IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2024, accepted 20 March 2024, date of publication 27 March 2024, date of current version 2 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381612

RESEARCH ARTICLE

An Improved VGG-19 Network Induced Enhanced Feature Pooling for Precise Moving Object Detection in Complex Video Scenes

PRABODH KUMAR SAHOO[1], (Member, IEEE), MANOJ KUMAR PANDA[2], UPASANA PANIGRAHI[3], GANAPATI PANDA[3], (Life Senior Member, IEEE), PRINCE JAIN[1], MD. SHABIUL ISLAM[4], (Senior Member, IEEE), AND MOHAMMAD TARIQUL ISLAM[5], (Senior Member, IEEE)

[1]Department of Mechatronics, Parul Institute of Technology, Parul University, Waghodia, Vadodara, Gujarat 391760, India
[2]Department of Electronics and Communication Engineering, GIET University, Gunupur, Rayagada, Odisha 765022, India
[3]Department of Electronics and Communication, C. V. Raman Global University, Bhubaneswar, Odisha 752054, India
[4]Faculty of Engineering (FOE), Multimedia University, Cyberjaya, Selangor 63100, Malaysia
[5]Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, UKM, Bangi, Selangor 43600, Malaysia

Corresponding authors: Mohammad Tariqul Islam (tariqul@ukm.edu.my), Md. Shabiul Islam (shabiul.islam@mmu.edu.my), and Prabodh Kumar Sahoo (sahooprabodhkumar@gmail.com)

ABSTRACT Background subtraction is a crucial stage in many visual surveillance systems. The prime objective of any such system is to detect local changes, and the system could be utilized to face many real-life challenges. Most of the existing methods have addressed the problems of moderate and fast-moving object detection. However, very few literature have addressed the issues of slow moving object detection and these methods need further improvement to enhance the efficacy of detection. Hence, within this article, our significant endeavor involved identifying moving objects in challenging videos through an encoder-decoder architectural design, incorporating an enhanced VGG-19 model alongside a feature pooling framework. The proposed algorithm has various folds of novelties: a pre-trained VGG-19 architecture is modified and is used as an encoder with a transfer learning mechanism. The proposed model learns the weights of the improved VGG-19 model by a transfer-learning mechanism which enhances the model's efficacy. The proposed encoder is designed using a smaller number of layers to extract crucial fine and coarse scale features necessary for detecting the moving objects. The feature pooling framework (FPF) employed is a hybridization of a max-pooling layer, a convolutional layer, and multiple convolutional layers with distinct sampling rates to retain the multi-scale and multi-dimensional features at different scales. The decoder network consists of stacked convolution layers projecting from feature to image space effectively. The developed technique's efficacy is demonstrated against thirty-six state-of-the-art (SOTA) methods. The outcomes acquired by the developed technique are corroborated using subjective as well as objective analysis, which shows superior performance against other SOTA techniques. Additionally, the proposed model demonstrates enhanced accuracy when applied to unseen configurations. Further, the proposed technique (MOD-CVS) attained adequate efficiency for slow, moderate, and fast-moving objects simultaneously.

INDEX TERMS Deep neural network, background subtraction, transfer learning, encoder-decoder architecture, feature pooling framework.

The associate editor coordinating the review of this manuscript and approving it for publication was Ganesh Naik.

I. INTRODUCTION

Visual surveillance is an essential technique for safety purposes and has two key steps: foreground separation followed by tracking. However, for an effective surveillance system local change detection is the primary assignment. For the last several decades, local change detection from challenging video scenes has been an arduous task and one of the diligent research areas in a visual surveillance system. Foreground segmentation from the image frames has numerous applications: activity recognition [1], traffic supervision [2], industrial monitoring [3], underwater surveillance [4] etc. The moving object detection process can retain the moving objects from the background in a sequence of complex video scenes. Therefore, the aforementioned procedure can be considered a binary classification task in which the pixels corresponding to the background are eliminated, while the pixels resembling moving objects are retained. Foreground separation from the complex video scene is challenging due to the dynamic background, camera jitter, missing information, slowly moving objects, etc. Background subtraction (BGS) approach [5] is the prominent way to partition the foreground from the background. The moving objects of the image frame are separated from the background in this approach. In the last few decades, several researchers developed various techniques across the globe for BGS. However, these existing techniques are well performed for specific challenges only. Also, the effectiveness of these BGS techniques is based on manual parameter tuning and handcrafted features. This gives rise to concerns regarding the development of more efficient and resilient techniques for detecting moving objects. Deep learning frameworks have been instrumental in advancing computer vision applications over the years. Also, for moving object detection deep neural networks are extensively used today, as they can retain low, mid, and high levels feature [6], [7], [8]. Further, the efficiency of the deep neural networks can be enhanced by utilizing a transfer learning strategy.

Several drawbacks have been identified in the deep neural network architecture for local change detection. The incorporation of deep learning frameworks in visual surveillance intensifies the complexity of the system. It has been observed that as the depth of the layers increases, the complexity of the model also escalates. Furthermore, it has been noted that training the deep neural network necessitates a larger number of sample frames. Additionally, it is rare to come across an end-to-end model for local change detection in existing techniques.

Therefore, a remarkable deep learning architecture in the form of an encoder-decoder model that effectively addresses multiple challenges encountered in complex as well as slow-moving video scenes is developed. An improved version of pre-trained VGG-19 deep learning framework as the encoder is adopted in the proposed methodology. The initial two blocks' weights were set as pre-trained weights, while the weights of the third block were trained specifically on challenging datasets, enhancing the model's resilience. With a transfer learning strategy, the proposed VGG-19 deep neural network preserves the appropriate features for moving object detection. Subsequently, the feature maps obtained from the encoder are fed into the feature-extracting framework, where features are pooled across different scales along the depth dimension. This is accomplished through the utilization of a max-pooling layer, a convolution layer, and multiple convolutional layers with distinct sampling rates. The decoder network in the proposed scheme effectively projects the feature label to the pixel label.

Therefore, the MOD-CVS contributes in five main ways:

1) A first and unique attempt for detecting local changes in challenging video datasets comprising of moderate, fast, and slowly-moving objects is made in this work using a feature pooling framework with the improved version of VGG-19 encoder-decoder type architecture.
2) The proposed algorithm provides better accuracy in four datasets with diverse challenges including slowly-moving objects, moderately and fast-moving objects, Indoor and outdoor image sequences, dynamic backgrounds, camera jitter, night video, low frame rate, thermal, etc.
3) The proposed model utilized fewer samples to train and attained better accuracy without extracting the temporal information from the challenging video scenes against current SOTA approaches.
4) Incorporating a transfer learning mechanism in the suggested scheme makes the model to learn the weights efficiently and enhances the efficiency.
5) A selected number of blocks in the proposed VGG-19 architecture is used to make the model less complex compared to the existing deep neural networks.

The efficiency of the developed model is substantiated by its validation on benchmark data sets, specifically designed for slowly-moving object detection [9], [10], fast and moderate moving object detection like *CD-Net 2014* dataset [11], *wallflower* dataset [12], and *Star* dataset [13]. To confirm our findings, the outcomes of the developed technique are compared to thirty-six SOTA techniques. To verify the efficacy of the developed technique, we conducted both visual and quantitative analyses, confirming its effectiveness.

The structure of the remaining sections in the article is as outlined below. Section II discusses the literature's presentations concerning local change detection. The proposed model in depth with a graphical illustration is discussed in Section III. Section IV describes the empirical outcomes analysis and ablation study. Section V provides the article's conclusions along with a glimpse into future work.

II. STATE-OF-THE-ART-TECHNIQUES

One of the most widely investigated topics in the field of computer vision involves the detection of local changes utilizing the background subtraction technique. Many researchers have worked for decades to build robust background subtraction algorithms that can detect objects in motion for

complex scenes including objects moving at a relatively low speed where the object motion is confined to a smaller region, objects moving at a relatively high speed where the subsequent frames have higher variation, variations in illumination, camera jitter, shadow, image captured at night time, low-frame rate, low contrast, low resolution, non-static background, etc. Taking into account the latest literature, the SOTA techniques are divided into two parts as follows:

A. SOTA TECHNIQUES FOR SLOW MOVING OBJECT DETECTION

Slowly moving object detection deals with the process of identifying and tracking objects that are moving at a relatively low speed, where the subsequent frames have lesser variation. In most cases, the spatial motion of the object is confined to a small area. While there are various techniques and approaches to object detection, detecting slowly moving objects can present specific challenges due to their reduced motion and potentially smaller visual cues. The most commonly used techniques for slowly moving object detection are frame differencing (FD) [14], optical flow (OF) [15], background subtraction (BGS) [16], Feature-based methods (FB) [17], [18], machine learning-based approaches (ML) [19]. The choice of method is decided based on the specific application, the characteristics of the slowly-moving objects, and the available computational resources. Combination or adaptation of multiple techniques may also be required to achieve accurate extraction of slowly-moving objects followed by tracking in various scenarios [20]. The BGS is an effective technique for detecting fast and moderately moving objects in a scene. It provides accurate segmentation of the foreground in real-time with low computational cost when the object moves slowly in a relatively static background. However, this method is sensitive to lighting changes, limited to static background, finds it difficult to handle occlusion, and requires background modeling if the background is not available. Again, it fails to extract the slow-moving object due to limited spatial change of pixels in object area [21], [22], [23]. Moving object detection using OF method use the motion vectors of pixels to determine the direction and magnitude of movement. It is very accurate at identifying and following fast-moving objects. As such, this method is highly adaptable to variations in texture, lighting, and other factors, making it ideal for tracking objects in real-time in video surveillance applications. Nevertheless, it is not able to effectively handle occlusion and is sensitive to image noise. It does not provide depth information about the object being tracked. It may not work well for objects that are not moving or moving slowly as it relies on the movement of objects [15], [24], [25]. FD is a common method for detecting moving objects in a video sequence. It is a fast method that can operate in real time, making it suitable for surveillance systems. It provides a cost-effective solution and can even detect objects that are partially occluded, by comparing changes between frames. However, it is sensitive to noise and small changes, such as camera shake or changes in lighting, which produce false positives and affect the accuracy of the result. It only detects moving objects that differ from the background, making it unsuitable for detecting objects having a similar color or texture as that of background [26], [27]. The FB methods can handle challenging scenarios where the appearance of the object varies due to different lighting or complex backgrounds. These methods can reduce the computational burden and process video in real time by extracting specific image features. On the contrary, the performance of these methods degrades significantly, if the features are not robustly detected or the features are affected by noise or occlusion. They also often require fine-tuning or retraining when dealing with new object classes or motion characteristics. These methods primarily pay attention to low-level image features like edges, corners, or texture patterns, without explicitly incorporating high-level semantic information. As a result, they may not be able to differentiate between objects with similar low-level features, leading to errors in object detection or tracking [28], [29], [30].

B. SOTA TECHNIQUES FOR MODERATELY AND FAST MOVING OBJECT DETECTION

Moderately and fast moving object detection deals with the process of identifying and tracking objects that are moving at a relatively high speed where the subsequent frames have higher variation.

Some of the SOTA ML and Deep-learning-based approaches have been discussed for moderate and fast-moving object detection in the literature. The object detection algorithm known as Single Shot Multi-Box Detector (SSD), introduced by Liu et al. [31] demonstrates efficient object detection in images by achieving a favorable balance between accuracy and speed. SSD applies non-maximum suppression (NMS) to filter out redundant bounding box prediction and produce the final set of object detection. The prime advantages of SSD are its simplicity, speed, and ability to detect objects at multi-scales. However, SSD sacrifices some accuracy for faster inference speed. It utilizes a predetermined set of anchor boxes to detect objects at various scales. Choosing the right scales and aspect ratios for these anchor boxes can be challenging. Objects that significantly deviate from these predefined anchor boxes may not be accurately detected. It can not also handle highly occluded objects. The faster R-CNN framework for object detection introduced by Ren et al. [32] become a popular and influential method in computer vision. This technique efficiently and accurately localizes objects with high precision using a region proposal network (RPN) to generate candidate object proposals. It allows the entire object detection system for end-to-end training towards optimizing the overall performance. However, it is more complex compared to previous object detection methods. It involves multiple components, including a region proposal network, a shared convolutional backbone, and an object-specific classifier.

