Unsupervised Deep Context Prediction for Background Estimation and Foreground Segmentation

Article *in* Machine Vision and Applications · October 2018

DOI: 10.1007/s00138-018-0993-0

CITATIONS
30

READS
895

4 authors:

Maryam Sultana
Kyungpook National University
13 PUBLICATIONS **111** CITATIONS
SEE PROFILE

Arif Mahmood
Information Technology University (ITU)
107 PUBLICATIONS **1,362** CITATIONS
SEE PROFILE

Sajid Javed
KUCARS UAE
53 PUBLICATIONS **960** CITATIONS
SEE PROFILE

Soon Ki Jung
Kyungpook National University
180 PUBLICATIONS **1,370** CITATIONS
SEE PROFILE

Machine Vision and Applications
https://doi.org/10.1007/s00138-018-0993-0

ORIGINAL PAPER

Unsupervised deep context prediction for background estimation and foreground segmentation

Maryam Sultana[1] · Arif Mahmood[2] · Sajid Javed[3] · Soon Ki Jung[1]

Received: 19 May 2018 / Revised: 8 October 2018 / Accepted: 14 November 2018

Abstract
Background estimation is a fundamental step in many high-level vision applications, such as tracking and surveillance. Existing background estimation techniques suffer from performance degradation in the presence of challenges such as dynamic backgrounds, photometric variations, camera jitters, and shadows. To handle these challenges for the purpose of accurate background estimation, we propose a unified method based on Generative Adversarial Network (GAN) and image inpainting. The proposed method is based on a context prediction network, which is an unsupervised visual feature learning hybrid GAN model. Context prediction is followed by a semantic inpainting network for texture enhancement. We also propose a solution for arbitrary region inpainting using the center region inpainting method and Poisson blending technique. The proposed algorithm is compared with the existing state-of-the-art methods for background estimation and foreground segmentation and outperforms the compared methods by a significant margin.

Keywords Background subtraction · Foreground detection · Context prediction · Generative adversarial networks

1 Introduction

Background estimation and foreground segmentation are fundamental steps in several computer vision applications, such as salient motion detection [63,78], video surveillance [3,16,31], visual object tracking [61,81], moving object detection [17,48,53,62], and object/image segmentation [22,50,51,66]. The goal of background modeling is to efficiently and accurately extract a model, which describes the scene in the absence of any foreground objects. On the other hand, foreground detection is the process of segmenting dynamic objects from prior knowledge of the background model. Recently, foreground detection-based algorithms have been referred to as co-saliency detection [23] methods. Saliency detection is currently one of the most active research topics in the development of many computer vision-based applications. The objective of saliency detection is similar to foreground detection, involving the identification of significant objects in the presence of complex background scenes. Hence, saliency detection algorithms are classified into various types, such as object segmentation [68] and video saliency detection [64].

First, in targeting the problem of background estimation, the primary issue is that background modeling in real-time applications becomes challenging in the presence of dynamic backgrounds, sudden illumination variations, and camera jitters, mainly as a result of sensors. A number of techniques have been proposed in the literature, which mostly address relatively simple scenarios for scene background modeling [4], because complex background modeling is highly challenging task. To solve the problem of background subtraction, Stauffer *et al.* [58] and Elgammal et al. [13] presented methods based on statistical background modeling. Such methods begin from an unreliable background mode

✉ Soon Ki Jung
skjung@knu.ac.kr

Maryam Sultana
maryam@vr.knu.ac.kr

Arif Mahmood
arif.mahmood@itu.edu.pk

Sajid Javed
s.javed.1@warwick.ac.uk

[1] Virtual Reality Laboratory, School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea

[2] Department of Computer Science, Information Technology University (ITU), Lahore, Pakistan

[3] Tissue Image Analytics Laboratory, Department of Computer Science, University of Warwick, Warwick, UK

Published online: 26 November 2018

and identify and correct initial errors during the background updating stage through the analysis of the extracted foreground objects from the video sequences. Other methods proposed in recent years have also solved background initialization as an optimal labeling problem [43,45,74]. These methods compute labels for each image region and provide the number of the best bootstrap sequence frame such that the region contains the background scene. Taking into account spatiotemporal information, the best frame is selected by minimizing a cost function. The background information contained in the frames selected for each region is then combined to generate the background model. Background model initialization methods based on missing data reconstruction have also been proposed [57]. These methods work when data is missing owing to foreground objects that occlude the bootstrap sequence. Thus, robust matrix and tensor completion algorithms [56] and inpainting methods [10] have been shown to be suitable for background initialization. More recently, deep neural networks have been introduced for image inpainting [46]. In particular, Chao Yang et al. [75] employed a trained convolutional neural network (CNN) as context encoder [46] with a combined reconstruction loss and adversarial loss [18] to directly estimate missing image regions. Then, a joint optimization framework updates the estimated inpainted region with fine texture details. This is achieved by hallucinating the missing image regions by modeling two kinds of constraints, the global context-based and the local texture-based, using CNNs. This framework is able to estimate missing image structures and is very fast to evaluate. Although the results are encouraging, it is unable to handle random region inpainting tasks with fine details.

In this study, we propose the prediction of missing image structures using the inpainting method, for the purpose of background scene initialization. We name our method *Deep Context Prediction* (DCP), because it has the ability to predict the context of a missing region via deep neural networks. A few visual results of the proposed DCP algorithm are shown in Fig. 1. Given an image, fast-moving foreground objects are removed using motion information leaving behind missing image regions [see Fig. 2 Step (1)]. We train a CNN to estimate the missing pixel values via the inpainting method. The CNN model consists of an encoder capturing the context of the whole image into a latent feature representation, and a decoder, which uses this representation to produce the missing content of the image. The model is closely related to auto-encoders [2,25], as it shares a similar architecture to an encoder–decoder. Our contributions are summarized as follows:

– We propose a novel background estimation and foreground detection model by using an image inpainting algorithm based on deep learning models.

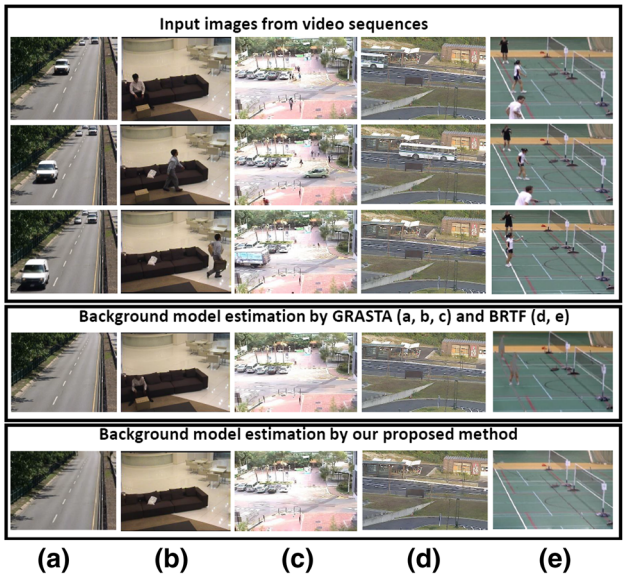

Fig. 1 Estimated background images from the SBM.net dataset: Sequences in **a** are from the category "Basic" named "Highway". **b** The sequence "Sofa" from the category "Intermittent Object Motion" **c** The sequence "Chuk Square" from the category "Very Short". **d** The sequence "Bus Station" from the category "Very Long". **e** The sequence "Badminton" is also from the category "Jitter." In almost all these cases, for accurate background estimation the average gray-level error (AGE) is lower for our proposed algorithm, as shown in Table 1

– We develop a deep learning-based end-to-end system, which is able to perform background estimation and foreground detection in the presence of various challenging real-time environments. In contrast to existing methods [37,38,70], our proposed method is completely unsupervised. We do not require any labeled data to train our deep learning model.
– Our proposed method DCP consists of two deep learning models: one for context prediction initialization and the other for fine texture optimization. In order to predict new contexts in missing image regions, we train a "Context Network" on scene-specific data. To improve the predicted context, local texture optimization is performed by using the "Texture Network." Both networks are trained in completely unsupervised fashion.
– The content estimated by context and texture networks is based on central region inpainting. We propose to transform this to random region inpainting using Modified Poisson Blending (MPB) technique.

The proposed DCP algorithm is based on context prediction, and therefore, it can predict homogeneous or blurry contexts more accurately compared with other background initialization algorithms. In case of background motion, DCP can still estimate the background by calculating motion masks by using the optical flow, as our target is to eliminate moving foreground objects only. DCP is also not affected by

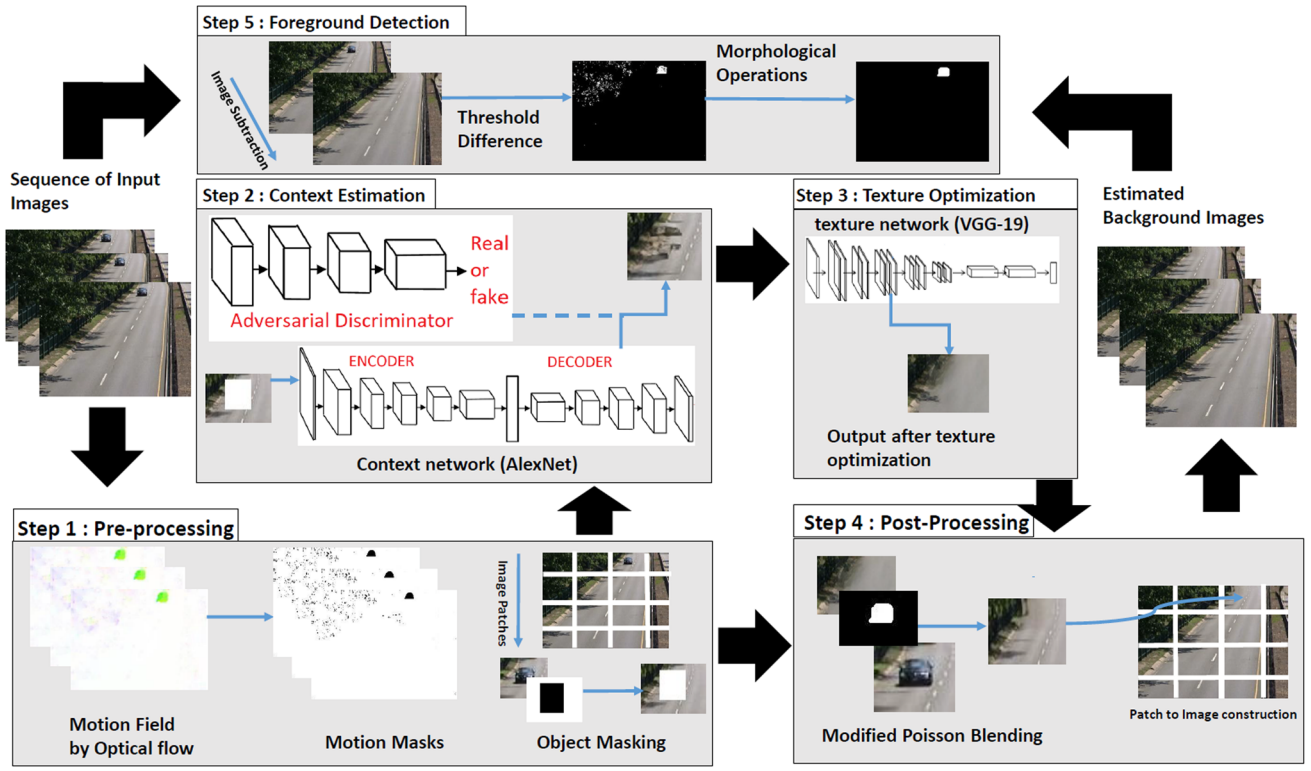

Fig. 2 Workflow of the proposed algorithm for background estimation. Step (1) describes the motion estimation via the dense optical flow, the creation of motion masks, image to patch conversion, and object masking for the center region inpainting task. Step (2) evaluates the prediction of the missing region using the context prediction hybrid GAN network. In step (3), to improve the fine texture details of the predicted context, the output of step (2) is fed to the texture network. It can be seen in step(3) that in this case some information of the road (white lines in the middle) is missed by texture network, and so in step(4), the MPB technique is applied to obtain the final results. In step (5), we threshold the difference between the background estimated via DCP and the current frame of the video sequence. Then, the thresholded difference is binarized and run through extensive morphological operations to extract the foreground moving object

intermittent object motion, for the same reason mentioned previously. In challenging weather conditions (rain, snow, or fog), the dense optical flow can identify foreground moving objects, so targeting only these objects to remove and inpaint them with background pixels makes DCP a good background estimator. In the case of difficult light conditions, DCP can estimate background accurately, owing to homogeneity in the contexts of scenes with low illumination.

