IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 September 2023, accepted 18 December 2023, date of publication 22 December 2023,
date of current version 28 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3345842

RESEARCH ARTICLE

MSF-NET: Foreground Objects Detection With Fusion of Motion and Semantic Features

JAE-YEUL KIM[1] **AND JONG-EUN HA**[2]

[1]Graduate School of Information and Communication Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, South Korea
[2]Department of Mechanical and Automotive Engineering, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Jong-Eun Ha (jeha@seoultech.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (2020R1A2C1013335).

ABSTRACT Visual surveillance requires robust detection of foreground objects under challenging environments of abrupt lighting variation, stationary foreground objects, dynamic background objects, and severe weather conditions. Most classical algorithms leverage background model images produced by statistical modeling of the change of brightness values over time. Since they have difficulties using global features, many false detections occur at the stationary foreground regions and dynamic background objects. Recent deep learning-based methods can easily reflect global characteristics compared to classical methods. However, deep learning-based methods still need to be improved in utilizing spatiotemporal information. We propose an algorithm for efficiently using spatiotemporal information by adopting a split and merge framework. First, we split spatiotemporal information on successive multiple images into spatial and temporal parts using two sub-networks of semantic and motion networks. Finally, separated information is fused in a spatiotemporal fusion network. The proposed network consists of three sub-networks, which we note as MSF-NET (Motion and Semantic features Fusion NETwork). Also, we propose a method to train the proposed MSF-NET stably. Compared to the latest deep learning algorithms, the proposed MSF-NET gives 9% and 13% higher FM in the LASIESTA and SBI datasets. Also, we designed the proposed MSF-NET to be lightweight to run in real-time on a desktop GPU.

INDEX TERMS Deep learning, foreground object detection, spatiotemporal information, visual surveillance.

I. INTRODUCTION

Before the advent of deep learning, most visual surveillance algorithms perform foreground object detection by processing the change in brightness values. They usually use background model images and update them periodically, denoted as background subtraction (BGS) algorithms. However, it has a limitation of a significant detection error in a challenging environment where stationary foreground and dynamic background objects exist. Information on the type of object in the spatial domain and whether things move in the temporal domain is required to classify a background and a foreground object in visual surveillance. Since these methods have difficulties extracting semantic information in the spatial domain, performance loss may occur.

Recent deep learning-based methods show superior foreground object detection performance. They can reflect global features on images better. However, they have difficulties reflecting temporal information due to memory size, while traditional approaches can reflect temporal information by updating the background model image. For this reason, deep learning-based methods show poor detection performance for a foreground object that has been stationary for a long time. FgSegNet-v2 [2], which records the best performance in the CDnet2014 [1] dataset, uses only the current image as input. Such a spatial network can consider semantic information well but is limited in not using temporal information. The spatial network generally shows a good detection performance in the same environment used for training. Still, the

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

detection performance in an environment not used for training is limited.

Deep learning-based methods treat temporal information using multiple images as input or background model images. They use various types of background model images by median [3], [4], [5], BGS algorithm [6], [7], manually acquired reference image [8], [9], separate background modeling process [10], and background image generation module [11], [12], [13]. However, these methods have a limitation in that a lot of performance degradation occurs when the background image contains many errors. This paper proposes a new integrated model that efficiently utilizes spatial and temporal domains to solve the limitations.

The contribution of this paper is as follows.

1) Motion and Semantic Fusion Network (MSF-NET) is proposed to effectively extract spatiotemporal information in multiple input images for visual surveillance. MSF-NET consists of three networks, separating spatial and temporal information through a semantic network (SN) and motion network (MN). Finally, a spatiotemporal fusion network (STFN) detects a foreground object by integrating information from both networks.

2) The proposed MSF-NET is mainly composed of three networks, which might cause difficulties in training them. A method for practical training of the proposed model is presented. A semantic network can be trained using ground-truth labels through the compact fusion module (CFM) without additional processing.

3) The proposed MSF-NET has an excellent performance in environments not used for training, and we show it using various datasets. In addition, the proposed structure is designed to be lightweight to enable real-time computation on desktop GPUs.

II. RELATED WORKS

We categorize visual surveillance algorithms into traditional approaches and deep learning approaches. Recent survey papers [60], [61] provide good reviews.

A. TRADITIONAL APPROACHES

Stauffer and Grimson [14] proposed Gaussian Mixture Models (GMM), a method based on multiple Gaussian distributions. In GMM [14], each pixel value is expressed as multiple Gaussian distributions. Elgammal et al. [15] proposed a probabilistic non-parametric algorithm using kernel density estimation. Barnich and Droogenbroeck [16] proposed the BGS algorithm ViBE. It detects foreground objects by calculating the Euclidean distance between the pixel value stored in the background model and the current pixel value. ViBE+ [17] uses the adapted distance measure and thresholding from the existing ViBE [16] and increases the detection performance by adding a process to detect blinking pixels.

St-Charles and Bilodeau [18] proposed LOBSTER using an LBSP descriptor [19]. LOBSTER [18] performs foreground object classification by calculating LBSP within a 5 × 5 mask, unlike ViBE [16], which compares pixel values. Haines and Xiang [20] proposed a foreground object detection algorithm based on Dirichlet process Gaussian mixture models. SuBSENSE [21] and PAWCS [22] have actively adjusted parameters, unlike ViBE [16] and LOBSTER [18], which have fixed hyperparameters. This has a more robust detection performance for dynamic and static foreground objects. Laugraud et al. [23] proposed a method of robustly generating a background image even when a foreground object exists in more than half of the observation time range. Panda and Meher [24] proposed a BGS algorithm using Color Difference Histogram (CDH) and Fuzzy Color Difference Histogram (FCDH) in a small local neighborhood.

Sajid and Cheung [25] proposed MBS, a BGS algorithm that can respond to light changes, dynamic backgrounds, and camera movements. Bianco et al. [26] proposed IUTIS, which improves detection performance by combining the results of other foreground object algorithms. Berjón et al. [27] proposed an algorithm that performs background and foreground modeling for robust foreground object detection. This method uses a particle filter in the tracking process and automatically selects an ROI to minimize the computational load. Ortego et al. [28] proposed a hierarchical post-processing framework that can improve the performance of BGS algorithms. With this method, a classical algorithm such as GMM showed a performance improvement of 10%. Still, there is a limit to showing a modest performance improvement of about 2% in a relatively newer algorithm such as PAWCS [22]. Garg et al. [29] proposed a background modeling technique that can be used in the traffic surveillance system. Compared to the existing BGS algorithm, the operation speed is several tens of times faster. Hossain et al. [30] proposed FAST-D, a BGS algorithm consisting of a segmentation strategy, dynamic threshold, and adaptive post-processing process. FAST-D [30] has the highest processing speed and detection performance among classical algorithms but is limited in that it performs poorly compared to deep learning-based methods.

Most traditional algorithms use background model images that are updated statistically. They can reflect temporal information well but need help considering the spatial information covering the whole image. In this paper, we reflect the spatial information on an entire image using the spatial network.

B. DEEP LEARNING APPROACHES

Braham and Droogenbroeck [3] proposed ConvNets based on LeNET [31]. It uses median computation to detect information in the temporal domain. Therefore, there is a limit that false detection may occur if a foreground object is in a stationary state for more than half of the observation range. Zhao et al. [11] proposed a two-stage network consisting of a background reconstruction and foreground segmentation network. Wang et al. [32] presented a new cascade architecture CNN with a more robust detection performance than the single CNN method.

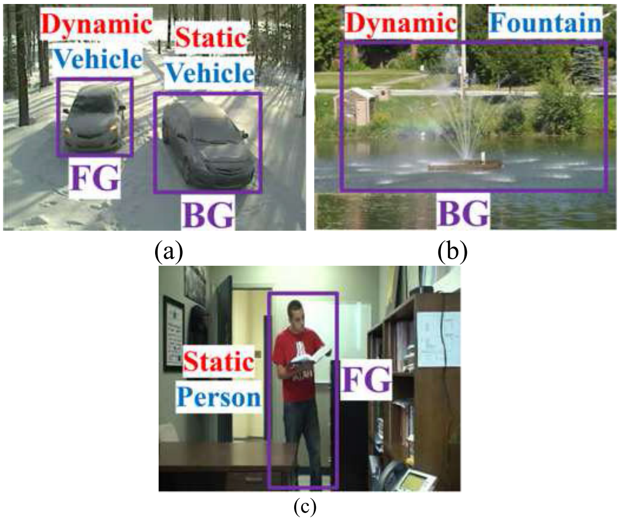
Zeng and Zhu [33] proposed a Multiscale Fully Convolutional Network (MFCN) using VGG-16 [34] as a backbone. MFCN is a structure that generates contrast features from each branch of VGG-16 and concatenates them in the decoder part to generate the final foreground map. Tezcan et al. [8] proposed DeepBS, which concatenates the background and current images to create a foreground probability map through CNN. A more sophisticated background image was created using the background pixel library proposed by SuBSENSE [21] and used for foreground object detection.

Lim and Keles [35] proposed FgSegNet, a spatial network using the current image as input data. Lin et al. [6] proposed an FCN that combines the current and SuBSENSE [21] background images as input data. Zeng et al. [36] suggested RTSS using the BGS algorithm and deep learning segmenter together. The deep learning segmenter used ICNET [37] and PSPNET [38]. Qui and Li [39] proposed a Fully Convolutional Encoder-decoder Spatial-temporal Network (FCESNet) composed of a feature encoder, a spatial-temporal information transmission module, and a feature decoder. FCESNet has a multi-input-multi-output structure and transmits information in the spatial and temporal domains using the ConvLSTM layer.

