appropriate for evaluating its generalizability in various situations. Therefore, this metric should be more noticeable in comparisons.

It can be seen from Table 2 that ORGRU achieves higher F-Measure values, demonstrating that robustness significantly improves accuracy under real-world

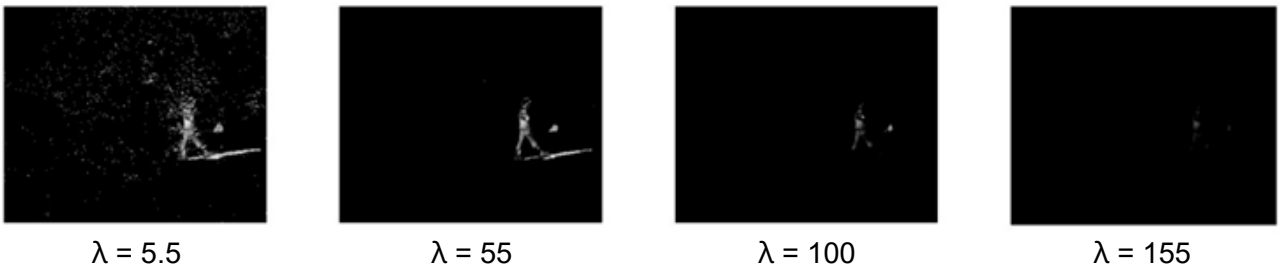


Figure 2. Effect of the Different Values for the Parameter λ.

Table 1. Criteria for Determining Parameters of the Confusion Matrix.

	Background (based on ground truth)	**Foreground (based on ground truth)**
Sparse (Foreground)	False Positive (FP)	True Positive (TP)
Low-rank (Background)	True Negative (TN)	False Negative (FN)

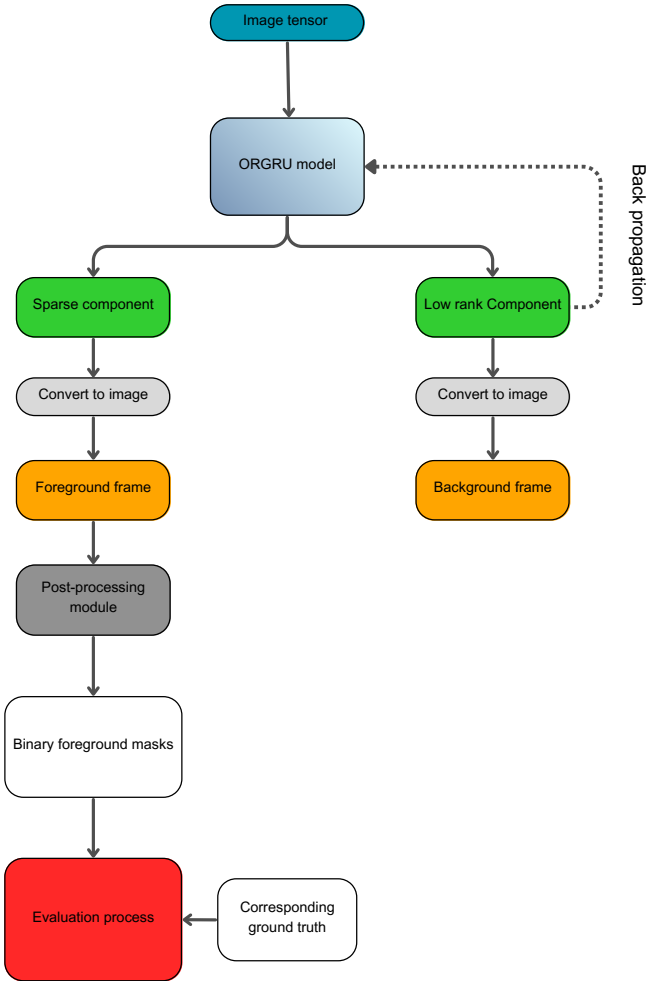


Figure 3. The Process of Producing Binary Foreground Masks and Evaluation.

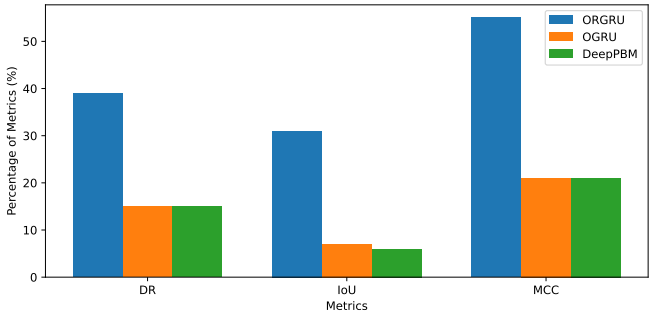


Figure 4. Percentage of the Overall Value of DR, IoU, and MCC Metrics.

challenges like illumination changes and occlusions. It is also important to note that the proposed method receives the frames one by one (it is not even window-based) and has the opportunity to decide on each frame only once.

Category-specific analysis reveals significant performance variations across challenging scenarios. ORGRU achieves maximum performance in the occlusion category due to its temporal modeling ability using GRU layers that effectively capture dependencies when foreground objects are temporarily occluded and reappear. Conversely, minimum performance occurs in snowy conditions because snowfall creates persistent noise from the first to the last frames. This challenge is amplified by the online nature of ORGRU, which processes frames sequentially without long-term background memory mechanisms needed to suppress such persistent dynamic noise. These results highlight that ORGRU performs best under structured motion conditions and faces limitations in highly dynamic, noisy environments.