2 Related work

In recent years, background subtraction and foreground detection [8,20,26–28,34,44] and scene background initialization [3,4,14,42,76] have remained the part of many key research studies. In the problem of background subtraction, the critical step is to improve the accuracy of the detection of the foreground. On the other hand, the task of estimating an image without any foreground is called scene background modeling. Many comprehensive studies have been conducted regarding this problem [3–5,35,41,42].

The Gaussian mixture model (GMM) [40,58,60,80,86] is a well-known technique for background modeling. It employs probability density functions as a mixture of Gaussians to model color intensity variations at the pixel level. Recent advances in GMM include the minimum spanning tree [9] and bidirectional analysis [54]. On the other hand, most GMM-based methods also suffer from performance degradation for complex and dynamic scenes. In the past, particularly for the problem of background modeling many research studies have been conducted using *Robust Principal Component Analysis* (RPCA). Wright et al. [71] presented the first proposal of RPCA-based method, which has the ability to handle outliers in the input data. Subsequently, Candes et al. [7] utilzed RPCA for background modeling and foreground detection. Beyond delivering a good performance, RPCA-based methods are not ideal for real-time applications, because these techniques involve a high computational complexity. Moreover, conventional RPCA-based methods process data in a batch manner. Batch methods are not suitable for real-time applications and mostly work offline. Some online and hybrid RPCA-based methods have also been pre-

sented in the literature to handle the batch problem [29], while global optimization remains a challenge with these approaches [24,26,73]. Xiaowei Zhou et al. [85] proposed an interesting technique known as *DEtecting Contiguous Outliers in the LOw-rank Representation* (DECOLOR). The limitation of no prior knowledge of spatial distribution of outliers in RPCA-based methods led to the development of this technique. Outlier information is modeled in this formulation using Markov Random Fields (MRFs).

Another online RPCA algorithm proposed by Jun He et al. [24] is the *Grassmannian Robust Adaptive Subspace Tracking Algorithm* (GRASTA). This is an online robust subspace tracking algorithm embedded with traditional RPCA. The algorithm operates on data that is highly sub-sampled. If the observed data matrix is corrupted by outliers, as in most cases of real-time applications, then an l^2-norm-based objective function is best-fit to the subspace. A hybrid approach employs a time window to obtain sufficient context information and then processes this like a small batch. Recently, Javed et al. [30] proposed a hybrid technique named *Motion-assisted Spatiotemporal Clustering of Low-rank* (MSCL) based on the RPCA approach. In this method, sparse coding is applied for each data matrix, and an estimation of the geodesic subspace-based Laplacian matrix is calculated. The normalized Laplacian matrices estimated over both the euclidean and geodesic distances are embedded into the basic RPCA framework. In 2015, Liu et al. [84] developed a technique called *Sparse Matrix Decomposition* (SSGoDec), which is capable of efficiently and robustly estimating the low-rank part $\mathbf{L}$ of the background and the sparse part $\mathbf{S}$ of an input data matrix $\mathbf{D} = \mathbf{L} + \mathbf{S} + \mathbf{G}$ with a noise factor $\mathbf{G}$. This technique alternatively assigns the low-rank approximation of the difference between the input data matrix and the sparse matrix **(D-S)** to $\mathbf{L}$. Similarly, it can assign the sparse approximation of **(D-L)** to $\mathbf{S}$. To overcome the batch constraint of RPCA-based methods Xu et al. [72] presented a method called *Grassmannian Online Subspace Updates with Structured-sparsity* (GOSUS). Although this method performs well for background estimation problems, global optimality remains a challenging issue in this approach. Qibin Zhao *et al.* [83] presented a method called *Bayesian Robust Tensor Factorization for Incomplete Multiway Data* (BRTF). This method is a generative model for robust tensor factorization in the presence of missing data and outliers. Guo et al. [19] presented a method called *Robust Foreground Detection Using Smoothness and Arbitrariness Constraints* (RFSA). In this method, the authors considered the smoothness and arbitrariness of a static background, thus formulating the problem in a unified framework from a probabilistic perspective.

Recently, CNN-based methods have also demonstrated significant performances for foreground detection by scene background modeling [6,70,82]. For instance, Wang et al. [70] proposed a simple yet effective supervised CNN-based method for detecting moving objects in static background scenes. CNN-based methods perform best for many complex scenes. However, our proposed method DCP is unsupervised, and therefore, it does not require any labeled data for training purposes. In addition to the mentioned conventional techniques, recent studies [77,79] have claimed that the co-saliency detection techniques can also address the problem of video foreground segmentation. For instance, Tsai et al. [59], recently presented a novel foreground detection algorithm based on co-saliency extraction via the locally adaptive fusion technique. Similarly, Fu *et al.* [15] also presented another co-saliency detection technique that works on the principal of the global correspondence between multiple images by learning them during the clustering process. Furthermore, to improve foreground detection via co-saliency detection under challenging conditions like camouflage of salient objects with the background, Han et al. [21] presented a solution. This consists of a unified metric learning-based algorithm to jointly learn discriminative feature representations.

3 Proposed method

Our proposed background foreground separation technique consists of five steps. (1) Motion mask evaluation via the dense optical flow. (2) Estimation of missing background pixels using a CE. (3) Improving the estimation of missing pixel textures using a multiscale neural patch synthesis. (4) A modified Poisson blending technique is applied to obtain the results. (5) The foreground objects are detected by applying a threshold on the difference between the background estimated from DCP and the current frame, which is later enhanced using MOs. The workflow diagram of DCP is presented in Fig. 2. A detailed description of the above steps follows.

3.1 Motion masks via optical flow

For the purpose of background estimation from video frames, we must first identify fast-moving foreground objects. These objects are recognized using the optical flow [39], which is then used to create a motion mask. The dense optical flow is calculated between each pair of consecutive frames in the given input video sequence S. The motion mask M is computed using motion information from a sequence of video frames. Let S_t and S_{t-1} denote two consecutive frames in S at the time instants t and $t-1$, respectively. Let $v^y_{t,p}$ and $u^x_{t,p}$ be the vertical and horizontal components, respectively, of the motion vector at position p computed between consecutive frames. Then, the corresponding motion mask $m_t \in \{0, 1\}$ is computed as follows:

$$m_{t,p} = \begin{cases} 1, & \text{if} \sqrt{(u^x_{t,p})^2 + (v^y_{t,p})^2} < t_h, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In the above equation, t_h is the threshold for the motion magnitude. This is computed by taking the average of all pixels in the motion field. The selection of the threshold t_h is adapted in such a manner that all pixels in S consisting of motions greater than t_h belong to the foreground. In order to avoid noise in the background, the threshold t_h is selected to be sufficiently large. A further explanation of the parameter t_h with a qualitative analysis is presented in Sect. 5.4.

3.2 Background pixels estimation via context prediction

Given image patches from a video with missing regions, such as foreground object regions, we predict the context via a CE [46]. The CE is a hybrid GAN model that is trained on the basis of a CNN to estimate the missing pixel values. This consists of two parts: an encoder that captures the context of a given image patch in a compact latent feature space and a decoder that uses an encoded representation to produce the missing image patch content. The overall architecture of the CE consists of a simple encoder–decoder pipeline. The CE is derived from AlexNet [32]. However, the network is not trained for classification, but rather is trained for context prediction. Because our objective is to generate the missing context, our network possesses channel-wise fully connected layers that help to create a link between the encoder and decoder for the purpose of context prediction. We selected AlexNet because it is relatively simple compared with the other deep CNN architectures [65,67], and therefore, it is easy to train and optimize. Training is performed on ImageNet, as well as using scene-specific video sequence patches. In order to learn an initial context prediction network, we train a regression network F to obtain a response $F(x_m)$, where x_m is the input image patch x pixel-wise multiplied by the object mask m_o: $x_m = x \odot m_o$. Here, m_o is a binary mask with a fixed central region that covers the whole object in the motion mask. The patch x_m with a missing region H is input to the context network. The response $F(x_m)$ of the trained context network is estimated via joint loss functions to estimate the background x_b in the missing region H. We have experimented with two joint loss functions, including the reconstruction loss L_{rec} and adversarial loss L_{adv} [46]. The reconstruction loss L_{rec} is defined as:

$$L_{\text{rec}}(x_m, x_b, H) = ||F(x_m) - C(x_b, H)||_2^2. \quad (2)$$

The adversarial loss is given as:

$$L_{\text{adv}}(x_m, x_b, H) = \max_D E_{x_m \in \chi_m}[\log(D(C(x_b, H))) + \log(1 - D(F(x_m)))], \quad (3)$$

where D is the adversarial discriminator, and $C(\cdot)$ defines the operation of extracting a sub-image in the central region during the inpainting process. The overall loss function is a linear combination of the reconstruction and adversarial losses.

$$L = \eta L_{\text{rec}}(x_m, x_b, H) + (1 - \eta) L_{\text{adv}}(x_m, x_b, H), \quad (4)$$

where η is the relative weight of each loss function.

3.3 Texture optimization of estimated background

In the last section, we estimated a background patch x_b via a CE. However, the estimated context still contains irregularities and blurry textures at a low resolution of the image patch. To resolve this blurry estimated context problem for high-resolution inpainting with fine details, we employ a texture network on a three-level pyramid of image patches. This network optimizes over three loss terms: the predicted context term initialized by the CE, the local texture optimization term, and the gradient loss term. The context prediction term captures the semantics, including global structures of the image patches. The texture term maps the local statistics of the input image patch texture, and the gradient loss term enforces the smoothness between the estimated and original contexts. In the three-level pyramid approach, the test image patch is assumed to always be cropped to 512×512, with a 256×256 hole in the center at a fine level. However, by downsizing to a coarse level using a step-size of two, a 128×128 size image patch with a 64×64 missing region is initialized by the CE. Subsequently, the context of the missing region is estimated in a coarse-to-fine manner. At each scale, joint optimization is performed to update the missing region, and then, upsampling is performed to initialize the joint optimization, which sets the context constraint for the next scale of the image patch. This process is repeated until the joint optimization is completed at the fine level of the pyramid. The texture optimization term is computed using $VGG - 19$ [55], which is pre-trained on ImageNet.

Once the context has been initialized by CE at the coarse scale, we use the output $F(x_m)$ and the original image as the initial context constraint for joint optimization. Let x_o denotes the original image patch with the missing region filled in by the CE. The upsampled version of x_o is adopted as the initialization for the joint optimization at fine scales. For the input image patch x_o, we would like to estimate the fine texture of the missing region. The region correspond-

ing to x_o in the feature map of the $VGG-19$ network is $\psi(x_o)$, and $\psi(H)$ denotes the feature map corresponding to the missing region. For the texture optimization, $C(\cdot)$ also defines the operation of extracting a sub-feature map in a rectangular region, i.e., the context of $\psi(x_o)$ within $\psi(H)$ is returned by $C(\psi(x_o), H)$. The optimal solution for the accurate reconstruction of the missing content is obtained by minimizing the following objective function at each scale $i = 1, 2, \ldots, n$:

$$\begin{aligned}\hat{x}^{i+1} = \arg\min\, & E_{CE}(C(x_o, H), C(x_o^i, H)) \\ & + \gamma E_T(\psi_T(x_o), \psi(H) + \delta \Pi(x_o),\end{aligned} \tag{5}$$

where $C(x_o^1, H) = F(x_o)$, $\psi_T(.)$ represents a feature map in the texture network T at an intermediate layer, and γ and δ are weighting reflecting parameters [75]. The first term E_{CE} in Eq. (5) is a context constraint, which is defined by the difference between the previous context prediction and the optimization result:

$$E_{CE}(C(x_o, H), C(x_o^i, H)) = ||C(x_o, H) - C(x_o^i, H)||_2^2. \tag{6}$$

The second term E_T in Eq. (5) handles the local texture constraint, which minimizes the inconsistency between the texture appearance outside and inside the missing region. We first select a single feature layer or a combination of different feature layers in the texture network T and then extract its feature map ψ_T. In order to perform the texture optimization, our target for each query local patch P of size $w \times w \times c$ in the missing region $\psi(H)$ is to find the most similar patch outside the missing region and calculate the loss as the mean of the queried local patch and its nearest neighbor distances.

$$\begin{aligned}& E_T(\psi_T(x_o), H) \\ & = \frac{1}{|\psi(H)|} \sum_{i \in \psi(H)} ||C(\psi_T(x_o), P_i) - C(\psi_T(x_o), P_{np(i)})||_2^2,\end{aligned} \tag{7}$$

In the above equation, the local neural patch centered at location i is denoted by P_i, the number of patches sampled in the region $\psi(H)$ is given by $|\psi(H)|$, and $np(i)$ is calculated as follows:

$$np_i = \arg \min_{j \in n(i) \wedge j \notin H^\psi} ||C(\psi_T(x), P_i) - C(\psi_T(x), P_j)||_2^2, \tag{8}$$

where $n(i)$ is the set of neighboring locations of i, excluding the overlap with the missing unknown region $\psi(H)$. We also add the gradient loss term to encourage smoothness in the texture optimization process [75]:

$$\begin{aligned}\Pi(x_o) = \sum_{j,k} & ((x_o(j, k+1) - x_o(j, k))^2 \\ & + (x_o(j+1, k) - x_o(j, k))^2),\end{aligned} \tag{9}$$

3.4 Blending of estimated and original textures

After the texture optimization, some information around the central region during the inpainting process is missed or removed owing to the rectangular-shaped region, as shown in Fig. 2. In order to change the predicted rectangular-shaped context to an irregular-shaped region, Modified Poisson Blending (MPB) technique [1] is applied. This is based on Poisson image editing for the purpose of seamless cloning. The MPB technique consists of three steps. The first step uses the source image, comprising the image inpainted via DCP, as a known region, and the target image, which consists of the original image containing the foreground, as an unknown region. Subsequently, it requires the motion mask obtained from the optical flow around an object of interest in the source image to solve the Poisson equation [47] under the gradient field and a predefined boundary condition. The MPB technique involves a few modifications to the Poisson image editing technique, which eliminate the bleeding problems in a composite image by applying Poisson blending with fair dependency on the source and target pixels, which consist of the inpainted context and original image pixels, respectively. In the next step, the MPB technique adopts the composite image as the unknown region and the target image with foreground objects as the known region. After applying the Poisson blending algorithm, we obtain another composite image, which will be used in third step. To reduce bleeding artifacts, the MPB technique generates an alpha mask, which is used to combine the two composite images from previous steps to obtain a final image that is free from color bleeding. In practice, this method helps to discard useless information that stems from the rectangular region inpainting process.