Lim and Keles [2] proposed FgSegNet-v2, a spatial network using the current image as input data. It has an encoder-decoder structure, and VGG-16 [33] is used as an encoder. FgSegNet-v2 currently records the best performance in the CDnet2014 dataset [1]. Patil and Murala [40] proposed a Motion Saliency Foreground Network (MSFgNet) composed of Submodules, Motion Saliency Network (MSNet), and Foreground Network (FgNet). It is challenging to generate a correct background model image because it is very lightly composed of two temporal pooling layers and one convolution layer in the background image estimation process.

Tezcan et al. [8] proposed a Background Subtraction algorithm for Unseen Videos (BSUV-Net). Although BSUV-NET shows excellent generalization performance even in a new environment that is not used for training, there is a limitation in that a certain level of user intervention is required to obtain an empty reference frame. Akilan et al. [41] proposed a framework to detect and track a vehicle in motion. Detection is performed using [3], and feature values are extracted in the relevant area.

Patil et al. [42] proposed a network composed of an Edge Extraction Mechanism (EEM) and Dense Residual Block (DRB) for the moving object segmentation task. Kim and Ha [7] proposed a framework for inputting the current image, several past images, and the background image of SuBSENSE [21] into a U-NET type network. Mandal and Vipparthi [4] proposed ChangeDet, which uses two types of background model images. Mandal et al. [13] proposed 3DCD that uses a background image generated by a Gradual Reduction Background Estimation (GRBE). GRBE consists of 3D convolution and 3D average pooling layers. 3DCD shows superior performance to the latest classical and deep learning methods in an environment not used for training.



(a) (b) (c)

FIGURE 1. Foreground and background object classification results according to spatial and temporal information (a) dynamic vehicle (b) dynamic fountain (c) stationary person.

Effective extraction of spatiotemporal information is important for visual surveillance. Chen et al. [64] proposed a spatiotemporal network that has a full spatial-temporal multi-scale interactions on the vanilla UNet [46] encoder-decoder structure. Perreault et al. [65] proposed FFAVOD to do feature fusion for video object detection. They use shared feature maps between nearby frames and feature fusion module for video object detection. Dong et al. [62] proposed an algorithm for universal moving object segmentation. First, they learn the distribution from temporal pixels with a defect iterative distribution learning network. Then, the stochastic Bayesian refinement network, which learns the spatial correlation, is applied to improve the binary mask. The proposed method consists of three networks to deal with spatiotemporal information. The spatial network uses a current image, the temporal network uses multiple difference images, and the last network integrates the outputs of two networks.

Deep learning-based approaches show improved results than traditional BGS algorithms due to their ability to consider spatial information. However, they have difficulties reflecting temporal information and show performance degradation in unseen environments. More details can be found in recent survey papers [60], [61]. In this paper, we propose MSF-NET, which effectively combines spatiotemporal information and performs better in unseen environments.

III. PROPOSED METHOD

Reflecting both the object's motion in the temporal domain and the semantic information of the object in the spatial domain is essential for robust foreground object detection in visual surveillance. As shown in Figure 1(a), we should detect a vehicle under motion as a foreground object while detecting the same vehicle as a background object when it is stationary for a long time. From an object detection viewpoint, a vehicle should be detected regardless of motion.

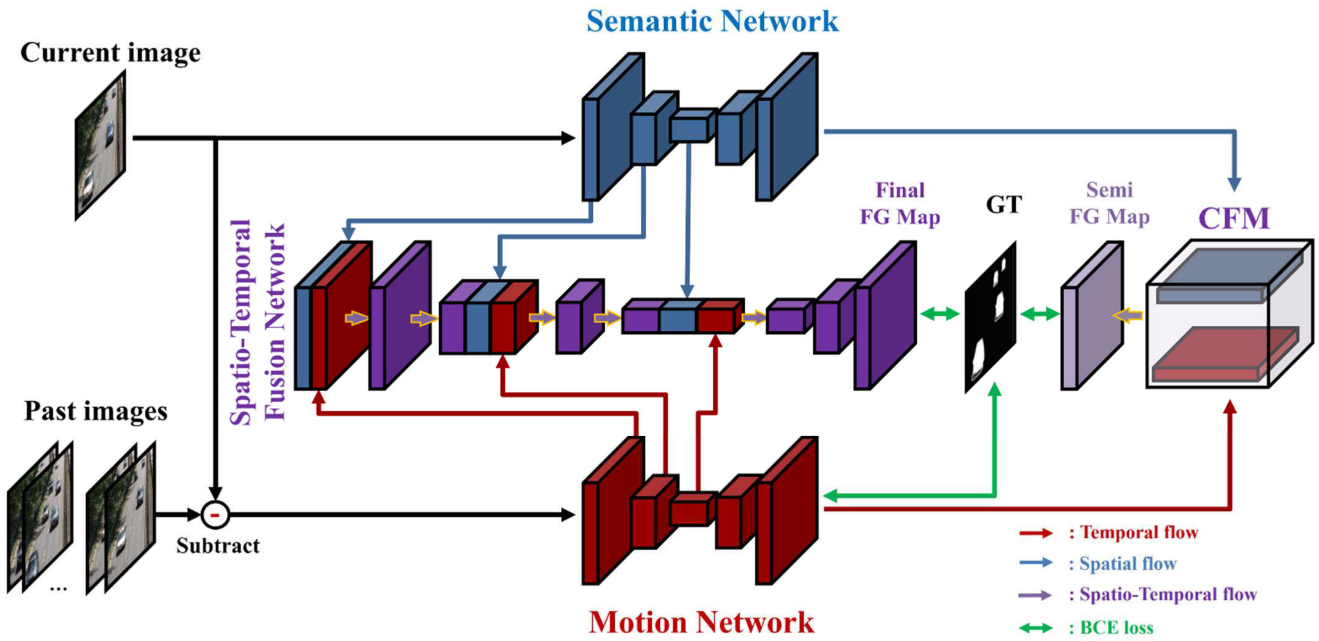


FIGURE 2. The structure of MSF-NET. MSF-Net consists of three networks: semantic, motion, and spatiotemporal fusion network. The output of three sub-networks is used for loss computation to train MSF-NET stably.

In visual surveillance, the same vehicle can be regarded as a foreground or background object according to the duration of the movement. As shown in Figure 1(b), objects such as fountains or grass should be classified as background objects regardless of their motion. In contrast, humans should be detected as foreground objects irrespective of action, as shown in Figure 1(c). In visual surveillance, foreground object detection has different aspects than object detection, as shown in previous cases. Therefore, temporal and spatial information should be considered for the robust detection of foreground objects in visual surveillance.

The proposed algorithm is based on a split and merge framework for effectively using spatiotemporal information. The final network consists of three sub-networks, as shown in Figure 2. First, a semantic network extracts spatial information on the current image. Motion network uses multiple difference images as input to process temporal information. Each past image is subtracted from the current image and used as input. We use difference images because they can easily detect temporal difference compared to multiple original images. Finally, features from two networks are integrated to obtain the final foreground map.

Figure 2 shows the proposed MSF-NET structure that consists of three sub-networks: semantic network, motion network, and spatiotemporal fusion network. Three networks follow the U-Net [46] structure that has an efficient design and has shown its performance in various applications. Since the proposed network consists of three sub-networks, we found a problem when we trained the network using the loss term, including only the final output. This problem is solved by considering the three networks' output in loss configuration. Generally, visual surveillance datasets provide foreground maps as ground truth labels. Therefore, the output of the semantic network in Figure 2 cannot be directly used in loss computation. In contrast, outputs of motion network and spatiotemporal fusion network can be directly used in loss computation. We generate semi-foreground maps through a compact fusion module that uses semantic network and motion network outputs as input. Through this process, we reflect the output of the semantic network in the loss terms. Since all outputs of three sub-networks are reflected in loss terms, it is possible to train the proposed network stably.

A. SEMANTIC NETWORK

We configure the semantic network by adopting U-NET [46], which extracts local and global features through pooling layers and skip connections. In the original U-NET [46], the sizes of the input and output layers are different. In visual surveillance, foreground object detection must consider the entire image area. Therefore, we design the semantic network with the same input and output size. Since three networks are used in the proposed method, real-time processing on the desktop GPU may be difficult when we use the model size of the original U-NET [46]. We reduce the model size of the semantic network by configuring layer depth to half compared to the U-NET [46].

We only use a current image as the input of the semantic network to prevent temporal information passes since we want temporal information to be independently processed in the motion network. As a result, the semantic network can concentrate on extracting spatial features of a current image without observing temporal information. We design the semantic network to extract only features of foreground objects regardless of motion.

Most deep-learning networks use the last layer as output. Still, we do not use the last layer of semantic and motion networks in computing the foreground map. The semantic map, the final result of the semantic network, acts as a constraint for training semantic features. The semantic network has six outputs: five intermediate semantic features and one final output, a semantic map. One from each of the five layers in the semantic network is used to extract semantic features at various sizes.

Figure 3(a) shows the structure of the semantic network. SF and SM stand for semantic features and semantic map. We use six tuples of output from the semantic network as follows.

$$(SF_1, SF_2, SF_3, SF_4, SF_5, SM) = f_{N_s}(I_c) \tag{1}$$

N_s represents the semantic network and I_c is a current image. SF_i represents the output of i-th layer of a semantic network, and SM represents the semantic map as the final output of the semantic network.

$SF_i (i = 1, \ldots, 5)$ is used as the input of spatiotemporal fusion network (STFN) as shown in Figure 2. We configure STFN to use the output of the left part not the right part, in U-Net, which corresponds to early fusion. For the motion network, the same process of early fusion is applied. We designed this early fusion configuration for SFTN to use raw information rather than processed information by semantic and motion networks. This conforms to our design rule of MSF-NET leveraging split and merge framework. In Figure 3(a), layers in blue correspond to $SF_i (i = 1, \ldots, 5)$. SM is used as input of compact fusion network (CFM).