The proposed method outperforms DeepPBM, which is the state-of-the-art among unsupervised methods, by 17% on the F-Measure. Based on their published code [72], DeepPBM requires 145M parameters while ORGRU uses only 16.3M parameters.

Table 2. The Quantitative Results of the Algorithms Evaluated with the LASIESTA Dataset.

Category	**Method**	**Recall**	**Precision**	**F-Measure**
Simple sequences	ORGRU	0.71	0.92	0.82
	OGRU	0.56	0.75	0.66
	DeepPBM	0.57	0.63	0.63
Camouflage	ORGRU	0.64	0.85	0.76
	OGRU	0.55	0.63	0.61
	DeepPBM	0.62	0.65	0.68
Occlusions	ORGRU	0.81	0.90	0.89
	OGRU	0.47	0.55	0.40
	DeepPBM	0.51	0.53	0.53
Illumination changes	ORGRU	0.70	0.85	0.79
	OGRU	0.52	0.55	0.55
	DeepPBM	0.48	0.52	0.50
Modified background	ORGRU	0.69	0.86	0.79
	OGRU	0.51	0.64	0.59
	DeepPBM	0.55	0.62	0.61
Bootstrap	ORGRU	0.59	0.83	0.69
	OGRU	0.54	0.66	0.63
	DeepPBM	0.55	0.60	0.60
Cloudy conditions	ORGRU	0.78	0.91	0.87
	OGRU	0.65	0.82	0.77
	DeepPBM	0.54	0.61	0.61
Rainy conditions	ORGRU	0.70	0.90	0.82
	OGRU	0.49	0.59	0.59
	DeepPBM	0.66	0.59	0.71
Snowy conditions	ORGRU	0.59	0.77	0.70
	OGRU	0.56	0.61	0.67
	DeepPBM	0.61	0.56	0.65
Sunny conditions	ORGRU	0.72	0.86	0.81
	OGRU	0.62	0.87	0.74
	DeepPBM	0.61	0.60	0.64
Average	ORGRU	0.69	0.87	0.79
	OGRU	0.55	0.67	0.62
	DeepPBM	0.57	0.59	0.62

Additionally, ORGRU uses a fixed parameter $\lambda = 55$ across all videos, while DeepPBM requires tuning its

latent dimension d based on video complexity and dynamics, making our approach more practical for deployment.

The overall F-measure value is also better than the non-robust mode of the proposed model by 17%. The role of λ and its effectiveness in making the algorithm more robust are determined.

In addition, the average F-measure comparison of the proposed ORGRU with existing state-of-the-art methods on the LASIESTA dataset is given in Table 3.

Although ORGRU has evaluated from the first frame of a video of the LASIESTA dataset, without considering the training phase, and receives frames one by one, the proposed method achieves an overall F-Measure of 0.79, which is the highest overall F-Measure. Table 3 demonstrates that the proposed method achieves the highest average performance on the LASIESTA database and it can handle unseen videos without needing ground truths. For a comprehensive evaluation, the proposed method is compared with new state-of-the-art supervised methods that are reported in [13, 85]. The results can be seen in Table 4.

The 5% gap between BSUVNet 2.0 [85] and 2% between ChangeDet [13] and 3DCD [86] methods, considering their supervised aspect, and using batches of data and corresponding ground truth for training should be noticed.

Addressing the time analysis concern, online ORGRU method processes each incoming frame ($H \times W$) with a computational complexity of $O(H \times W \times (m \times d + m^2))$, where d is the feature dimension and m is the hidden state dimension of the GRU. This per-frame processing occurs only once in an online manner.

In contrast, offline approaches typically involve processing the entire video dataset (with N frames) in batches across multiple training epochs (I). The complexity of these methods is significantly higher, often involving terms like $O(I \times (BN) \times f(H, W))$, where B is the batch size and $f(H, W)$ represents the computational cost per frame within a batch, which can be substantial due to operations across multiple frames or complex feature extraction repeated over the entire dataset in each epoch.

The inherent difference in processing paradigms—online, single-pass per frame versus offline, multi-pass over the entire dataset—highlights the efficiency of the ORGRU method. The architectural advantage of online processing for each frame only once offers a clear reduction in computational time compared to the iterative and batch-based nature of offline techniques.

4.6 Qualitative Results

The categories that have been studied include most of the challenges in the background subtraction task, such as bad weather conditions, shadows, illumination changes, bootstrap, dynamic backgrounds, camouflage, and occlusions. Table 5 provides numerical comparison results with the three mentioned methods.

Visually, it can be seen that the results of the ORGRU look much better than the corresponding baseline method, DeepPBM, which shows good agreement with the above quantitative evaluation results. A simple sequence has no camouflage, occlusion, lighting changes, continuous background changes, bootstrapping, or camera movement. Background and foreground separation appears well in visual results. Objects in the camouflage category have an appearance similar to some areas in the background. Furthermore, these moving objects remain static on such background regions for short periods of time. The weakest performance is in this category.

The short-term memory of the network and the online nature of the method may cause weakness in the efficiency of the proposed algorithm. This should be investigated and improved in the future. In the occlusion sequences, although the moving objects are removed from the scene for a while, ORGRU also detects them immediately after appearing again and shows good robustness to use in real time. Modified background sequences show situations in which some background elements are subtracted or abandoned.