3.5 Foreground detection

In this work, we mainly focus on the problem of background initialization. However, in this section we also extend our work to foreground detection. Thus, we are able to compare our work with foreground detection algorithms, in addition to the work on background initialization. For the purpose of foreground detection, we threshold the difference between the background estimated via DCP and the current frame of the video sequence. The difference is thresholded and binarized, and then processed through MOs with suitable structuring elements SEs. Thus, the work presented in this section may be considered as post-processing.

These operations consist of an opening operation on an image, followed by erosion and then dilation with the same SE:

$$I \circ \text{SE} = (I \ominus \text{SE}) \oplus \text{SE}, \tag{10}$$

where I is the binarized difference, and $\ominus$ and $\oplus$ denote erosion and dilation, respectively. Subsequently, a closing operation is performed on the image. In the reverse manner, dilation can be followed by erosion with same SE, but with a different SE used in the opening operation.

$$I' \bullet \text{SE} = (I' \oplus \text{SE}) \ominus \text{SE}. \tag{11}$$

Here, I' denotes the difference image from Eq. (10), and $\oplus$ and $\ominus$ denote dilation and erosion, respectively. Successive opening and closing of the binarized difference frame with a suitable SE lead us to foreground objects separated from the background. The choice of SE is highly crucial in the successive opening and closing of the binarized difference frame, as it may lead to false detection if not selected according to the shapes of the objects in the video frames. The MO not only fills the missing regions in the thresholded difference, but also removes the unconnected pixel values of the background, which are considered to represent noise in the foreground detection process.

4 Implementation details

Our background estimation and foreground detection techniques are based on the inpainting model, similar to [75]. The CE network is already trained on the ImageNet dataset with 1.2M images for 110 epochs on a Titan-X GPU. This required one month to train, with a constant learning rate of 10^{-3} for the center region inpainting [46]. We performed fine tuning on this trained network by additionally training it with scene-specific data in terms of patches of size 128×128 for three epochs. In the object masking step, illustrated in step 1 of Fig. 2, the missing region is filled using the constant mean value. A stochastic gradient descent (SGD) solver performs the optimization of the CE. The encoder in the CE is pool-free, meaning that all pooling layers are replaced with convolutions of the same stride and kernel size. However, the overall stride of the context network remains the same. Instinctively, there is not much reason to use pooling layers for reconstruction-based deep networks. The texture optimization is performed by a $VGG-19$ network pre-trained on ImageNet for classification. Texture optimization requires approximately 1 min to fill the highest level of the pyramid, which consists of the 256×256 missing region of a 512×512 image patch, in a coarse-to-fine manner using a Titan-X GPU. The frame selection for inpainting the background is performed by the summation of pixel values in the forward frame difference process. If the sum of the difference pixels is small, then the current frame is selected for inpainting. The weighting factors in our proposed method have empirical values of 0.999 for η and $5e^{-6}$ for γ and δ in the implementation of our model, as shown in Eqs. (4) and (5), respectively.

5 Experiments

We evaluate our proposed approach on two different datasets, including Scene Background Modeling (SBM.net)[1] for background estimation and the Change Detection 2014 Dataset (CDnet2014) [69] for foreground detection. On both datasets, our proposed algorithm outperforms existing state-of-the -art algorithms by a significant margin.

5.1 Evaluation of deep context prediction for background estimation

We selected all videos from seven categories of the SBM.net dataset, as shown in Table 1. Each category in the SBM.net dataset contains challenging video sequences for background modeling. In this experiment, the results are compared with those of six state-of-the-art methods, including RFSA [19], GRASTA [24], BRTF [83], GOSUS [72], SSGoDec [84], and DECOLOR [85], using the implementations of the original authors. The background estimation models are compared in terms of the AGE, percentage of error pixels (pEPs), percentage of clustered error pixels (pCEPs), multiscale structural similarity index (MSSSIM), color image quality measure (CQM), and peak signal-to-noise ratio (PSNR) [42]. For the best performance, the aim is to minimize the AGE, pEPs, and pCEPs, while maximizing the MSSSIM, PSNR, and CQM (see Fig. 5). A detail description of results with respect to each category follows.

Category: Background Motion contains six video sequences. In this category, the proposed DCP algorithm achieved the best performance among all the compared methods. The performances of DECOLOR, SSGoDec, RFSA, GRASTA, and BRTF remained quite similar, with the minimal difference in the AGE shown in Table 1. GOSUS achieved the highest AGE among all the compared methods. Targeting only foreground objects to be eliminated and filling background pixel values via the inpainting method allow DCP to perform better in this category compared to the other methods. The visual results are presented in the first row of Fig. 3.

[1] http://scenebackgroundmodeling.net/.

Table 1 The AGE scores over the SBM.net dataset for six state-of-the-art methods compared with DCP for background subtraction

Category	Videos	AGE						
		DCP	RFSA [19]	GRASTA [24]	BRTF [83]	GOSUS [72]	SSGoDec [84]	DECOLOR [85]
Background Motion	Canoe	*6.3250*	14.8805	14.9438	14.8798	14.9677	14.9464	13.6732
	Advertisement Board	*2.3378*	3.4762	3.4812	3.4640	3.4733	3.4742	3.6604
	Fall	*19.0737*	24.3364	24.6026	24.4283	24.5935	24.5702	24.8117
	Fountain 01	9.6775	*5.7150*	5.7539	5.7383	5.7750	5.7442	6.2959
	Fountain 02	14.0579	7.3288	7.0811	7.3307	7.0867	7.0801	*6.4137*
	Overpass	*6.4089*	14.7162	14.7489	14.7183	14.7614	14.7369	8.6909
Average AGE		**9.6468**	11.7422	11.7686	11.7599	12.1183	11.5934	11.6340
Basic	511	*3.5786*	5.0972	6.1220	4.9151	5.2681	6.6025	7.5225
	Blurred	*2.1041*	4.9735	47.3112	4.9527	105.1528	51.9253	4.7345
	Camouflage FG Objects	*2.5789*	4.7951	5.0457	4.7411	4.3364	5.9418	4.4703
	Complex Background	6.3453	6.8593	6.2202	6.8868	6.1947	6.1828	*5.6215*
	Hybrid	*4.2021*	6.1795	6.5101	5.9201	6.4777	5.8003	6.6420
	IPPR2	*6.8575*	6.9256	6.9256	6.9249	6.9256	6.9256	6.9258
	I_SI_01	4.1119	3.2955	3.3333	3.2895	3.2895	3.3518	*2.5187*
	Intelligent Room	5.9152	3.3890	3.4871	3.3546	3.5144	3.4951	*3.3934*
	Intersection	*2.6911*	13.9704	13.9690	13.9752	13.9720	13.9726	13.1152
	MPEG4_40	3.9292	4.3346	5.6052	4.2712	5.5649	5.6711	*3.7329*
	PETS2006	6.5818	*4.7506*	5.5686	4.7573	5.4968	5.6115	5.5221
	Fluid Highway	*4.3549*	12.1362	9.3360	10.1921	9.3345	9.2739	10.1913
	Highway	4.8638	4.0454	4.1048	*4.0381*	4.0901	4.0941	4.0762
	Skating	*5.855*	26.0429	25.9610	26.1047	26.0922	25.9509	25.7092
	Street Corner at Night	*9.5308*	10.2057	10.1120	10.1791	10.0807	10.1170	12.9509
	wetSnow	*12.3658*	37.6461	37.7272	38.1130	37.7126	37.7054	38.6056

Table 1 continued

Category	Videos	AGE						
		DCP	RFSA [19]	GRASTA [24]	BRTF [83]	GOSUS [72]	SSGoDec [84]	DECOLOR [85]
Average AGE		**5.3666**	9.6654	12.3337	9.5385	15.8439	12.6639	9.7332
Intermittent Motion	AVSS2007	*7.3008*	21.3837	21.3776	21.3957	21.3896	21.3746	35.5689
	CaVignal	13.9885	1.6927	1.7240	1.7131	1.7379	1.7182	*1.3504*
	Candela m1.10	8.6512	*3.8845*	3.8889	3.9102	3.8977	3.9043	5.4697
	I_CA_01	16.8939	15.4821	15.4496	15.4985	15.4312	15.4297	*14.6558*
	I_CA_02	13.4803	9.9255	6.6146	9.8810	*6.2029*	7.1204	9.8810
	I_MB_01	9.2338	8.1882	7.3860	8.0478	*7.1827*	7.6550	11.5584
	I_MB_02	9.5397	8.6324	8.6360	8.6361	8.6307	8.6353	*3.6324*
	Teknomo	*4.8436*	6.7690	6.7382	6.7388	6.7315	6.7312	6.7310
	UCF-traffic	*4.1126*	33.0448	33.0449	33.0464	33.0426	33.0432	32.9837
	Uturn	*7.4448*	23.4947	23.5190	23.4939	23.5187	23.5163	21.2872
	Bus Station	8.9723	3.5451	*3.5409*	3.5513	3.5525	3.5474	6.5359
	Copy Machine	7.3156	8.1650	8.2640	8.1819	8.2836	8.2483	*4.9248*
	Office	16.6488	9.2656	9.1716	9.2710	9.1694	9.2024	*3.3454*
	Sofa	4.9927	4.2697	4.2711	4.2637	4.2708	4.2616	*4.1817*
	Street Corner	8.9535	*7.6411*	7.7734	7.6425	7.8462	7.6832	27.5613
	Tramstop	7.1293	2.4173	2.4268	2.4282	2.4483	2.4153	*2.4079*
Average AGE		**9.3438**	10.4876	10.2392	10.4812	10.2085	10.2804	12.0047
Jitter	CMU	8.1714	7.3476	6.9292	7.3197	7.7878	7.6034	*6.8975*
	I_MC_02	*9.0549*	15.7418	13.9334	15.4235	15.4017	15.6302	15.9440
	I_SM_04	4.5583	3.3464	*2.5355*	3.0923	3.7768	4.3339	4.1406
	O_MC_02	12.6371	16.3119	17.3914	16.6375	16.0443	16.4781	*12.3657*
	O_SM_04	*7.7459*	12.0224	12.0262	13.2998	15.6505	13.9053	15.6806
	Badminton	14.2284	16.9398	17.1787	16.4044	16.6059	14.2486	*6.6003*
	Boulevard	*11.5450*	19.4259	15.4555	16.6356	20.0932	16.9604	23.8209
	Sidewalk	*14.9378*	24.7621	24.1964	22.8313	16.5027	15.8949	18.4447
	Traffic	21.3232	7.5524	24.5624	8.6431	7.5449	*6.7522*	26.5434