This complexity can make it more challenging to understand and implement. Lin et al. [33] introduced a novel loss function called Focal Loss, specifically designed for dense object detection tasks like object detection and instance segmentation. The Focal Loss addresses the issue of class imbalance and the overwhelming number of easy negative examples that can hinder the training of object detectors. Focal Loss introduces an additional hyper-parameter, called the focusing parameter, which controls the rate at which the loss is down-weighted for easy negatives. Choosing an appropriate value for this parameter requires careful tuning, and an improper setting can affect the performance of the model. Zhou et al. [34] propose a novel method for object detection called ''Objects as Points'', which aims at achieving efficient real-time object detection. This method demonstrates impressive real-time performance, enabling fast object detection in videos and live-streaming applications. While the method achieves high accuracy in detecting objects, the localization accuracy may not be as precise as some other object detection methods that rely on bounding boxes. This limitation might affect tasks that require precise localization, such as object tracking or fine-grained object recognition. Hei Law and Jia Deng. present a novel object detection framework called CornerNet, which detects objects by treating them as paired key points. CornerNet represents objects as key points and models the object's spatial information, which helps in precise localization and reduces false positives. This approach treats objects as key points. Hence, it may struggle with objects that have complex or highly variable poses. Since the model primarily focuses on detecting corners, it may not be as effective in cases where key points are not prominent or informative [35]. Zhu et al. [36] proposed Generative Adversarial Networks (GANs) which is a popular class of deep learning models used in generative modeling tasks, such as image synthesis and data generation. Images, audio, and text can all be produced by GANs in a realistic manner while still closely resembling the training set. This makes them valuable in various applications, including art generation, data augmentation, and synthetic data creation for training other models. These networks are prone to a phenomenon known as mode collapse, wherein the generator generates a restricted range of samples, thereby not capturing the full distribution of data. This results in generated samples that lack diversity and fail to cover all the modes present in the training data. MotionRec [37] is composed of a temporal depth reductionist (TDR) block, a motion saliency estimation (MoSENet) network, and regression and classification blocks. This represents the initial endeavor to concurrently localize and classify dynamic entities within a video referred to as MOR (Moving object recognition) using a unified deep learning framework in a single stage *CDNet-2014* dataset. Unified frameworks may sometimes be computationally expensive. In AE-NE [38] is entirely unsupervised. It operates with a fixed set of hyperparameters, and the architecture of the autoencoder is dynamically determined based on image size and background complexity, devoid of manual supervision. The autoencoder is additionally trained to anticipate background noise, enabling the calculation of a pixel-dependent threshold for foreground segmentation in each frame. This model is ill-suited for processing night videos, as indicated by the low score it achieved in this category on the *CDNet-2014* dataset. The model is not recommended for scenarios where the video is anticipated to depict substantial stationary objects over an extended duration. TSS [39] has made significant contributions to computer vision and video analysis. This method can learn hierarchical features from data, enabling them to discern patterns and variations in motion more effectively than traditional methods. It may be highly specialized and may not generalize well across different domains or environmental conditions. Fine-tuning or retraining might be necessary for optimal performance in diverse settings. A real time multiple object tracking [40] method is based on a modified version of deep simple online and real time tracking (Deep SORT) algorithm. Deep learning methods can handle a large number of objects and complex scenes simultaneously, making them suitable for tracking multiple objects in crowded environments. Training deep learning models for multiple object tracking requires large annotated datasets, which can be time-consuming and expensive to create, particularly for diverse scenarios. Jiawei et al. [41] proposed a 3D video object detection framework emphasizing enduring temporal visual correlation, termed BA-Det. BA-Det operates as a two-stage object detector, proficient in concurrently acquiring knowledge in object detection and temporal feature correspondence through the introduced feature metric object bundle adjustment (OBA) loss. The method exclusively concentrates on objects, such as cars, trucks, and trailers. The effectiveness of flexible objects like pedestrians has not been explored. Further, the related works on background subtraction techniques using explainable deep learning frameworks, outlined in a Table 1 while emphasizing the principal contributions, advantages and disadvantages.

It is found that all the above SOTA methods discussed in the literature related to slow moving object detection are capable of identifying the objects when the variation among the consecutive frames is much less. However, the SOTA schemes addressed for moderately and fast-moving object detection detect the object when there is a higher variation among the successive frames. Hence, from the above discussion, it may be concluded that a single method can not detect all types of moving objects at various speeds. This motivated us to develop a moving object detection framework using a VGG-19 architecture with structural modification-induced FPF to detect moving objects at slow, moderate, and fast speeds. In the proposed design the improved VGG-19 architecture can retain details at various levels. The proposed VGG-19 architecture-induced FPF module capable of preserving the details of objects at

TABLE 1. Summary of the existing background subtraction techniques using explainable deep learning frameworks.

Approaches	Contribution	Advantages	Disadvantages
xDNN [42]	Application of explainable deep neural networks in real-world scenario, in specific domains and industries.	Increases interpretability of model decisions. Facilitate better debugging and improvement of deep neural networks.	Finding the optimal equilibrium between interpretability and model accuracy can pose a difficulty.
DNNs [43]	This could involve developing methods for generating human-understandable explanations for model predictions.	Users who may not have a deep understanding of machine learning can benefit from explanations that help them comprehend the model's decisions.	It may come with a computational cost, potentially impacting the overall performance and efficiency of the model.
EfctResDet50 [44]	Applying object detection techniques to radiography images could expand the scope of automated analysis, providing valuable insights for medical professionals.	The ability to detect and classify multiple objects in radiography images can enhance the system's utility, allowing it to assess various aspects of image quality and identify multiple anomalies simultaneously.	Real-time, explainable, and multiclass object detection systems tend to be computationally intensive, requiring powerful hardware and significant resources.
Act Recogn [45]	Practical insights provided in the paper can be beneficial for practitioners looking to implement explainable deep learning models in real-time action recognition systems.	Focusing on real-time action recognition adds practical significance to the research, as real-time applications are crucial in various domains such as surveillance, robotics, and human-computer interaction.	Suffer from limitations in terms of the dataset used or the scope of actions considered, potentially affecting the generalizability of the findings.

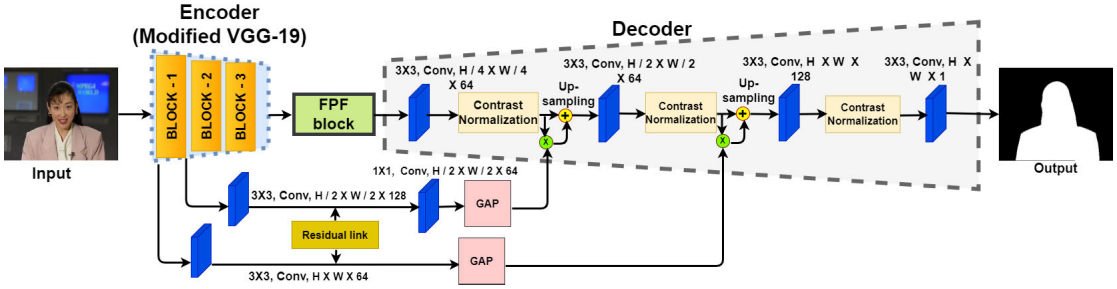

FIGURE 1. Representation of the developed BGS model in the block diagram.

several scales and different speeds. The designed decoder architecture can effectively project features to image space. Section III of the paper focuses on the proposed methodology.

III. THE PROPOSED ALGORITHM

This article presents a unique and durable deep-learning model for foreground segmentation from a complex video scene for various challenging scenarios. Here, we have developed a deep learning model in which a modified VGG-19 network is used as an encoder integrated with a feature pooling framework (FPF) to effectively detect objects at diverse sizes from the video scenes. The FPF block can retain the sparse and dense features from image frames that are suitable for local change detection. The decoder network learns a mapping from the feature label into a pixel label effectively. Fig.1 represents the developed network with the dimensions of each layer of the feature map in detail.

A. ENCODER NETWORK

The design in this work has improved the pretained VGG-19 network and adhered as an encoder network. A typical VGG-19 network is used for several image-processing applications. Nonetheless, the said framework has yet to be explored for foreground segmentation. Here, we have used the abilities of the VGG-19 network for foreground separation. The original VGG-19 network [46] has five blocks, each with stacked convolutional layers, and the activation function is the rectified linear unit (ReLU). Convolutional layers can retain the input image's spatial information and the ReLU function in the proposed model activates the required neurons that boost the efficiency of the architecture.

The proposed model capitulated with an altered form of deep VGG-19 network, which comprises the starting three blocks. Where the weights of the first two blocks are the same as the weights of the original VGG-19 architecture [46], and the weights of the third block are accomplished by using the transfer-learning (T-L) strategy for the challenging dataset. T-L, as a mechanism, assimilates information from the input to the output domain. In the developed technique, applying the T-L strategy investigates novel tasks built upon the foundation of tasks previously learned by the original deep VGG-19 network. Also, the T-L strategy enhances the model's speed and robustness, particularly when training on a limited number of samples. To optimize the utilization of high spatial resolution and frequency details, the fourth and fifth blocks of the original VGG-19 network have been omitted in the MOD-CVS. A detailed description of the altered VGG-19 deep learning model with dimensions of each layer of the feature map is presented in Fig. 2. The high spatial frequency features are retained at the first block of the encoder by using 3×3 convolution layers with 64 and 128 filters. It is found that the 3×3 kernel allows for the learning of hierarchical features. Also, the 3×3 convolutional layers with increased filter numbers (64 and 128) allow the network to capture intricate and diverse low-level features. This can

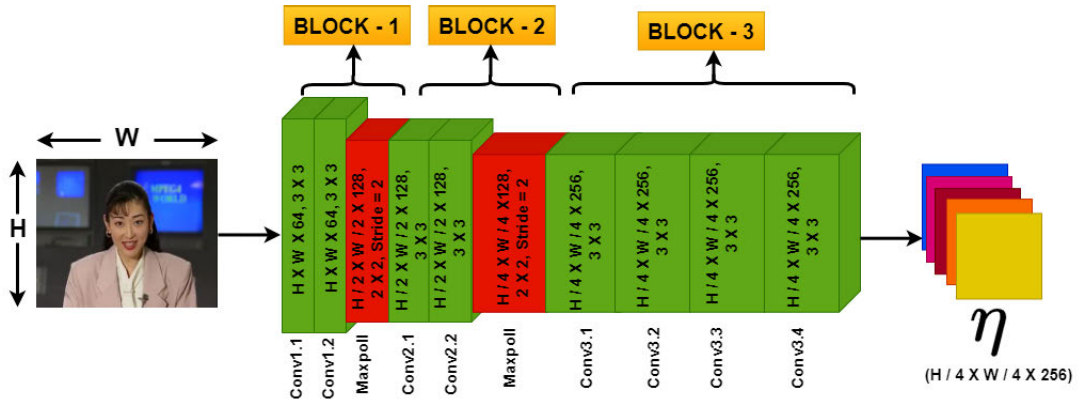

FIGURE 2. Detailed description of the altered VGG-19 deep learning model in the block diagram.

