Performance analysis of U-Net with hybrid loss for foreground detection

Rudrika Kalsotra1 · Sakshi Arora1

Received: 6 May 2022 / Accepted: 10 October 2022 / Published online: 8 November 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
With the latest developments in deep neural networks, the convolutional neural network (CNN) has made considerable progress in the area of foreground detection. However, the top-rank background subtraction algorithms for foreground detection still have many shortcomings. It is challenging to extract the true foreground against complex background. To tackle the bottleneck, we propose a hybrid loss-assisted U-Net framework for foreground detection. A proposed deep learning model integrates transfer learning and hybrid loss for better feature representation and faster model convergence. The core idea is to incorporate reference background image and change detection mask in the learning network. Furthermore, we empirically investigate the potential of hybrid loss over single loss function. The advantages of two significant loss functions are combined to tackle the class imbalance problem in foreground detection. The proposed technique demonstrates its effectiveness on standard datasets and performs better than the top-rank methods in challenging environment. Moreover, experiments on unseen videos also confirm the efficacy of proposed method.

Keywords Background subtraction · Class imbalance · Detection · Foreground · Hybrid loss · U-Net

1 Introduction
Deep learning has transformed the field of computer vision and holds the ability to upgrade the performance of video analytics in complex scenarios. The primary goal of video analytics is to extract the region of interests (ROI) from videos. Foreground detection is the primary task in video analytics and forms the basis of various high-level applications including automated visual surveillance [1, 2], human–computer interaction [3], behavior analysis [4, 5], autonomous driving [6], smart parking [7], traffic-flow statistics [8, 9], and object tracking [10]. The post-Covid-19 situation highlights the need of effective solutions of video analytics for real-world applications like crowd-control management, patient tracking, face-mask detection, and contact tracing [11–14]. The methods for foreground detection are broadly classified into three categories: frame differencing, background subtraction, and optical flow [15]. At present, the most reliable and extensively used method for foreground detection is background subtraction. Various significant surveys have been published to provide the comprehensive account of background subtraction process [16–24]. The challenges that stand in the way of efficacious detection of foreground objects in videos are dynamic background, shadows, illumination variations, camera jitter, camouflage effects, moving cameras, bad weather conditions, video noise, occlusion, and intermittent object motion [25]. The factors that change from frame to frame are background scene, illumination, object motion, viewpoint, and environmental conditions. These factors degenerate the background subtraction method’s performance to a great extent. The conventional background subtraction methods have made tremendous progress in extracting foreground objects from the videos [26–29]. However, these conventional methods rely on human-engineered features and usually result in false detections or fragmented foreground objects. Deep neural networks have largely superseded the conventional approaches. The CNN-based background subtraction method yield better performance in scenarios with highly dynamic background, variable illumination, moving cameras, hard shadows, camera jitter, and night videos.

For foreground detection, the deep neural networks still need improvements in certain challenging scenarios. A first
An attempt to employ the CNN for background subtraction was made in 2016 [30]. Thereafter, several supervised background subtraction algorithms have attained the state-of-the-art results. FgSegNet_v2 [31] has achieved an average F-score above 0.98 and is the top-ranked background subtraction method among other supervised deep learning-based methods in the changedetection.net dataset [32]. However, the performance of FgSegNet_v2 drops significantly in case of unseen video scenes. The current deep learning-based methods have many shortcomings. From computational complexity to scene-specific models, the researchers are trying to fix the flaws of current deep neural networks. Moreover, the performance of deep learning-based methods has been adversely affected by the class imbalance problem. In foreground detection, the background class is overrepresented in the training as it has more samples than the foreground objects. This leads to foreground–background imbalance in case of foreground detection [33]. Such class imbalance problem is tackled by utilizing appropriate loss function for training the deep neural networks [34]. Various prominent background subtraction methods employ the binary cross-entropy (BCE) loss function and ignore the class imbalance problem [35–38]. Some recent works utilize the variants of BCE loss like weighted BCE [31, 39] and focal loss [40–42] to deal with the class imbalance problem. However, the traditional loss functions are still vulnerable to the foreground–background imbalance problem. It has been proven that the integration of two or more appropriate loss functions can handle the severe class imbalance problem and outperform the variants of BCE loss in medical segmentation tasks [43, 44]. Hence, there is a need to extend the use of hybrid loss function in foreground detection to tackle the severe class imbalance problem.

In this paper, we propose a hybrid loss-assisted background subtraction method to extract foreground from videos. An efficient U-Net framework with VGG16 encoder is adopted to empirically investigate the effectiveness of hybrid loss in foreground detection. Unlike the popular FgSegNet_v2 method [31] trained with traditional loss, the proposed model utilizes the hybrid loss to tackle the class imbalance problem and for the precise detection of small foreground objects. To alleviate the limitations of appearance-based single frame foreground detection, the proposed network incorporates spatiotemporal information through the fusion of background model and change detection mask with the current frame. Table 1 outlines the overview of the background subtraction methods and their comparison with our proposed method.

The paper has the following key contributions:

- Extraction of semantically complete region of interests from videos with the incorporation of transfer learning in U-Net-based framework.
- A hybrid loss-assisted learning to mitigate the class imbalance problem for foreground detection and to improve the convergence speed of the network.
- Inclusion of background information and change detection mask in the network model to improve the segmentation performance in complex scenarios.
- Empirical investigations of single loss function with the proposed network during learning process to justify the robustness of hybrid loss for highly imbalanced data.

Table 1 Comparative overview of the recent methods and the proposed method

| Methods            | Year | Model   | Input                      | Pre-trained weights | Loss function        |
|--------------------|------|---------|----------------------------|--------------------|----------------------|
| DeepBS [35]        | 2017 | CNN     | Current image, Background image | No                 | Binary cross-entropy |
| FgSegNet_v2 [31]   | 2018 | CNN     | Current image              | Yes                | Weighted binary cross-entropy |
| TMT-GAN [39]       | 2018 | GAN     | Low and High brightness image pair | Yes                | Weighted binary cross-entropy |
| X-net [40]         | 2019 | CNN     | Two consecutive current images | Yes                | Focal loss           |
| MS-ST [41]         | 2019 | CNN     | Current image              | Yes                | Focal loss           |
| 3DCD [36]          | 2020 | CNN     | Current image, Background image | No                 | Binary Cross-entropy |
| EFDNet [37]        | 2021 | CNN     | Current image              | Yes                | Binary Cross-entropy |
| STPNet [42]        | 2021 | CNN     | Three consecutive images   | No                 | Focal loss           |
| BSUV-Net 2.0 [45]  | 2021 | U-Net   | Current image, Two background images | No                 | Jaccard loss         |
| U-Net with multiple difference images [38] | 2021 | U-Net   | Ten difference images      | No                 | Binary Cross-entropy |
| Ours (proposed)    | –    | U-net   | Current image, Background image, Change detection mask | Yes | Focal loss + Dice loss |
The remaining sections of the paper are structured as follows: In Sect. 2, we briefly introduce the existing background subtraction methods. The proposed methodology which includes U-Net-based architecture and hybrid loss function is explained in Sect. 3. Sect. 4 covers the training details, evaluation datasets, experimental results and performance comparisons. Sect. 5 validates the robustness of the proposed method on unseen videos. The qualitative results are presented in Sect. 6. Sect. 7 demonstrates the impact of hybrid loss and different input structures on foreground detection. The final section provides important insights into the strength and weakness of the proposed method and also highlights the future work.