Table 1 continued

Category	Videos	AGE						
		DCP	RFSA [19]	GRASTA [24]	BRTF [83]	GOSUS [72]	SSGoDec [84]	DECOLOR [85]
Average AGE		**11.5780**	13.7167	14.9121	13.3652	13.2675	12.4230	14.4931
Very Short	CUHK Square	*2.8429*	5.4994	4.8949	5.8176	5.2220	5.0429	6.2694
	Dynamic Background	13.7524	7.7233	7.8492	7.5747	7.9276	7.3880	*7.3760*
	MIT	*3.5838*	4.9527	5.7849	4.4991	5.8378	5.2764	4.9524
	Noisy Night	*3.9116*	6.1301	5.5040	6.3509	5.3378	5.6906	5.4483
	Toscana	11.5422	8.7331	6.4773	7.4022	6.8142	*6.3869*	7.4014
	Town Center	4.1427	4.4226	4.4247	4.2329	*3.8596*	3.9657	4.4225
	Two Leave Shop1cor	10.0183	4.0515	4.0172	4.2124	*3.9300*	3.8685	4.0503
	Pedestrians	5.0736	5.0318	*4.9441*	4.9996	4.9974	4.9682	5.0225
	People In Shade	6.9680	9.0900	6.5455	10.7783	*3.6842*	9.3889	10.7812
	SnowFall	*5.2768*	32.8871	31.0542	31.2511	31.8320	30.3902	34.2603
Average AGE		**6.7112**	8.8522	8.1496	8.7119	7.9443	8.2366	8.9984
Illumination Changes	Camera Parameter	6.2206	75.1204	6.1471	*6.1126*	6.1389	6.1475	45.2837
	Dataset3 Camera1	14.5708	23.3046	22.0816	22.5116	22.0816	22.0816	*2.8850*
	Dataset3 Camera2	18.7047	6.5041	5.7156	5.8965	5.7156	5.7156	*3.7555*
	I_IL_01	*7.4329*	8.3048	23.6585	23.5775	23.6585	23.6585	22.4594
	I_IL_02	19.3833	8.4842	7.5423	7.4007	7.5423	7.5423	*5.1225*
	Cubicle	*11.4636*	26.1490	19.4842	31.2116	19.4842	19.4842	13.0519
Average AGE		**12.9627**	24.6445	14.1049	16.1184	14.1035	14.1049	15.4263
Very Long	Bus Stop Morning	*3.1641*	5.6652	5.7055	5.6396	5.6739	5.6794	5.7419
	Dataset4 Camera1	6.7405	3.1857	3.1886	3.1876	3.1794	3.1948	*3.1661*
	Ped And Storrow Drive	8.5110	5.5780	5.0913	5.4323	5.3057	5.2445	*4.5065*
	Ped And Storrow Drive3	*2.8661*	3.5503	3.6693	3.5531	3.6100	3.5598	3.9688
	Terrace	*6.0016*	19.9480	18.9514	19.1109	19.0254	19.0258	10.2339
Average AGE		*5.4567*	7.5854	7.3212	7.3847	7.3589	7.3409	5.5234
Average AGE of all categories		**8.7237**	13.2359	11.9362	11.9229	12.1183	11.5934	11.6340

The best AGE score for each video sequence is shown in italics, and the best average AGE score for each category is shown in bold

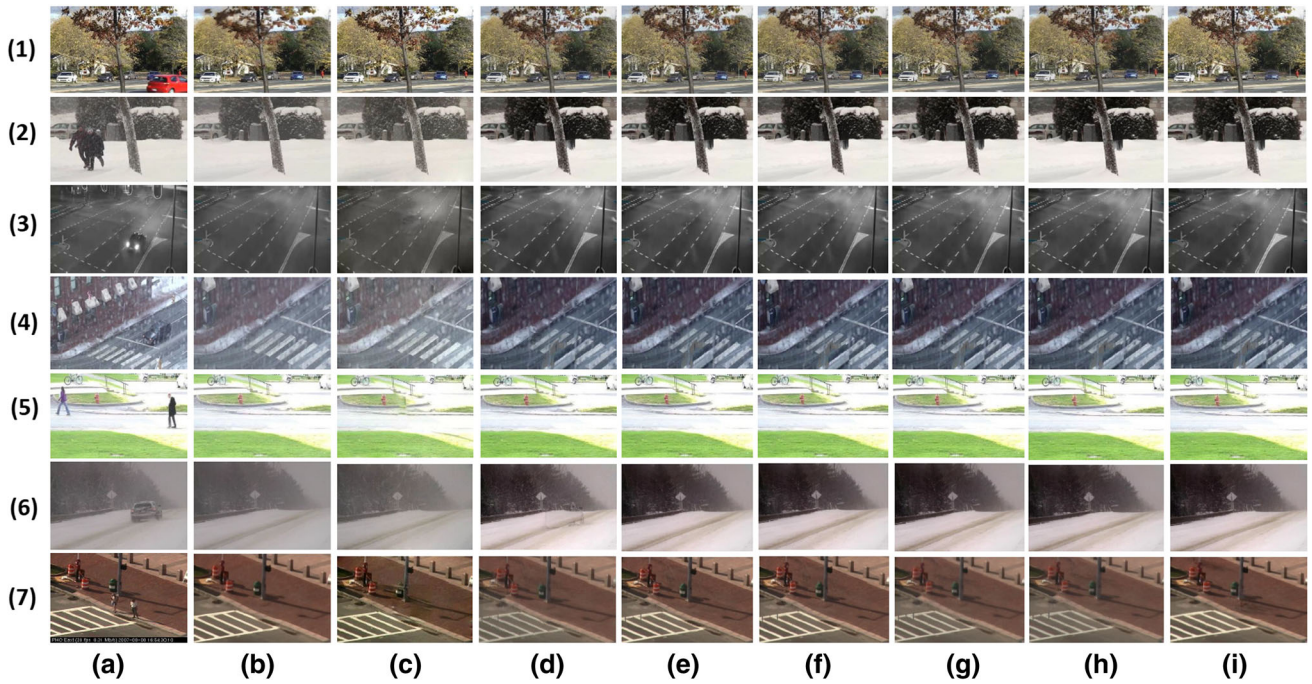

Fig. 3 Qualitative results for the proposed method. **a** Seven images from the input video sequences, **b** ground truth, background models estimated by **c** the proposed DCP method, **d** RFSA, **e** GRASTA, **f** BRTF, **g** GOSUS, **h** SSGoDec, and **i** DECOLOR. Each input sequence is selected from the following different categories from top to bottom: (1) Sequence "Fall" from "Background Motion," (2) "Skating" from "Basic," (3) "StreetCornerAtNight" from "Basic," (4) "WetSnow" from "Basic," (5) "Pedestrians" from "Very Short," (6) "Snowfall" from "Very Short," and (7) "SideWalk" from "Jitter"

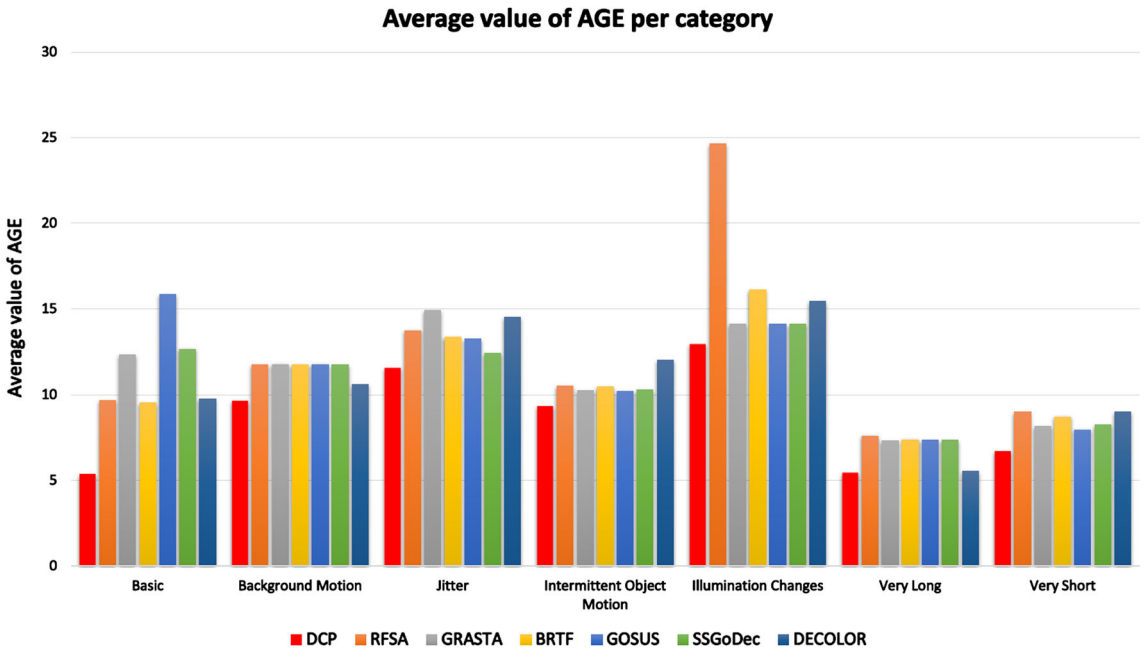

Fig. 4 Performance comparison of each method on the basis of the AGE according to each category on the SBM.net dataset

Category: Basic contains 16 video sequences (Table 1). Our proposed approach DCP performed effectively for almost all video sequences. DCP achieved an average AGE of 5.367 (visual results are shown in Fig. 4), because this category contains relatively simple scenes for background estimation. It can be observed in Table 1 that RFSA, BRTF,

and DECOLOR achieved an almost equal second lowest score for the AGE, while GOSUS and GRASTA achieved slightly higher values. GOSUS suffered from the largest performance degradation among the compared methods. In terms of a qualitative analysis, DCP estimated a better background compared to the other methods, and the results are shown in Fig. 3c–e. The reason for this is that the contexts for the "Wet-snow," "Skating," and "Street Corner at Night" video sequences are homogeneous in the whole frame as background pixel values. This key aspect is favorable for our proposed method.

Category: Intermittent Motion contains 16 video sequences (Table 1). This category contains video sequences that contain ghosting artifacts in the detected motion. DCP performed well in this category, achieving the lowest AGE score among all compared techniques of 9.344. The RFSA, GRASTA, BRTF, GOSUS, and SSGoDec methods achieved almost equal higher scores for the AGE (Fig. 4 and Table 1). DECOLOR exhibited the highest error rate for the background estimation on this category. The ghosting artifacts pose a significant challenge for all the algorithms as the foreground becomes the part of background, resulting in a failure in accurate background recovery.

Category: Jitter contains nine video sequences (Table 1). DCP achieved the lowest AGE among all the compared methods, owing to the fact that this category contains video sequences with blurry contexts, and such contexts are easy to predict using our proposed method. RFSA, BRTF, and GOSUS achieved higher AGE scores in this category, while GRASTA and DECOLOR exhibited the largest performance degradations among the compared methods. It can be observed in Fig. 1e that GRASTA was not able to recover a clean background, while DCP estimated it accurately. SSGoDec was also able to recover a clean background with a low AGE score, as shown in Fig. 3h.

Category: Very Short contains 10 video sequences, each with only few frames (Table 1). DCP again achieved the lowest AGE score in this category. GOSUS also performed well, achieving the second lowest AGE score as shown in Table 1. However, RFSA, GRASTA, BRTF, DECOLOR, and SSGoDec achieved almost equal AGE scores. In terms of a qualitative analysis, it can be seen in Fig. 3c that on the video sequence "SnowFall," for instance, DCP achieved the lowest AGE score. This is because in the case of bad weather, such as snow or rain, the contexts of videos become blurry and are reasonably easy for DCP to estimate.

Category: Illumination Changes contains six video sequences (Table 1). This category poses a great deal of challenges for all the methods. DCP managed to achieve the lowest AGE score among the compared methods, owing to the fact that context prediction in low light and with less sharp details comprises a reasonably favorable condition for our proposed method. GRASTA, GOSUS, and SSGoDec also performed well, achieving the second lowest AGE score among the compared methods. BRTF and DECOLOR obtained almost equal AGE scores. In addition, RFSA had the highest error rate, as shown in Table 1, because of the spatiotemporal smoothness of foreground and the correlation of the background constraint.