ORGRU foreground detection results are very accurate and fast after adding or subtracting an object. The bootstrap challenge is formed by sequences containing moving objects from the first frame. After a few frames, the proposed method has learned the background and foreground well. Illumination change sequences contain sudden illumination changes, which ORGRU showed remarkable accuracy and performed correctly even after light was added to the scene.

In cloudy, rainy, snowy, and sunny conditions, which contain dynamic backgrounds due to the tree branches moving with the wind, snowing, raining, and hard shadows, the proposed method can extract foreground objects and is more robust to the background troubles. The ability of models to reconstruct background scenes is another aspect that can be evaluated. Since there is no ground truth for this work, quantitative evaluation is not possible and only visual experiments are shown in Figure 5.

Table 3. Category-wise comparison of the proposed method with the state-of-the-art methods in terms of F-measure on the LASIESTA dataset published in LASIESTA's paper.

Method	I_SI_01	I_SI_02	I_CA_01	I_CA_02	I_OC_01	I_OC_02	I_IL_01	I_IL_02	I_MB_01	I_MB_02	I_BS_01	I_BS_02	O_CL_01	O_CL_02	O_RA_01	O_RA_02	O_SN_01	O_SN_02	O_SU_01	O_SU_02	AVG
Wren et al. [78]	0.83	0.80	0.80	0.71	0.95	0.82	0.52	0.46	0.81	0.67	0.46	0.49	0.87	0.85	0.89	0.82	0.74	0.45	0.68	0.83	0.72
Stauffer and Grismon [79]	0.84	0.82	0.88	0.77	0.95	0.83	0.35	0.24	0.83	0.69	0.36	0.37	0.89	0.85	0.74	0.83	0.74	0.47	0.62	0.83	0.70
Zivkovic and Van der Heijden [80]	0.91	0.90	0.91	0.75	0.99	0.91	0.16	0.31	0.93	0.80	0.55	0.52	0.93	0.82	0.86	0.90	0.52	0.24	0.54	0.88	0.72
Maddalena and Petrosino [81]	0.89	0.85	0.95	0.74	0.98	0.85	0.85	0.38	0.85	0.68	0.40	0.45	0.90	0.85	0.83	0.86	0.70	0.46	0.75	0.86	0.75
Maddalena and Petrosino [82]	0.96	0.94	0.84	0.87	0.96	0.95	0.19	0.23	0.97	0.85	0.40	0.40	0.97	0.98	0.84	0.86	0.91	0.71	0.87	0.88	0.78
Cuevas and Garcia [83]	0.81	0.76	0.84	0.63	0.83	0.88	0.80	0.79	0.78	0.68	0.51	0.66	0.93	0.90	0.75	0.87	0.82	0.09	0.65	0.81	0.74
Haines and Xiang [84]	0.96	0.81	0.92	0.87	0.89	0.95	0.89	0.81	0.98	0.71	0.63	0.73	0.69	0.96	0.82	0.96	0.31	0.04	0.81	0.90	0.78
Proposed method	**0.87**	**0.76**	**0.76**	**0.75**	**0.91**	**0.86**	**0.82**	**0.75**	**0.78**	**0.79**	**0.66**	**0.72**	**0.89**	**0.84**	**0.83**	**0.80**	**0.78**	**0.61**	**0.77**	**0.85**	**0.79**

Table 4. Reported F-measures for the Best Supervised Methods on LASIESTA.

Method	I_SI_02	I_CA_02	I_OC_02	I_IL_02	I_MB_02	I_BS_02	O_CL_02	O_RA_02	O_SN_02	OO_U_02	AVG
BSUV-Net 2.0 [85]	0.89	0.60	0.95	0.89	0.76	0.69	0.89	0.93	0.70	0.91	0.82
3DCD [86]	0.86	0.49	0.93	0.85	0.79	0.87	0.87	0.87	0.49	0.83	0.79
FgSegNet_S [28]	0.20	0.60	0.53	0.25	0.60	0.28	0.19	0.16	0.05	0.18	0.30
FgSegNet_M [28]	0.56	0.55	0.65	0.42	0.56	0.19	0.28	0.18	0.01	0.33	0.37
MSFS [87]	0.53	0.58	0.25	0.41	0.63	0.25	0.54	0.54	0.05	0.29	0.41
ChangeDet [13]	0.83	0.66	0.89	0.87	0.77	0.82	0.90	0.85	0.51	0.81	0.79
Proposed method	0.76	0.75	0.86	0.75	0.79	0.72	0.84	0.80	0.61	0.85	0.77