2 Related works

Different concepts for background subtraction have been exploited to improve the foreground detection in challenging scenarios. Numerous background subtraction algorithms are available for extracting the foreground objects from video sequence, each with their own advantages and limitations. Some of the established background subtraction methods are summarized in this section.

Conventional methods generally follow the three stage process of background subtraction, which includes: background estimation, foreground extraction, and background maintenance [46]. The primitive models that employed arithmetic mean, median, and histograms for background modeling and the model based on single Gaussian [47] are unable to tackle the challenging scenes. A multi-modal model based on the mixture of Gaussians was developed to detect the foreground in videos [26]. This method is suitable for slow illumination variations and dynamic background motion. A sample-based algorithm, known as ViBe [48], was propounded for foreground detection. The proposed random update scheme and spatial information propagation method is integrated with conservation update scheme for background maintenance. This method is not robust against challenging scenarios like shadows and highly dynamic backgrounds. An advanced algorithm based on the multi-feature fusion was presented to tackle the problem of camouflage and illumination variations in foreground detection [27]. This method adopts a feedback mechanism to automatically tune the internal parameters and shows high performance among the other background subtraction methods. In another work, St-Charles et al. [28] integrates a word-dictionary scheme with multi-feature fusion to enhance the robustness against challenging scenarios. A genetic programming-based approach, called IUTIS-5, was proposed to merge the existing background subtraction algorithms to generate more reliable foreground masks [29]. This method enables the automatic selection and merging of various existing algorithms as well as the post-processing of foreground masks. In comparison to other conventional methods, the IUTIS-5 produces better results.

Deep neural networks have attracted great research attention after the success of CNN for background subtraction. Babaei et al. [35] proposed a deep CNN framework named DeepBS for foreground detection. An advanced algorithm based on the integration of two traditional background subtraction methods was proposed for background modeling. The network is trained with small image regions obtained from current frame, background reference frame, and ground truth mask. Then, the network outputs are post-processed for the enhanced foreground masks. An encoder–decoder architecture centered on triplet CNN and transposed CNN was proposed for background subtraction by Lim and Keles [49]. This architecture is known as FgSegNet-M. The other two variants of FgSegNet, [50] and [31] produce better results in challenging scenarios. Yang et al. [41] presented the fusion of CNN and ConvLSTM for background subtraction to segment the ROI from videos. This spatiotemporal algorithm utilized CNN to extract the spatial features and ConvLSTM with temporal sampling to extract the temporal information. Mandal et al. [36] proposed the 3D-CNN-based framework for background subtraction and scene-independent evaluation set-up for performance comparison in unseen videos. To deal with the complex scenes in foreground detection, Zheng et al. [51] designed a Bayesian generative adversarial network to detect moving objects in videos. Furthermore, the authors proposed an advanced algorithm called BSPVGAN [52] to enhance the potential of background subtraction in unseen videos.

In recent years, U-Net has attained great success in many image segmentation tasks [53, 54]. Inspired by the achievements of U-Net in segmentation, few attempts have been made by researchers to extend the use of U-Net for foreground detection. Tezcan et al. [55] introduced the U-Net-based model called BSUV-Net which utilizes current image and two background images along with their semantic information to detect moving objects in unseen videos. An advanced version of BSUV-Net, named BSUV-Net 2.0 [45], introduced a series of data augmentation approaches to enhance the segmented foreground in complex videos. Another U-Net with different input structure was designed to handle the challenges of background subtraction [56]. A set of six images which includes current image, background image and four past images are stacked and utilized as an input to the U-Net. Rahmon et al. [57] proposed a U-Net-based model called MU-Net for moving object detection. A proposed deep neural network accepts a single-stream RGB input which includes current image and two change detection masks. The spatiotemporal information is incorporated into U-Net by extracting motion and change cues from two unsupervised background subtraction methods [58, 59].
To achieve the better generalization ability, a U-Net-based architecture with multiple difference images as input was designed to segment the moving objects from videos [38]. A fine-tuning of U-Net is done to use ten difference images as an input. The purpose of employing multiple difference images is to incorporate the temporal information along with the spatial information in the network.

U-Net has proven to possess great potential in the extraction of foreground objects from videos. It is demonstrated in [55, 56] that incorporation of spatiotemporal information in the deep neural network improves the performance of background subtraction in challenging scenarios. The current deep neural networks for foreground detection exploit multiple input images for temporal reasoning. In this research, we follow the similar idea of incorporating spatiotemporal data into the U-Net but adopt a different input structure. Instead of stacking multiple images, we integrate the details of reference background model and change detection mask with the current image to improve the generalization ability.

3 Proposed method

Inspired by the recent successful application of U-Net in segmentation tasks, we propose a U-Net framework integrated with VGG-16 for computational efficiency and precise foreground detection. The proposed approach derives the benefits of both transfer learning and encoder–decoder structure. The inputs to the proposed network include current frame, reference background model, and change detection mask.

The RGB frames are converted into the grayscale. Then, all the three input frames are merged into a 3-channel image. The information of the background in foreground detection plays a key role during training process of the network. In the proposed method, temporal median filter is applied over each pixel for 150 initial frames to generate the background frame. The inclusion of change detection mask refines the segmented foreground in complex scenes. The change is estimated by applying a nonparametric background subtraction method called SuBSENSE [27]. This unsupervised method detects changes by integrating spatiotemporal binary features with color features. The graphical representation of flow of the proposed scheme is illustrated in Fig. 1. This section introduces the all-inclusive description of the network architecture and the hybrid loss function.