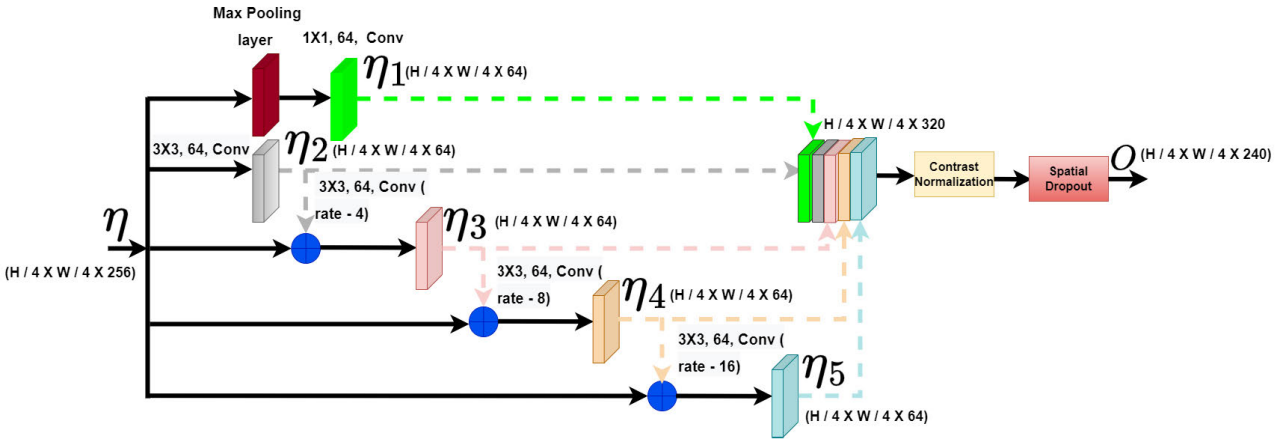

FIGURE 3. Proposed feature pooling framework.

enhance the model's ability to discriminate between classes and improve its overall efficacy in tasks such as segmentation, object detection, or recognition. These fine-scale features are transferred for the decoder network via skip connections and global average pooling (GAP), which enhances the feature presentation.

B. FEATURE POOLING FRAMEWORK

To effectively preserve objects of different scales from challenging video scenes, this work presents a feature pooling framework (FPF) between the encoder and decoder networks which is shown in Fig. 3. Also, the dimensions of each layer of the feature map of the FPF module are shown in Fig. 3. The max-pooling layer is hybridized in the FPF module with 64, 1×1 filter size convolutional (conv.) layer, 64, 3×3 filter size conv. layer, and atrous conv. layers with different dilation rates of 4, 8, and 16, respectively. The suggested approach uses atrous conv. layers with a 64, 3×3 filter size. Atrous conv. layers are valuable in certain contexts for enlarging the receptive field without increasing the number of parameters or computational cost. They are crucial in complex scenarios to capture broader context information without significantly inflating the model's complexity. The max-polling layer can retain the maximum information η_1 for taking window size 2×2 from the encoder output η. The conv. layer and different atrous conv. layers of the FPF block, which are effectively represented as η_2, η_3, η_4, and η_5, can anticipate sparse and dense feature space from the high-dimensional feature space η. Then, η_1, η_2, η_3, η_4, and η_5 features are concatenated along the channels and processed through contrast normalization (CN) followed by a spatial dropout layer with a rate of 0.25 to produce the FPF block output of 320 feature maps. Observations indicate that the proposed model demonstrates improved performance with the utilization of the CN layer instead of the batch normalization layer. Also, the choice of the dropout rate (0.25 in this case) is often based on experimentation and hyperparameter tuning. A rate of 0.25 implies that 25% of the features will be randomly dropped out during training, which is chosen to strike a balance between preventing over-fitting and allowing the network to learn from a variety of features. Additionally, the inclusion of a spatial dropout layer effectively preserves spatial information while simultaneously reducing redundant information.

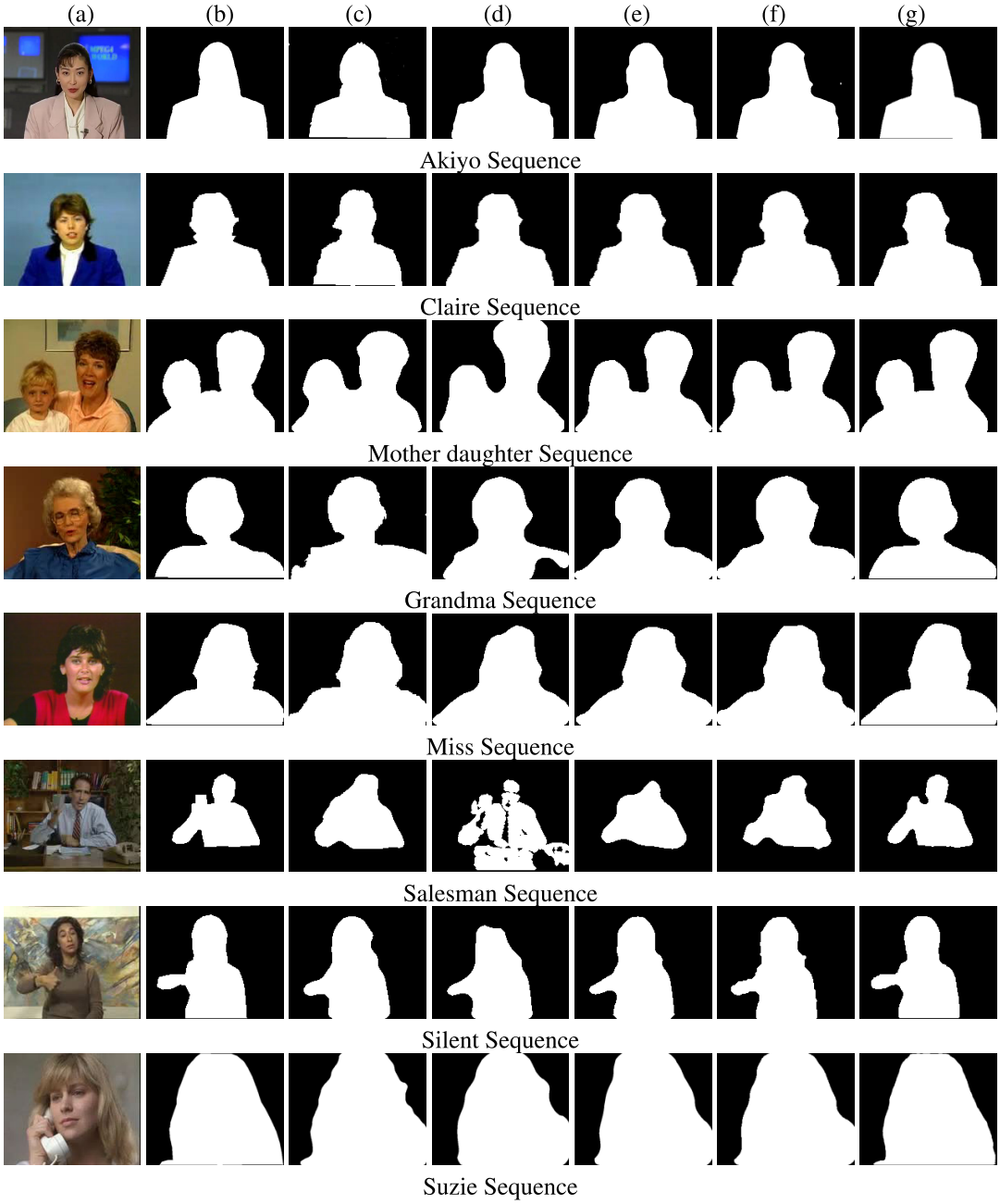

FIGURE 4. Foreground segmentation for various sequences: (a) original frame (b) ground-truth image, outcomes attained by background subtraction technique based on (c) Badri et al. [47], (d) Zhu et al. [48], (e) Sahoo et al. [20], (f) Sahoo et al. [49] and (g) MOD-CVS.

C. DECODER NETWORK

Spatial information of the complex video scene is essential for effective moving object detection. Therefore, the developed decoder network comprises a stack of conv. layers in the proposed model that preserve spatial information efficiently. The initial conv. layer consists of 64 filters with a 3×3 size, projecting the 240 feature maps obtained from the FPF block into 64 feature maps. These features are succeeded by the CN layer and the ReLU function is fused with the fine-scale features retained at the end of the first block of the encoder, followed by the GAP layer. The feature fusion is achieved using the coefficients obtained through the application of the GAP layer on the features that are extracted at the end of encoder BLOCK - 1 using 3×3 convolution layers with 128 filters perform element-wise multiplication (X) with the feature maps of the initial conv. layer of the decoder network. Subsequently, the resulting features are added (+) to the outputs of the initial conv. layer of the decoder network. Using the GAP layer in the decoder framework enhances the performance of the proposed model.

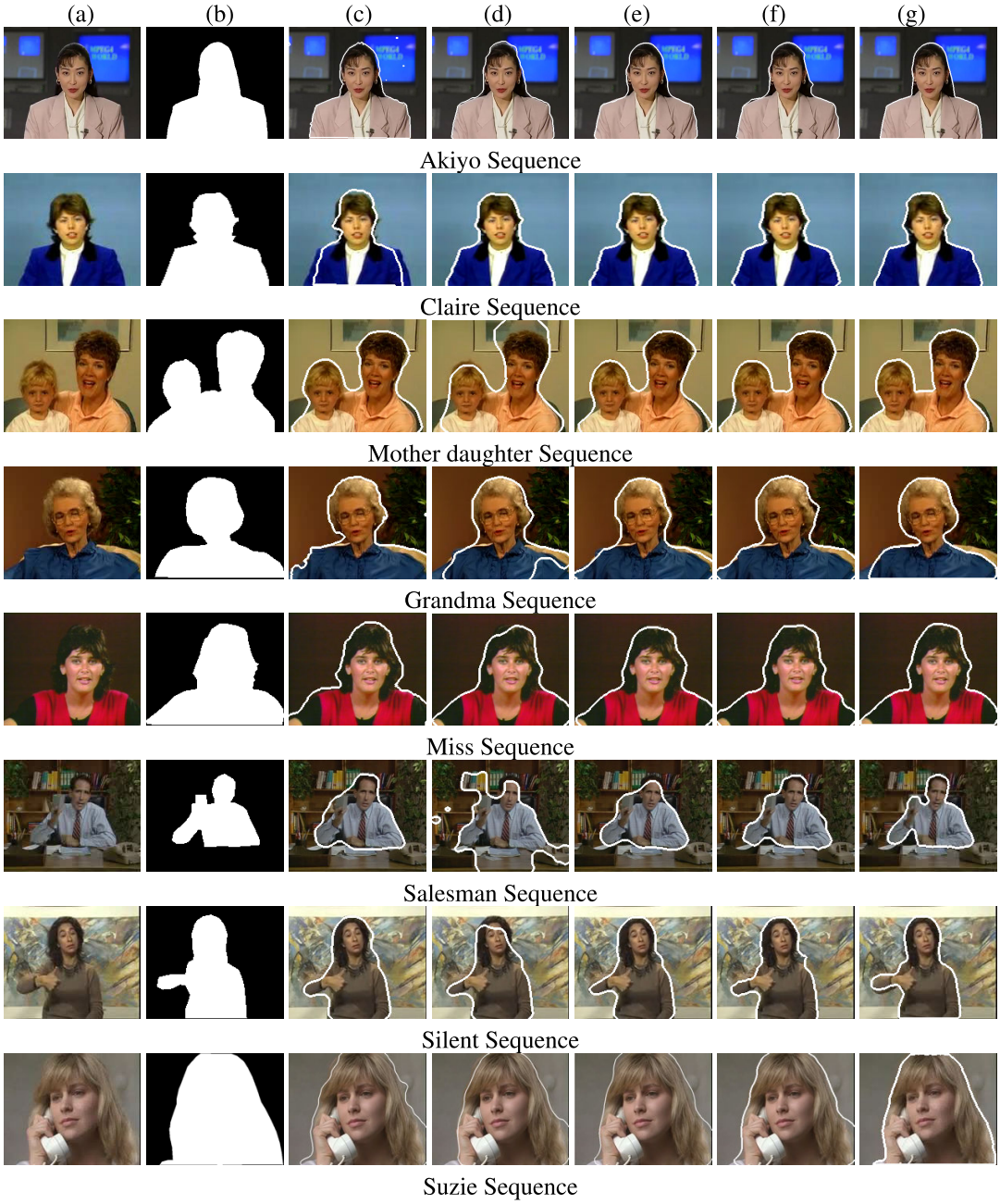

FIGURE 5. Foreground segmentation for various sequences: (a) original frame (b) ground-truth image, outcomes attained by background subtraction technique based on (c) Badri et al. [47], (d) Zhu et al. [48], (e) Sahoo et al. [20], (f) Sahoo et al. [49] and (g) MOD-CVS.

