symmetry

Article

Dynamic-Aware Network for Moving Object Detection

Hongrui Zhang [1], Luxia Yang [1,*] and Xiaona Du [2]

[1] School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China; zhanghongrui@tynu.edu.cn
[2] School of Computer, Henan University of Engineering, Zhengzhou 451190, China; dxna1125@126.com
* Correspondence: luxiayang@tynu.edu.cn

Abstract: Moving object detection (MOD) plays an important role in many applications that aim to identify regions of interest in videos. However, most existing MOD methods ignore the variability brought by time-varying information. Additionally, many network frameworks primarily focus on low-level feature learning, neglecting the higher-level contextual understanding required for accurate detection. To solve the above issues, we propose a symmetric Dynamic-Aware Network (DAN) for MOD. DAN explores the interactions between different types of information via structural design and feature optimization. To locate the object position quickly, we build a Siamese convolutional network to emphasize changes in the scene. Subsequently, a Change-Aware Module (CAM) is designed, which can maximize the perception of object change cues by exploiting complementary depth-varying features and different levels of disparity information, thereby enhancing the feature discrimination capability of the network. Moreover, to reinforce the effective transfer between features, we devise a Motion-Attentive Selection Module (MASM) to construct an autonomous decoder for augmenting detail representation. Experimental results on benchmark datasets indicate the rationality and validity of the proposed approach.

Keywords: moving object detection; feature selection; time-varying information; complementary feature

check for updates

Citation: Zhang, H.; Yang, L.; Du, X. Dynamic-Aware Network for Moving Object Detection. *Symmetry* **2024**, *16*, 1620. https://doi.org/10.3390/sym16121620

Academic Editor: Zhixun Su

Received: 20 October 2024
Revised: 25 November 2024
Accepted: 2 December 2024
Published: 6 December 2024

1. Introduction

Moving object detection is a fundamental task in computer vision, which aims to segment foreground pixels from the background. In the past few decades, MOD has received continuous attention and plays an important role in many fields, e.g., target recognition [1–3], autonomous driving [4–6], anomaly detection [7–10], video analysis [11–15], and sports [16,17]. Therefore, high-quality object detection results are crucial in the above applications. However, real-world scenarios present considerable challenges for moving object detection due to the presence of complex and variable environmental factors.

Initially, traditional MOD methods were proposed to address the challenges posed by complex scenarios. Most of them used hand-crafted features to obtain prediction results. Nevertheless, hand-crafted features often lacked the high-level semantic information necessary for accurate target detection. Moreover, many traditional MOD approaches were designed to tackle a single challenge and performed poorly when faced with scenarios involving multiple challenges [18,19]. Recently, the development of deep learning has overcome the limitations of manual features in traditional methods, and the detection performance has been significantly improved over traditional techniques. However, there are still some key issues that need to be addressed in the existing deep learning-based MOD techniques.

(1) **Reasonable utilize spatio-temporal information.** In the design of network structure, some methods [20–22] focus on extracting spatial features and do not fully utilize the continuity of temporal information, which is a relatively stable clue in video analysis. In addition, there are also some methods that combine spatio-temporal information

to obtain moving objects [23,24]. Yet, the method ignores the variability brought by time-varying information, which is an important feature in moving object detection.

(2) **Mining deep features for more meaningful clues.** Deep features contain abundant semantic abstract information, facilitating the acquisition of accurate target details. Many methods, however, directly feed unprocessed deep information into the decoder without fully exploiting the value of deep features. Some other approaches obtain multiscale features by pyramid pooling, but the strategy cannot establish correlations among different types of features [21,25].

(3) **Optimizing the transfer of information between encoder and decoder.** As the network layers become deeper, there is a certain degree of loss in object features. The conventional approach involves passing encoding features to the decoder via a skip connection, but the low-level features contain more coarse information [26]. It is unwise to completely ignore all low-level information that can supply rich spatial structure characteristics to the network. And yet, the direct use of these features introduces interference, which will affect detection accuracy.

Based on the above analysis, we propose a new dynamic-aware network (DAN) to cope with the above issues. It utilizes multi-level change information to explore the internal connections of spatio-temporal features through dynamic perception. Considering the prominence of change information in moving object detection, we design a Siamese convolutional network (SCN) to extract different levels of object change information. To learn more valuable cues from the deep features, we employ features in different states to further exploit the dynamic properties of deep change information. Additionally, to alleviate the degradation of detection accuracy due to the increasing depth of the network, a selection mechanism is designed to reinforce the learning of motion features.