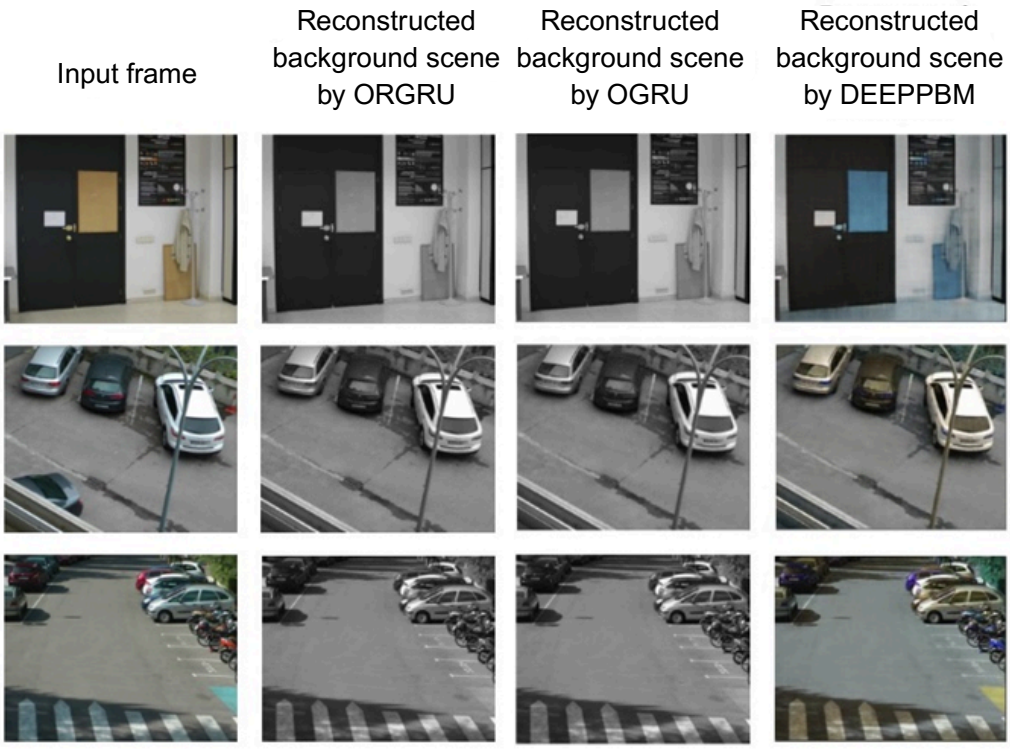


Figure 5. Visual Comparison of ORGRU, OGRU, and DeepPBM Methods in Reconstructing Background Scenes.

5 Conclusions

In this paper, a novel deep learning method is proposed for the background subtraction task by gener-

Table 5. Qualitative Comparison of ORGRU, OGRU, and DeepPBM.

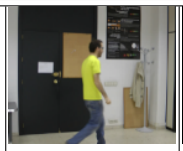
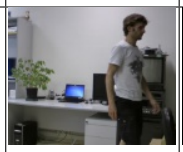
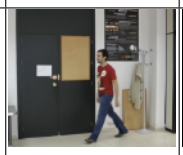
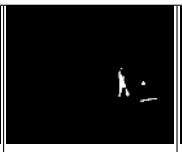
Category	Video Name	Input frame	Ground Tr.	ORGRU	OGRU	DeepPBM
Simple sequences	I_SI_01-183.bmp					
Camouflage	I_CA_02-470.bmp					
Occlusions	I_OC_01-207.bmp					
Illumination changes	I_IL_01-153.bmp					
Modified background	I_MB_01-254.bmp					
Bootstrap	I_BS_02-126.bmp					
Cloudy conditions	O_CL_02-283.bmp					
Rainy conditions	O_RA_02-237.bmp					
Snowy conditions	O_SN_01-346.bmp					

Table 5. Qualitative Comparison of ORGRU, OGRU, and DeepPBM (Continued..).

Category	Video Name	Input frame	Ground Tr.	ORGRU	OGRU	DeepPBM
Sunny conditions	O_SU_01-199.bmp					

alizing the $X = L + S$ framework to the deep GRU-based neural networks. Our approach offers an unsupervised and robust solution within the framework of deep recurrent networks, suitable for real-time background subtraction applications. By utilizing a simple architecture, our model reconstructs the background and foreground frames of an input video frame using a deep recurrent network. The proposed model learns the underlying patterns of the video sequence using backpropagation, receiving frames of the video sequence one at a time, and reconstructing the background image using GRUs without supervision. This learning process continues until the end of the task, resulting in enhanced effectiveness. The evaluation of our proposed model was conducted on the LASIESTA database, utilizing video sequences captured by static cameras. The experimental results demonstrate that our online and robust GRU-based method outperformed the non-robust mode by 17% in terms of F-measure and also surpassed DeepPBM and several other advanced supervised and unsupervised methods. Our method achieves superior performance with fewer parameters than DeepPBM, making it more suitable for resource-constrained environments.

The unsupervised nature of our proposed method, along with its frame-by-frame operation, makes it suitable for real-time applications. Additionally, its robustness enables it to handle changes, noise, and outlier data effectively. Therefore, the proposed method can be employed for pre-processing video data in complex video processing problems such as traffic monitoring, action recognition, intelligent surveillance, and anomaly detection. The method achieves low computational cost, high processing speed, operates without requiring complete datasets, and enables real-time deployment. In future work, we aim to enhance the ORGRU to support PTZ and moving camera categories, further expanding the applicability of our method.

Conflict of Interest

The authors declare that they have no conflict of interest. The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Data Availability

No new data were generated in this study. Data sharing does not apple to this article.

References

[1] Murari Mandal and Santosh Kumar Vipparthi. An empirical review of deep learning frameworks for change detection: Model design, experimental frameworks, challenges and research needs. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6101–6122, 2021. doi:10.1109/TITS.2021.3077883.

[2] Xinyue Zhao, Guangli Wang, Zaixing He, and Huilong Jiang. A survey of moving object detection methods: A practical perspective. *Neurocomputing*, 503:28–48, 2022. doi:https://doi.org/10.1016/j.neucom.2022.06.104.