3.1 U-net architecture

A U-shaped architecture which contains encoder–decoder network with long skip connections is proposed to segment the foreground objects from videos. The overall architecture of the hybrid loss-assisted U-Net is demonstrated in Fig. 2. The dimensions of the input images are 320 (W) × 240 (H) × 3 (C). The encoder network comprises of four downsampling blocks based on the VGG-16 to capture the context. Each downsampling block consists of the Convolutional layers combined with ReLU function and the Max Pooling (Max Pool) layer. The proposed network incorporates 13 convolutional layers from the VGG-16. The corresponding upsampling blocks in decoder network adopt
the combination of Upsampling layer, Convolutional layers, Batch Normalization (BN) layer, and ReLU function. In decoder part, BN layer with ReLU function is applied after each convolution layer. The fine-scale encoder features are fused with the coarse-scale decoder features via long skip connections to retrieve the lost spatial information during downsampling. Therefore, skip connections are utilized in the proposed encoder–decoder network for better network training and precise segmentation masks. The main task of labeling pixels into foreground and background is guided by the last 1 × 1 convolutional layer with a sigmoid function. A thresholding is applied to the outcome of last layer to attain the binary masks and the value of threshold θ is set to 0.5. The proposed model is executed in Keras framework with Tensorflow as the backend. The proposed network has 25,862,849 total parameters and the number of trainable parameters is 25,859,009. The detailed configuration and specifications of the hybrid loss-assisted U-Net are illustrated in Table 2.

### 3.2 Hybrid loss function

In video surveillance data, especially the frames with small ROI and without foreground lead to severe class imbalance problem. The class imbalance data adversely affect the performance of foreground detection. A large number of background regions have the dominant effect on loss function during learning process. The commonly used loss function in image segmentation is binary cross-entropy (BCE). However, the results produced by the BCE loss are inclined toward the background class and it fails to handle the class imbalance issue. Moreover, a variant of BCE known as focal loss [60] is mainly opted to handle the class imbalance in the recent background subtraction task. Focal loss (FL) is defined as:

\[
FL(pr, gt) = \begin{cases} 
-\alpha(1 - pr)^\gamma \log(pr), & \text{if } gt = 1 \\
-(1 - \alpha)pr^\gamma \log(1 - pr), & \text{otherwise}
\end{cases}
\]

Here, \(pr \in [0, 1]\) and \(gt \in (0, 1)\) specify the predicted probability and ground truth, respectively. In case of foreground detection, the values of \(pr\) and \(gt\) are either 0 or 1, where 0 represents the background and 1 is the foreground. \(\alpha\) and \(\gamma\) are the hyperparameters in FL, where \(\alpha \in (0, 1]\) represents a class weighting parameter and \(\gamma \in [0, 5]\) is a tunable focusing parameter. In our study, we set \(\alpha = 0.25\) and \(\gamma = 2.0\).

The dice loss (DL) has shown better performance for the medical segmentation tasks with extreme class imbalance and small ROIs [61]. It is defined as follows:

\[
DL(pr, gt) = 1 - \frac{2 \sum_{i=1}^{N} pr_i gt_i + \delta}{\sum_{i=1}^{N} pr_i^2 + \sum_{i=1}^{N} gt_i^2 + \delta}
\]

where \(\delta \in [0, 1]\) is a smoothing factor to prevent infinite values and the value of \(\delta\) is set as 1.
Based on the above two standard loss functions, we employ the hybrid loss function to handle the severe class imbalance for foreground detection. The integration of focal loss and dice loss also improves the convergence speed [44]. Therefore, we define the hybrid loss function (HBL) as:

\[
HBL(pr, gt) = (1 \times FL(pr, gt)) + DL(pr, gt)
\]

4 Experimental results

The components of the evaluation process including evaluation datasets, training settings, and results are presented in this section. The proposed hybrid loss-assisted U-Net is compared to the top-rank methods based on the qualitative as well as quantitative results.

4.1 Evaluation datasets

Two benchmark video datasets, namely changedetection.net dataset [32] and LASIESTA dataset [62], are utilized to evaluate the potential of hybrid loss-assisted U-Net for foreground detection. The changedetection.net dataset comprises 53 realistic videos, split in 11 categories covering diverse set of challenges like intermittent object motion, low framerate, night videos, PTZ, shadow, thermal, and turbulence. The pixel-level manually annotated ground truth with five labels: static, moving, hard shadow, unknown, and outside ROI were provided for evaluation purpose. The changedetection.net dataset is employed for training as well as evaluation purpose in this paper. Another video dataset, LASIESTA contains 48 videos recorded in both indoor and outdoor environment. The indoor videos are further categorized into seven sets: simple sequences, occlusions, modified background, illumination variations, bootstrap, camouflage, and moving camera. The set of outdoor videos includes cloudy days, rainy days, snowy conditions, sunny days, and moving camera. This dataset also contains two set of videos with simulated motion: set of indoor videos (I_SM) and outdoor videos (O_SM). The LASIESTA dataset is used to examine the generalization potential of the proposed U-Net.

4.2 Training and evaluation details

The encoder network of the proposed U-Net is initialized with the weights of pre-trained VGG-16 model to upgrade the training efficiency. A random frame selection strategy is adopted for training purpose. The training data involve 200 video frames (3-channel input) and their corresponding ground truth from each 53 videos of changedetection.net dataset. It is clearly demonstrated in [63] that 200 training frames are sufficient to attain good results in complex videos. The training data are shuffled and split into two parts: 75% for training and 25% for validation. Thus, we have 10,600 processed video frames and their corresponding