Overall, the contributions of our method can be summarized as follows.

(1) We propose a Dynamic-Aware Network (DAN) that fully utilizes spatio-temporal information and salient target features for moving object detection, which can effectively explore the intrinsic connection between features to obtain accurate predictions.

(2) We design a Change-Aware Module (CAM) using all change information of different layers and high-level salient features, which can fully leverage the value of deep information and maximize the perception of object change information.

(3) We devise a Motion-Attentive Selection Module (MASM) to alleviate the target blur caused by partial loss of detail, which can acquire discriminative features.

2. Related Work

With this subsection, we briefly summarize the research on MOD. For introduction, we classify the previous approaches into traditional methods and deep learning-based methods, as shown below.

2.1. Traditional Methods

In the past decades, scholars have proposed many methods for moving object detection based on traditional machine learning techniques due to their wide application prospects. The core processes mainly involved in traditional methods are background model building, comparison of different video frames, and foreground extraction.

Zhu et al. [27] first differentiated the current frame from the previous frame and next frame, respectively. After that, the obtained difference result undergoes a summation operation. Following this, the difference between the result of the previous difference operation and the subsequent frame is calculated using the dissimilarity operation. At last, the result of the difference operation is compared with the difference image of the previous frame to acquire the final detection target. This method can reduce the interference of clutter and capture the precise target boundary.

Huang et al. [28] investigated a frame difference method based on a self-updating averaged background model. The goal of this method is to identify the moving object by averaging techniques on the background, as well as performing difference and logic operations on the current frame. Also, a neighborhood binary discriminant filtering method is proposed to reduce the effect of isolated noise. To solve the challenge of incomplete objects caused by object overlap in images, Luo et al. [29] proposed a two-layer, three-frame differential method to fill the empty regions. Meanwhile, a statistical analysis algorithm is explored for historical location data to eliminate noise in the event of noise interference during detection.

Sandeep et al. [30] presented a novel approach for detecting moving objects utilizing the concept of block three-frame difference, thereby effectively mitigating camera jitter and object size variability. The distinctive element of this method lies in the selection of the maximum disparity between two difference values, followed by their partitioning into non-overlapping blocks. Subsequently, the average intensity value of each block is computed, enabling the identification of foreground and background pixels based on a predefined threshold and the average intensity value. Building upon their previous research, Sandeep et al. [31] pursued a comprehensive investigation and put forth an advanced moving object detection method that integrates the frame difference technique with the W4 algorithm. This integration serves to partially alleviate the impact of variations in illumination and noise to a certain extent.

To compute the difference image, Oussama et al. [32] subtracted two input correction frames on each pixel position and then employed the OTSU algorithm to refine the foreground. Zeng et al. [33] proposed a general sample-based background differencing method that constructs a background model using both color features and Haar features. In addition, the background model is updated from the spatial and temporal domains using a stochastic strategy.

To enhance the accuracy of MOD at night, Pan et al. [34] first recognized the scene information by extracting the Weber and texture features from the object. Subsequently, they implemented a dedicated light detection module to compensate for the challenges posed by nighttime illumination. Cioppa et al. [35] investigated a background subtraction method combined with asynchronous semantic segmentation, namely Asynchronous Semantic Background Subtraction (ASBS). The ASBS analyzes the temporal changes in pixel characteristics and incorporates the results of semantic-based segmentation to update the background model. To improve the performance of moving object detection in environments with illumination change and noise, Kalli et al. [36] used a fuzzy C-mean algorithm based on a partial illumination field to model the background.

2.2. Deep Learning-Based Methods

The deep learning-based approach improves the feature discrimination capability of the network by acquiring high-level semantic information about objects.

Initially, Braham et al. [37] attempted to build a convolutional neural network (CNN) to implement background subtraction. In [38], a multi-resolution CNN with a cascade architecture is integrated into a semi-automatic moving object detection framework. The framework used images of different resolutions and foreground masks to acquire moving objects. Based on the research of [38], Lim et al. [25] designed a feature pooling module and a dilation convolution unit to obtain multiscale information on moving targets.

Fully Convolutional Network (FCN) is a popular choice in computer vision due to its computational efficiency and compatibility with image inputs of different sizes [39]. Midhula et al. [40] designed a background subtraction method that incorporates WeSamBE and optical flow algorithms for effective background modeling. Further, the method utilizes full-residual connectivity to efficiently fuse fine and coarse features.