B. MOTION NETWORK

Motion network also uses U-NET [46] structure like a semantic network. But, the motion network uses multiple difference images as input to focus on temporal information, not spatial information. We split spatial and temporal information on multiple successive images throughout this design. The proposed algorithm uses one current image and 49 past images as input data. In the motion network, 49 difference images generated by subtracting the current and 49 past images are used as input. Information loss in similar regions may occur when we use different images, not original ones. Nevertheless, we use difference images as the input of the motion network to focus on temporal information.

Since spatial information is already reflected in the semantic network, information loss related to the spatial domain in the motion network can be compensated. Most spatial domain information can be removed through subtraction, but a small amount of spatial domain information can still flow into the motion network. Unlike the semantic network, where direct loss computation using a foreground label is impossible, the motion network can perform loss computation with a foreground label.

Like a semantic network, we use six outputs from the motion network. It generates five motion features and the final output of a motion map. Figure 3(b) shows the structure of the motion network. MF and MM stand for motion features and motion maps, respectively. We use six tuples of output from the motion network as follows.

$$(MF_1, MF_2, MF_3, MF_4, MF_5, MM) = f_{N_m}(I_1^d, \ldots, I_N^d) \tag{2}$$

N_m represents motion network, and I_i^d represents a difference image between a current image and an i-th past image. MF_i represents the output of i-th layer of motion network. MM represents a motion map that is the final output of the motion network. Later, $MF_i (i = 1, \ldots, 5)$ is used as input of spatiotemporal fusion network (STFN) as shown in Figure 2. Similar to semantic network, we use outputs from the left part of the motion network as inputs of STFN. In Figure 3(b), layers in red correspond to $MF_i (i = 1, \ldots, 5)$. MM is used as input of compact fusion network (CFM).

C. COMPACT FUSION NETWORK

The proposed method first divides spatiotemporal information on multiple images into spatial and temporal domains. The semantic network extracts features of foreground objects in the spatial domain regardless of movement. For example, a vehicle at rest is not a foreground object. Still, since the vehicle belongs to candidates of foreground objects, the semantic network needs to detect the features of the corresponding vehicle. However, in most visual surveillance datasets, labels of object types are not provided. They provide labels for foreground and background objects. Therefore, it isn't easy to train a semantic network with the capability of classifying into multiple classes using the provided dataset. We present a method to indirectly train the semantic network using the semi-foreground map (SFM) by compact fusion module rather than directly training the semantic network using foreground labels provided by the dataset.

Figure 4 shows the structure of the proposed compact fusion module (CFM). The internal operation in CFM is as follows.

$$SFM = tanh\left(K * SM * up_{(16,16)}\left(mp_{(16,16)}\left(K * MM\right)\right)\right) \tag{3}$$

K represents a constant multiplied by a semantic and motion map set to 1.5. SM represents a semantic map and corresponds to the output of the semantic network in Figure 3(a). MM represents the motion map and corresponds to the output of the motion network in Figure 3(b). $up_{(16,16)}$ represents up-sampling of 16 × 16 size and $mp_{(16,16)}$ represents max pooling of 16 × 16 size.

The compact fusion module uses the semantic and motion network outputs as inputs. Each input layer multiplies every pixel by K. Then, for the motion map, we perform max pooling with a 16 × 16 filter size, and finally, we apply 16 × 16 up-sampling to have an equal size with the semantic

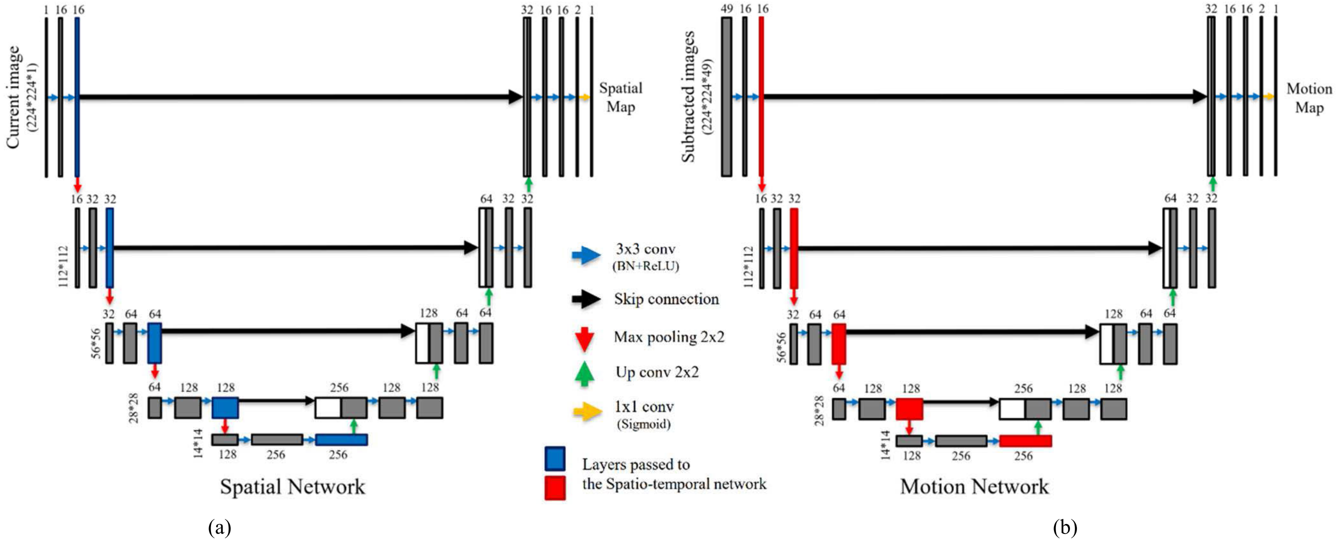


FIGURE 3. The structure of semantic and motion network (a) semantic network (b) motion network.

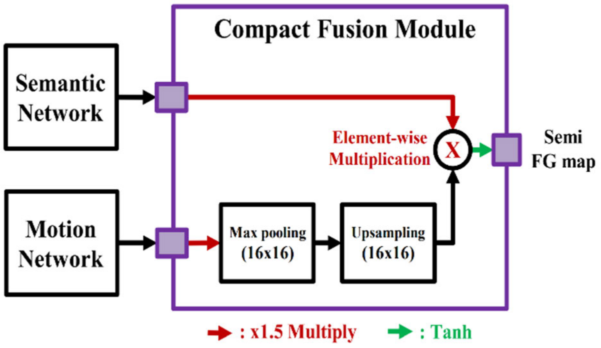


FIGURE 4. The structure of the compact fusion module.

map. Finally, tanh follows by elementwise multiplication. Since SM and MM have a value between 0 and 1, results after elementwise multiplication would have a value between 0 and K^2. We denote the final result obtained by applying tanh as a semi-foreground map. In the final stage, we select tanh as the activation function because it has a sharp gradient compared to sigmoid in the range between 0 and 1, which is advantageous for backpropagation during training. We design the output of CFM to act like a foreground probability map. Throughout this, we can use the output of the CFM in loss computation by comparing it with the ground truth foreground map.

The introduction of CFM into the proposed MSF-NET is based on the following considerations, and we show the validity of the CFM module in experimental results.

First, MSF-NET consists of three sub-networks: a semantic network, a motion network, and a spatiotemporal fusion network. For stable training of the proposed MSF-NET, each output of three sub-networks is needed to participate in loss computation. The output of the semantic network is similar to the feature map of object classification. Therefore, comparing it with the ground truth foreground map is meaningless. We convert the output of the semantic map into a semi-foreground map by fusing it with the output of the motion map using CFM. We can use the semi-foreground map in loss computation since it has a similar tendency to a ground truth foreground map.

Second, we design a semantic network to focus on the foreground object regions with an additional margin of nearby areas of foreground objects. To this end, the motion map is amplified using max pooling and up-sampling followed by elementwise multiplication with a semantic map. The introduction of the previous step is based on the following considerations. When candidates of foreground objects are stationary, the output of the motion network will be close to 0, while the semantic network detects features on the object's part and gives output with a high value. If we multiply the semantic map and motion map without max pooling and up-sampling in the motion map, the result would always be 0.

Third, we configure CFM to have no trainable parameters. When there are parameters in CFM, a semi-foreground map having low cost is possible after training regardless of the quality of the semantic map and motion map. To cope with this problem, we configure CFM with no trainable parameters. CFM acts as an auxiliary module to train the semantic network.

D. SPATIOTEMPORAL NETWORK

We configure the semantic and motion networks in the proposed MSF-NET to extract spatial and temporal features. For the following reasons, more than semantic and motion networks are required for robust foreground object detection.