| Layer | Output dimensions | Block | Layer | Output dimensions | Block |
|-------|------------------|-------|-------|------------------|-------|
| Input | 240×320×3        | –     | Upsampling | 30×40×512 | –     |
| Conv + ReLU | 240×320×64 | Downsampling B1 | Concatenate | 30×40×1024 | –     |
| Conv + ReLU | 240×320×64 |             | Conv + BN + ReLU | 30×40×512 | –     |
| Max Pool | 120×160×64 |             | Conv + BN + ReLU | 30×40×512 | –     |
| Conv + ReLU | 120×160×128 | Downsampling B2 | Upsampling | 60×80×256 | –     |
| Conv + ReLU | 120×160×128 |             | Concatenate | 60×80×512 | –     |
| Max Pool | 60×80×128     |             | Conv + BN + ReLU | 60×80×256 | –     |
| Conv + ReLU | 60×80×256      | Downsampling B3 | Upsampling | 120×160×128 | –     |
| Conv + ReLU | 60×80×256      |             | Concatenate | 120×160×256 | –     |
| Conv + ReLU | 60×80×256      |             | Conv + BN + ReLU | 120×160×128 | –     |
| Max Pool | 30×40×256     |             | Conv + BN + ReLU | 120×160×128 | –     |
| Conv + ReLU | 30×40×512      | Downsampling B4 | Upsampling | 240×320×64 | –     |
| Conv + ReLU | 30×40×512      |             | Concatenate | 240×320×128 | –     |
| Conv + ReLU | 30×40×512      |             | Conv + BN + ReLU | 240×320×64 | –     |
| Max Pool | 15×20×512     |             | Conv + BN + ReLU | 240×320×64 | –     |
| Conv + ReLU | 15×20×512      | Bridge | Conv | 240×320×1 | Last layer |
| Conv + ReLU | 15×20×512      |             | Output | 240×320×1 | –     |
ground truth images for training and validation. The testing data comprise all the video frames of changedetection.net dataset. This training-testing set-up follows a video-group-optimized evaluation strategy and falls into the category of scene-dependent evaluation [20]. For scene-independent evaluation, the proposed U-Net is also tested on completely unseen videos of LASIESTA dataset. Furthermore, the ground truth images are relabeled before training to ignore the labels other than static and moving pixels. The original ground truth images from changedetection.net dataset and their respective relabeled ground truth used during the development are presented in Fig. 3. During the training, the Adam optimizer with an initial learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ is employed to update the parameters of network. We set minibatch size to 32 and epochs to 100 for training. A hybrid loss function consisting of focal loss and dice loss is utilized to train the proposed U-Net. If the validation loss does not decrease in four consecutive epochs, the learning rate is dropped by 0.1. Moreover, an early stopping with a patience of 10 is applied to monitor the performance for each epoch on a validation dataset and to terminate the learning process if validation loss does not improve.

4.3 Quantitative analysis

The proposed method is evaluated quantitatively by using multiple metrics in this research. These evaluation metrics include recall, precision, specificity, percentage of wrong classification (PWC), and F-measure. Quantitative evaluation on changedetection.net dataset is reported in Table 3. From Table 3, it is apparent that the proposed U-Net achieves above 0.90 F-measure in all categories except in low frame rate and PTZ categories.

The proposed hybrid loss-assisted U-Net model is compared against the state-of-the-art methods to prove its effectiveness. Comparison with unsupervised background subtraction methods, namely SuBSENSE [27], IUTIS-5 [29], WeSamBE [64], and WisenetMD [65] is reported in Table 4. The unsupervised algorithms are outperformed.
by the proposed method on all challenges as shown in Table 4. The proposed method is also tested against the supervised deep learning-based methods for a fair comparison, and the results are provided in Table 5. These supervised methods include cascade CNN [63], Deep BS [35], and MU-Net2 [57]. Table 5 indicates the proposed method’s higher performance in low framerate, night videos, turbulence, bad weather, and dynamic background categories. The proposed method holds second rank after cascade CNN method in PTZ category. In the categories of camera jitter and intermittent object motion, the proposed technique yields less F-measure than MU-Net2, although it is still competitive with current methods. An overall performance of proposed method in comparison to both unsupervised and supervised background subtraction methods is illustrated in Fig. 4. On the changedetection.net dataset, the proposed hybrid loss-assisted U-Net has the highest F-measure of 0.9372 and recall of 0.9584. The value of precision for proposed method is 0.9229, second highest after the MU-Net2 method.

5 Cross-dataset evaluation

The generalization ability of the proposed model is tested via a cross-dataset evaluation. A set of complete unseen videos from the LASIESTA dataset is utilized to evaluate the generalization aspect of proposed hybrid loss-assisted U-Net model. Quantitative results of proposed method on LASIESTA dataset are reported in Table 6. Out of 48 videos under different video categories, 44 videos and their corresponding labeled masks are available for evaluation purpose. The proposed method is compared with the state-of-the-art background subtraction methods and the results

| Categories               | SuBSENSE [27] | IUTIS-5 [29] | WeSamBE [64] | WisenetMD [65] | Proposed |
|--------------------------|---------------|--------------|--------------|----------------|----------|
| Bad weather              | 0.8619        | 0.8248       | 0.8608       | 0.8616         | 0.9588   |
| Low framerate            | 0.6445        | 0.7743       | 0.6602       | 0.6404         | 0.8966   |
| Night videos             | 0.5599        | 0.5290       | 0.5929       | 0.5701         | 0.9091   |
| PTZ                      | 0.3476        | 0.4282       | 0.3844       | 0.3367         | 0.8475   |
| Turbulence               | 0.7792        | 0.7836       | 0.7737       | 0.8304         | 0.9332   |
| Baseline                 | 0.9503        | 0.9567       | 0.9413       | 0.9487         | 0.9836   |
| Dynamic background       | 0.8177        | 0.8902       | 0.7440       | 0.8376         | 0.9897   |
| Camera jitter            | 0.8152        | 0.8332       | 0.7976       | 0.8228         | 0.9577   |
| Intermittent object Motion | 0.6569      | 0.7296       | 0.7392       | 0.7264         | 0.9032   |
| Shadow                   | 0.8986        | 0.9084       | 0.8999       | 0.8984         | 0.9678   |
| Thermal                  | 0.8171        | 0.8303       | 0.7962       | 0.8152         | 0.9618   |
| Overall                  | 0.7408        | 0.7717       | 0.7446       | 0.7535         | 0.9372   |

Bold entries indicate the best results

| Categories               | Cascade CNN [63] | DeepBS [35] | MU-Net2 [57] | Proposed |
|--------------------------|------------------|-------------|--------------|----------|
| Bad weather              | 0.9431           | 0.8301      | 0.9343       | 0.9588   |
| Low framerate            | 0.8370           | 0.6002      | 0.8706       | 0.8966   |
| Night videos             | 0.8965           | 0.5835      | 0.8362       | 0.9091   |
| PTZ                      | **0.9168**       | 0.3133      | 0.8185       | 0.8475   |
| Turbulence               | 0.9108           | 0.8455      | 0.9272       | **0.9332**|
| Baseline                 | 0.9786           | 0.9580      | 0.9892       | **0.9897**|
| Dynamic background       | 0.9658           | 0.8761      | 0.9824       | 0.9577   |
| Camera jitter            | 0.9758           | 0.8990      | 0.9845       | **0.9678**|
| Intermittent object Motion | 0.8505         | 0.6098      | **0.9894**   | 0.9032   |
| Shadow                   | 0.9593           | 0.9304      | **0.9845**   | 0.9678   |
| Thermal                  | 0.8958           | 0.7583      | **0.9842**   | 0.9618   |
| Overall                  | 0.9209           | 0.7458      | 0.9369       | **0.9372**|

Bold entries indicate the best results
Performance analysis of U-Net with hybrid loss for foreground detection