[3] Yizhong Yang, Tingting Xia, Dajin Li, Zhang Zhang, and Guangjun Xie. A multi-scale feature fusion spatial–channel attention model for background subtraction. *Multimedia Systems*, 29(6):3609–3623, 2023. doi:https://doi.org/10.1007/s00530-023-01139-1.

[4] Kevin D McCay, Edmond SL Ho, Claire Marcroft, and Nicholas D Embleton. Establishing pose based features using histograms for the detection of abnormal infant movements. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5469–5472. IEEE, 2019. doi:https://doi.org/10.1109/EMBC.2019.8857680.

[5] Hazar Mliki, Fatma Bouhlel, and Mohamed Hammami. Human activity recognition from uav-captured video sequences. *Pattern Recognition*, 100:107140, 2020. doi:https://doi.org/10.1016/j.patcog.2019.107140.

[6] Z Mortezaie, H Hassanpour, and A Beghdadi. Re-identification in video surveillance systems considering appearance changes. *Scientia Iranica*, 31(2):103–117, 2024. doi:https://doi.org/10.24200/sci.2023.57617.5331.

[7] Daniel Berjón, Carlos Cuevas, Francisco Morán, and Narciso García. Real-time nonparametric background subtraction with tracking-based foreground update. *Pattern Recognition*, 74:156–170, 2018. doi:https://doi.org/10.1016/j.patcog.2017.09.009.

[8] Shuai Yi, Xiaogang Wang, Cewu Lu, Jiaya Jia, and Hongsheng Li. l_0 regularized stationary-time estimation for crowd analysis. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):981–994, 2016. doi:https://doi.org/10.1109/TPAMI.2016.2560807.

[9] Thangarajah Akilan, QM Jonathan Wu, and Wandong Zhang. Video foreground extraction using multi-view receptive field and encoder–decoder dcnn for traffic and surveillance applications. *IEEE Transactions on Vehicular Technology*, 68(10):9478–9493, 2019. doi:https://doi.org/10.1109/TVT.2019.2937076.

[10] Hui-Shyong Yeo, Byung-Gook Lee, and Hyotaek Lim. Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware. *Multimedia Tools and Applications*, 74(8):2687–2715, 2015. doi:https://doi.org/10.1007/s11042-013-1501-1.

[11] Belmar Garcia-Garcia, Thierry Bouwmans, and Alberto Jorge Rosales Silva. Background subtraction in real applications: Challenges, current models and future directions. *Computer Science Review*, 35:100204, 2020. doi:https://doi.org/10.1016/j.cosrev.2019.100204.

[12] Maryam Sultana, Arif Mahmood, and Soon Ki Jung. Unsupervised moving object detection in complex scenes using adversarial regularizations. *IEEE Transactions on Multimedia*, 23:2005–2018, 2020. doi:https://doi.org/10.1109/TMM.2020.3006419.

[13] Murari Mandal and Santosh Kumar Vipparthi. Scene independency matters: An empirical study of scene dependent and scene independent evaluation for cnn-based change detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):2031–2044, 2020. doi:https://doi.org/10.1109/TITS.2020.3030801.

[14] Thierry Bouwmans, Sajid Javed, Maryam Sultana, and Soon Ki Jung. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. *Neural Networks*, 117:8–66, 2019. doi:https://doi.org/10.1016/j.neunet.2019.04.024.

[15] Donald R Cantrell, Leon Cho, Chaochao Zhou, Syed HA Faruqui, Matthew B Potts, Babak S Jahromi, Ramez Abdalla, Ali Shaibani, and Sameer A Ansari. Background subtraction angiography with deep learning using multi-frame spatiotemporal angiographic input. *Journal of Imaging Informatics in Medicine*, 37(1):134–144, 2024. doi:https://doi.org/10.1007/s10278-023-00921-x.

[16] Carlos Cuevas, Eva María Yáñez, and Narciso García. Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta. *Computer Vision and Image Understanding*, 152:103–117, 2016. doi:https://doi.org/10.1016/j.cviu.2016.08.005.

[17] Rudrika Kalsotra and Sakshi Arora. A comprehensive survey of video datasets for background subtraction. *IEEE access*, 7:59143–59171, 2019. doi:https://doi.org/10.1109/ACCESS.2019.2914961.

[18] Thierry Bouwmans and El Hadi Zahzah. Robust pca via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014. doi:https://doi.org/10.1016/j.cviu.2013.11.009.

[19] Srivatsa Prativadibhayankaram, Huynh Van Luong, Thanh Ha Le, and André Kaup. Compressive online video background–foreground separation using multiple prior information and optical flow. *Journal of Imaging*, 4(7):90, 2018. doi:https://doi.org/10.3390/jimaging4070090.

[20] Namrata Vaswani, Thierry Bouwmans, Sajid Javed, and Praneeth Narayanamurthy. Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery. *IEEE signal processing magazine*, 35(4):32–55, 2018. doi:https://doi.org/10.1109/MSP.2018.2826566.

[21] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Spatiotemporal low-rank modeling for complex scene background initialization. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1315–1329, 2016. doi:https://doi.org/10.1109/TCSVT.2016.2632302.

[22] Rudrika Kalsotra and Sakshi Arora. Background subtraction for moving object detection: explorations of recent developments and challenges. *The Visual Computer*, 38(12):4151–4178, 2022. doi:https://doi.org/10.1007/s00371-021-02286-0.

[23] Praneeth Narayanamurthy and Namrata Vaswani. Provable dynamic robust pca or robust subspace tracking. *IEEE Transactions on Information Theory*, 65(3):1547–1577, 2018. doi:https://doi.org/10.1109/TIT.2018.2872023.