The semantic network cannot observe temporal domain information since we use only a current image as input. The

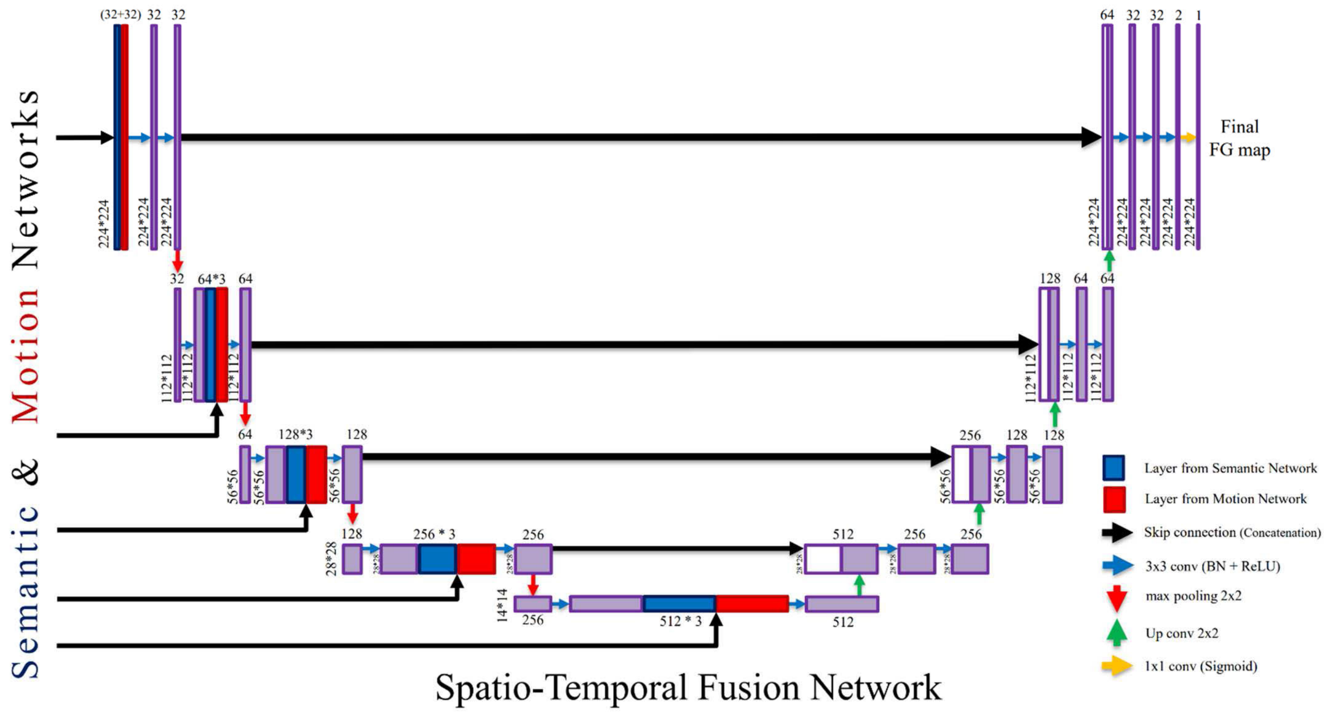


FIGURE 5. The structure of spatiotemporal fusion network (STFN).

motion network has limitations in catching spatial domain information since we use multiple difference images. Also, the compact fusion module comprises multiplication, max pooling, up-sampling, and non-linear functions without trainable parameters, so it has limitations in generating a stable foreground probability map. Therefore, an additional network that integrates spatial and temporal information is required. We propose a spatiotemporal fusion network (STFN) for this purpose.

Figure 5 shows the structure of STFN. Its design follows U-NET [46] and provides a foreground map by processing features from semantic and motion networks. Unlike semantic and motion networks that receive a current image and multiple difference images, STFN uses ten outputs from semantic and motion networks to effectively integrate spatiotemporal information as input.

The process of generating a foreground map by STFN is as follows.

$$SMF_i = Concatenate(SF_i, MF_i) \tag{4}$$

$$FFM = f_{N_{stfn}}(SMF_1, SMF_2, SMF_3, SMF_4, SMF_5) \tag{5}$$

SMF_i represent i-th semantic motion feature. SF_i and MF_i is i-th semantic feature and motion feature and their locations are displayed in blue and red in Figure 3. N_{stfn} represents a spatiotemporal fusion network. FFM represents a final foreground map.

In Figure 5, blue layers correspond to semantic features received from the semantic network, and red layers represent motion features obtained from the motion network. Layers marked in purple represent the fusion layers of semantic and motion features. STFN uses concatenated features from semantic and motion networks as input. After convolution twice, max pooling is used to reduce the layer size.

We can effectively use the spatial information received from the semantic network and temporal information obtained from the motion network. The expanding path in the STFN is similar to semantic and motion networks.

E. LOSS TERM

The proposed MSF-NET comprises three networks: semantic, motion, and spatiotemporal fusion. The configuration of cost term only using the final foreground map of the spatiotemporal fusion network does not guarantee stable training of the proposed MSF-NET, found by experiments. In the proposed method, we reflect the results of three sub-networks directly or indirectly in cost computation to train the proposed MSF-NET stably. The configuration of cost terms is as follows.

$$L_{train} = L_{FFM} + L_{SFM} + L_{MM} \tag{6}$$

$$L_{val} = L_{FFM} \tag{7}$$

L_{train} represents the cost term used for training. L_{FFM} is a loss term by a final foreground map generated by STFN. L_{SFM} is a loss item using a semi-foreground map obtained by a compact fusion module (CFM). L_{MM} is a loss item using a motion map, which is the output of the motion network. All these three loss terms use binary cross entropy. The final loss is the sum of three loss terms, and we equally reflect three terms. L_{val} is validation loss, and it is used for the adjustment of the learning rate during training.

IV. EXPERIMENTAL RESULTS

The proposed method comprises three sub-networks: semantic, motion, and spatiotemporal fusion network. Each network is designed following the U-NET [46], and batch normalization [47] is used after all convolution layers to improve training speed and stability. He et al. initialization [48] was used for the initialization of each layer. In the last convolution layer of each network, the sigmoid function was used to set the layer output to [0, 1], and ReLU was used as the activation function except for the output layer.

The proposed MSF-NET has 29,619,895 parameters, among which 29,599,467 parameters are trainable. Adam optimizer [49] was used for training. The batch size is set to 4, and the initial learning rate is set to 0.001. When validation loss does not decrease more than five times, we reduce the learning rate by half. In addition, if the validation loss dropped less than ten times, we stopped training.

The proposed method uses 50 images from the past to the present as input of MSF-NET. Most foreground object detection methods use a frame interval of 1. However, there is a limitation in observing data over a long period when we use a frame interval of 1. In the proposed method, the frame interval is set to 10. Therefore, observation is possible for a total range of 490 frames. By adjusting frame intervals during training, we could obtain robust detection results in an environment where a foreground object has been stationary for a long time. This training method can be applied without additional data augmentation. In ablation studies, experiments show performance improvement by the proposed method.

Training a model using samples from a test environment gives the best performance, but preparing labels requires a lot of time and cost. For this reason, it is necessary to design a model that can operate well in unseen environments. Therefore, it is required to evaluate the detection performance in a completely different environment not used for training. Mandal and Vipparthi [4] proposed two evaluation methods of Scene Independent Evaluation (SIE) and Scene Dependent Evaluation (SDE) for visual surveillance. The SIE environment refers to an environment that separates training data and evaluation data into separate scenes, and the SDE environment refers to an environment in which specific scenes are internally divided into training data and evaluation data.

We use three datasets of CDnet2014 [1], LASIESTA [50], and SBI [51] in experiments. We divide experiments into two cases. Firstly, a model is trained with the CDnet2014 [1] dataset, which has the largest number of samples among the three datasets, and then evaluated on the LASIESTA [50] and SBI [51] datasets. Secondly, we internally divide the CDnet 2014 [1], LASIESTA [50], and SBI [51] datasets into training and evaluation data, respectively. All experiments were performed with a scene-independent evaluation (SIE) method.

The CDnet2014 dataset [1] provides five labels: background, shadow, unknown motion, out of ROI, and foreground. Since the shadow is regarded as a background object in most visual surveillance datasets, we treat shadows as background labels during training. Unknown motion is a label that exists on the outline of an object. Since the distinction between foreground and background objects is ambiguous for unknown motion and out of ROI label, we exclude them in loss computation.

For evaluation metrics, we use Recall, Precision, FM, and PWC, defined as follows.

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{FM} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$\text{PWC} = \frac{FP + FN}{TP + TN + FP + FN} \times 100$$

TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

A. SIE EXPERIMENTS USING DIFFERENT DATASETS

We first train the proposed model using the CDnet2014 [1] dataset, and evaluation is done using the LASIESTA [50] and SBI [51] datasets. Since night and thermal images in the CDnet2014 do not exist in the LASIESTA [50] and SBI [51] datasets, we exclude them in training. Finally, we train the proposed model using 23 scenes from five categories in the CDnet2014 [1] dataset. For each scene, 80% was used as train data, and 20% was used as validation data. A total of 42,345 images were used, of which 33,868 were used as training data, and 8,477 were used as validation data.

Firstly, evaluation is done using the LASIESTA [50] dataset. In the LASIESTA [50] dataset, we evaluate using 20 scenes from 10 categories except scenes containing camera movement.

Table 1 shows the comparison results of FM values by the proposed method and other algorithms. In Table 1, four algorithms are done in the SIE setup, and different algorithms' results are based on training using the LASIESTA dataset.

Figure 6 compares foreground maps with the proposed and other methods. Among comparison methods, 3DCD [13] and FgSegNet-v2 [2] do training under the same conditions as the proposed method. Fast-D [30], SuBSENSE [21], Berjón et al. [27], and Haines and Xiang [20] is a classic foreground object detection methods. SuBSENSE [21] was evaluated using BGSlibrary [52]. Results of Fast-D [30], Berjón et al. [27], and Haines and Xiang [20] are ones noted in each paper.

The proposed method offers outstanding detection performance with an average FM of 0.9487 and a mean false detection rate (PWC) of 0.2336 in the LASIESTA dataset. It amounts to a 9% higher FM than the latest deep learning-based algorithm, 3DCD [13]. It amounts to 13% higher FM than Fast-D [30], the latest classical algorithm.

Since the proposed method uses images from a wide range of 490 frames as input, the model takes a little longer to adapt to the light change environment where an instantaneous light

TABLE 1. FM score comparison to other algorithms on the LASIESTA dataset.