Fig. 4 Overall analysis of proposed method and state-of-the-art methods on changedetection.net dataset

Table 6 Quantitative evaluation of proposed hybrid loss-assisted U-Net on LASIESTA dataset

| Categories                | Recall  | Specificity | Precision | PWC    | F-measure |
|---------------------------|---------|-------------|-----------|--------|-----------|
| Simple sequences (I_SI)   | 0.9945  | 0.9978      | 0.9365    | 0.2184 | 0.9647    |
| Camouflage (I_CA)         | 0.9954  | 0.9966      | 0.9575    | 0.3415 | 0.9761    |
| Occlusions (I_OC)         | 0.9893  | 0.9982      | 0.9104    | 0.1885 | 0.9482    |
| Illumination changes (I_IL)| 0.9919  | 0.9977      | 0.9159    | 0.2407 | 0.9524    |
| Modified background (I_MB)| 0.9923  | 0.9954      | 0.9242    | 0.4651 | 0.9570    |
| Bootstrap (I_BS)          | 0.9872  | 0.9981      | 0.8662    | 0.2003 | 0.9224    |
| Moving camera (I_MC)      | 0.9827  | 0.9988      | 0.9250    | 0.1378 | 0.9529    |
| Rainy conditions (O_RA)   | 0.9408  | 0.9985      | 0.9088    | 0.2380 | 0.9246    |
| Sunny conditions (O_SU)   | 0.9726  | 0.9981      | 0.8340    | 0.2046 | 0.8979    |
| Moving camera (O_MC)      | 0.9876  | 0.9978      | 0.8719    | 0.2314 | 0.9261    |
| Overall                   | 0.9834  | 0.9977      | 0.9050    | 0.2466 | 0.9422    |

Table 7 F-measure values to compare the proposed hybrid loss-assisted U-Net with the state-of-the-art methods on LASIESTA dataset

| Categories                | Haines [66] | Cuevas [67] | MSFgNet [68] | STPNet [42] | BSUV-Net 2.0 [45] | Proposed |
|---------------------------|-------------|-------------|--------------|-------------|--------------------|----------|
| Simple sequences (I_SI)   | 0.89        | 0.88        | 0.93         | 0.90        | 0.92               | 0.96     |
| Camouflage (I_CA)         | 0.89        | 0.84        | 0.92         | 0.81        | 0.68               | 0.98     |
| Occlusions (I_OC)         | 0.92        | 0.78        | 0.92         | 0.92        | 0.96               | 0.95     |
| Illumination changes (I_IL)| 0.85        | 0.65        | 0.90         | 0.74        | 0.88               | 0.95     |
| Modified background (I_MB)| 0.84        | 0.93        | 0.91         | 0.93        | 0.81               | 0.96     |
| Bootstrap (I_BS)          | 0.68        | 0.66        | 0.72         | 0.86        | 0.77               | 0.92     |
| Rainy conditions (O_RA)   | 0.89        | 0.87        | 0.87         | 0.89        | 0.94               | 0.92     |
| Sunny conditions (O_SU)   | 0.86        | 0.72        | 0.79         | 0.86        | 0.79               | 0.90     |
| Overall                   | 0.85        | 0.79        | 0.87         | 0.86        | 0.84               | 0.94     |

Bold entries indicate the best results
are reported in Table 7. It is evident from the results that proposed method outperforms all the state-of-the-arts in complex video scenes. For occlusions and rainy conditions video categories, the proposed method shows slightly lower performance and ranks second after the BSUV-Net 2.0 [45] method. The videos under moving camera and simulated motion categories in LASIESTA dataset are divided into four groups by the BSUV-Net 2.0 method and its performance has been evaluated under three different versions as mentioned in Table 8. Here, different versions of BSUV-Net 2.0 imply different combinations of data augmentations. The proposed method outperforms all the three versions of BSUV-Net 2.0 method and shows remarkable performance in all the four video categories.

### 6 Visual results

The visual results offer a qualitative evaluation of the proposed technique as well as a comparison to the state-of-the-art methods. The qualitative comparison on various challenges of foreground detection is presented in Fig. 5 and Fig. 6.

It is apparent from the visual results that proposed hybrid loss-assisted U-Net achieves precise foreground detection, keeping the ROI intact. Figure 5 and Fig. 6 demonstrate that the conventional background subtraction methods like SuBSENSE method and Wisenet MD method generate either fragmented foregrounds or ghost artifacts along with the ROI. In contrast, the CNN-based methods such as Cascade CNN, MU-Net2, and DeepBS add faulty pixels around the ROI and produce distorted foreground objects in certain challenging conditions.

**Table 8** Comparison of proposed method with BSUV-Net 2.0 [45] in terms of F-measure on Moving camera and Simulated motion videos of LASIESTA dataset

| Categories          | BSUV-Net 2.0-A | BSUV-Net 2.0-B | BSUV-Net 2.0-C | Proposed |
|---------------------|----------------|----------------|----------------|-----------|
| Indoor-pan & tilt   | 0.48           | 0.52           | 0.58           | 0.97      |
| Outdoor-pan & Tilt  | 0.56           | 0.42           | 0.58           | 0.93      |
| Indoor-jitter       | 0.81           | 0.88           | 0.50           | 0.91      |
| Outdoor-jitter      | 0.75           | 0.85           | 0.50           | 0.91      |

(BSU-Net 2.0-A: spatially aligned crop augmentation, BSU-Net 2.0-B: randomly shifted crop, BSU-Net 2.0-C: PTZ camera crop) Bold entries indicate the best results.

**Fig. 5** Qualitative results obtained on test data from six video categories of change-detection.net dataset: a Original frame, b Ground truth foreground maps, c SuBSENSE [27], d Wisenet MD [65], e Cascade CNN [63], f MU-Net2 [57], g DeepBS [35], h Proposed method.
Performance analysis of U-Net with hybrid loss for foreground detection

A qualitative evaluation on the LASIESTA dataset is illustrated in Fig. 7. The visual results obtained on LASIESTA dataset indicate impeccable foreground detection in both simple and complex video sequences.

7 Ablation study

The impact of hybrid loss supervision and different input structures on foreground detection is thoroughly investigated in this section. The category-wise comparative analysis of proposed U-Net model with single loss functions (focal loss and dice loss) and hybrid loss (focal + dice) is reported in Table 9. The hybrid loss-assisted U-Net model developed for foreground detection leads to significant improvement in case of complex videos. The superiority of hybrid loss supervision over single losses is clearly accessible from Table 10 as it demonstrates the overall analysis on changedetection.