Afterward, the fused features are Up-sampled and passed through the second conv. layer consisting of 64 filters with a 3×3 size followed by the CN layer and ReLU function to generate the 64 feature maps. Again these feature maps are fused with the fine-scale features extracted at the beginning of the first block of the encoder, followed by the GAP layer. The feature fusion is achieved using the coefficients obtained through the application of the GAP layer on the features that are extracted at the beginning of encoder BLOCK - 1 using 3×3 convolution layers with 64 filters perform element-wise multiplication (X) with the feature maps of the second conv. layer of the decoder network. Subsequently, the resulting features are added (+) to the outputs of the second conv. layer of the decoder network. The fused features are Up-sampled and projected into 128 feature maps by utilizing a third conv. layer consisting of 128 filters with a 3×3 size. It is observed that these features provide a better presentation of the object and background pixels and boost the performance of the developed model. Eventually, a final conv. layer contains 1 filter with a

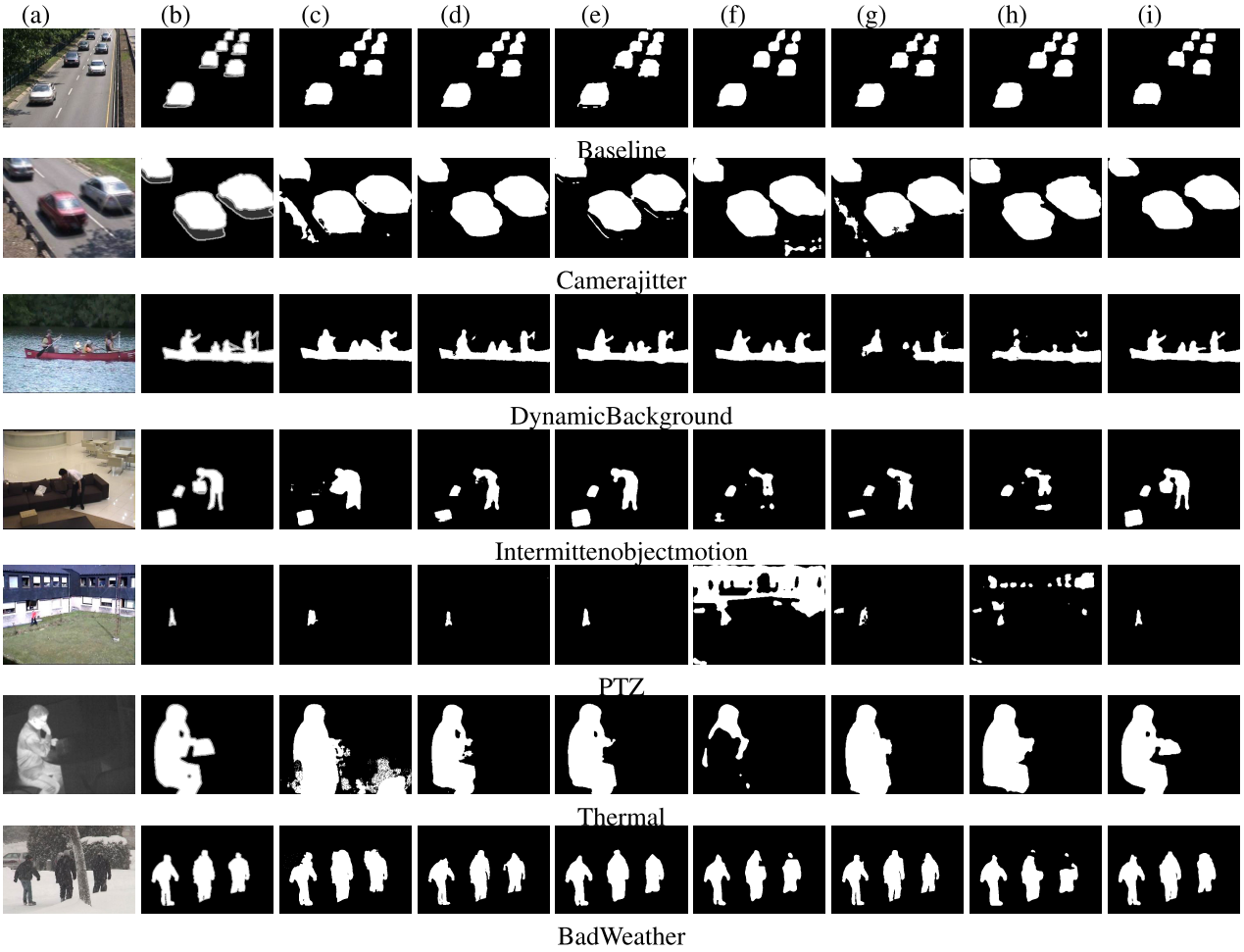

FIGURE 6. Foreground segmentation for various sequences: (a) original frame (b) ground-truth image, outcomes attained by BGS technique dependent on (c) BSUV-Net _SemanticBGS [50], (d) BSUV-Net 2.0 [51], (e) Cascaded CNN [52], (f) DeepBS [53], (g) Fast BSUV-Net 2.0 [51], (h) WisenetMD [54], and (i) MOD-CVS.

1×1 size preceding a sigmoid activation function that accurately projects the feature space into image space. A threshold value of 0.9 provides the mask effectively for the corresponding RGB input image. It is found that the threshold value of 0.9 provides better accuracy for challenging video scenes.

IV. ANALYSIS OF SIMULATION BASED EXPERIMENTAL RESULTS

The developed model is running on a Windows 10 operating system with 8GB RAM with Python programming. The proposed work is trained and tested over the *NVIDIA* Tesla *T4* GPU given by the Google Co-laboratory pro version. The proposed work is implemented by utilizing the TensorFlow backend with the Keras library. The significance of the presented model is tested on the challenging data sets [9], [10], [11], [12], [13]. The efficiency of our developed algorithm is corroborated by resembling its results with the outcomes acquired by thirty-six SOTA techniques using subjective and objective analysis.

A. PARAMETER SETTINGS AND TRAINING DETAILS

A *NVIDIA* Tesla *T4* GPU system with a batch size of 2 is used to train the model from beginning to end. The developed model's reduced batch size can have a special regularisation effect and help the model converge more quickly. There are P pixels in each frame and N = 25 frames are used to train this model. Furthermore, we train the model using the binary cross entropy loss (BCEL) function. This compares each pixel's actual and predicted class labels.

To train the proposed approach, we used the RMSProp optimizer with $\rho = 0.9$ and $\epsilon = 1e - 08$. Comparatively speaking to other traditional optimizers, this offers a faster convergence rate. The learning rate is initially set to 0.0001. The learning rate is subsequently scaled down by 10 if, after 5 consecutive epochs, the validation loss does not reduce. To train the model, we preserved a maximum of 100 epochs.However, if the validation loss did not decrease for ten consecutive epochs, an early stopping strategy was used. Sequential feeding of the training frames to the model could lead to biased learning weights. Because successive

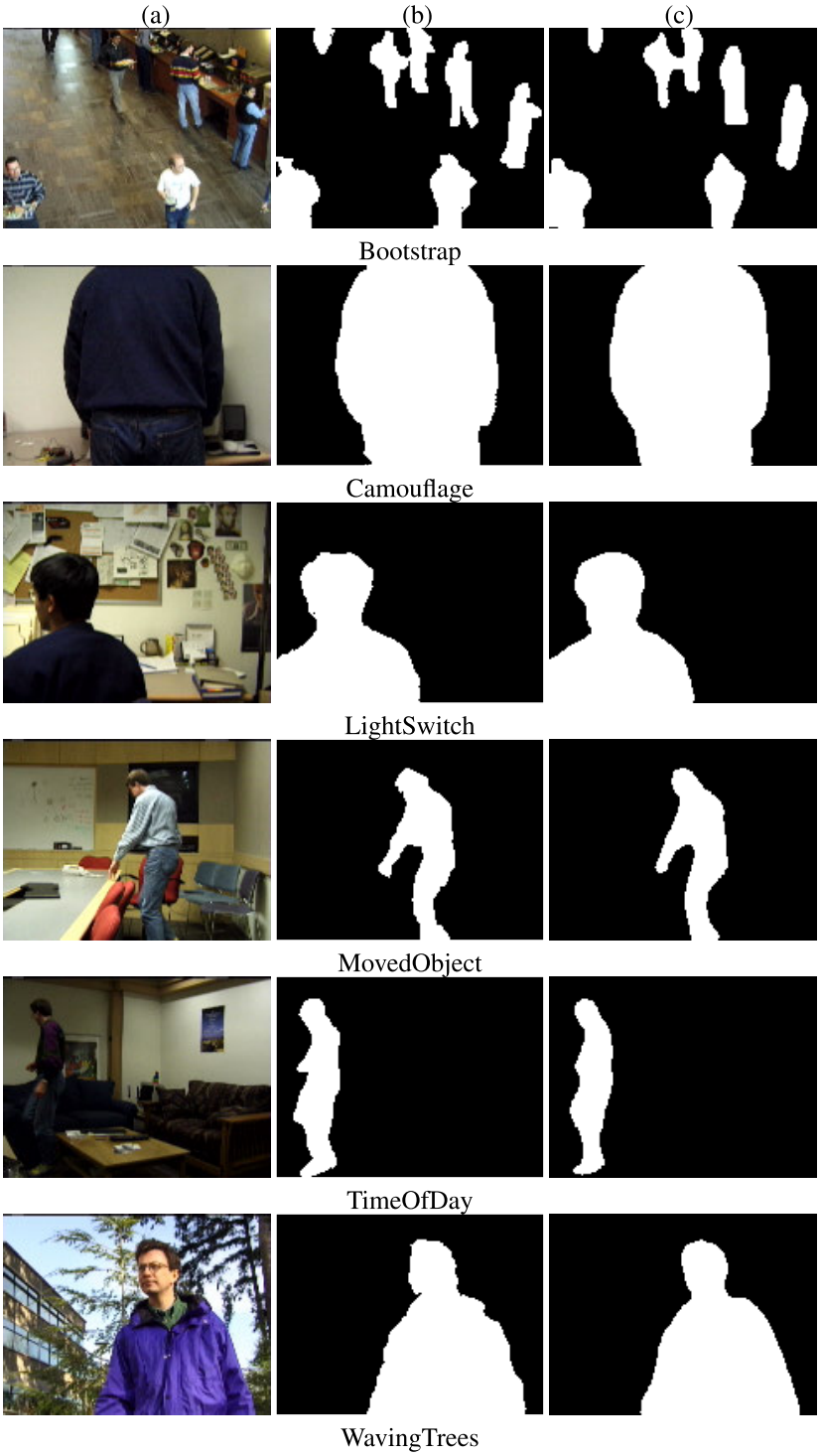

FIGURE 7. Foreground segmentation for various sequences:(a) original frame (b) ground-truth image, (c) outcomes attained by MOD-CVS for *wallflower* dataset.

frames have a strong correlation with one another, this issue occurs. As a result, we randomly select the training frames to train the model initially. These frames are split into 20% for validation and 80% for training. To solve the issue of imbalanced data classification during model training, we provide the foreground class with more weights and the background class with fewer weights.