5.1.1 Overall performance comparison of DCP on background estimation

Upon averaging the results over all seven categories, DCP achieved an AGE of 8.724, which is smallest among the compared methods, as shown in Fig. 5a. For a fair comparison and broader evaluation than the AGE, the results for five other metrics have also been calculated. In Fig. 5b, it can be observed that DCP achieves the smallest pCEPS among the compared methods. This is followed by BRTF, GOSUS, and SSGoDec, and the GRASTA, RFSA, and DECOLOR methods exhibited the highest pCEPs scores. A minimum score for the pEPs metric indicates an accurate background estimation (Fig. 5c). Among the compared methods, only DCP achieved the minimum score, while the other compared methods all achieved similar pEPs scores. In Fig. 5d, it is shown that DCP achieved the maximum (best) score for the CQM metric. It can also be observed in the visual results (Fig. 3d–i) that the color quality for some background images extracted by the compared methods differs from the input images, ground truths, and backgrounds estimated by DCP. For this reason, all the compared methods have different scores for the CQM metric. In Fig. 5e, f, for the PSNR and MSSSIM metrics, respectively, the aim is to achieve the highest value for the best performance. The proposed DCP algorithm achieved

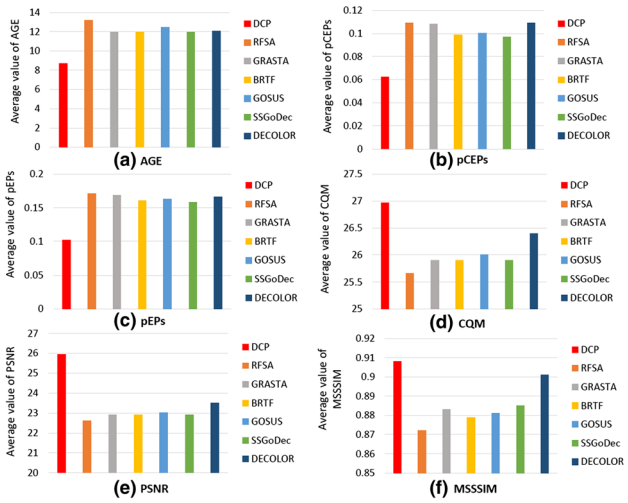

Fig. 5 Average performance comparison between DCP and six state-of-the-art methods on each metric on the seven categories of the SBM.net dataset. **a** AGE, **b** pCEPs, and **c** pEPs (minimum is best). **d** CQM, **e** PSNR, and **f** MSSSIM (maximum is best)

the highest values in both cases. Overall, DCP achieved the best scores among the compared methods for all considered metrics.

5.2 Evaluation of deep context prediction for foreground detection

We selected seven categories from the CDnet2014 [69] dataset. The results are compared with six state-of-the-art methods, namely MSSTBM [40], GMM-Zivkovic [86], CP3-Online [36], GMM-Stauffer [58], KDE-Elgammal [13], and RMoG [60], using the implementations of the original authors. The foreground detection performance is compared using average F measure across all the video sequences within each category. The metrics used to calculate F measure are as follows:

$$\text{Re} = \frac{T_p}{T_p + F_n}, \tag{12}$$

$$\text{Sp} = \frac{T_n}{T_n + F_p}, \tag{13}$$

$$\text{FNR} = \frac{F_n}{T_p + F_n}, \tag{14}$$

$$\text{PWC} = 100 \times \left(\frac{F_n + F_p}{T_p + F_n + F_p + T_n} \right), \tag{15}$$

$$\text{Pre} = \frac{T_p}{T_p + F_p}, \quad and \tag{16}$$

$$F = \frac{2(Pre \times Re)}{Pre + Re}, \tag{17}$$

where T_p indicates true positives, T_n is true negatives, F_p is false positives, F_n is false negatives, Re is recall, Sp is specificity, FNR is the false negative rate, PWC is the percentage of wrong classifications, Pre is precision, and F is the F-measure. Detailed explanations of the results on the seven categories of the CDnet2014 dataset follow.

Category: Baseline in CDnet2014 dataset contains four video sequences. The average F-measure scores across all four video sequences are presented in Table 2. All the compared methods, including DCP, achieved scores of over 0.8 for this category (Table 2). However, KDE-ElGammal obtained the highest score among the compared methods, with CP3-Online in second position. Although DCP achieved an F-measure score of over 0.8, it was unable to outperform the KDE-ElGammal method, owing to the fact that successive opening and closing on noisy video frames lead to false detections. The visual results are presented in the first row of Fig. 6.

Category: Camera Jitter also contains four video sequences. DCP achieved the highest F-measure score among the compared methods, as shown in Table 2. The is because the blurry contexts resulting from camera jitters are easy to predict using our proposed method to obtain an accurate background estimation. Subsequently, the binarized thresholded difference between the estimated background and current frame erodes the noisy pixels of the background in successive opening and closing operations. This leads us to obtain an accurate foreground detection performance with fewer missing pixel values for foreground objects (third row of Fig. 6). RMOG also performed well in this category, achieving the second best score among the compared methods.

Category: Shadow contains six video sequences. MSSTBM achieved the highest score among the compared methods, with RMOG the second best. This category posed challenges to our proposed method, as sometimes shadows became replicated in the context prediction algorithm, which generated errors in the background estimation as well as the foreground detection. For our proposed method, the opening and closing of the binarized thresholded difference frame successfully filled the missing values in the foreground compared to the other methods, as shown in the sixth row of Fig. 6. This led to DCP achieving the third best F-measure in this category.

Category: Dynamic Background also contains six video sequences. Here, as shown in Table 2, DCP achieved the highest averaged F-measure score among the compared methods. The homogeneous contexts in the video sequences of this category represent a favorable condition for our proposed method. RMOG also performed well, achieving the second best F-measure score. The qualitative results are presented in Fig. 6. It can be seen from the visual results that the successive opening and closing with a suitable SE removed the noisy pixel values for a moving background.

Category: Thermal contains five video sequences, which have been captured by a far-infrared camera. DCP achieved the highest averaged F-measure score among the compared methods, with CP3-Online being the second best. This is for the same reason as explained for the previous category. A homogeneous context is one of the major factors for an accurate background estimation using DCP, and this leads to a noiseless foreground detection performance. The seventh row of Fig. 6 shows that all methods, including DCP, accurately detected the foreground object, except for RMOG, which exhibits missing pixel values within the detected foreground object.

Category: Intermittent Object Motion contains six video sequences, involving scenarios known to cause ghosting artifacts in detected motion, i.e., objects move, then stop for a short while, after which they begin moving again. DCP achieved the highest averaged F-measure score in this category, with RMOG being the second best among the compared methods. The main reason behind this is that our proposed approach does not contain any motion-based constraints for moving foreground objects. The compared all methods contain constraints on the motion of foreground objects, which

Table 2 Comparison of six state-of-the-art methods with the proposed DCP algorithm using the F measure on the CDnet2014 dataset

Categories	MSSTBM [40]	GMM-Zivkovic [86]	CP3-Online [36]	GMM-Stauffer [58]	KDE-ElGammal [13]	RMOG [60]	DCP
Baseline	0.8450	0.8382	*0.8856*	0.8245	**0.9092**	0.7848	0.8187
Camera Jitter	0.5073	0.5670	0.5207	0.5969	0.5720	*0.7010*	**0.8376**
Shadow	**0.8130**	0.7232	0.6539	0.7156	0.7660	*0.8073*	0.7665
Dynamic Background	0.5953	0.6328	0.6111	0.6330	0.5961	*0.7352*	**0.7757**
Thermal	0.5103	0.6548	*0.7917*	0.6621	0.7423	0.4788	**0.8212**
Intermittent Object Motion	0.4497	0.5325	**0.6177**	0.5207	0.4088	0.5431	*0.5979*
Bad Weather	0.6371	0.7406	0.7485	0.7380	*0.7571*	0.6826	**0.8212**
Average	0.6225	0.6736	*0.7010*	0.6771	0.6833	0.6761	**0.7620**

The first and second highest scores for each category are shown in bold and italics, respectively

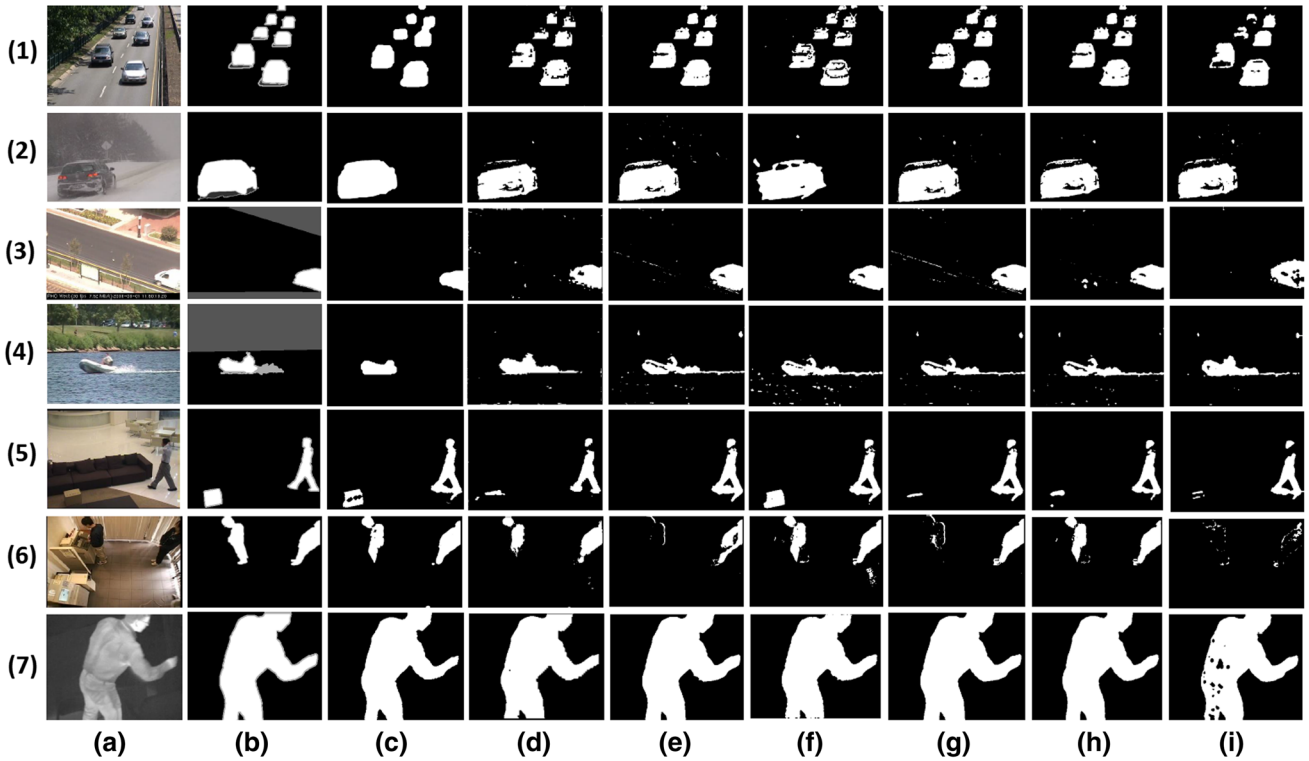

Fig. 6 Qualitative results of the proposed DCP method: **a** Seven images from the input video sequences of Cdnet2014 dataset, **b** ground truth, **c** foreground detected by the proposed DCP method, **d** MSSTBM, **e** GMM-Zivkovic, **f** CP3-Online, **g** GMM-Stauffer, **h** KDE-ElGammal, **i** RMOG. From top to bottom: each input sequence is selected from different category: (1) Sequence "Highway" from "Baseline," (2) "Snowfall" from "Bad Weather," (3) "Boulevard" from "Camera Jitter," (4) "Boats" from "Dynamic Background," (5) "Sofa" from "Intermittent Object Motion," (6) "Copy Machine" from "Shadow," and (7) "Library" from "Thermal"

if violated lead to false detection and a low F-measure score. The visual results shown in the fifth row of Fig. 6 demonstrate that the foreground objects vanish if motion-based constraints are violated.

Category: Bad Weather contains four video sequences captured in challenging winter weather conditions, i.e., snow storms, snow on the ground, and fog. DCP achieved highest averaged F-measure score among the compared methods, with KDE-ELGammal being the second best method. This category is another example of homogeneous contexts in video sequences. It can be observed in the visual results in the second row of Fig. 6 that DCP accurately estimated the foreground object with almost no unconnected noisy pixels of the background compared to the other methods.