[24] Thangarajah Akilan, QM Jonathan Wu, Wei Jiang, Amin Safaei, and Jie Huo. New trend in video foreground detection using deep

learning. In *2018 IEEE 61St international midwest symposium on circuits and systems (MWSCAS)*, pages 889–892. IEEE, 2018. doi:https://doi.org/10.1109/MWSCAS.2018.8623825.

[25] Islam Osman, Mohamed Abdelpakey, and Mohamed S Shehata. Transblast: Self-supervised learning using augmented subspace with transformer for background/foreground separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 215–224, 2021. doi:https://doi.org/10.1109/ICCVW54120.2021.00029.

[26] Tsubasa Minematsu, Atsushi Shimada, and Rin-ichiro Taniguchi. Simple background subtraction constraint for weakly supervised background subtraction network. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. doi:https://doi.org/10.1109/AVSS.2019.8909896.

[27] Nhat Minh Chung and Synh Viet-Uyen Ha. Bgsubnet: Robust semisupervised background subtraction in realistic scenes. *IEEE Sensors Journal*, 24(6):9172–9186, 2024. doi:https://doi.org/10.1109/JSEN.2024.3358181.

[28] Long Ang Lim and Hacer Yalim Keles. Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognition Letters*, 112:256–262, 2018. doi:https://doi.org/10.1016/j.patrec.2018.08.002.

[29] Yi Wang, Zhiming Luo, and Pierre-Marc Jodoin. Interactive deep learning method for segmenting moving objects. *Pattern Recognition Letters*, 96:66–75, 2017. doi:https://doi.org/10.1016/j.patrec.2016.09.014.

[30] Víctor Manuel Mondéjar-Guerra, José Rouco, Jorge Novo, and Marcos Ortega. An end-to-end deep learning approach for simultaneous background modeling and subtraction. In *BMVC*, page 266, 2019.

[31] Dongdong Zeng and Ming Zhu. Background subtraction using multiscale fully convolutional network. *IEEE Access*, 6:16010–16021, 2018. doi:https://doi.org/10.1109/ACCESS.2018.2817129.

[32] Gani Rahmon, Filiz Bunyak, Guna Seetharaman, and Kannappan Palaniappan. Motion u-net: Multi-cue encoder-decoder network for motion segmentation. In *2020 25th International conference on pattern recognition (ICPR)*, pages 8125–8132. IEEE, 2021. doi:https://doi.org/10.1109/ICPR48806.2021.9413211.

[33] Rezaei Behnaz, Farnoosh Amirreza, and Sarah Ostadabbas. Deeppbm: Deep probabilistic background model estimation from video sequences. In *International Conference on Pattern Recognition*, pages 608–621. Springer, 2021. doi:https://doi.org/10.1007/978-3-030-68790-8_47.

[34] Keita Takeda and Tomoya Sakai. Unsupervised deep learning of foreground objects from low-rank and sparse dataset. *Computer Vision and Image Understanding*, 240:103939, 2024. doi:https://doi.org/10.1016/j.cviu.2024.103939.

[35] Thierry Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer science review*, 11:31–66, 2014. doi:https://doi.org/10.1016/j.cosrev.2014.04.001.

[36] Moein Shakeri and Hong Zhang. Moving object detection in time-lapse or motion trigger image sequences using low-rank and invariant sparse decomposition. In *Proceedings of the IEEE international conference on computer vision*, pages 5123–5131, 2017. doi:https://doi.org/10.1109/ICCV.2017.548.

[37] Praneeth Narayanamurthy and Namrata Vaswani. Nearly optimal robust subspace tracking. In *International Conference on Machine Learning*, pages 3701–3709. PMLR, 2018. URL `https://proceedings.mlr.press/v80/narayanamurthy18a.html`.

[38] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A deep convolutional neural network for video sequence background subtraction. *Pattern recognition*, 76:635–649, 2018. doi:https://doi.org/10.1016/j.patcog.2017.09.040.

[39] Xueying Wang, Lei Liu, Guangli Li, Xiao Dong, Peng Zhao, and Xiaobing Feng. Background subtraction on depth videos with convolutional neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. doi:https://doi.org/10.1109/IJCNN.2018.8489230.

[40] Fateme Bahri, Moein Shakeri, and Nilanjan Ray. Online illumination invariant moving object detection by generative neural network. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2018. doi:https://doi.org/10.1145/3293353.3293369.

[41] Sakti Saurav, Pankaj Malhotra, Vishnu TV, Narendhar Gugulothu, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Online anomaly detection with concept drift adaptation using recurrent neural networks. In *Proceedings of the acm india joint international conference on data science and management of data*, pages 78–87, 2018. doi:https://doi.org/10.1145/3152494.3152501.

[42] Pei Xu, Mao Ye, Qihe Liu, Xudong Li, Lishen Pei, and Jian Ding. Motion detection via a couple of auto-encoder networks. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

doi:https://doi.org/10.1109/ICME.2014.6890140.

[43] Pei Xu, Mao Ye, Xue Li, Qihe Liu, Yi Yang, and Jian Ding. Dynamic background learning through deep auto-encoder networks. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 107–116, 2014. doi:https://doi.org/10.1145/2647868.2654914.

[44] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017. doi:https://doi.org/10.1145/3097983.3098052.

[45] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999. doi:https://doi.org/10.1109/CVPR.1999.784637.