LASIESTA DATASET	Proposed (SIE)	LTS-D [62]	LTS-U [62] (SIE)	ADNN [63]	3DCD [13] (SIE)	Fast-D [30]	SuBSENSE [21]	Berjón et al. [27]	Haines et al. [20]	FgSegNet - v2 [2](SIE)
I_SI	0.9680	0.9869	0.9229	0.9536	0.9076	0.9287	0.9086	0.8806	0.8876	0.7568
I_CA	0.9463	0.9310	0.7851	0.9504	0.6721	0.8924	0.8702	0.8444	0.8938	0.7194
I_OC	0.9739	0.9895	0.9447	0.9759	0.9559	0.9194	0.9249	0.7807	0.9223	0.3563
I_IL	0.9198	0.9888	0.5460	0.7661	0.9365	0.5021	0.4685	0.6488	0.8491	0.4628
I_MB	0.9623	0.9921	0.9400	0.9802	0.8776	0.9430	0.9173	0.9374	0.8440	0.6312
I_BS	0.9496	0.9861	0.8898	0.9707	0.8370	0.6182	0.6480	0.6644	0.6809	0.3825
O_CL	0.9523	0.9901	0.9435	0.9814	0.9004	0.9368	0.9155	0.9277	0.8267	0.4343
O_RA	0.9256	0.9870	0.9356	0.9868	0.8378	0.9378	0.8756	0.8670	0.8908	0.4912
O_SN	0.9253	0.9891	0.8451	0.9647	08461	0.8789	0.7925	0.7787	0.1750	0.0777
O_SU	0.9643	0.9773	0.8949	0.9300	0.8975	0.8710	0.7919	0.7222	0.8568	0.1861
Average	0.9487	0.9806	0.8648	0.9460	0.8668	0.8428	0. 8113	0.8051	0.7826	0.4498

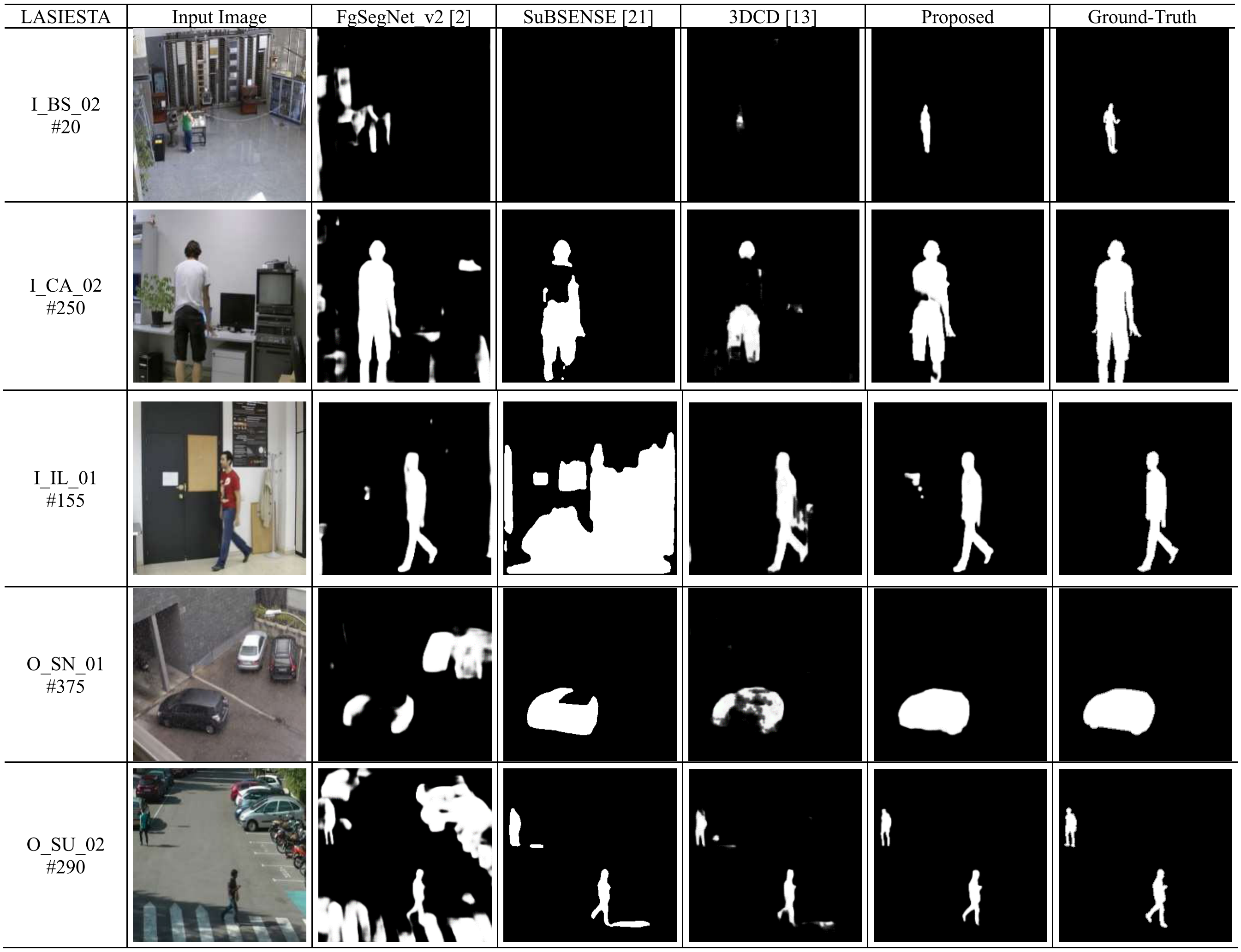


FIGURE 6. Qualitative evaluation of the LASIESTA dataset.

change occurs. Therefore, there was a slight drop in the light change environment of 'I_IL' in Table 1, but it showed the second-best performance among other algorithms. In addition, the proposed method offers robust performance in environments where objects have been stopped for a long time, in bootstrap and challenging weather environments.

Next, we evaluate the proposed method using the SBI dataset. We use the same trained model to evaluate the LASIESTA [50] dataset. The SBI dataset consists of a total of 14 scenes. Snellen and Foliage, which classify moving leaves as foreground objects contrary to the convention of visual surveillance, were excluded from evaluation. Also, the

TABLE 2. Comparison of FM score with other algorithms on the SBI dataset.

SBI DATASET	Proposed (SIE)	Yang et al. [53]	3DCD [13] (SIE)	SuBSENSE [21]	ReProCS [54]	FgSegNet-v2 [2] (SIE)	MEROP [55]
Board	0.9161	0.91	**0.9187**	0.5777	0.69	0.6942	0.48
CIVIAR1	**0.9726**	0.95	0.9401	0.9144	0.70	0.7920	0.71
CIVIAR2	**0.9414**	0.84	0.8735	0.8714	0.48	0.0600	0.51
CaVignal	0.8188	0.83	0.5275	0.3980	0.58	**0.9191**	0.62
Candela	0.8988	**0.93**	0.5223	0.5356	0.81	0.3335	0.58
Hall & Monitor	**0.9749**	0.83	0.8993	0.7758	0.79	0.7592	0.80
Highway1	**0.9424**	0.72	0.7212	0.5523	0.61	0.7759	0.61
Highway2	**0.9707**	0.95	0.9310	0.8937	0.47	0.9174	0.73
Human Body2	0.9130	0.88	**0.9157**	0.8346	0.61	0.6068	0.62
IBM TEST2	**0.9692**	0.89	0.8967	0.9390	0.66	0.5357	0.65
People & Foliage	**0.7942**	-	0.7746	0.2660	-	0.6845	-
Average	**0.9193**	-	0.8110	0.6871	-	0.6435	-
Average(Except P&F scene)	**0.9318**	0.87	0.8146	0.7293	0.64	0.6394	0.63