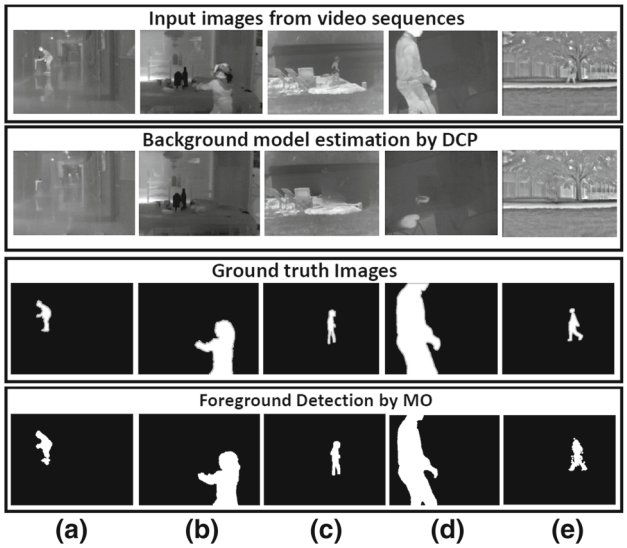

Fig. 7 Estimated background examples from the CDnet2014 dataset: All sequences are taken from the category "Thermal." **a** "Corridor" sequence, **b** "Dining Room" sequence, **c** "Lake Side" sequence, **d** "Library" sequence, and **e** "Park" sequence. For all of these video sequences, DCP estimated an accurate background, which also leads to better foreground detection

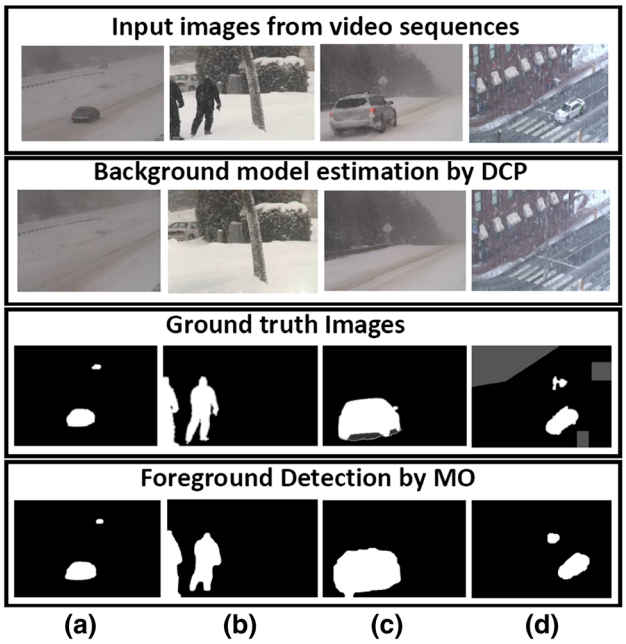

Fig. 8 Estimated background examples from the CDnet2014 dataset: The sequences are **a** "Blizzard," **b** "Skating," **c** "SnowFall," and **d** "Wet-Snow," all from the "Bad Weather" category. In all of these video sequences, DCP estimated an accurate background, which also leads to better foreground detection

5.2.1 Overall performance comparison of DCP for foreground detection

Table 2 shows that DCP achieved the highest averaged F-measure score across all categories. CP3-Online achieved the second best performance. GMM-Stauffer, GMM-Zivkovic, KDE-ElGammal, and RMOG achieved almost equal F-measure scores, and MSSTBM achieved the lowest score among the compared methods (Table 2). For a better foreground detection, the aim for the metrics [defined in (12), (13), (14), (15), and (16)] is to maximize the values of Re, Sp, and Precision and to minimize the values of FNR and PWC. The proposed DCP algorithm achieved the top scores for the Re and FNR among the compared methods, of 0.809 and 0.191, respectively. This means that our proposed method achieved more correct detection and less incorrect detection for foreground objects. Moreover, for the PWC, Sp, and Precision metrics, DCP achieved the best scores of 2.671, 0.977, and 0.773, respectively, which are higher than most of the other methods.

5.3 Performance of DCP on the basis of a homogeneous context

As explained in Sect. 3, our proposed method estimates the background on the basis of context prediction, and so in this section we discuss the key aspects of DCP concerning the types of context present in video sequences containing different scenes, specifically for application to background estimation. Table 1 shows that the AGE score is different for all categories and even between individual videos in the same category, for all the compared methods including DCP. The reason for this is that the context of every video is different, with different kinds of indoor and outdoor scenes. Therefore, for the compared methods, including DCP, the average gray-level score is different, which presents quite a challenge in some cases. For convenience, we target the discussion of homogeneous contexts on a few video sequences in the SBM.net and CDnet2014 datasets. We selected two categories from the CDnet2014 dataset on the basis of their homogeneous contexts in the video sequences. A category-wise discussion follows.

Category: Bad Weather is an example from the CDnet2014 dataset with similar contexts. Figure 8 presents the visual results for the video sequence "blizzard." The other three video sequences are the "skating," "wetsnow," and "snowfall" sequences from the "Basic" category in the SBM.net dataset. These video sequences exhibit the minimum AGE score, and their visual results are presented in the second, fourth, and sixth rows, respectively, of Figs. 3c and 8.

Category:Thermal is another challenging category in CDnet2014 dataset, which includes videos that have been captured by far-infrared cameras. The interesting fact concerning this category is it includes video sequences with thermal artifacts such as heat stamps, heat reflection on floors, windows, camouflage effects, and moving objects that have

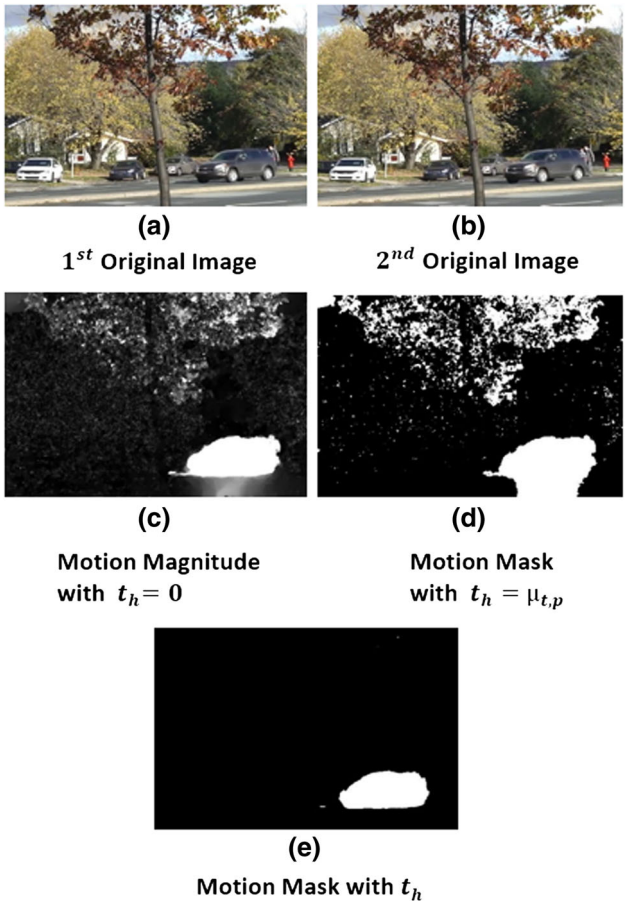

Fig. 9 Motion masks estimated by the optical flow. The threshold t_h (in Eq. (18)) is kept large to eliminate this kind of noise while identifying fast-moving objects via the optical flow method

the same temperature as the surrounding region.[2] This is a highly favorable environment for context prediction using DCP. The visual results for all five video sequences in this category are presented in Fig. 7.

5.4 Ablation study

Because our proposed framework consists of five steps, in this section we present a discussion of each step of our proposed method, with reference to its effects on the background estimation and foreground detection.

5.4.1 Selection of t_h for motion mask estimation

The first step of our proposed method involves motion mask estimation via the optical flow, as explained in Sect. 3.1. It can be seen in equation (1) that t_h plays an important role in thresholding the noisy pixels of the background while estimating the moving foreground pixels. For challenging

[2] http://jacarini.dinf.usherbrooke.ca/datasetOverview/.

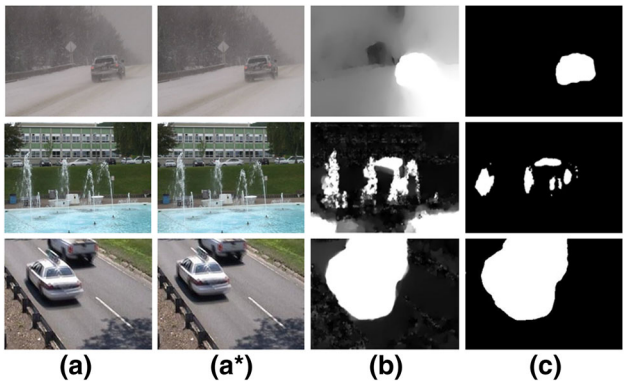

Fig. 10 Estimated motion mask examples: **a** first original image (S_t), (a*) second original image (S_{t-1}), as mentioned in Sect. 3.1. **b** Motion masks calculated without t_h. **c** Motion masks calculated with t_h. Row (1) "SnowFall" sequence, row (2) "fountain01" sequence, and row (3) "traffic" sequence. All three sequences presented in this figure are included in the SBM.net and CDnet2014 datasets under the following categories: row (1) "Very Short" and "Bad Weather," row (2) "Background Motion" and "Dynamic Background," and row (3) "Jitter," and "Camera Jitter"

backgrounds, where the assumption of a static background and dynamic foreground is violated, a higher value of t_h will remove the moving background pixels. For instance, it can be seen in Fig. 9 that in the dynamic background case the motion mask estimated with $t_h = 0$ contains considerable noise. Figure 9d shows that noisy pixels are still present if $t_h = \mu_{t,p}$, which is the mean value of the motion magnitude. Therefore, we propose setting t_h as follows:

$$t_h = \frac{\max(m_{t,p}) + \mu_{t,p}}{2} \tag{18}$$

which turns out to be effective in removing noise for dynamic backgrounds (see Fig. 10). Selecting $t_h = \max(m_{t,p})$, the maximum value of the motion magnitude could potentially lead to hard thresholding on fast-moving objects, resulting in a loss of valuable information.

5.4.2 Comparison of background samples estimated by context prediction and texture optimization in reference to MPB

It is shown in Fig. 2 that we use AlexNet for context prediction, because of its simple architecture and efficient optimization. Because our first objective is to initialize the context prediction term with a less deep network, AlexNet represents the best choice compared to other deeper networks [65,67]. AlexNet is trained from scratch on ImageNet for context prediction rather than image classification, as explained in Sect. 3.2. The initialized context prediction term helps the texture optimization stage to achieve better inpainting results in a coarse-to-fine manner. For the case of background estimation via DCP, it can be seen from

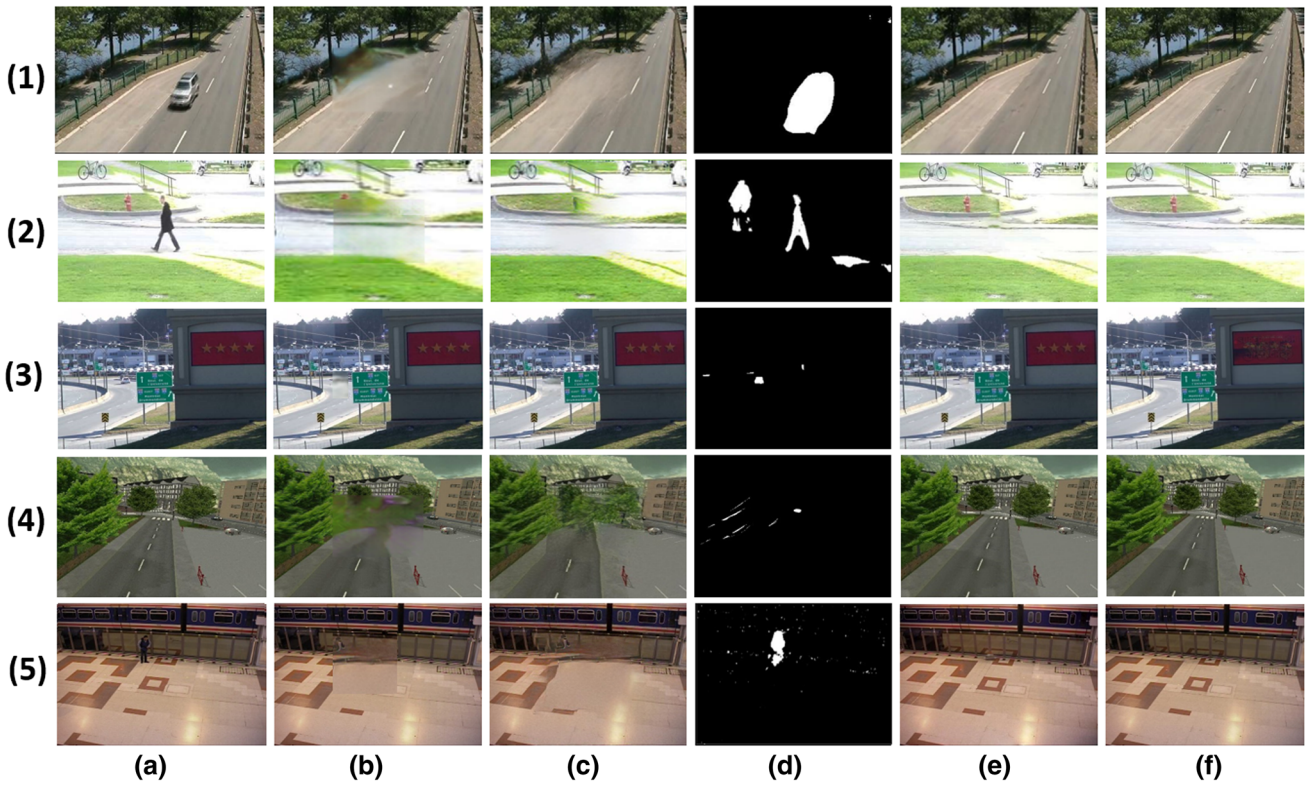

Fig. 11 Estimated background examples from the SBM.net dataset: **a** original images, **b** background pixels estimated by context network, **c** background estimated by texture network, **d** motion masks, **e** post-processing by MPB, and **f** ground truths. Each input sequence is selected from the following different categories from the SBM.net dataset from top to bottom: (1) "PedAndStorrowDrive3" from "Very Long," (2) "Pedestrians" from "Very Short," (3) "advertisementBoard" from "Background Motion," (4) "511" from "Basic," and (5) "$PETS2006$" from "Basic"