[46] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 28–31. IEEE, 2004. doi:https://doi.org/10.1109/ICPR.2004.1333992.

[47] Sriram Varadarajan, Paul Miller, and Huiyu Zhou. Spatial mixture of gaussians for dynamic background modelling. In *2013 10th IEEE international conference on advanced video and signal based surveillance*, pages 63–68. IEEE, 2013. doi:https://doi.org/10.1109/AVSS.2013.6636617.

[48] Ahmed Elgammal, David Harwood, and Larry Davis. Non-parametric model for background subtraction. In *European conference on computer vision*, pages 751–767. Springer, 2000. doi:https://doi.org/10.1007/3-540-45053-X_48.

[49] Shengcai Liao, Guoying Zhao, Vili Kellokumpu, Matti Pietikäinen, and Stan Z Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1301–1306. IEEE, 2010. doi:https://doi.org/10.1109/CVPR.2010.5539817.

[50] Hanzi Wang and David Suter. A consensus-based method for tracking: Modelling background scenario and foreground appearance. *Pattern recognition*, 40(3):1091–1105, 2007. doi:https://doi.org/10.1016/j.patcog.2006.05.024.

[51] Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin. A self-adjusting approach to change detection based on background word consensus. In *2015 IEEE winter conference on applications of computer vision*, pages 990–997. IEEE, 2015. doi:https://doi.org/10.1109/WACV.2015.137.

[52] Olivier Barnich and Marc Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6):1709–1724, 2010. doi:https://doi.org/10.1109/TIP.2010.2101613.

[53] Sajid Javed, Seon Ho Oh, Andrews Sobral, Thierry Bouwmans, and Soon Ki Jung. Or-pca with mrf for robust foreground detection in highly dynamic backgrounds. In *Asian conference on computer vision*, pages 284–299. Springer, 2014. doi:https://doi.org/10.1007/978-3-319-16811-1_19.

[54] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Background–foreground modeling based on spatiotemporal sparse subspace clustering. *IEEE Transactions on Image Processing*, 26(12):5840–5854, 2017. doi:https://doi.org/10.1109/TIP.2017.2746268.

[55] Alessandra Stagliano, Nicoletta Noceti, Alessandro Verri, and Francesca Odone. Online space-variant background modeling with sparse coding. *IEEE Transactions on Image Processing*, 24(8):2415–2428, 2015. doi:https://doi.org/10.1109/TIP.2015.2421435.

[56] Hyomin Ahn and Myungjoo Kang. Dynamic background subtraction with masked rpca. *Signal, Image and Video Processing*, 15(3):467–474, 2021. doi:https://doi.org/10.1007/s11760-020-01766-5.

[57] Moein Shakeri and Hong Zhang. Corola: A sequential solution to moving object detection using low-rank approximation. *Computer Vision and Image Understanding*, 146:27–39, 2016. doi:https://doi.org/10.1016/j.cviu.2016.02.009.

[58] Marc Braham, Sébastien Piérard, and Marc Van Droogenbroeck. Semantic background subtraction. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4552–4556. Ieee, 2017. doi:https://doi.org/10.1109/ICIP.2017.8297144.

[59] Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin. Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, 24(1):359–373, 2014. doi:https://doi.org/10.1109/TIP.2014.2378053.

[60] Rui Wang, Filiz Bunyak, Guna Seetharaman, and Kannappan Palaniappan. Static and moving object detection using flux tensor with split gaussian models. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 414–418, 2014. doi:https://doi.org/10.1109/CVPRW.2014.68.

[61] Xiao Bian, Ser Nam Lim, and Ning Zhou. Multiscale fully convolutional network with

application to industrial inspection. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016. doi:https://doi.org/10.1109/WACV.2016.7477595.

[62] Yuan Dai and Long Yang. Background subtraction for video sequence using deep neural network. *Multimedia Tools and Applications*, 83(35):82281–82302, 2024. doi:https://doi.org/10.1007/s11042-024-18843-3.

[63] Dongdong Zeng, Xiang Chen, Ming Zhu, Michael Goesele, and Arjan Kuijper. Background subtraction with real-time semantic segmentation. *IEEE access*, 7:153869–153884, 2019. doi:https://doi.org/10.1109/ACCESS.2019.2899348.

[64] Mohamed Chafik Bakkay, Hatem A Rashwan, Houssam Salmane, Louahdi Khoudour, D Puig, and Yassine Ruichek. Bscgan: Deep background subtraction with conditional generative adversarial networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4018–4022. IEEE, 2018. doi:https://doi.org/10.1109/ICIP.2018.8451603.

[65] Maryam Sultana, Arif Mahmood, Sajid Javed, and Soon Ki Jung. Unsupervised deep context prediction for background estimation and foreground segmentation. *Machine Vision and Applications*, 30(3):375–395, 2019. doi:https://doi.org/10.1007/s00138-018-0993-0.

[66] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. doi:https://doi.org/10.48550/arXiv.1606.05908.

[67] Terence Broad and Mick Grierson. Autoencoding video frames. Technical report, Technical Report (London: Goldsmiths, 2016), available at¡ http://research . . . , 2016.

[68] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in Order Restricted Statistical Inference: Proceedings of the Symposium on Order Restricted Statistical Inference held in Iowa City, Iowa, September 11–13, 1985*, pages 28–47. Springer, 1986. doi:https://doi.org/10.1007/978-1-4613-9940-7_3.

[69] Panos M Pardalos. Convex optimization theory, 2010.