To better understand the contribution of reference background frame and change detection mask in the learning network, we adopt three different input structures with the changedetection.net dataset. To prove the effectiveness of selected hybrid loss (focal + dice) over other hybrid loss functions, a comparative analysis is reported in Table 11. A comparison is made among the three hybrid loss functions: binary cross-entropy + dice loss, binary cross-entropy + Jaccard loss, and focal + dice loss. An overall analysis of proposed U-Net model with various hybrid loss functions is provided in Table 12. A combination of focal and dice loss achieves the lowest PWC of 0.1730, which indicates that the hybrid loss performs better than the other loss functions. Figure 8 presents the qualitative comparison among various loss functions. A combination of focal and dice loss addresses the problem of small ROIs effectively in comparison to other loss functions. The results presented in Fig. 8 illustrates that the other two hybrid loss functions results into fragmented foregrounds.
proposed U-net framework for experimental study. Comparative analysis of proposed method with different input structures is reported in Table 13. Initially, the current RGB frame is used as an input. Then, a combination of current frame and change detection mask is used as an input. For the final experiment, combination of three different frames (current frame, background frame, and change detection mask) is used as an input.

Table 9 F-measure values for comparative analysis of proposed model with single loss (focal and dice) and hybrid loss (focal + dice) on the changedetection.net dataset

| Categories                | Focal loss | Dice loss | Hybrid loss |
|---------------------------|------------|-----------|-------------|
| Bad weather               | 0.9420     | 0.9439    | 0.9588      |
| Low framerate             | 0.7237     | 0.7721    | 0.8966      |
| Night videos              | 0.8230     | 0.8511    | 0.9091      |
| PTZ                       | 0.6708     | 0.7580    | 0.8475      |
| Turbulence                | 0.9061     | 0.8329    | 0.9332      |
| Baseline                  | 0.9071     | 0.8889    | 0.9836      |
| Dynamic background        | 0.9320     | 0.8713    | 0.9897      |
| Camera jitter             | 0.9285     | 0.9224    | 0.9577      |
| Intermittent object motion| 0.8608     | 0.8529    | 0.9032      |
| Shadow                    | 0.9639     | 0.8728    | 0.9678      |
| Thermal                   | **0.9667** | 0.8727    | 0.9618      |
| Overall                   | 0.8750     | 0.8581    | **0.9372**  |

Bold entries indicate the best results

Table 10 Overall analysis of proposed model with single loss (focal and dice) and hybrid loss (focal + dice) on the changedetection.net dataset

| Loss functions | Recall | Specificity | Precision | PWC | F-measure |
|----------------|--------|-------------|-----------|-----|-----------|
| Focal loss     | 0.8836 | 0.9987      | 0.8936    | 0.2517 | 0.8750    |
| Dice loss      | 0.8089 | 0.9988      | 0.9282    | 0.5174 | 0.8581    |
| Hybrid loss    | **0.9584** | **0.9989** | 0.9229    | 0.1730 | **0.9372** |

Bold entries indicate the best results
mask) is adopted as an input to the network. It is clear from the comparative analysis that the use of only current frame as an input produces poor results. A combination of current frame and change detection mask yields better results than the single frame input but its performance drops in turbulence, intermittent object motion, and shadow categories. The results obtained on different challenging scenes manifested the significance of inclusion of background information and change detection mask along with current frame.

### 8 Conclusions

In this paper, a hybrid loss-assisted U-Net was proposed for precise foreground detection in complex scenes. The proposed encoder–decoder architecture incorporates reference background frame, change detection mask along with current frame in the learning network to segment the ROI from the video scene. Compared with the only current frame input, the proposed method with merged frame as input achieved better results. Integration of transfer learning in

---

**Table 11** F-measure values for comparative analysis of proposed model with different hybrid loss functions on the changedetection.net dataset

| Categories          | Binary cross-entropy + Dice loss | Binary cross-entropy + Jaccard loss | Focal + Dice loss |
|---------------------|----------------------------------|-------------------------------------|-------------------|
| Bad weather         | 0.9283                           | 0.9350                              | **0.9588**        |
| Low framerate       | 0.7843                           | 0.7839                              | **0.8966**        |
| Night videos        | 0.8072                           | 0.7997                              | **0.9091**        |
| PTZ                 | 0.7116                           | 0.7394                              | **0.8475**        |
| Turbulence          | 0.8307                           | 0.8044                              | **0.9332**        |
| Baseline            | 0.8724                           | 0.8689                              | **0.9836**        |
| Dynamic background  | 0.8250                           | 0.8482                              | **0.9897**        |
| Camera jitter       | 0.8864                           | 0.9023                              | **0.9877**        |
| Intermittent object motion | 0.7037                        | 0.6770                              | **0.9032**        |
| Shadow              | 0.8857                           | 0.8784                              | **0.9678**        |
| Thermal             | 0.8326                           | 0.8223                              | **0.9618**        |
| Overall             | 0.8244                           | 0.8236                              | **0.9372**        |

Bold entries indicate the best results

**Table 12** Overall analysis of proposed model with different hybrid loss functions on the changedetection.net dataset

| Hybrid loss functions | Recall       | Specificity  | Precision  | PWC          | F-measure   |
|-----------------------|--------------|--------------|------------|--------------|-------------|
| Binary cross-entropy + Dice loss | 0.7799 | 0.9984 | 0.9120 | 0.6128 | 0.8244 |
| Binary cross-entropy + Jaccard loss | 0.7907 | 0.9983 | 0.8988 | 0.6203 | 0.8236 |
| Focal + Dice loss     | **0.9584**   | 0.9989       | **0.9229** | **0.1730**  | **0.9372** |

Bold entries indicate the best results

---

Fig. 8 Qualitative results obtained on test data from four video categories of changedetection.net dataset: a Original frame, b Ground truth foreground maps, c Focal + Dice loss, d Binary cross-entropy + Dice loss, e Binary cross-entropy + Jaccard loss.
the encoder induced better feature representation and generalization ability. The performance evaluation on unseen videos verified the effectiveness of proposed method. The class imbalance problem which adversely affects the detection of foreground objects is handled through the hybrid loss supervision. Experimental results on benchmark dataset demonstrated the advantages of hybrid loss over single loss for foreground detection. Besides, compared with the current background subtraction methods, the proposed method produced considerably more precise results on challenging scenes and well avoid faulty pixels. The results showed that the proposed technique has achieved relatively higher F-measure.

To improve the proposed U-Net performance in IOM, camera jitter, and PTZ video categories, we will incorporate temporal information in learning network in the future. We will also explore the deep neural networks for the background estimation in our future work.