Fig. 11b that the context estimated by CE is rather blurry, with less details texture-wise. However, the results presented in Fig. 11c show the performance of texture optimization for fine context prediction, hence achieving a better background model. After texture optimization, the next step is to transform this square-shaped context to an irregular shape, where MPB plays an important role. In our proposed method, the role of MPB is only to transfer the regular-shaped predicted context to an irregular one using motion masks, thus making it different from previously proposed methods such as [33,49]. The importance of post-processing by the MPB method can be seen in Fig. 11, where the context optimized by the texture network in Fig. 11c in a regular shape is transformed to an irregular shape (Fig. 11e) using motion masks, as shown in Fig. 11d. This step helps to discard extra information, which is introduced in the square-shaped region inpainting process. Similarly, for the case of foreground detection Fig. 12 presents the results for each step of our proposed method. It can be seen from Fig. 12b that after texture optimization the background pixels are estimated in a coarse-to-fine manner by our texture network. However, the extra information is discarded by the MPB technique. For instance, the first row in Fig. 12a shows a sequence from the CDnet2014 dataset, in which a foreground moving object is identified by the motion mask presented in the first row of Fig. 12c. It is clearly visible in this figure that the running fountain in the scene is part of the background information that is removed during the context prediction in the square-shaped inpainting process (first row of Fig. 12b). The MPB technique efficiently recovers this lost information of the background by discarding the extra predicted context with the help of motion masks. Thus, this shows the importance of post-processing with the help of the MPB technique.

5.5 Background estimation via DCP in cluttered/occluded scenes

"Occlusion" in background scenes means that some scene information is hidden by foreground objects. The proposed method can handle occlusions at some level, with many challenging conditions in real-time background estimation scenarios. Nevertheless, occlusions could be at a severe level, hiding most of the background information. For instance, as mentioned in Sect. 5, for the purpose of background esti-

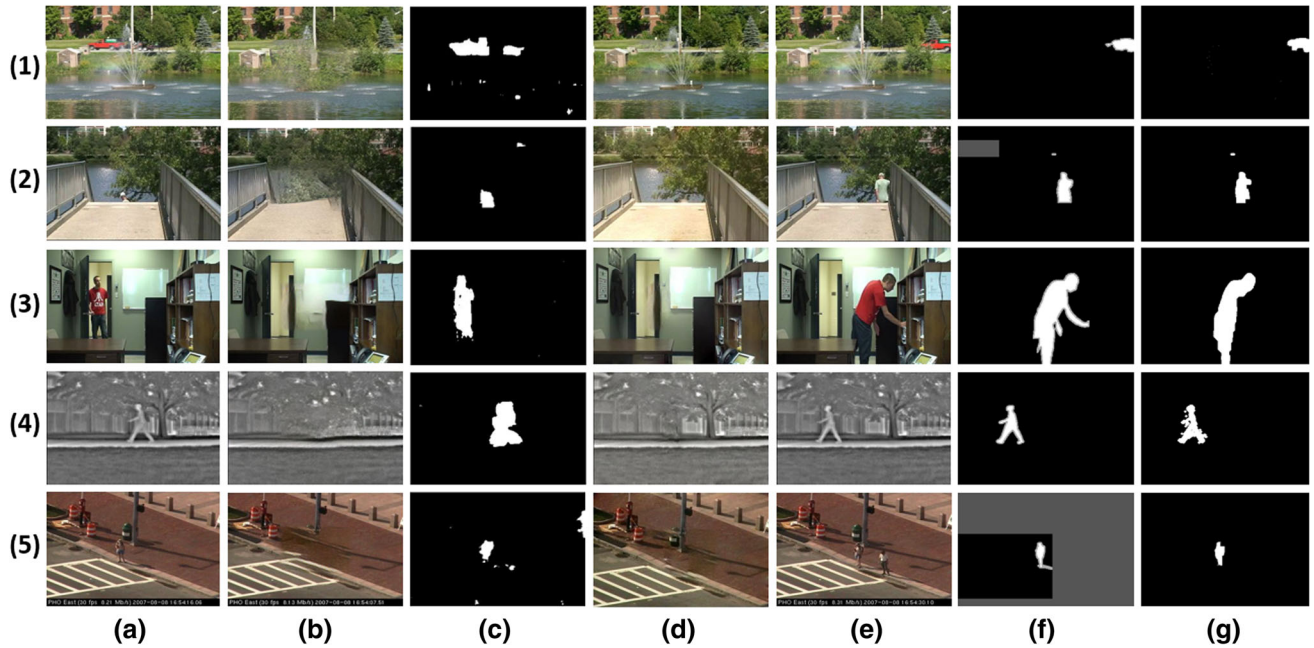

Fig. 12 Estimated foreground examples from the CDnet2014 dataset: **a** original images, **b** background pixels estimated by texture network, **c** motion masks via optical flow, **d** estimated background images, **e** randomly selected original images, **f** ground truths, and **g** detected foregrounds. Each input sequence is selected from the following different categories from the CDnet2014 dataset from top to bottom: (1) "fountain02" from "Dynamic Background," (2) "overpass" from "Dynamic Background," (3) "office" from "Baseline," (4) "park" from "Thermal," and (5) "sidewalk" from "Camera Jitter"

mation we selected seven out of eight categories from the SBM.net dataset. The eighth category in this dataset is "Clutter," which is also known as "Occlusion," representing video sequences in which most of the background information is occluded. This kind of scenario is quite challenging for our method, because if most of the background information is not available for the missing region prediction, then inpainting models, including our proposed method, suffer from performance degradation. The qualitative results are presented in Fig. 13. However, occlusion in visual object tracking applications, such as [11,52], is quite different from occlusion in foreground object detection. The main reason is that object tracking occlusion can comprise either self-occlusion or inter-object occlusion [12], which is totally different from cluttered background scenarios.

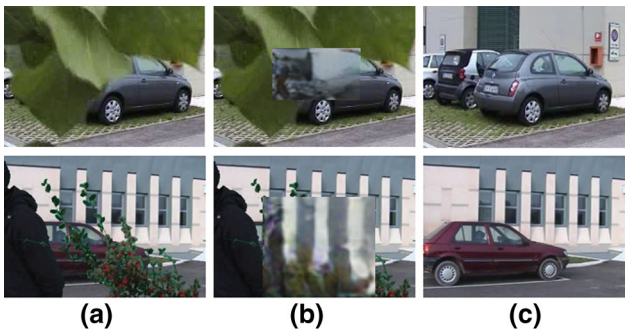

Fig. 13 Estimated background example from the SBM.net dataset: Row (1) Sequence "Foliage" from category "Clutter," Row (2) sequence "Foliage&People" also from category "Clutter" (**a**) original images (**b**) estimated background by DCP (**c**) Ground truths

5.6 Failure cases for DCP

Although DCP achieved a good performance in most cases, it still has some limitations and failure cases. The estimation of complex background structures (Fig. 14) and large-scale foreground objects is quite challenging. The limitation of the proposed method involves large foreground objects to be accurately inpainted. In these cases, the network is not able to correctly fill the region in an irregular shape. We used the Poisson blending technique to transform the center region of the inpainting context to an irregular region.

6 Conclusion

In this work, a unified method called DCP is proposed for background estimation and foreground segmentation using GAN and image inpainting. The proposed method is based on an unsupervised visual feature learning-based hybrid GAN model for context prediction, along with a seman-

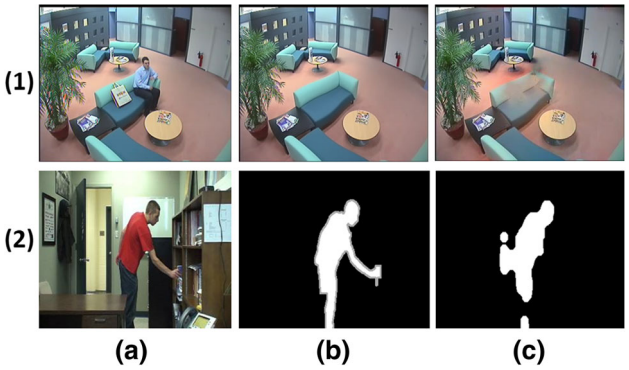

Fig. 14 Estimated background example from the SBM.net dataset: row (1) **a** the sequence *Candela m*1.10, **b** ground truth, **c** estimated background by DCP. Table 1 shows that for the category "Intermittent Motion" the AGE of DCP for this video sequence is larger than for all the compared methods. Estimated foreground example from the CDnet2014 dataset: row (2) **a** the sequence "office" from the category "Baseline," **b** ground truth, **c** estimated foreground by DCP. Table 2 shows that for the category "Baseline," the F-measure of DCP is smaller than for all the compared methods

tic inpainting network for texture optimization. A solution for random region inpainting is also proposed using center region inpainting and Poisson blending. The proposed DCP algorithm is compared with 12 existing algorithms: six for background estimation on the SBM.net dataset and six for foreground detection on the CDnet2014 dataset. The proposed algorithm outperforms the compared methods by a significant margin. Our experiments demonstrate the effectiveness of the proposed approach compared to the existing algorithms. The proposed algorithm demonstrates excellent results in poor weather and thermal imaging categories, where most of the existing algorithms suffer from performance degradation. Nevertheless, in order to mitigate the failure cases for our proposed method, we consider that adapting deeper networks such as [65,67] instead of using AlexNet will be an important direction for future work.

Acknowledgements This research was supported by Development project of leading technology for future vehicle of the business of Daegu metropolitan city (No. 20171105).

References

1. Afifi, M., Hussain, K.F.: Mpb: a modified poisson blending technique. Comput. Vis. Media **1**(4), 331–341 (2015)
2. Bengio, Y., et al.: Learning deep architectures for ai. Found. Trends Mach. Learn. **2**(1), 1–127 (2009)
3. Bouwmans, T., Zahzah, E.H.: Robust pca via principal component pursuit: a review for a comparative evaluation in video surveillance. Comput. Vis. Image Underst. **122**, 22–34 (2014)
4. Bouwmans, T., Maddalena, L., Petrosino, A.: Scene background initialization: a taxonomy. Pattern Recognit. Lett. **96**, 3–11 (2017)
5. Bouwmans, T., Javed, S., Zhang, H., Lin, Z., Otazo, R.: On the applications of robust pca in image and video processing. Proc. IEEE **106**(8), 1427–1457 (2018)
6. Braham, M., Van Droogenbroeck, M.: Deep background subtraction with scene-specific convolutional neural networks. In: 2016 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–4. IEEE (2016)
7. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11 (2011)
8. Cao, X., Yang, L., Guo, X.: Total variation regularized rpca for irregularly moving object detection under dynamic background. IEEE Trans. Cybern. **46**(4), 1014–1027 (2016)
9. Chen, M., Wei, X., Yang, Q., Li, Q., Wang, G., Yang, M.H.: Spatiotemporal GMM for background subtraction with superpixel hierarchy. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1518–1525 (2017)
10. Colombari, A., Cristani, M., Murino, V., Fusiello, A.: Exemplar-based background model initialization. In: Proceedings of the third ACM International Workshop on Video Surveillance & Sensor Networks, pp. 29–36. ACM (2005)
11. Dong, X., Shen, J., Yu, D., Wang, W., Liu, J., Huang, H.: Occlusion-aware real-time object tracking. IEEE Trans. Multimed. **19**(4), 763–771 (2017)
12. Dong, X., Shen, J., Wang, W., Liu, Y., Shao, L., Porikli, F.: Hyperparameter optimization for tracking with continuous deep q-learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
13. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: European Conference on Computer Vision, pp. 751–767. Springer, Berlin (2000)
14. Erichson, N.B., Donovan, C.: Randomized low-rank dynamic mode decomposition for motion detection. Comput. Vis. Image Underst. **146**, 40–50 (2016)
15. Fu, H., Cao, X., Tu, Z.: Cluster-based co-saliency detection. IEEE Trans. Image Process. **22**(10), 3766–3778 (2013)
16. Gao, Z., Cheong, L.F., Wang, Y.X.: Block-sparse RPCA for salient motion detection. IEEE T-PAMI **36**(10), 1975–1987 (2014)
17. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2016)
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
19. Guo, X., Wang, X., Yang, L., Cao, X., Ma, Y.: Robust foreground detection using smoothness and arbitrariness constraints. In: European Conference on Computer Vision, pp 535–550. Springer, Berlin (2014)
20. Haines, T.S., Xiang, T.: Background subtraction with dirichletprocess mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4), 670–683 (2014)
21. Han, J., Cheng, G., Li, Z., Zhang, D.: A unified metric learning-based framework for co-saliency detection. In: IEEE Transactions on Circuits and Systems for Video Technology (2017)
22. Han, J., Quan, R., Zhang, D., Nie, F.: Robust object co-segmentation using background prior. IEEE Trans. Image Process. **27**(4), 1639–1651 (2018a)
23. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process. Mag. **35**(1), 84–100 (2018b)
24. He, J., Balzano, L., Szlam, A.: Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1568–1575. IEEE (2012)

25. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
26. Javed, S., Oh, S.H., Bouwmans, T., Jung, S.K.: Robust background subtraction to global illumination changes via multiple features-based online robust principal components analysis with markov random field. J. Electron. Imaging **24**(4), 043011 (2015)
27. Javed, S., Jung, S.K., Mahmood, A., Bouwmans, T.: Motion-aware graph regularized RPCA for background modeling of complex scenes. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 120–125. IEEE (2016)
28. Javed, S., Mahmood, A., Bouwmans, T., Jung, S.K.: Spatiotemporal low-rank modeling for complex scene background initialization. In: IEEE Transactions on Circuits and Systems for Video Technology (2016)
29. Javed, S., Mahmood, A., Bouwmans, T., Jung, S.K.: Background-foreground modeling based on spatiotemporal sparse subspace clustering. IEEE Trans. Image Process. **26**(12), 5840–5854 (2017a)
30. Javed, S., Mahmood, A., Bouwmans, T., Jung, S.K.: Background-Foreground Modeling Based on Spatiotemporal Sparse Subspace Clustering. IEEE T-IP (2017)
31. Javed, S., Mahmood, A., Al-Maadeed, S., Bouwmans, T., Jung, S.K.: Moving object detection in complex scene using spatiotemporal structured-sparse RPCA. IEEE Trans. Image Process. **28**, 1007–1022 (2018)
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Bartlett, P.L, Pereira, F.C.N, Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 1097–1105. Neural Information Processing Systems Conference (2012)
33. Kwok, T.H., Sheung, H., Wang, C.C.: Fast query for exemplar-based image completion. IEEE Trans. Image Process. **19**(12), 3106–3115 (2010)
34. Li, X., Zhao, B., Lu, X.: A general framework for edited video and raw video summarization. IEEE Trans. Image Process. **26**(8), 3652–3664 (2017)
35. Li, X., Zhao, B., Lu, X.: Key frame extraction in the summary space. IEEE Trans. Cybern. **48**(6), 1923–1934 (2018)
36. Liang, D., Hashimoto, M., Iwata, K., Zhao, X., et al.: Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes. Pattern Recognit. **48**(4), 1374–1390 (2015)
37. Lim, L.A., Keles, H.Y.: Foreground Segmentation Using a Triplet Convolutional Neural Network for Multiscale Feature Encoding (2018). arXiv preprint arXiv:1801.02225
38. Lim, L.A., Keles, H.Y.: Learning Multi-scale Features for Foreground Segmentation (2018). arXiv preprint arXiv:1808.01477
39. Liu, C., et al.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis. PhD thesis, Massachusetts Institute of Technology (2009)
40. Lu, X.: A multiscale spatio-temporal background model for motion detection. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 3268–3271. IEEE (2014)
41. Lu, X., Li, X.: Group sparse reconstruction for image segmentation. Neurocomputing **136**, 41–48 (2014)
42. Maddalena, L., Petrosino, A.: Towards benchmarking scene background initialization. In: International Conference on Image Analysis and Processing, pp. 469–476. Springer, Berlin (2015)
43. Nakashima, Y., Babaguchi, N., Fan, J.: Automatic generation of privacy-protected videos using background estimation. In: 2011 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2011)
44. Ortego, D., SanMiguel, J.C., Martínez, J.M.: Rejection based multipath reconstruction for background estimation in video sequences with stationary objects. Comput. Vis. Image Underst. **147**, 23–37 (2016)
45. Park, D., Byun, H.: A unified approach to background adaptation and initialization in public scenes. Pattern Recognit. **46**(7), 1985–1997 (2013)
46. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
47. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. **22**(3), 313–318 (2003)
48. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
49. Shen, J., Jin, X., Zhou, C., Wang, C.C.: Gradient based image completion by solving the poisson equation. Comput. Graph. **31**(1), 119–126 (2007)
50. Shen, J., Hao, X., Liang, Z., Liu, Y., Wang, W., Shao, L.: Real-time superpixel segmentation by dbscan clustering algorithm. IEEE Trans. Image Process. **25**(12), 5933–5942 (2016)
51. Shen, J., Peng, J., Dong, X., Shao, L., Porikli, F.: Higher order energies for image segmentation. IEEE Trans. Image Process. **26**(10), 4911–4922 (2017a)
52. Shen, J., Yu, D., Deng, L., Dong, X.: Fast online tracking with detection refinement. IEEE Trans. Intell. Transp. Syst. **19**, 162–173 (2017b)
53. Shen, J., Peng, J., Shao, L.: Submodular trajectories for better motion segmentation in videos. IEEE Trans. Image Process. **27**(6), 2688–2700 (2018)
54. Shimada, A., Nagahara, H., Taniguchi, R.I.: Background modeling based on bidirectional analysis. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1979–1986. IEEE (2013)
55. Simonyan, K., Zisserman, A: Very Deep Convolutional Networks for Large-Scale Image Recognition (2014). arXiv preprint arXiv:1409.1556
56. Sobral, A., Zahzah, Eh: Matrix and tensor completion algorithms for background model initialization: a comparative evaluation. Pattern Recognit. Lett. **96**, 22–33 (2017)
57. Sobral, A., Bouwmans, T., Zahzah, E.H.: Comparison of matrix completion algorithms for background initialization in videos. In: International Conference on Image Analysis and Processing, pp 510–518. Springer, Berlin (2015)
58. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246–252. IEEE (1999)
59. Tsai, C.C., Qian, X., Lin, Y.Y.: Segmentation guided local proposal fusion for co-saliency detection. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 523–528. IEEE (2017)
60. Varadarajan, S., Miller, P., Zhou, H.: Spatial mixture of gaussians for dynamic background modelling. In: 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 63–68. IEEE (2013)
61. Vaswani, N., Bouwmans, T., Javed, S., Narayanamurthy, P.: Robust PCA and Robust Subspace Tracking (2017). arXiv preprint arXiv:1711.09492
62. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001 (CVPR 2001), vol. 1, pp. I-I. IEEE (2001)
63. Wang, W., Shen, J., Shao, L.: Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans. Image Process. **24**(11), 4185–4196 (2015)
64. Wang, W., Shen, J., Sun, H., Shao, L.: Vicos2: video co-saliency guided co-segmentation. IEEE Trans. Circuits Syst. Video Technol. **28**, 1727–1736 (2017)

65. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. IEEE Trans. Pattern Anal. Mach. Intell. 1–16 (2018)
66. Wang, W., Shen, J., Porikli, F., Yang, R.: Semi-supervised video object segmentation with super-trajectories. IEEE Trans. Pattern Anal. Mach. Intell. (2018)
67. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. **27**(1), 38–49 (2018c)
68. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **1**, 20–33 (2018d)
69. Wang, Y., Jodoin, P.M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P.: Cdnet 2014: an expanded change detection benchmark dataset. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 393–400. IEEE (2014)
70. Wang, Y., Luo, Z., Jodoin, P.M.: Interactive deep learning method for segmenting moving objects. Pattern Recognit. Lett. **96**, 66–75 (2017b)
71. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: Advances in Neural Information Processing Systems, pp. 2080–2088. Curran Associates, Inc (2009)
72. Xu, J., Ithapu, V., Mukherjee, L., Rehg ,J,. Singh, V.: Gosus: Grassmannian online subspace updates with structured-sparsity. In: ICCV (2013)
73. Xu, J., Ithapu, V.K., Mukherjee, L., Rehg, J.M., Singh, V.: Gosus: Grassmannian online subspace updates with structured-sparsity. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3376–3383. IEEE (2013)
74. Xu, X, Huang, T.S.: A loopy belief propagation approach for robust background estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008 (CVPR 2008), pp. 1–7. IEEE (2008)
75. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-Resolution Image Inpainting Using Multi-scale Neural Patch Synthesis (2016). arXiv preprint arXiv:1611.09969
76. Ye, X., Yang, J., Sun, X., Li, K., Hou, C., Wang, Y.: Foreground-background separation from video clips via motion-assisted matrix restoration. IEEE Trans. Circuits Syst. Video Technol. **25**(11), 1721–1734 (2015)
77. Zhang, D., Han, J., Li, C., Wang, J., Li, X.: Detection of co-salient objects by looking deep and wide. Int. J. Comput. Vis. **120**(2), 215–232 (2016)
78. Zhang, D., Meng, D., Han, J.: Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Trans. Pattern Anal. Mach. Intell. **39**(5), 865–878 (2017)
79. Zhang, D., Fu, H., Han, J., Borji, A., Li, X.: A review of co-saliency detection algorithms: fundamentals, applications, and challenges. ACM Trans. Intell. Syst. Technol. **9**(4), 38 (2018)
80. Zhang, T., Liu, S., Xu, C., Lu, H.: Mining semantic context information for intelligent video surveillance of traffic scenes. IEEE Trans. Ind. Inform. **9**(1), 149–160 (2013)
81. Zhang, T., Liu, S., Ahuja, N., Yang, M.H., Ghanem, B.: Robust visual tracking via consistent low-rank sparse learning. Int. J. Comput. Vis. **111**(2), 171–190 (2015a)
82. Zhang, Y., Li, X., Zhang, Z., Wu, F., Zhao, L.: Deep learning driven blockwise moving object detection with binary scene modeling. Neurocomputing **168**, 454–463 (2015b)
83. Zhao, Q., Zhou, G., Zhang, L., Cichocki, A., Amari, S.I.: Bayesian robust tensor factorization for incomplete multiway data. IEEE Trans. Neural Netw. Learn. Syst. **27**(4), 736–748 (2016)
84. Zhou, T., Tao, D.: Godec: randomized low-rank and sparse matrix decomposition in noisy case. In: ICML. Omnipress (2011)
85. Zhou, X., Yang, C., Yu, W.: Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE T-PAMI **35**(3), 597–610 (2013)
86. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004 (ICPR 2004), vol. 2, pp. 28–31. IEEE (2004)

Maryam Sultana is a Ph.D. student at Virtual Reality Laboratory, School of Computer Science and Engineering, Kyungpook National University Republic of Korea. She received her M.Sc. and M.Phil. degrees in electronics from Quaid-i-Azam university Pakistan in 2013 and 2016, respectively. Her research interests include background modeling, foreground object detection and generative adversarial networks.

Arif Mahmood is currently an Associate Professor with Information Technology University (ITU), Lahore, Pakistan. Before that he was Post-Doctoral Researcher with the Department of Computer Science and Engineering, Qatar University, Doha. He received his Masters and Ph.D. degrees in computer science from the Lahore University of Management Sciences in 2003 and 2011 respectively. After Ph.D., he joined The University of Western Australia (UWA) as Research Assistant Professor with the School of Computer Science and Software Engineering and School of Mathematics and Statistics. In UWA he mainly worked on object and action recognition. He also worked on characterizing structure of complex networks using sparse subspace clustering. His broad areas of interest include data clustering, classification, action, and object recognition. His major research interests are in computer vision and pattern recognition, action detection and person segmentation in crowded environments, and background-foreground modeling in complex scenes.

Sajid Javed is currently a Post-doctoral research fellow in the Department of Computer Science, University of Warwick, UK. He obtained his B.Sc. (hons) degree in Computer Science from University of Hertfordshire, UK, in 2010. He joined the Virtual Reality Laboratory of Kyungpook National University, Republic of Korea, in 2012 where he completed his combined Master's and Doctoral degrees in Computer Science. His research interests include background modeling and foreground object detection, robust principal component analysis, matrix completion, and subspace clustering.

Soon Ki Jung is a professor in the School of Computer Science and Engineering at Kyungpook National University, Republic of Korea. He received his M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1992 and 1997, respectively. He has been a visiting professor at University of Southern California, USA, in 2009. He has been an active executive board member of Human Computer Interaction, Computer Graphics, and Multimedia societies in Korea. Since 2007, he has also served as executive board member of IDIS Inc. His research areas include a broad range of computer vision, computer graphics, and virtual reality topics.