B. SUBJECTIVE ANALYSIS

For slow moving objects the visual demonstration of the detected results achieved by the existing techniques and our developed algorithm is presented in Fig. 4. Fig. 4 (a) and (b) depict the original frames and associated ground-truth images, respectively. The results obtained by the Badri et al. [47] technique are presented in Fig. 4 (c)

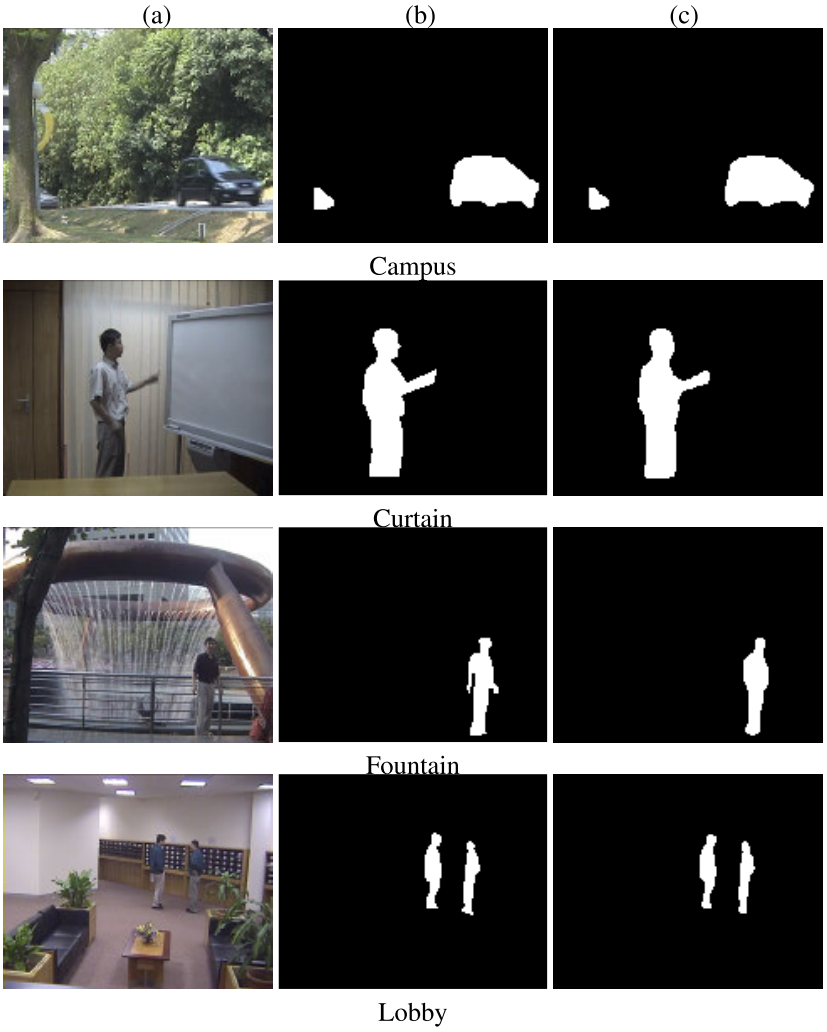

FIGURE 8. Foreground segmentation for various sequences: (a) original frame (b) ground-truth image, (c) outcomes attained by MOD-CVS for *Star* dataset.

where the said technique detected background pixels as foreground pixels for various slow moving image sequences. Fig. 4 (d) represents the detected outcomes obtained by the Zhu et al. [48] scheme where the missed alarm rate is high. Fig. 4 (e) and Fig. 4 (f) denote the outcomes achieved by the Sahoo et al. [20], and Sahoo et al. [49] techniques respectively, where a high false negative rate is observed. The outcomes attained by the developed model are illustrated in Fig. 4 (g) where the background and foreground pixels are classified accurately. Fig. 5 (a) and Fig. 5 (b) indicate the input images and the associated ground-truth frames. From Fig. 5 (g), it is evident that the developed technique accurately captured the moving object shape, demonstrating lower false negative and false positive rates against the Badri et al. [47], Zhu et al. [48], Sahoo et al. [20], and Sahoo et al. [49] existing techniques presented in 5 (c), (d), (e), and (f) respectively.

The change detection output is visually analyzed using seven sequences chosen from the *CD-Net 2014* dataset. The challenging effects on video scenes include low contrast, non-static background, low frame rate, noise, shadow, poor resolution, low signal-to-noise ratio, lack of object shape and textural details in the images, etc. The developed technique's performance is visually compared with that of six established deep learning methods, including BSUV-Net _SemanticBGS [50], BSUV-Net 2.0 [51], Cascaded CNN [52], DeepBS [53], Fast BSUV-Net 2.0 [51], WisenetMD [54]. Fig. 6 (a) and (b) represent input images and their associated ground-truth frames respectively. The object detection outcomes achieved by BSUV-Net_ SemanticBGS as demonstrated in Fig. 6 (c), where it can be seen that the background is identified as the foreground. Fig. 6 (d) represents BSUV-Net 2.0 [51] technique outcomes where numerous false alarms are present in the target scene. The segmented outcome of the Cascaded CNN [52] method is showcased in Fig. 6 (e), where the said technique is unable to detect a few information of the object in motion. Fig. 6 (f) shows outcomes of the DeepBS [53] method where numerous edge pixels are absent due to imbalanced pixel values across

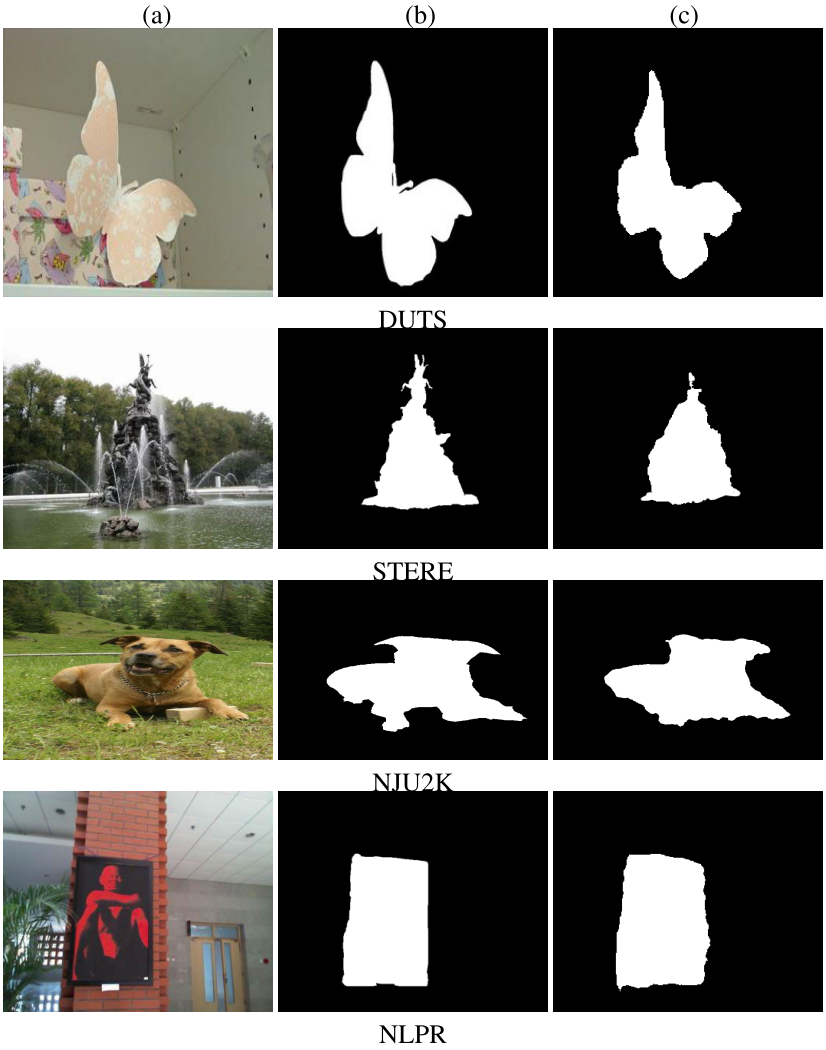

FIGURE 9. Foreground segmentation for various sequences: (a) original frame (b) ground-truth image, (c) outcomes attained by MOD-CVS for DUTS [55], STERE [56], NJU2K [57], and NLPR [58] datasets.

various video frames, this leads to a significant number of missed alarms in the detected outcomes. The outcomes of the Fast BSUV-Net 2.0 [51] technique are represented in Fig. 6 (g), where the mentioned technique incorrectly categorized certain pixels of an object as background. Fig. 6 (h) represents the WisenetMD [54] algorithm's results, where this method encounters difficulty in discerning subtle variations in grey values, resulting in the generation of ghost. In contrast, the MOD-CVS showcased in Fig. 6 (i), gives better performance against the existing SOTA techniques as well as precisely classifying background and foreground accurately. In complex video scenes, the developed technique can successfully determine the shapes of moving objects. The MOD-CVS is further tested in *wallflower* datasets as shown in Fig. 7. Original frame and associated ground truth images are represented in Fig. 7 (a) and (b). Fig. 7 (c) illustrate the result attained by the developed technique. From Fig. 7 (c), it is found that the developed technique attained better results for *wallflower* dataset. Again, the MOD-CVS is validated on *Star* dataset. The input frame and their corresponding ground-truth image are presented in Fig. 8 (a) and (b). Fig. 8 (c) illustrates the proposed method's results where it is noted that the MOD-CVS has the capability to accurately classifying the foreground as well as background pixels with lesser noise.

C. OBJECTIVE ANALYSIS

To assess the efficacy of the developed technique, we have made a quantitative distinction between the developed technique and the prevailing SOTA techniques for slow moving objects including average F-measure (AF) and average miss classification error (AMCE) are outlined in Table 2, and Table 3. From these Tables, It is found that the developed algorithm attained a greater value of AF with a reduced value of AMCE against the Badri et al. [47], Zhu et al.

TABLE 2. Comparison of average F measure in percentage of proposed scheme with different SOTA techniques.

Sample Video	Methods				
	Badri *et al.* [47]	Zhu *et al.* [48]	Sahoo *et al.* [20]	Sahoo *et al.* [49]	MOD-CVS
Akiyo	96.01	96.06	98.37	98.49	**99.26**
Mother daughter	89.21	90.38	96.24	97.09	**99.25**
Grandma	83.24	83.69	95.04	95.31	**98.73**
Miss	90.16	90.77	96.94	97.14	**99.11**
Silent	90.83	91.72	95.55	95.77	**97.97**
Suzie	88.20	88.64	97.50	97.84	**98.54**
Salesman	69.56	74.42	92.77	93.79	**97.94**
Claire	97.03	96.65	98.50	98.88	**98.68**
Teleprompter	97.23	97.16	98.63	98.96	**98.82**
Speech	57.63	58.99	94.46	96.34	**98.83**
Average	85.91	86.85	96.40	96.96	**98.71**

TABLE 3. Comparison of average miss classification error (AMCE) in percentage of proposed scheme with different SOTA techniques.

Sample Video	Methods				
	Badri *et al.* [47]	Zhu *et al.* [48]	Sahoo *et al.* [20]	Sahoo *et al.* [49]	MOD-CVS
Akiyo	2.02	2.81	1.40	1.22	**0.54**
Mother daughter	4.55	12.43	3.58	2.94	**0.73**
Grandma	5.74	13.04	4.38	4.26	**1.07**
Miss	3.04	7.96	2.88	1.75	**0.83**
Silent	3.66	4.08	3.19	2.63	**1.09**
Suzie	4.47	5.68	2.92	2.92	**1.78**
Salesman	12.22	12.45	4.15	2.94	**0.90**
Claire	3.23	1.92	0.80	0.84	**0.81**
Teleprompter	3.47	2.56	0.91	0.71	**0.69**
Speech	4.12	2.68	2.24	2.14	**0.67**
Average	4.46	6.56	2.64	2.46	**0.91**

[48], Sahoo et al. [20], and Sahoo et al. [49] contemporary methods.

To further justify the efficiency of this proposed algorithm, the developed model is tested on the *CD-Net 2014* dataset with various challenging sequences, including averages for Precision (AP), Recall (AR), F-measure (AF), and Percentage of Wrong Classification (APWC). The objective is to simultaneously reduce the percentage of wrong classifications (PWC) and increase F-measure, Precision, Recall [59]. We compared the result obtained by *CD-Net 2014* datasets against eighteen existing BGS SOTA methods, including eight deep learning techniques: DeepBS [53], WisenetMD [54], Fast BSUV-Net 2.0 [51], SemanticBGS [60], BSUV-Net [50], BSUV-Net + SemanticBGS [50], IUTIS-5 [61], and BMN-BSN [62]. Table 4 shows that the proposed model achieves superior values for AP, AR, and AF while exhibiting a lower APWC compared to all SOTA deep learning techniques. Also, the MOD-CVS compared with ten non-deep learning existing techniques: SWCD [63], CVABS [64], PAWCS [65], WiSARDrp [66], Multimode Background [67], BMOG [68], WeSAMBE [69], RT-SBS-v1 [70], M4CD Version 2.0 [71], and CL-VID [72]. In Table 4, it is evident that the developed technique shows higher values for AP, AR, and AF, while also presenting a lower APWC compared to SOTA techniques that are not based on deep learning.