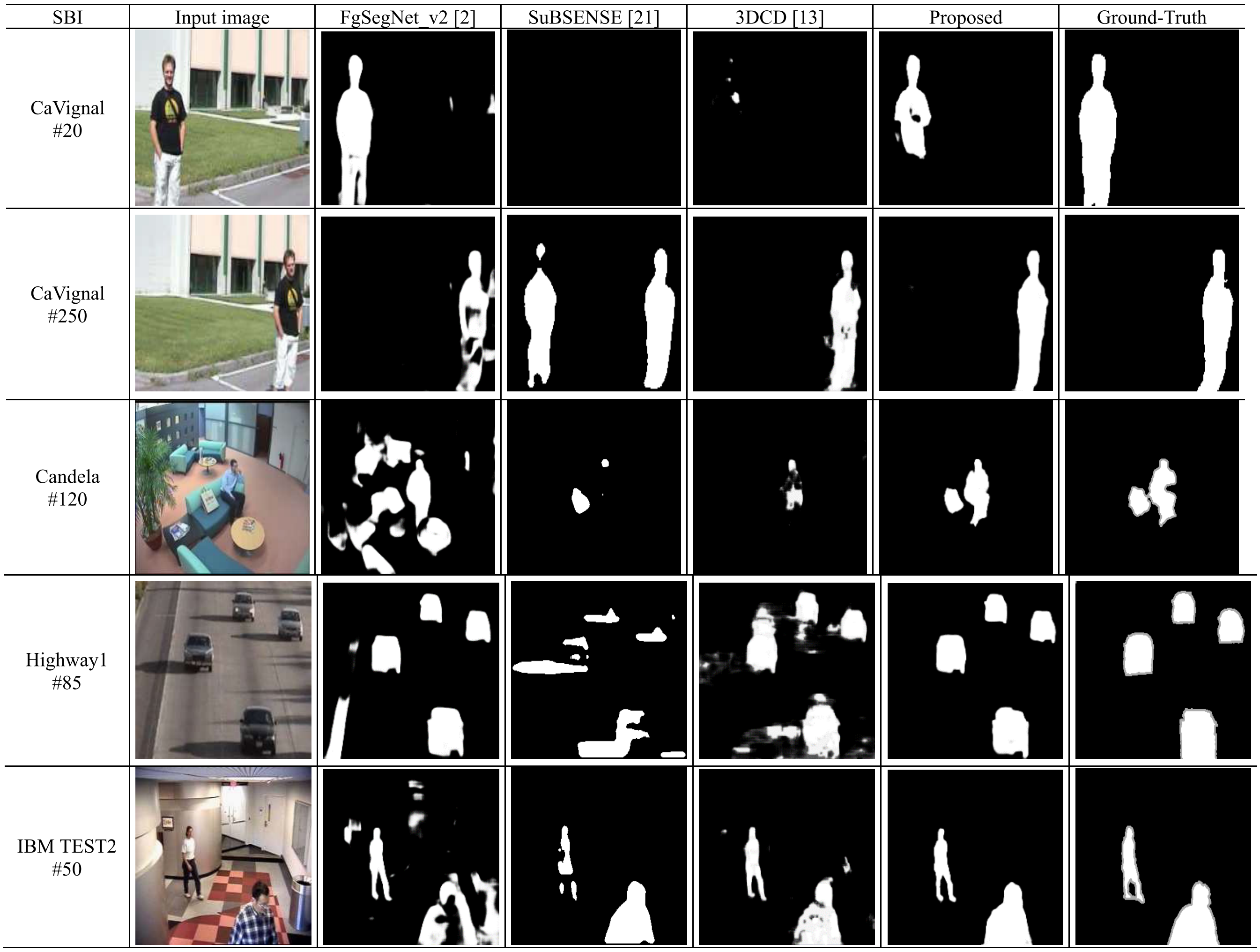


FIGURE 7. Qualitative evaluation on the SBI dataset.

Toscana scene comprising six non-consecutive images was excluded from evaluation since the proposed method uses 50 images as input.

Table 2 compares FM scores by the proposed method and other algorithms. Figure 7 shows the result of foreground object detection on the SBI dataset by the proposed and other methods. Like experiments on the LASIESTA dataset, 3DCD [13] and FgSegNet-v2 [2] were trained using the same data as the proposed method. Yang et al. [53], SuBSENSE [21], ReProCS [54], and MEROP [55] belong to classic foreground object detection methods.

TABLE 3. Comparison of performance with other algorithms in CDne2014 internal conflict environment.

CDnet2014	BL	PE	SW	BO	PA	TP	TS	BS	CO	T1	Avg	Avg(Except C.M)
FgSegNet-S [35]	0.74	0.65	0.12	0.42	0.17	0.57	0.41	0.52	0.74	0.17	0.45	0.53
FgSegNet-v2 [2]	0.70	0.33	0.22	0.62	0.52	0.74	0.43	0.53	0.77	0.12	0.50	0.58
PAWCS [22]	0.66	0.95	0.74	0.88	0.21	0.91	0.86	0.86	0.65	0.68	0.74	0.75
WeSamBe [56]	0.86	0.97	0.85	0.64	0.41	0.91	0.86	0.86	0.89	0.71	0.80	0.80
IUTIS-5 [26]	0.80	0.97	0.81	0.75	0.65	0.89	0.87	0.87	0.90	0.63	0.81	0.84
MBS [25]	0.86	0.96	**0.90**	0.90	0.62	0.89	0.87	0.87	**0.92**	0.54	0.83	0.86
BMN-BSN [57]	0.84	0.96	0.63	**0.95**	0.77	0.72	0.82	0.92	0.90	0.56	0.81	0.86
3DCD [13]	**0.94**	0.93	0.83	0.88	0.84	**0.92**	0.75	0.79	**0.92**	**0.82**	**0.86**	0.87
SemBGS [58]	0.84	0.98	0.85	0.98	0.69	0.88	0.92	0.92	0.82	0.30	0.82	0.88
BSUV-NET [8]	0.82	0.97	0.69	0.89	0.91	0.91	0.80	**0.94**	0.83	0.66	0.84	0.88
BSUV-NET+SemBGS	0.82	0.97	0.71	0.91	0.91	0.91	0.80	**0.96**	0.82	0.65	0.85	0.89
Proposed(MSF-NET)	0.87	**0.98**	0.63	0.92	**0.96**	0.84	**0.94**	0.92	0.86	0.59	0.85	**0.91**

TABLE 4. Comparison of performance with other algorithms in LASIESTA internal conflict environment.

LASIESTA	I_SI_2	I_CA_2	I_OC_2	I_IL_2	I_MB_ 2	I_BS_2	O_CL_ 2	O_RA_ 2	O_SN_ 2	O_SU_ 2	Avg
FgSegNet-S [35]	0.20	0.60	0.53	0.25	0.60	0.28	0.19	0.16	0.05	0.18	0.30
FgSegNet-M [35]	0.56	0.55	0.65	0.42	0.56	0.19	0.28	0.18	0.01	0.33	0.37
FgSegNet-v2 [2]	0.53	0.58	0.25	0.41	0.63	0.25	0.54	0.54	0.05	0.29	0.41
3DCD [13]	0.86	0.49	0.93	0.85	0.79	0.87	0.87	0.87	0.49	0.83	0.79
ChangeDet [4]	0.83	0.66	0.89	**0.87**	0.77	0.82	0.90	0.85	0.51	0.81	0.79
Proposed(MSF-NET)	**0.98**	**0.94**	**0.97**	0.65	**0.88**	**0.88**	**0.95**	**0.98**	**0.76**	**0.95**	**0.89**

The proposed method gives 13% higher FM than 3DCD [13], the latest deep learning method. It has a 7% higher FM than Yang et al. [53]. The proposed method shows a good detection performance in most environments except the Cavignal scene. The Cavignal scene is a challenging scenario belonging to a bootstrap environment where stationary foreground objects exist from the start frame. In the Cavignal scene, FgSegNet-v2 [2] shows the best performance. The proposed method gives miss-detection results at an early stage of the Cavignal scene, but it can correctly detect later, as shown in Figure 7.

The proposed method enables robust detection when the object of interest is slightly moving. Still, detection performances are degraded in the combined case of the Cavignal scene and the bootstrap environment. However, as shown in Cavignal #250 in Figure 7, the proposed method provides accurate detection when the object of interest moves. The proposed method performs well in challenging environments like the Candela scene, where a foreground object is stationary for a long time, and Highway 1 and 2 scenes, where the camera is slightly shaken.

B. SIE EXPERIMENTS USING THE SAME DATASET

In this section, we provide experimental results following the SIE setup. CDnet2014 [1], LASIESTA [50], and SBI [51] datasets are internally divided into training and evaluation data and used for experiments. We follow the method in 3DCD [13] for internal division.

First, the CDnet2014 [1] dataset was internally divided, and an evaluation was done. Among 12 categories in CDnet2014 [1], 11 categories were used for experiments, excluding the PTZ category obtained under camera motion. One scene from each category is used for evaluation data, and the remaining scenes are used for training data.

Table 3 shows the comparison results of the FM score by the proposed method and other algorithms. Results of different algorithms are used as ones noted in 3DCD [13]. The proposed method gives an average FM score of 0.85, which is 1% lower than 3DCD [13], having an average FM score of 0.86 on CDnet2014 [1]. The proposed method shows inferior performance in the sidewalk (SW) scene of the camera jitter category, and in the turbulence 1 (T1) scene, which involves camera shakes.

On the other hand, for scenes excluding ones caused by camera movement, the proposed method gives an average FM score of 0.91 and a PWC of 0.57, and it amounts to a 5% higher FM and 25% lower PWC score than 3DCD [13].

Next, the LASIESTA [50] dataset was internally divided, and evaluation was done. LASIESTA dataset [50] consists of 10 categories and 20 scenes. For each category, scene 1 was used for training, and scene 2 was used for evaluation. Table 4 compares FM scores by the proposed method and other deep-learning algorithms. Results of different algorithms are used, ones noted in 3DCD [13]. The proposed method shows a 13% higher FM score than the latest deep learning method of 3DCD [13].

Finally, the SBI [51] dataset was internally divided, and evaluation was done. Four scenes of candela, CAVIAR2, cavignal, and highway2 were used for evaluation data, and the remaining nine scenes were used for training data. Table 5 shows the comparison results of FM scores by the proposed method and other algorithms. The proposed method offers a dramatically 32% higher FM score than 3DCD [13].

TABLE 5. Comparison of performance with other algorithms in SBI internal conflict environment (FM).

Method	Candela	CAVIAR2	CaVignal	Highway2	Avg
FgSegNet-S [35]	0.23	0.11	0.68	0.24	0.31
FgSegNet-M [35]	0.15	0.14	0.72	0.21	0.30
FgSegNet- v2 [2]	0.27	0.10	0.63	0.58	0.40
ChangeDet [4]	0.61	0.56	0.48	0.64	0.57
3DCD [13]	0.67	0.62	0.53	0.59	0.60
Proposed(MSF-NET)	**0.75**	**0.86**	**0.74**	**0.82**	**0.79**

C. ABLATION STUDIES

We show the validity of sub-modules in the proposed MSF-NET through ablation studies. We present experimental results without using critical components of subtraction in the motion network and CFM module. Also, we investigate the effect of training MSF-NET under a frame interval of one, not ten, which is used for the test. Experiments are done after training using the CDnet2014 dataset [1], and then evaluations are done using LASIESTA [50] and SBI [51] datasets.

Tables 6 and 7 show comparison results of FM scores for not using critical components in the proposed MSF-NET on LASIESTA [50] and SBI [51] datasets. Figure 8 shows some representative results according to ablation studies.