Author contributions RK contributed to the data collection, design and implementation of the research, to the analysis of results and to the writing of the manuscript. SA reviewed the results and approved the final version of the manuscript.

Funding No funding or other support was received from any organization for this study.

Availability of data and materials The datasets analyzed during this study are available from the public data repositories at the website of http://www.changedetection.net/ and https://www.gti.ssr.upm.es/data/lasiesta_database.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Li, L., Wang, Z., Hu, Q., Dong, Y.: Adaptive nonconvex sparsity based background subtraction for intelligent video surveillance. IEEE Trans. Industr. Inf. 17(6), 4168–4178 (2020)
2. Kim, C., Lee, J., Han, T., Kim, Y.M.: A hybrid framework combining background subtraction and deep neural networks for rapid person detection. J. Big Data 5(1), 1–24 (2018)
3. Penciuc, D., El Baf, F., & Bouwmans, T. (2006). Comparison of background subtraction methods for an interactive learning space. NETTIES 2006.
4. Perrett, T., Mirmehdi, M., Dias, E.: Visual monitoring of driver and passenger control panel interactions. IEEE Trans. Intell. Transp. Syst. 18(2), 321–331 (2016)
5. Tu, G.J., Hansen, M.K., Kryger, P., Ahrendt, P.: Automatic behaviour analysis system for honeybees using computer vision. Comput. Electron. Agric. 122, 10–18 (2016)
6. Ravi Kiran, B., Roldao, L., Irastorza, B., Verastegui, R., Suss, S., Yogamani, S., & Trehard, G. (2018). Real-time dynamic object detection for autonomous driving using prior 3d-maps In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0.
7. Ke, R., Zhuang, Y., Pu, Z., Wang, Y.: A smart, efficient, and reliable parking surveillance system with edge artificial intelligence on IoT devices. IEEE Trans. Intell. Transp. Syst. 22(8), 4962–4974 (2020)
8. Unzueta, L., Nieto, M., Cortés, A., Barandiaran, J., Otaegui, O., Sánchez, P.: Adaptive multicycle background subtraction for robust vehicle counting and classification. IEEE Trans. Intell. Transp. Syst. 13(2), 527–540 (2011)
9. Quesada, J., & Rodriguez, P. (2016, September). Automatic vehicle counting method based on principal component pursuit background modeling. In 2016 IEEE International conference on image processing (ICIP), pp. 3822–3826. IEEE.
10. Hadi, R.A., George, L.E., Mohammed, M.J.: A computationally economic novel approach for real-time moving multi-vehicle detection and tracking toward efficient traffic surveillance. Arab. J. Sci. Eng. 42(2), 817–831 (2017)
11. Jayashri, S.: Video analytics on social distancing and detecting mask. Turk. J. Comput. Math. Educ. (TURCOMAT) 12(9), 2916–2921 (2021)
12. Zuo, F., Gao, J., Kurkcu, A., Yang, H., Ozbay, K., Ma, Q.: Reference-free video-to-real distance approximation-based urban social distancing analytics amid COVID-19 pandemic. J. Transp. Health 21, 101032 (2021)
13. El Gemayel, C., Chaccour, K., & El Gemayel, J. (2021, September). Automated face detection and control system using computer vision based video analytics to avoid the spreading of Covid-19. In 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 87–92). IEEE.
14. Pi, Y., Nath, N.D., Sampathkumar, S., Behzadan, A.H.: Deep learning for visual analytics of the spread of COVID-19 infection in crowded urban environments. Nat. Hazard. Rev. 22(3), 04021019 (2021)
15. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 34(3), 334–352 (2004)
16. Garcia-Garcia, B., Bouwmans, T., Silva, A.J.R.: Background subtraction in real applications: Challenges, current models and future directions. Comput. Sci. Rev. 35, 100204 (2020)
17. Bouwmans, T., Maddalena, L., Petrosino, A.: Scene background initialization: A taxonomy. Pattern Recogn. Lett. 96, 3–11 (2017)
18. Bouwmans, T.: Traditional and recent approaches in background modeling for foreground detection: An overview. Comput. Sci. Rev. 11, 31–66 (2014)
19. Bouwmans, T., Javed, S., Sultana, M., Jung, S.K.: Deep neural network concepts for background subtraction: A systematic review and comparative evaluation. Neural Netw. 117, 8–66 (2019)
20. Mandal, M., & Vipparthi, S. K. (2021). An empirical review of background subtraction algorithms by genetic programming. IEEE Trans. Evol. Comput. 25(1), 359–373 (2014)
21. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: SuBSENSE: A universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. 24(1), 246–252.
22. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: SubSENSE: A universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. 24(1), 177–180 (2015)
23. Bianco, S., Ciocca, G., Schettini, R.: Combination of video change detection algorithms by genetic programming. IEEE Trans. Evol. Comput. 21(6), 914–928 (2017)
24. Braham, M., & Van Droogenbroeck, M. (2016, May). Deep background subtraction with scene-specific convolutional neural networks. In 2016 international conference on systems, signals and image processing (IWSSIP) (pp. 1–4). IEEE.
25. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: SuBSENSE: A universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. 24(1), 359–373 (2014)
26. St-Charles, P.L., Bilodeau, G.A., Bergevin, R.: SuBSENSE: A universal change detection method with local adaptive sensitivity. IEEE Trans. Image Process. 24(1), 177–180 (2015)
27. Bianco, S., Ciocca, G., Schettini, R.: Combination of video change detection algorithms by genetic programming. IEEE Trans. Evol. Comput. 21(6), 914–928 (2017)
28. Braham, M., & Van Droogenbroeck, M. (2016, May). Deep background subtraction with scene-specific convolutional neural networks. In 2016 international conference on systems, signals and image processing (IWSSIP) (pp. 1–4). IEEE.
29. Lim, L.A., Keles, H.Y.: Learning multi-scale features for foreground segmentation. Pattern Anal. Appl. 23(3), 1369–1380 (2020)
30. Wang, Y., Jodoin, P. M., Porikli, F., Konrad, J., Benezeth, Y., & Ishwar, P. (2014). CDNet 2014: An expanded change detection benchmark dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 387–394).
31. Hbij K. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance problems in object detection: A review. IEEE Trans. Pattern Anal. Mach. Intell. 43(10), 3388–3415 (2020)
32. Jaffari, R., Hashmani, M.A., Reyes-Aldasoro, C.C.: A novel focal phi loss for power line segmentation with auxiliary classifier U-Net. Sensors 21(8), 2803 (2021)
33. Babaei, M., Dinh, D.T., Rigoll, G.: A deep convolutional neural network for video sequence background subtraction. Pattern Recogn. 76, 635–649 (2018)
34. Mandal, M., Dhar, V., Mishra, A., Vipparthi, S.K., Abdel-Motaleb, M.: 3DCD: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos. IEEE Trans. Image Process. 30, 546–558 (2020)
35. Shahbaz, A., Jo, K.H.: Dual camera-based supervised foreground detection for low-end video surveillance systems. IEEE Sens. J. 21(7), 9359–9366 (2021)
36. Kim, J.Y., Ha, J.E.: Foreground objects detection by U-Net with multiple difference images. Appl. Sci. 11(4), 1807 (2021)
37. Sukkos, D., Ho, E.S., Shum, H.P.: Illumination-aware multi-task GANs for foreground segmentation. IEEE Access 7, 10976–10986 (2019)
38. Zhang, J., Li, Y., Chen, F., Pan, Z., Zhou, X., Li, Y., Jiao, S.: X-net: A binocular summation network for foreground segmentation. IEEE Access 7, 71412–71422 (2019)
39. Yang, Y., Zhang, T., Hu, J., Xu, D., Xie, G.: End-to-end background subtraction via a multi-scale spatio-temporal model. IEEE Access 7, 97949–97958 (2019)
40. Yang, Y., Ruan, J., Zhang, Y., Cheng, X., Zhang, Z., & Xie, G. (2021). STPNet: A Spatial-Temporal Propagation Network for Background Subtraction. IEEE Transactions on Circuits and Systems for Video Technology.
41. Song, L., Lin, J., Wang, Z.I., Wang, H.: An end-to-end multi-task deep learning framework for skin lesion analysis. IEEE J. Biomed. Health Inform. 24(10), 2912–2921 (2020)
42. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: U-DUNet: a dimension-fusion U-shape network for chronic stroke lesion segmentation. IEEE/ACM Trans. Comput. Biol. Bioinf. 18(3), 940–950 (2019)
43. Tezcan, M.O., Ishwar, P., Konrad, J.: BSUV-Net 2.0: spatio-temporal data augmentations for video-agnostic supervised background subtraction. IEEE Access 9, 53849–53860 (2021)
44. Sobral, A., Vacavant, A.: A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. Comput. Vis. Image Underst. 122, 4–21 (2014)
45. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfifnder: real-time tracking of the human body. IEEE Trans. Pattern Anal. Mach. Intell. 19(7), 780–785 (1997)
46. Barnich, O., & Van Droogenbroeck, M. (2009, April). ViBe: a powerful random technique to estimate the background in video sequences. In 2009 IEEE international conference on acoustics, speech and signal processing (pp. 945–948). IEEE.
47. Lim, L.A., & Keles, H.Y.: Foreground segmentation using a triple convolutional neural network for multiscale feature encoding. arXiv 2018. arXiv preprint arXiv:1801.02225.
48. Lim, L.A., Keles, H.Y.: Foreground segmentation using convolutional neural networks for multiscale feature encoding. Pattern Recogn. Lett. 112, 256–262 (2018)
51. Zheng, W., Wang, K., Wang, F.: Background subtraction algorithm based on Bayesian generative adversarial networks. Acta Automatica Sinica 44(5), 878–890 (2018)