Further, to check the efficiency of the MOD-CVS, the experiment has been done on *Star* datasets, which consists of image sequences of challenging video scenes: noise in the video scene, non-static background, changes in lighting

TABLE 4. Quantitative comparison of MOD-CVS on *CD-Net 2014* dataset with different deep learning and non-deep learning based methods.

Approaches		AP	AR	AF	APWC
Deep learning	DeepBS [53]	0.8332	0.7545	0.7458	1.9920
	WisenetMD [54]	0.7668	0.8179	0.7535	1.6136
	Fast BSUV-Net 2.0 [51]	0.8425	0.8181	0.8039	0.9054
	SemanticBGS [60]	0.8305	0.7890	0.7892	1.0722
	BSUV-Net [50]	0.8113	0.8203	0.7868	1.1402
	BSUV-Net + SemanticBGS [50]	0.8319	0.8179	0.7986	1.1326
	IUTIS-5 [61]	0.8087	0.7849	0.7717	1.1986
	BMN-BSN [62]	0.7032	0.8250	0.7188	2.9060
Non-deep learning	SWCD [63]	0.7527	0.7839	0.7583	1.3414
	CVABS [64]	0.7782	0.7818	0.7701	1.3090
	PAWCS [65]	0.7857	0.7718	0.7403	1.1992
	WiSARDrp [66]	0.7880	0.7062	0.7095	1.7197
	Multimode Background [67]	0.7382	0.7389	0.7288	1.2614
	BMOG [68]	0.6981	0.7265	0.6543	2.9757
	WeSAMBE [69]	0.7679	0.7955	0.7446	1.5105
	RT-SBS-v1 [70]	0.7743	0.7406	0.7153	1.5098
	M4CD Version 2.0 [71]	0.7423	0.7885	0.7038	2.3011
	CL-VID [72]	0.5529	0.8316	0.5813	7.4772
	MOD-CVS	**0.8895**	**0.8283**	**0.8269**	**0.5134**

TABLE 5. Average similarity measure for *Star* dataset (In this table AP, BT, CA, CU, ES, FO, LO, and ST indicates the Airport, Bootstrap, Campus, Curtain, Escalator, Fountain, Lobby, and Station respectively).

Approaches	AP	BT	CA	CU	ES	FO	LO	ST
GMM [73]	0.3335	0.3838	0.0757	0.7580	0.1388	0.6854	0.6519	0.5363
DPGMM [74]	0.5676	0.6496	0.7876	0.8411	0.5522	0.7424	0.6665	0.6733
Feature bags [75]	0.6011	0.6238	0.8011	0.8963	0.5610	0.7672	0.8868	0.6674
Video plane [13]	0.1135	0.3079	0.1596	0.1841	0.1294	0.0999	0.1554	0.5209
Self-organizing [76]	0.5943	0.6019	0.6960	0.8178	0.5770	0.6554	0.6489	0.6677
MOD-CVS	**0.7792**	**0.9209**	**0.9858**	**0.9702**	**0.9278**	**0.8795**	**0.9337**	**0.9355**

conditions, and shadow. we have compared with five SOTA techniques: GMM [73], DPGMM [74], Feature bags [75], Video plane [13], and Self-organizing [76]. We employed the average similarity measure [76] to assess the effectiveness of the developed technique. The average similarity measure attained by the proposed approach compared to different SOTA methods is shown in Table 5. The results in Table 5 indicate that the MOD-CVS exhibits higher accuracy in the average similarity measure on the *Star* datasets in comparison to other current SOTA techniques considered.

Eventually, to evaluate the efficacy of the developed model, a well-known *Wallflower* dataset is used for testing that contains indoor and outdoor video scenes captured by a CCD camera on a non-static background, illumination variations, and video noise. The effectiveness of the developed technique is validated through a comparative analysis with nine established SOTA techniques: Fuzzy Mode [77], ViBe [78], BRPCA [79], GMM [73], Codebook [80], DeepBS [53], Triplet CNN [81], MsEDNet [82], and STAM [83]. The evaluation metric employed for this database is AF. Analysis of Table 6 reveals that the proposed algorithm achieves the highest AF values compared to all the considered SOTA techniques.

D. UNSEEN VIDEO SETUP

In an unseen video arrangement, the training, as well as the testing set, contains different videos. The proposed framework is trained with the Claire, Mother daughter,

TABLE 6. Average F measure for *wallflower* dataset (In this table BT, LS, CM, MO, WT, and TD denotes the Bootstrap, Light switch, Camouflage, Moved object, Waving tree, and Time of day).

Approaches	BT	LS	CM	MO	WT	TD
ViBe [78]	0.5433	0.1888	0.9006	0.3967	0.5513	0.7271
Fuzzy Mode [77]	0.7920	0.7842	0.9478	0.8967	0.7806	0.9515
Codebook [80]	0.4727	0.6135	0.9418	0.5132	0.6943	0.9301
GMM [73]	0.6054	0.1189	0.8524	0.4001	0.5626	0.8363
BRPCA [79]	0.8278	0.4489	0.8764	0.8969	0.5867	0.8929
STAM [83]	0.7414	0.9090	0.7369	0.8392	0.5325	0.3429
DeepBS [53]	0.7479	0.6114	0.9857	0.6583	0.9546	0.5494
MsEDNet [82]	0.8754	0.8625	0.9493	0.8778	0.8196	0.5703
Triplet CNN [81]	0.5494	0.6321	0.9688	0.8895	0.9552	0.6048
MOD-CVS	**0.9340**	**0.9668**	**0.9915**	**0.9654**	**0.9839**	**0.9199**

TABLE 7. Average F-measure of the MOD-CVS in unseen video setup on slow moving object, *wallflower*, *Star* dataset.

Dataset	Videos used for training	Video used for testing	MOD-CVS
Slow moving object	Claire, Mother daughter, Grandma	Akiyo	0.86
	Claire, Mother daughter, Grandma	Teleprompter	0.94
	Claire, Mother daughter, Grandma	Speech	0.77
	Salesman, Teleprompter, Speech	Miss	0.79
	Salesman, Teleprompter, Speech	Suzie	0.91
wallflower	BT, CM, LS	MO	0.83
	BT, CM, LS	WT	0.87
	CM, LS, TD	BT	0.90
	CM, LS, TD	WT	0.75
	CM, LS, TD	MO	0.85
Star	AP, BT, CA, CU	ES	0.89
	AP, BT, CA, CU	FO	0.84
	AP, BT, CA, CU	LO	0.97
	AP, BT, CA, CU	ST	0.80
	CA, ES, FO, LO	BT	0.78

TABLE 8. Average F-measure comparison of the MOD-CVS in unseen setup on *CDNet-2014* dataset with different techniques.

Approaches	Bl	Pe	Sw	Bo	Pa	Tp	Ts	Bs	Co	T1
SuBSENSE [84]	**0.85**	0.95	0.81	0.69	0.48	0.85	0.86	0.86	**0.91**	0.79
PAWCS [65]	0.66	0.95	0.74	0.88	0.21	**0.91**	0.86	0.86	0.65	0.68
IUTIS-5 [61]	0.80	0.97	0.81	0.75	0.65	0.89	0.87	0.87	0.90	0.63
BSUV-Net [50]	0.82	0.97	0.69	0.89	0.91	**0.91**	0.80	**0.94**	0.83	0.66
SematicBGS [60]	0.84	**0.98**	0.85	**0.98**	0.88	0.88	0.92	0.92	0.82	0.30
MOD-CVS	0.82	0.90	**0.87**	0.89	**0.94**	0.80	**0.95**	0.87	**0.91**	**0.85**

TABLE 9. Ablation study of MOD-CVS on slowly-moving data sets without and with global average pooling in terms of average F measure comparison.

Name of the Video	MOD-CVS without Global Average Pooling	MOD-CVS with Global Average Pooling
Akiyo	98.85	**99.26**
Mother daughter	99.09	**99.25**
Grandma	98.27	**98.73**
Miss	98.93	**99.11**
Silent	95.44	**97.97**

and Grandma image sequences, and for testing Akiyo, Teleprompter, and Speech image frames are used. Similarly, the model is trained using the Salesman, Teleprompter, and Speech image sequences, and for testing Miss and Suzie image frames are used. From Table 7, it is observed that the designed model attained a better average F-measure value for the unseen setup. Similarly, we have investigated the efficacy of the MOD-CVS in unseen setup for the *wallflower*, and *Star* databases. Table 7 indicate that the developed model exhibited better AF values for the *wallflower* and *Star* databases in an unseen configuration. Additionally, the

TABLE 10. Ablation study of the MOD-CVS on slowly-moving data sets without and with feature pooling framework in terms of average F measure comparison.

Name of the Video	MOD-CVS without Feature Pooling Framework	MOD-CVS with Feature Pooling Framework
Akiyo	99.05	**99.26**
Mother daughter	98.79	**99.25**
Grandma	97.31	**98.73**
Miss	99.03	**99.11**
Silent	96.24	**97.97**

TABLE 11. Run-time of different schemes on *CD-Net 2014* dataset.

Methods	Processing time(frames per second(FPS))
DeepBS [53]	10
BSUV-Net [50]	6
BSPVGAN [85]	5
Semantic BGS [60]	7
WisenetMD [54]	12
MOD-CVS	**21**

TABLE 12. Assessment of the MOD-CVS's average F-measure via an ablation study, utilizing a k-fold cross-validation training approach on the change *CDNet-2014* dataset.

k-fold cross validation	Minimum	Maximum	Average	Standard Deviation
k=5	0.8004	0.8112	0.8105	0.0122
k=10	0.8018	0.8136	0.8157	0.0212

TABLE 13. Comparison of average F-measure of MOD-CVS with different swin transformer based method (- indicates non-availability of the result).

Approaches	NJU2K [57]	STERE [56]	NLPR [58]	DUTS [55]	SIP [86]
DFTR [87]	0.923	0.914	0.934	0.900	0.913
DTMINet [88]	0.933	0.920	0.929	**0.950**	**0.929**
SwinNet [89]	0.908	0.922	0.893	-	0.912
MOD-CVS	**0.936**	**0.928**	**0.937**	0.921	0.927

effectiveness of the developed algorithm is assessed in unseen setups for the *CD-Net 2014* dataset. As shown in Table 8, the proposed model demonstrated satisfactory accuracy compared to established BGS techniques. In this table, Bl, Pe, Sw, Bo, Pa, Tp, Ts, Bs, Co, and T1 depicts the blizzard (from BadWeather), pedestrian (from Baseline), sidewalk (from Camera Jitter), boats (from Dynamic Background), parking (from Intermittent Object Motion), turnpike05fps (from Low Framerate), tramstation (from Night Videos), busstation (from Shadow), corridor (from Thermal), and turbulence1(from Turbulence), respectively.

E. ABLATION STUDY

To analyse the importance of each element in the developed BGS deep-learning based framework, an ablation study is performed. Table 9 demonstrates the efficacy the developed algorithm with and without the GAP layer. It is found that, the inclusion of the GAP layer in the proposed model consistently yields a higher AF value when compared to the version without the GAP layer across all challenging videos. Likewise, an ablation study of the proposed architecture is conducted, exploring its performance both without and with the integration of a feature pooling framework (FPF). From

TABLE 14. Comparison of average F-measure of auto encoder based method with MOD-CVS method on *CDNet-2014*dataset.