[70] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. doi:http://dx.doi.org/10.1561/2200000016.

[71] Ofir Lindenbaum, Yariv Aizenbud, and Yuval Kluger. Probabilistic robust autoencoders for outlier detection. *arXiv preprint arXiv:2110.00494*, 2021. doi:https://doi.org/10.48550/arXiv.2110.00494.

[72] Deepppbm. `https://github.com/ostadabbas/DeepPBM`. Accessed: 2024-07-29.

[73] Arezoo Sedghi, Esmat Rashedi, Maryam Amoozegar, and Fatemeh Afsari. Online vehicle detection using gated recurrent units. In *international conference on Artificial intelligence and smart Vehicle*, 2023.

[74] Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *IEEE international conference on neural networks*, pages 586–591. IEEE, 1993. doi:https://doi.org/10.1109/ICNN.1993.298623.

[75] Xin Liu, Guoying Zhao, Jiawen Yao, and Chun Qi. Background subtraction based on low-rank and structured sparse decomposition. *IEEE Transactions on Image Processing*, 24(8):2502–2514, 2015. doi:https://doi.org/10.1109/TIP.2015.2419084.

[76] Antoine Vacavant, Thierry Chateau, Alexis Wilhelm, and Laurent Lequievre. A benchmark dataset for outdoor foreground/background extraction. In *Asian Conference on Computer Vision*, pages 291–300. Springer, 2012. doi:https://doi.org/10.1007/978-3-642-37410-4_25.

[77] Bingxin Hou, Ying Liu, Nam Ling, Yongxiong Ren, and Lingzhi Liu. A survey of efficient deep learning models for moving object segmentation. *APSIPA Transactions on Signal and Information Processing*, 12(1), 2023. doi:http://dx.doi.org/10.1561/116.00000140.

[78] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):780–785, 1997. doi:https://doi.org/10.1109/34.598236.

[79] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):747–757, 2002. doi:https://doi.org/10.1109/34.868677.

[80] Zoran Zivkovic and Ferdinand Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006. doi:https://doi.org/10.1016/j.patrec.2005.11.005.

[81] Lucia Maddalena and Alfredo Petrosino. A self-organizing approach to background subtraction for visual surveillance applications. *IEEE Transactions on image processing*, 17(7):1168–1177, 2008. doi:https://doi.org/10.1109/TIP.2008.924285.

[82] Lucia Maddalena and Alfredo Petrosino.

The sobs algorithm: What are the limits? In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 21–26. IEEE, 2012. doi:https://doi.org/10.1109/CVPRW.2012.6238922.

[83] Carlos Cuevas and Narciso García. Improved background modeling for real-time spatio-temporal non-parametric moving object detection strategies. *Image and Vision Computing*, 31(9):616–630, 2013. doi:https://doi.org/10.1016/j.imavis.2013.06.003.

[84] Tom SF Haines and Tao Xiang. Background subtraction with dirichletprocess mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):670–683, 2013. doi:https://doi.org/10.1109/TPAMI.2013.239.

[85] M Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Bsuv-net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction. *IEEE Access*, 9:53849–53860, 2021. doi:https://doi.org/10.1109/ACCESS.2021.3071163.

[86] Murari Mandal, Vansh Dhar, Abhishek Mishra, Santosh Kumar Vipparthi, and Mohamed Abdel-Mottaleb. 3dcd: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos. *IEEE Transactions on Image Processing*, 30:546–558, 2020. doi:https://doi.org/10.1109/TIP.2020.3037472.

[87] Long Ang Lim and Hacer Yalim Keles. Learning multi-scale features for foreground segmentation. *Pattern Analysis and Applications*, 23(3):1369–1380, 2020. doi:https://doi.org/10.1007/s10044-019-00845-9.

Arezoo Sedghi received her M.Sc. in Electrical Engineering–Telecommunication Systems from the Graduate University of Advanced Technology in Kerman, Iran. She is currently pursuing a Ph.D. at IT:U Interdisciplinary Transformation University Austria, where she is part of the GeoSocial Artificial Intelligence research group. Her research focuses on analyzing physiological signals to investigate human behavior and emotional responses.

Maryam Amoozegar received the Ph.D. degree in Software Engineering from the Iran University of Science and Technology, Tehran, Iran, in 2021. She is now an assistant professor at the Computer and Information Technology department of the Institute of Science and High Technology and Environmental Sciences at the Graduate University of Advanced Technology in Kerman. Her research interests are anomaly detection, streaming data processing, data mining, machine learning, and Swarm intelligence.

Esmat Rashedi received her B.Sc. degree in Electrical Engineering, and her M.Sc. and Ph.D. degrees in Communication Engineering from Shahid Bahonar University of Kerman, Iran, in 2004, 2007, and 2013, respectively. Now she is an associate Professor at Kerman Graduate University of Advanced Technology. Her research interests include heuristic optimization algorithms, soft computing, image processing, pattern recognition, and deep learning.

Fatemeh Afsari received her B.Sc. and M.Sc. degrees in Computer Science and Engineering from Shiraz University, specializing in Artificial Intelligence and Robotics. She earned her Ph.D. from Shahid Bahonar University of Kerman and served as an Assistant Professor from 2014. Her research focuses on intelligent computing, medical imaging, and AI-driven solutions for disease diagnosis and tissue analysis. She currently leads several projects, with a particular interest in deep learning, domain adaptation, and interpretable models in computational pathology.