We get a relative 1% and 2% improvement in the LASIESTA and SBI datasets when we use train interval adjustment. That improvement can be regarded as one that can be obtainable when initial values of parameters are randomly chosen. Therefore, we can conclude that train interval adjustment has negligible effects.

In the LASIESTA dataset, a clear performance improvement can be noticed. In the SBI dataset, a clear performance improvement like the LASIESTA dataset cannot be seen, which requires further investigation.

In Figure 8, red boxes correspond to miss detection that regards stationary background objects as foreground objects. Using multiple difference images as input of motion network and training the semantic network using CFM prevent erroneous detection of stationary background objects as foreground objects, as shown in Figure 8. In Figure 8, blue boxes correspond to regions not detected as foregrounds, which occurs when foreground objects are stationary for a long time. We could partially solve this problem by adjusting frame intervals during training and test time, as shown in Figure 8.

TABLE 6. The evaluation result of the FM score in the LASIESTA dataset according to whether or not the proposed methods are applied.

LASIESTA DATASET	Prop osed	W/O Subtract	W/O CFM	W/O Subtract & CFM	W/O Train interval Adjustme nt
I_SI	0.97	0.87	0.51	0.67	0.97
I_CA	0.95	0.85	0.80	0.84	0.88
I_OC	0.97	0.62	0.74	0.62	0.98
I_IL	0.92	0.87	0.52	0.48	0.96
I_MB	0.96	0.91	0.88	0.79	0.97
I_BS	0.95	0.93	0.83	0.79	0.93
O_CL	0.95	0.94	0.82	0.74	0.88
O_RA	0.93	0.93	0.67	0.86	0.95
O_SN	0.93	0.93	0.87	0.66	0.90
O_SU	0.96	0.89	0.89	0.65	0.97
Average	0.95	0.87	0.75	0.71	0.94

TABLE 7. The evaluation result of the FM score in the SBI dataset according to whether or not the proposed methods are applied.

SBI DATASET	Prop osed	W/O Subtrac t	W/O CFM	W/O Subtract & CFM	W/O Train interval Adjustment
Board	0.92	0.96	0.94	0.96	0.87
CIVIAR1	0.97	0.97	0.97	0.91	0.98
CIVIAR2	0.94	0.91	0.92	0.80	0.94
CaVignal	0.82	0.81	0.83	0.83	0.72
Candela	0.90	0.87	0.92	0.83	0.88
Hall & Monitor	0.97	0.93	0.96	0.93	0.97
Highway1	0.94	0.95	0.88	0.89	0.94
Highway2	0.97	0.95	0.96	0.96	0.95
Human Body2	0.91	0.89	0.83	0.75	0.96
IBM TEST2	0.97	0.97	0.96	0.88	0.97
People & Foliage	0.79	0.79	0.77	0.79	0.72
Average	0.92	0.91	0.90	0.86	0.90

D. LAYER VISUALIZATION

Figure 9 shows the output of some inner layers by the proposed MSF-NET. The proposed method is based on a split and merge framework comprising three sub-networks of semantic, temporal, and spatiotemporal fusion networks. Semantic network targets extract spatial information from a current image. The temporal network wants to extract temporal information on multiple difference images. These two networks correspond to a split part. Finally, the foreground map is obtained by a spatiotemporal fusion network corresponding to a merging part. As shown in Figure 9, outputs of the motion network and semantic network contain significant false positives. Motion network causes an error that regards a stationary foreground object as background. A semantic network generates a mistake that reacts to background objects. However, we can obtain decent detection results through STFN regardless of the erroneous detection of each network. This shows the validity of the proposed MSF-NET designed by a split and merge framework.

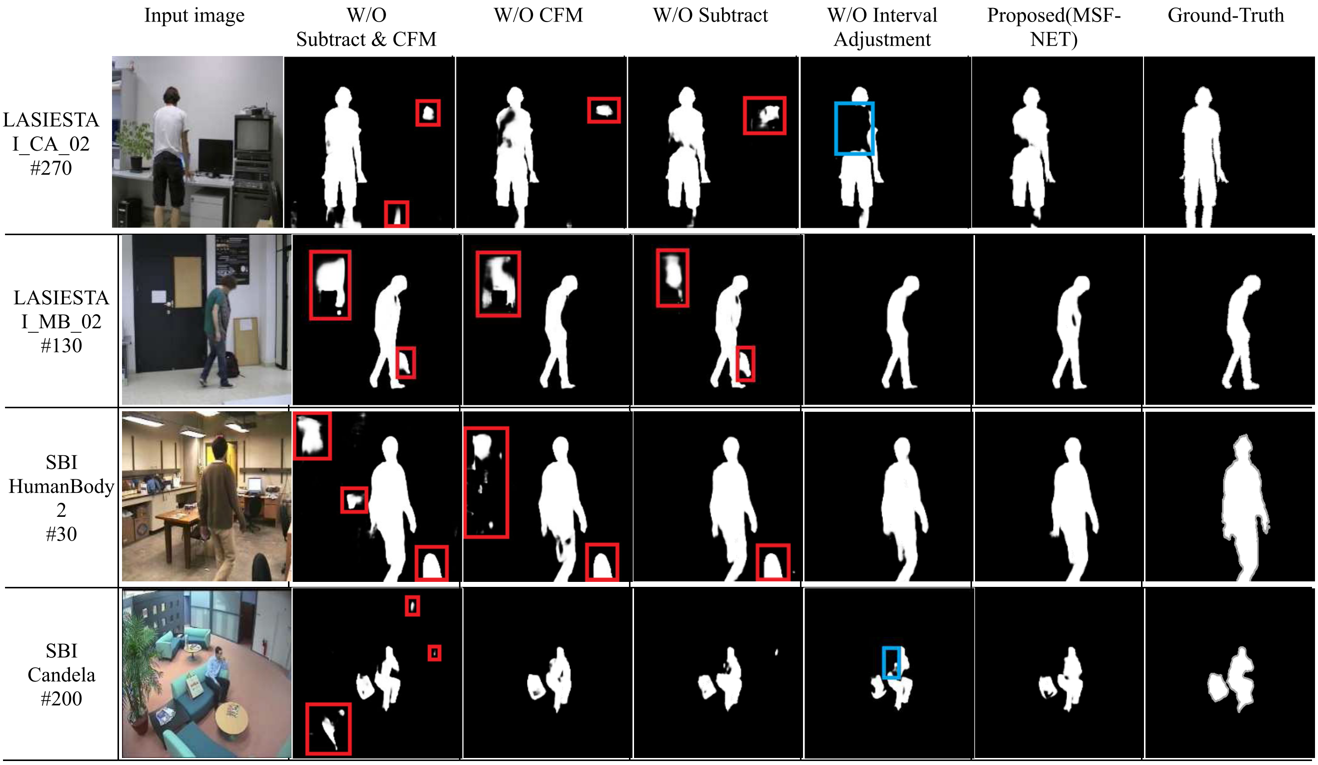


FIGURE 8. Ablation study qualitative evaluation result (red box: when the background is incorrectly classified as foreground, blue box: when the foreground is incorrectly classified as background).

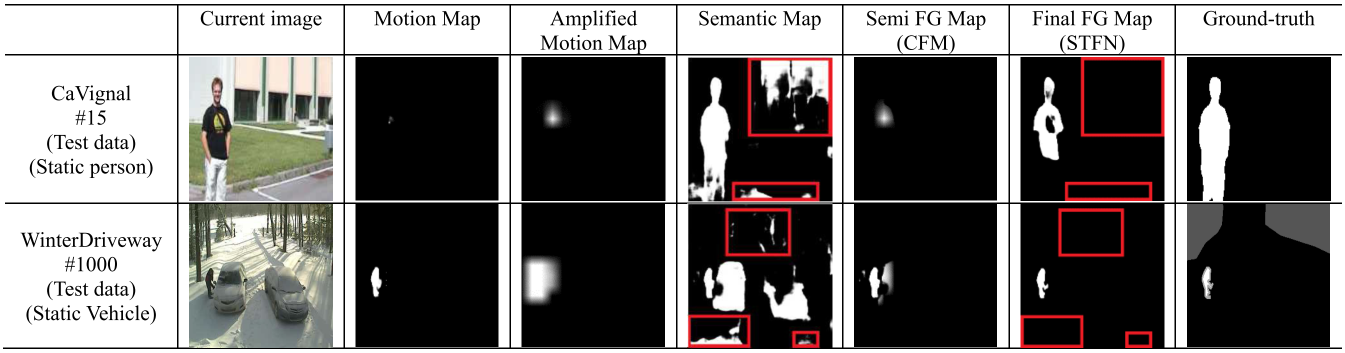


FIGURE 9. Visualization of inner layers of the proposed algorithm.

TABLE 8. Comparison of the number of parameters and computation time with other algorithms.

Method	Number of Param	fps
BSUV-Net 2.0 [43]	-	6 (Tesla P100)
Fast BSUV-Net 2.0 [43]	-	29 (Tesla P100)
3DCD [13]	0.13M	30 (RTX 2080Ti)
MTPA [59]	-	37 (Tesla V100)
FgSegNet-v2 [2]	9.23M	111 (RTX 2080Ti)
ADNN [63]	-	0.33 (GTX 1080)
LTS [62]	-	6.7 (RTX 3090)
Proposed(MSF-NET)	29.62M	60 (RTX 2080Ti)

E. COMPUTATIONAL COST

We compare the computation time of the proposed method with the latest deep learning algorithms of 3DCD [13] and FgSegNet-v2 [2]. The three models were implemented using Keras, and experiments were done using a computer with NVIDIA RTX2080Ti 11GB. Table 8 compares the number of parameters and the number of frames processed per second by the proposed and comparing methods. The proposed method can process 60fps on RTX 2080Ti and has twice the processing speed compared to 3DCD [13]. FgSegNet-v2 [2] has twice the operation speed of the proposed method but shows low foreground object detection performance.