52. Zheng, W., Wang, K., Wang, F.: A novel background subtraction algorithm based on parallel vision and Bayesian GANs. Neurocomputing 394, 178–200 (2020)

53. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M. (eds.) International Conference on Medical image computing and computer-assisted intervention, pp. 234–241. Springer, Cham (2015)

54. Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-Net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9, 82031–82057 (2021)

55. Tezcan, O., Ishwar, P., & Konrad, J. (2020). BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2774–2783).

56. Kim, J.Y., Ha, J.E.: Foreground objects detection using a fully convolutional network with a background model image and multiple original images. IEEE Access 8, 159864–159878 (2020)

57. Rahmon, G., Bunyak, F., Seetharaman, G., & Palaniappan, K. (2021, January). Motion U-Net: multi-cue encoder-decoder network for motion segmentation. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 8125–8132). IEEE.

58. Bunyak, F., Palaniappan, K., Nath, S.K., Seetharaman, G.: Flux tensor constrained geodesic active contours with sensor fusion for persistent object tracking. J. Multimed. 2(4), 20 (2007)

59. Zivkovic, Z., Van Der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recogn. Lett. 27(7), 773–780 (2006)

60. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal Loss for Dense Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 42(2), 318–327 (2020)

61. Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565–571). IEEE.

62. Cueva, C., Yu, E.M., Garcia, N.: Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA. Comput. Vis. Image Underst. 152, 103–117 (2016)

63. Wang, Y., Luo, Z., Jodoin, P.M.: Interactive deep learning method for segmenting moving objects. Pattern Recogn. Lett. 96, 66–75 (2017)

64. Jiang, S., Lu, X.: WeSamBE: A weight-sample-based method for background subtraction. IEEE Trans. Circuits Syst. Video Technol. 28(9), 2105–2115 (2017)

65. Lee, S.H., Lee, G.C., Yoo, J., Kwon, S.: Wisenetmd: Motion detection using dynamic background region analysis. Symmetry 11(5), 621 (2019)

66. Haines, T.S., Xiang, T.: Background subtraction with dirichlet-process mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 36(4), 670–683 (2013)

67. Berjón, D., Cueva, C., Morán, F., Garcia, N.: Real-time nonparametric background subtraction with tracking-based foreground update. Pattern Recogn. 74, 156–170 (2018)

68. Patil, P.W., Murala, S.: MSFgNet: A novel compact end-to-end deep network for moving object detection. IEEE Trans. Intell. Transp. Syst. 20(11), 4066–4077 (2018)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rudrika Kalsotra received the bachelor’s and master’s degree in computer applications from University of Jammu, Jammu and Kashmir, India, in 2013 and 2016, respectively. She has been awarded with gold medal for academic excellence in bachelor’s and master’s degree. She is currently pursuing the Ph.D. degree in computer science at Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India. Her research interest includes video analytics and image processing. Currently she is doing her research work on Moving Object Detection for automated video surveillance. She is working on Background Subtraction technique and deep learning algorithms.

Sakshi Arora received bachelor’s degree in Sciences and master’s degree in Computer Applications from University of Jammu. She has received her Ph.D. degree in computer science from Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India. She has over ten years’ experience in teaching. She is currently an Assistant Professor in School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Jammu and Kashmir, India. Her research interests include image processing, pattern recognition and metaheuristic techniques for real world optimization. She has authored several research papers, published at reputed journals and conferences. She has been associated with various international conferences as paper referee/ program technical committee member.