Approaches	Base-line	Bad weather	Dynamic backgr.	Int. objct. motio.	Night	Low framerate	Shadow	Camera jitter	Turbu-lence	Thermal	PTZ	Overall
AE-NE [38]	0.8959	**0.8337**	0.6225	0.8231	0.5172	**0.6771**	0.8947	0.9230	**0.8382**	0.7999	0.8000	0.7841
MOD-CVS	**0.9838**	0.5607	**0.9340**	**0.9445**	**0.8343**	0.5823	**0.9750**	**0.9307**	0.4733	**0.9455**	**0.9314**	**0.8269**

Table 10, it is observed that the proposed algorithm with the FPF module is capable of attaining higher accuracy as compared to without the FPF module. The FPF module between the encoder and decoder effectively learns a mapping from high-dimensional feature to a multi-dimensional feature.

Additionally, the ablation study culminated in a run-time comparison of the proposed approach compared to various SOTA techniques using the *CDNet-2014* dataset. Table 11 reveals that the processing time of the developed architecture is 21 frames per second, underscoring the comparatively lower computational complexity of the MOD-CVS compared to many existing SOTA methods.

The proposed method is tested for effectiveness with k-fold cross-validation. The results of the MOD-CVS's performance with k-fold cross-validation (k = 5 and k = 10) are outlined in Table 12. In this study, when k equals 5, we partitioned the entire set of 159,278 frames from *CDNet-2014* dataset into 5 folds. Initially, we utilised the first fold for testing and the remaining folds for training during the training of the proposed model. Subsequently, the second fold served as the testing set, with the remaining folds employed for training, and this process continued iteratively. Similarly, k = 10 was used to train the suggested model, and 159,278 frames on *CDNet-2014* are divided into 10 folds. Subsequently, testing is carried out with one fold while training is conducted with the remaining folds in a sequential manner. According to the data in Table 12, the suggested approach utilizing a k-fold cross-validation with values of k equal to 5 and 10 demonstrates outcomes with an average F-Measure of 0.8105 and 0.8157 for k = 5 and 10, respectively. However, the developed MOD-CVS technique without cross-fold validation training mechanism attains a higher value of average F-Measure equals to 0.8269.

Further, the efficacy of the proposed MOD-CVS model is verified in Table 13which illustrates a comparison of the average F-Measure between the proposed MOD-CVS and various Swin Transformer-based methods. It is found that the proposed MOD-CVS method got comparatively higher value than the existing methods. Furthermore, the consistency can be demonstrated for NJU2K [57], STERE [56], NLPR [58], and DUTS [55] datasets. Fig. 9 (a) and (b) depict the input frame and its corresponding ground-truth image, respectively. In Fig. 9 (c), the outcomes of the proposed method are portrayed, emphasizing the MOD-CVS's proficiency in precisely categorizing both foreground and background pixels.

In Table 14 evaluation of the average F-Measure between the autoencoder-based AE-NE [38] approach and the proposed MOD-CVS method on *CDNet-2014* dataset is verified. It is clearly demonstrated that the MOD-CVS exhibits a higher F-measure value in comparison to alternative methods.

V. CONCLUSION

This research work tackles the task of detecting moving objects in challenging video scenes by employing a deep-learning architecture with an encoder-decoder design. The proposed model detects moving objects in complex video scenes including objects moving at different speeds, low contrast, non-static background, low frame rate, noise, image capture at night time, shadow, poor resolution, low signal-to-noise ratio, lack of object shape and textural details in the images, low contrast, etc. To extract diverse features accurately at multiple levels, we have used an improved version of pre-trained VGG-19 deep learning network as an encoder. Also, the transfer learning mechanism in the encoder network enhances the efficacy of the MOD-CVS model. Further, various layers in the proposed VGG-19 deep neural network are capable of preserving the low, mid, and high-level features that are essential for local change detection. The feature pooling framework (FPF) between the encoder and decoder networks efficiently preserves objects of various scales from challenging video frames. In the proposed algorithm, the FPF model effectively learns a mapping from higher-dimensional feature space to a multi-scale as well as multi-dimensional feature space that can classify the foreground and background pixels with simple decision boundaries. The decoder network in the MOD-CVS model contains a stack of convolutional layers that effectively project feature space to image space. The effectiveness of the MOD-CVS algorithm is corroborated using subjective and objective analysis against thirty-six SOTA techniques. It is observed that the MOD-CVS model retains the shape of the moving object accurately with a reduced amount of pores and holes as compared to the SOTA techniques. Also, the MOD-CVS provides adequate accuracy for unseen video setups. However, the performance of the MOD-CVS work is reduced when the moving object size is small. Also, the proposed work provides frontier outcomes when there is a higher variation in the scene. In the future, we aim to improve the accuracy of the MOD-CVS by investigating a robust hybridized deep neural architecture.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.

[2] J.-W. Hsieh, S.-H. Yu, Y.-S. Chen, and W.-F. Hu, "Automatic traffic surveillance system for vehicle tracking and classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 175–187, Jun. 2006.

[3] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. Syst. Man, Cybern., Part C (Appl. Reviews)*, vol. 34, no. 3, pp. 334–352, Aug. 2004.
[4] D. K. Rout, B. N. Subudhi, T. Veerakumar, and S. Chaudhury, "Spatio-contextual Gaussian mixture model for local change detection in underwater video," *Expert Syst. Appl.*, vol. 97, pp. 117–136, May 2018.
[5] B. N. Subudhi, M. K. Panda, T. Veerakumar, V. Jakhetiya, and S. Esakkirajan, "Kernel-induced possibilistic fuzzy associate background subtraction for video scene," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 3, pp. 1–12, Jan. 2022.
[6] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Netw.*, vol. 117, pp. 8–66, Sep. 2019.
[7] M. K. Panda, A. Sharma, V. Bajpai, B. N. Subudhi, V. Thangaraj, and V. Jakhetiya, "Encoder and decoder network with ResNet-50 and global average feature pooling for local change detection," *Comput. Vis. Image Understand.*, vol. 222, Sep. 2022, Art. no. 103501.
[8] M. K. Panda, B. N. Subudhi, T. Bouwmans, V. Jakheytiya, and T. Veerakumar, "An end to end encoder–decoder network with multi-scale feature pulling for detecting local changes from video scene," in *Proc. 18th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2022, pp. 1–8.
[9] M. Reisslein, L. Karam, P. Seeling, and F. Fitzek. (2000). *Yuv Video Sequences*. [Online]. Available: http://trace.eas.asu.edu/yuv/
[10] C. Montgomery. (2004). *Xiph.org Video Test Media [Derf's Collection]*. [Online]. Available: https://media.xiph.org/video/derf/
[11] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 393–400.
[12] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 255–261.
[13] J. Li and X.-M. Zhang, "Video object plane extraction for surveillance applications," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2004, pp. 3928–3931.
[14] P. K. Sahoo, P. Kanungo, and K. Parvathi, "Three frame based adaptive background subtraction," in *Proc. Int. Conf. High Perform. Comput. Appl. (ICHPCA)*, Dec. 2014, pp. 1–5.
[15] J. H. Duncan and T.-C. Chou, "On the detection of motion and the computation of optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 3, pp. 346–352, Mar. 1992.
[16] S. K. Choudhury, P. K. Sa, S. Bakshi, and B. Majhi, "An evaluation of background subtraction for object detection vis-a-vis mitigating challenging scenarios," *IEEE Access*, vol. 4, pp. 6133–6150, 2016.
[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, vol. 1, Dec. 2001, pp. 1511–1518.
[18] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[20] P. K. Sahoo, P. Kanungo, and S. Mishra, "A fast valley-based segmentation for detection of slowly moving objects," *Signal, Image Video Process.*, vol. 12, no. 7, pp. 1265–1272, Oct. 2018.
[21] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.
[22] P. Kanungo, A. Narayan, P. K. Sahoo, and S. Mishra, "Neighborhood based codebook model for moving object segmentation," in *Proc. 2nd Int. Conf. Man Mach. Interfacing (MAMI)*, Dec. 2017, pp. 1–6.
[23] P. KaewTraKulPong and R. Bowden, *An Improved Adaptive Background Mixture Model for Real-time Tracking With Shadow Detection*. Cham, Switzerland: Springer, 2002, pp. 135–144.
[24] J. Huang, W. Zou, Z. Zhu, and J. Zhu, "An efficient optical flow based motion detection method for non-stationary scenes," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2019, pp. 5272–5277.
[25] L. Fan, T. Zhang, and W. Du, "Optical-flow-based framework to boost video object detection performance with object enhancement," *Expert Syst. Appl.*, vol. 170, May 2021, Art. no. 114544.
[26] J. Guo, J. Wang, R. Bai, Y. Zhang, and Y. Li, "A new moving object detection method based on frame-difference and background subtraction," in *Proc. IOP Conf. Series, Mater. Sci. Eng.*, Sep. 2017, vol. 242, no. 1, p. 012115.
[27] S. S. Sengar and S. Mukhopadhyay, "Moving object detection based on frame difference and w4," *Signal, Image Video Process.*, vol. 11, no. 7, pp. 1357–1364, Oct. 2017.
[28] X. Huang, F. Wu, and P. Huang, "Moving-object detection based on sparse representation and dictionary learning," *AASRI Proc.*, vol. 1, pp. 492–497, Jan. 2012.
[29] M. F. Savaş, H. Demirel, and B. Erkal, "Moving object detection using an adaptive background subtraction method based on block-based structure in dynamic scene," *Optik*, vol. 168, pp. 605–618, Sep. 2018.
[30] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, May 2021.
[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision—ECCV 2016*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.
[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1–9.
[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[34] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
[35] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[36] B. Zhu, J. Jiao, and D. Tse, "Deconstructing generative adversarial networks," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 7155–7179, Nov. 2020.
[37] M. Mandal, L. K. Kumar, M. Singh Saran, and S. K. Vipparthi, "MotionRec: A unified deep framework for moving object recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2723–2732.
[38] B. Sauvalle and A. de La Fortelle, "Autoencoder-based background reconstruction and foreground segmentation with background noise estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3243–3254.
[39] T. Sahoo, B. Mohanty, and B. K. Pattanayak, "Moving object detection using deep learning method," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 9s, pp. 282–290, 2024.
[40] D. Meimetis, I. Daramouskas, I. Perikos, and I. Hatzilygeroudis, "Real-time multiple object tracking using deep learning methods," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 89–118, Jan. 2023.
[41] J. He, Y. Chen, N. Wang, and Z. Zhang, "3D video object detection with learnable object-centric global optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5106–5115.
[42] P. Angelov and E. Soares, "Towards explainable deep neural networks (XDNN)," *Neural Netw.*, vol. 130, pp. 185–194, Oct. 2020.
[43] L. Hiley, A. Preece, and Y. Hicks, "Explainable deep learning for video recognition tasks: A framework recommendations," 2019, *arXiv:1909.05667*.
[44] S. Naddaf-Sh, M.-M. Naddaf-Sh, H. Zargarzadeh, M. Dalton, S. Ramezani, G. Elpers, V. S. Baburao, and A. R. Kashani, "Real-time explainable multiclass object detection for quality assessment in 2-Dimensional radiography images," *Complexity*, vol. 2022, pp. 1–17, Aug. 2022.
[45] S. A. Mahmoudi, O. Amel, S. Stassin, M. Liagre, M. Benkedadra, and M. Mancas, "A review and comparative study of explainable deep learning models applied on action recognition in real time," *Electronics*, vol. 12, no. 9, p. 2027, Apr. 2023.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[47] B. N. Subudhi and P. K. Nanda, "Detection of slow moving video objects using compound Markov random field model," in *Proc. TENCON—IEEE Region 10 Conf.*, Nov. 2008, pp. 1–6.
[48] Z. Zhu and Y. Wang, "A hybrid algorithm for automatic segmentation of slowly moving objects," *AEU—Int. J. Electron. Commun.*, vol. 66, no. 3, pp. 249–254, Mar. 2012.

[49] P. K. Sahoo, P. Kanungo, S. Mishra, and B. P. Mohanty, "Entropy feature and peak-means clustering based slowly moving object detection in head and shoulder video sequences," *J. King Saud Univ.—Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5296–5304, Sep. 2022.