V. CONCLUSION

In this paper, we proposed MSF-NET for robust foreground object detection in visual surveillance. The proposed MSF-NET consists of three sub-networks: a semantic network, a motion network, and a spatiotemporal fusion network. It is designed by the following split and merge framework. First, we split spatial and temporal information on multiple successive images. A semantic network extracts spatial features from a current image. A motion network focuses on extracting temporal information by using multiple difference images. Finally, a spatiotemporal fusion network integrates spatial and temporal information by adopting early fusion from the semantic and motion networks. Also, we propose a method for stably training the proposed MSF-NET. We make all outputs of three sub-networks involved in cost computation by introducing a semi-foreground map through a compact fusion module having no trainable parameters. Experimental results and ablation studies show the validity of the proposed MSF-NET. However, the proposed method could have better detection performance in an environment where the camera is moving. For further research, we will focus on solving this problem.

REFERENCES

[1] 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.

[2] L. A. Lim and H. Y. Keles, "Learning multi-scale features for foreground segmentation," *Pattern Anal. Appl.*, vol. 23, no. 3, pp. 1369–1380, Aug. 2020.

[3] M. Braham and M. Van Droogenbroeck, "Deep background subtraction with scene-specific convolutional neural networks," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2016, pp. 1–4.

[4] M. Mandal and S. K. Vipparthi, "Scene independency matters: An empirical study of scene dependent and scene independent evaluation for CNN-based change detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2031–2044, Mar. 2022.

[5] I. Osman, A. Eltantawy, and M. S. Shehata, "Task-based parameter isolation for foreground segmentation without catastrophic forgetting using multi-scale region and edges fusion network," *Image Vis. Comput.*, vol. 113, Sep. 2021, Art. no. 104248.

[6] C. Lin, B. Yan, and W. Tan, "Foreground detection in surveillance video with fully convolutional semantic network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4118–4122.

[7] J.-Y. Kim and J.-E. Ha, "Foreground objects detection using a fully convolutional network with a background model image and multiple original images," *IEEE Access*, vol. 8, pp. 159864–159878, 2020.

[8] 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. 2763–2772.

[9] R. Huang, M. Zhou, Y. Xing, Y. Zou, and W. Fan, "Change detection with various combinations of fluid pyramid integration networks," *Neurocomputing*, vol. 437, pp. 84–94, May 2021.

[10] 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.

[11] X. Zhao, Y. Chen, M. Tang, and J. Wang, "Joint background reconstruction and foreground segmentation via a two-stage convolutional neural network," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2017, pp. 343–348.

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

[13] M. Mandal, V. Dhar, A. Mishra, S. K. Vipparthi, and M. Abdel-Mottaleb, "3DCD: Scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos," *IEEE Trans. Image Process.*, vol. 30, pp. 546–558, 2021.

[14] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 246–252.

[15] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.

[16] 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, Jun. 2011.

[17] M. Van Droogenbroeck and O. Paquot, "Background subtraction: Experiments and improvements for ViBe," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 32–37.

[18] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 509–515.

[19] G.-A. Bilodeau, J.-P. Jodoin, and N. Saunier, "Change detection in feature space using local binary similarity patterns," in *Proc. Int. Conf. Comput. Robot Vis.*, May 2013, pp. 106–112.

[20] T. S. F. Haines and T. Xiang, "Background subtraction with DirichletProcess mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 670–683, Apr. 2014.

[21] 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.

[22] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Universal background subtraction using word consensus models," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4768–4781, Oct. 2016.

[23] B. Laugraud, S. Piérard, M. Braham, and M. Van Droogenbroeck, "Simple median-based method for stationary background generation using background subtraction algorithms," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2105, pp. 477–484.

[24] D. K. Panda and S. Meher, "Detection of moving objects using fuzzy color difference histogram based background subtraction," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 45–49, Jan. 2016.

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

[26] 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.

[27] D. Berjón, C. Cuevas, F. Morán, and N. García, "Real-time nonparametric background subtraction with tracking-based foreground update," *Pattern Recognit.*, vol. 74, pp. 156–170, Feb. 2018.

[28] D. Ortego, J. C. Sanmiguel, and J. M. Martínez, "Hierarchical improvement of foreground segmentation masks in background subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1645–1658, Jun. 2019.

[29] K. Garg, N. Ramakrishnan, A. Prakash, and T. Srikanthan, "Rapid and robust background modeling technique for low-cost road traffic surveillance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2204–2215, May 2020.

[30] Md. A. Hossain, Md. I. Hossain, Md. D. Hossain, N. T. Thu, and E.-N. Huh, "Fast-D: When non-smoothing color feature meets moving object detection in real-time," *IEEE Access*, vol. 8, pp. 186756–186772, 2020.

[31] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[32] 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.

[33] D. Zeng and M. Zhu, "Background subtraction using multiscale fully convolutional network," *IEEE Access*, vol. 6, pp. 16010–16021, 2018.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[35] L. A. Lim and H. Y. Keles, "Foreground segmentation using convolutional neural networks for multiscale feature encoding," *Pattern Recognit. Lett.*, vol. 112, pp. 256–262, Sep. 2018.

[36] D. Zeng, X. Chen, M. Zhu, M. Goesele, and A. Kuijper, "Background subtraction with real-time semantic segmentation," *IEEE Access*, vol. 7, pp. 153869–153884, 2019.

[37] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 405–420.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[39] M. Qiu and X. Li, "A fully convolutional encoder–decoder spatial–temporal network for real-time background subtraction," *IEEE Access*, vol. 7, pp. 85949–85958, 2019.

[40] P. W. Patil and S. Murala, "MSFgNet: A novel compact end-to-end deep network for moving object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4066–4077, Nov. 2019.

[41] T. Akilan, Q. J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3D CNN-LSTM-based image-to-image foreground segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 959–971, Mar. 2020.

[42] P. W. Patil, K. M. Biradar, A. Dudhane, and S. Murala, "An end-to-end edge aggregation network for moving object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8146–8155.

[43] 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.

[44] M. Ellenfeld, S. Moosbauer, R. Cardenes, U. Klauck, and M. Teutsch, "Deep fusion of appearance and frame differencing for motion segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4334–4344.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Sep. 2015, pp. 234–241.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," 2015, *arXiv:1502.01852*.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Comput. Vis. Image Understand.*, vol. 152, pp. 103–117, Nov. 2016.

[51] L. Maddalena and A. Petrosino, "Towards benchmarking scene background initialization," in *Proc. Int. Conf. Image Anal. Process.*, Sep. 2015, pp. 469–476.

[52] A. Sobral and T. Bouwmans, "BGS library: A library framework for algorithm's evaluation in foreground/background segmentation," 2014, doi: 10.1201/b17223-29.

[53] J. Yang, W. Shi, H. Yue, K. Li, J. Ma, and C. Hou, "Spatiotemporally scalable matrix recovery for background modeling and moving object detection," *Signal Process.*, vol. 168, Mar. 2020, Art. no. 107362.

[54] P. Narayanamurthy and N. Vaswani, "Provable dynamic robust PCA or robust subspace tracking," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1547–1577, Mar. 2019.

[55] P. Narayanamurthy and N. Vaswani, "A fast and memory-efficient algorithm for robust PCA (MEROP)," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4684–4688.

[56] 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.

[57] 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.

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

[59] P. W. Patil, A. Dudhane, S. Murala, and A. B. Gonde, "Deep adversarial network for scene independent moving object segmentation," *IEEE Signal Process. Lett.*, vol. 28, pp. 489–493, 2021.

[60] M. Mandal and S. K. Vipparthi, "An empirical review of deep learning frameworks for change detection: Model design, experimental frameworks, challenges and research needs," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6101–6122, Jul. 2022.

[61] X. Zhao, G. Wang, Z. He, and H. Jiang, "A survey of moving object detection methods: A practical perspective," *Neurocomputing*, vol. 503, pp. 28–48, Sep. 2022.

[62] G. Dong, C. Zhao, X. Pan, and A. Basu, "Learning temporal distribution and spatial correlation for universal moving object segmentation," 2023, *arXiv:2304.09949*.

[63] C. Zhao, K. Hu, and A. Basu, "Universal background subtraction based on arithmetic distribution neural network," *IEEE Trans. Image Process.*, vol. 31, pp. 2934–2949, 2022.

[64] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang, and H. Qin, "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3995–4007, 2021.

[65] H. Perreault, G.-A. Bilodeau, N. Saunier, and M. Héritier, "FFAVOD: Feature fusion architecture for video object detection," *Pattern Recognit. Lett.*, vol. 151, pp. 294–301, Nov. 2021.

JAE-YEUL KIM received the B.S. and M.E. degrees in mechanical and automotive engineering from the Graduate School of Automotive Engineering, Seoul National University of Science and Technology, Seoul, South Korea, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in information and communication engineering with the Graduate School, Daegu Gyeongbuk Institute of Science and Technology (DGIST). His current research interests include visual surveillance using deep learning and scene understanding for autonomous navigation.

JONG-EUN HA received the B.S. and M.E. degrees in mechanical engineering from Seoul National University, Seoul, South Korea, in 1992 and 1994, respectively, and the Ph.D. degree in mechanical engineering from KAIST, Daejeon, South Korea, in 2000. From February 2000 to August 2002, he was with Samsung Corning, developing an algorithm for machine vision systems. From 2002 to 2005, he was with the Department of Multimedia Engineering, Tongmyong University. Since 2005, he has been a Professor with the Department of Mechanical and Automotive Engineering, Seoul National University of Science and Technology. His current research interests include deep learning, intelligent robots, and vehicles.