[50] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net: A fully-convolutional neural network for background subtraction of unseen videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2774–2783.

[51] M. O. Tezcan, P. Ishwar, and J. Konrad, "BSUV-Net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction," *IEEE Access*, vol. 9, pp. 53849–53860, 2021.

[52] Y. Wang, Z. Luo, and P. M. Jodoin, "Interactive deep learning method for segmenting moving objects," *Pattern Recognit. Lett.*, vol. 96, pp. 66–75, Sep. 2017.

[53] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.

[54] S. Lee, G. Lee, J. Yoo, and S. Kwon, "WisenetMD: Motion detection using dynamic background region analysis," *Symmetry*, vol. 11, no. 5, pp. 1–15, 2019.

[55] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7253–7262.

[56] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[57] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.

[58] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Computer Vision—ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 92–109.

[59] M. K. Panda, B. N. Subudhi, T. Veerakumar, and V. Jakhetiya, "Modified ResNet-152 network with hybrid pyramidal pooling for local change detection," *IEEE Trans. Artif. Intell.*, pp. 1–14, Jul. 2023, doi: 10.1109/TAI.2023.3299903.

[60] M. Braham, S. Piérard, and M. Van Droogenbroeck, "Semantic background subtraction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4552–4556.

[61] S. Bianco, G. Ciocca, and R. Schettini, "Combination of video change detection algorithms by genetic programming," *IEEE Trans. Evol. Comput.*, vol. 21, no. 6, pp. 914–928, Dec. 2017.

[62] V. M. Mondéjar-Guerra, J. Rouco, J. Novo, and M. Ortega, "An end-to-end deep learning approach for simultaneous background modeling and subtraction," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 266.

[63] Ş. Işık, K. Özkan, S. Günal, and Ö. N. Gerek, "SWCD: A sliding window and self-regulated learning-based background updating method for change detection in videos," *J. Electron. Imag.*, vol. 27, no. 2, p. 1, Mar. 2018.

[64] Ş. Işık, K. Özkan, and Ö. N. Gerek, "CVABS: Moving object segmentation with common vector approach for videos," *IET Comput. Vis.*, vol. 13, no. 8, pp. 719–729, Dec. 2019.

[65] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 990–997.

[66] M. De Gregorio and M. Giordano, "WiSARDrp for change detection in video sequences," in *Proc. ESANN*, 2017, pp. 453–458.

[67] H. Sajid and S. S. Cheung, "Universal multimode background subtraction," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3249–3260, Jul. 2017.

[68] I. Martins, P. Carvalho, L. Corte-Real, and J. L. Alba-Castro, "BMOG: Boosted Gaussian mixture model with controlled complexity," in *Proc. Pattern Recognit. Image Anal., 8th Iberian Conf.*, 2017, pp. 50–57.

[69] S. Jiang and X. Lu, "WeSamBE: A weight-sample-based method for background subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2105–2115, Sep. 2018.

[70] A. Cioppa, M. V. Droogenbroeck, and M. Braham, "Real-time semantic background subtraction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3214–3218.

[71] K. Wang, C. Gou, and F.-Y. Wang, "M^{4}CD: A robust change detection method for intelligent visual surveillance," *IEEE Access*, vol. 6, pp. 15505–15520, 2018.

[72] E. López-Rubio, M. A. Molina-Cabello, R. M. Luque-Baena, and E. Domínguez, "Foreground detection by competitive learning for varying input distributions," *Int. J. Neural Syst.*, vol. 28, no. 5, Jun. 2018, Art. no. 1750056.

[73] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, May 2000.

[74] T. S. Haines and T. Xiang, "Background subtraction with Dirichlet process mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 670–683, Dec. 2014.

[75] B. N. Subudhi, S. Ghosh, S. C. K. Shiu, and A. Ghosh, "Statistical feature bag based background subtraction for local change detection," *Inf. Sci.*, vol. 366, pp. 31–47, Oct. 2016.

[76] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.

[77] B. N. Subudhi, T. Veerakumar, S. Esakkirajan, and A. Ghosh, "Kernelized fuzzy modal variation for local change detection from video scenes," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 912–920, Apr. 2020.

[78] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Dec. 2011.

[79] C. Guyon, T. Bouwmans, and E. Zahzah, "Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis," *Principal Compon. Anal.*, vol. 10, pp. 223–238, Mar. 2012.

[80] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground–background segmentation using codebook model," *Real-Time Imag.*, vol. 11, no. 3, pp. 172–185, Jun. 2005.

[81] T. P. Nguyen, C. C. Pham, S. V. Ha, and J. W. Jeon, "Change detection by training a triplet network for motion feature extraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 433–446, Feb. 2019.

[82] P. W. Patil, S. Murala, A. Dhall, and S. Chaudhary, "MsEDNet: Multi-scale deep saliency learning for moving object detection," in *Proc. IEEE Int. Conf. Syst. Man, Cybern. (SMC)*, Oct. 2018, pp. 1670–1675.

[83] D. Liang, J. Pan, H. Sun, and H. Zhou, "Spatio-temporal attention model for foreground detection in cross-scene surveillance videos," *Sensors*, vol. 19, no. 23, p. 5142, Nov. 2019.

[84] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.

[85] W. Zheng, K. Wang, and F.-Y. Wang, "A novel background subtraction algorithm based on parallel vision and Bayesian GANs," *Neurocomputing*, vol. 394, pp. 178–200, Jun. 2020.

[86] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[87] H. Zhu, X. Sun, Y. Li, K. Ma, S. Kevin Zhou, and Y. Zheng, "DFTR: Depth-supervised fusion transformer for salient object detection," 2022, *arXiv:2203.06429*.

[88] C. Zeng, S. Kwong, and H. Ip, "Dual Swin-transformer based mutual interactive network for RGB-D salient object detection," *Neurocomputing*, vol. 559, Nov. 2023, Art. no. 126779.

[89] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, Jul. 2022.

PRABODH KUMAR SAHOO (Member, IEEE) received the M.E. degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India, in 2005, and the Ph.D. degree from the Centurion University of Technology and Management, Odisha, India, in 2019. Currently, he is an Associate Professor with the Mechatronics Department, Parul University, Vadodara, Gujarat, India. He has contributed significantly to his field, having published six peer-reviewed journal articles, presented five international conference papers, and secured two international patents. His primary research interests include image processing, computer vision, and cyber-physical systems.

MANOJ KUMAR PANDA received the B.Tech. degree in electronics and telecommunication from the Sanjay Memorial Institute of Technology, Odisha, India, in 2007, the M.Tech. degree in electronics and communication from the National Institute of Science and Technology, Odisha, in 2011, and the Ph.D. degree from the Indian Institute of Technology Jammu, India, in 2022. He is an Assistant Professor with the Department of Electronics and Communication Engineering, GIET University, Gunupur, Rayagada, Odisha. His current research interest includes image and video processing.

UPASANA PANIGRAHI received the M.Tech. degree from the Gandhi Institute of Engineering and Technology, Gunupur, Odisha, and BPUT, Bhubaneswar, India, in 2011. She is currently pursuing the Ph.D. degree with C. V. Raman Global University, Bhubaneswar. She has six years of experience as an Assistant Professor with the Gandhi Engineering College, BPUT. Her research areas include digital image processing and digital signal processing.

GANAPATI PANDA (Life Senior Member, IEEE) received the Ph.D. degree in electronics and communication engineering from IIT Kharagpur, in 1981. He did a postdoctoral research work with The University of Edinburgh, U.K., from 1984 to 1986. Currently, he is a Professor and a Research Advisor with C. V. Raman Global University, Bhubaneswar, India. He is also a Professorial Fellow with the Indian Institute of Technology, Bhubaneswar. Prior to this, he was a Professor, the Dean, and the Deputy Director with the School of Electrical Sciences, IIT Bhubaneswar. He has guided 46 Ph.D. students in the fields of signal processing, communication, and machine learning.

PRINCE JAIN is currently an Assistant Professor with the Mechatronics Engineering Department, Parul Institute of Technology, Parul University, Vadodara, India. He received the Visvesvaraya Ph.D. Scheme Fellowship to complete the Ph.D. dissertation with Punjab Engineering College (Deemed to be University), Chandigarh, India. He is the author or coauthor of about 16 research journal articles, 20 conference papers, and a few book chapters on various topics related to antennas, machine learning, and metamaterials. His research interests include machine learning, artificial intelligence, optimization techniques, metamaterial absorbers/antennas at RF, THz and visible frequencies, material science, nanotechnology, and biomedical signal processing. He is serving as a Topical Advisory Panel Member for Micromachines and Materials (MDPI). He has contributed as a peer reviewer of prestigious publishers, including IEEE, Elsevier, IOPscience, Wiley, PIER, Emerald, and PLOS. He currently serving as an Academic Editor for *Journal of Electrical and Computer Engineering*, Hindawi, and *PLOS One*.

MD. SHABIUL ISLAM (Senior Member, IEEE) received the B.Sc. and first M.Sc. degrees from the Department of Applied Physics and Electronics, Rajshahi University, Bangladesh, in 1985 and 1986, respectively, the second M.Sc. degree in micro controller based system design from the Department of Electrical, Electronics and System Engineering (EESE), Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia, in 1997, and the Ph.D. degree in VLSI design from the Faculty of Engineering (FOE), Multimedia University (MMU), Cyberjaya, Selangor, Malaysia, in 2008. He has been a Professor with FOE, since 2016. From 2009 to 2016, he was an Associate Professor with The Institute of Microengineering and Nanoelectronics (IMEN), UKM, where he led the Micro and Nano System Laboratory. From 1999 to 2009, he was a Lecturer with FOE, where he was a Research Assistant with EESE, from 1993 to 1999. From 1991 to 1993, he was a Scientific Officer with the Institute of Electronics and Material Science (IEMS), Bangladesh Atomic Energy Commission (BAEC), Saver, Dhaka, Bangladesh. His research expertise covers a wide range of engineering disciplines. They include micro/nano system design, VLSI design, micro-powering harvesting, embedded system design and interfacing using microcontroller, FPGA realization based on fuzzy logic (FL) algorithm, and 5G communication. He has authored more than 224 peer-reviewed publications, five research books, one book chapter, and three patents. He has received a few awards, the gold medal, the silver medal, and the best paper award.

MOHAMMAD TARIQUL ISLAM (Senior Member, IEEE) is currently a Professor with the Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia (UKM); and a Visiting Professor with Kyushu Institute of Technology, Japan. He is the author or coauthor of about 600 research journal articles, nearly 250 conference papers, and a few book chapters on various topics related to antennas, metamaterials, and microwave imaging with 25 inventory patents filed. His publications have been cited 14,000 times and an H-index is 53 (Source: Scopus). He has supervised about 50 Ph.D. theses, 30 M.Sc. theses, and has mentored more than ten postdoctoral researchers and a visiting scholars. He has developed the Antenna Measurement Laboratory, which includes antenna design and measurement facility till 40 GHz. His Google scholar citation is 22500 and H-index is 62. He was a recipient of more than 40 research grants from the Malaysian Ministry of Science, Technology and Innovation; Ministry of Education; UKM Research Grant; and International Research Grants from Japan, Saudi Arabia, and Kuwait. His research interests include communication antenna design, metamaterial, satellite antennas, and microwave imaging. He has been serving as an Executive Committee Member for IEEE AP/MTT/EMC Malaysia Chapter, from 2019 to 2020; the Chartered Professional Engineer (C.Eng.); a fellow of IET, U.K.; and a Senior Member of IEICE, Japan. He received several International Gold Medal Awards; the Best Invention in Telecommunication Award for his Research and Innovation; the Best Researcher Awards at UKM; the 2018, 2019, and 2020 IEEE AP/MTT/EMC Malaysia Chapter Excellent Award; the Best Innovation Award; the Best Researcher Award by UKM; and the Publication Award from Malaysian Space Agency. He was an Associate Editor of *IET Electronics Letter*. He also serves as the Guest Editor for *Sensors* and *Nanomaterials* and an Associate Editor for IEEE Access.