Lin et al. [41] first acquired the background image using the SuBSENSE [42] algorithm. Then, the background image is stitched with the current frame, and this result is entered into the designed deep FCN, which can learn the global discrepancy between the background and video frame. Qiu et al. [43] designed a Fully Convolutional Encoder-Decoder Spatial-Temporal Network (FCESNet) for moving object detection. In FCESNet, the spatio-temporal correlation between frames is obtained by the constructed spatio-temporal information transmission module.

In recent years, the effectiveness of attention mechanisms in image-processing tasks has been widely recognized [44–46]. Minematsu et al. [47] incorporated an attention module into the designed moving object detection network to obtain positional cues. Zhang et al. [48] introduced a moving object detection method that utilizes a dual correlation attention director, which designed a dual correlation attention module (DCAM) to fuse features of the same scale.

Numerous studies have demonstrated the effectiveness of 3D convolution in capturing characteristics in both spatial and temporal dimensions in videos. Sakkos et al. [49] employed 3D convolution to simultaneously capture changes in the temporal and spatial aspects of objects.

In [50], background subtraction was implemented using a 3D convolutional neural network (CNN). Specifically, the constructed network has six layers, including alternating 3D convolutional and pooling layers and fully connected layers. Yu et al. [51] designed a 3D-CNN based on spatio-temporal attention for detecting moving objects.

Furthermore, there are many methods that utilize generative adversarial networks to obtain moving targets. Zheng et al. [52] presented a method that combines parallel vision and Bayesian generative adversarial networks (BGANs) for moving object detection. Concretely, the approach involves obtaining the background image through median filtering and performing background subtraction using BGANs. Additionally, parallel vision theory is employed to enhance the accuracy of detection. Bahri et al. [53] designed an online incremental moving object detection model using generative adversarial networks. In this way, the impact of illumination changes and shadows on detection accuracy is alleviated.

To provide a clearer understanding of the details of various methods, we summarize the above approaches, as shown in Table 1.

Table 1. Relevant data for different methods.

Classification	Method	Backbone	Dataset	Running Time	GPU	F1			
						CDnet 2014	Wallflower	USCD	SBI 2015
Traditional methods	Zhu [27]	—	—	5 FPS	—	—	—	—	—
	Huang [28]	—	—	—	—	—	—	—	—
	Luo [29]	—	—	—	—	—	—	—	—
	Sandeep [30]	—	CAVIAR	—	—	—	—	—	—
	Sandeep [31]	—	CAVIAR	—	—	—	—	—	—
	Oussama [32]	—	CDnet2014	3.02 s/frame	—	—	—	—	—
	Zeng [33]	—	CDnet2014	3 FPS	—	0.69	—	—	—
	Pan [34]	—	CDnet2014	—	—	0.70	—	—	—
	Cioppa [35]	—	CDnet2014	—	—	0.75	—	—	—
	Kalli [36]	—	Wallflower	—	—	—	0.78	—	—
Deep learning-based methods	Braham [37]	—	CDnet2014	—	—	0.90	—	—	—
	Wang [38]	—	CDnet2014	—	GTX 970	0.84	—	—	—
	Lim [25]	VGG16	CDnet2014+SBI2015	—	—	0.95	—	—	0.98
	Midhula [40]	—	CDnet2014	—	GTX 970	0.94	—	—	—
	Lin [41]	VGG16	CDnet2014	—	GTX 1080Ti	0.69	—	—	—
	Qiu [43]	ConvLSTM	CDnet2014	112 FPS	Titan X	0.86	—	—	—
	Minematsu [47]	VGG16	CDnet2014	134 FPS	GTX 1080Ti	0.85	—	—	—
	Sakkos [49]	3DCNN	CDnet2014	—	Titan X	0.95	—	—	—
	Gao [50]	3DCNN	CDnet2012	—	—	0.95 (CDnet2012)	—	—	—
	Zheng [52]	GAN	CDnet2014+USCD+SBI2015	23 FPS	GTX 970	0.95	—	0.92	0.92
	Bahri [53]	—	CDnet2014+Wallflower	4.9 FPS	GTX 1080Ti	0.83	0.85	—	—

3. Methodology

This section describes the presented Dynamic-Aware Network (DAN) in detail. Firstly, we provide an overview of the structure of DAN. After that, we give detailed analyses of the designed change-aware module and motion-attentive selection module, respectively.

3.1. Overview

The previous approach fails to properly incorporate spatiotemporal information and neglects the dynamic cues provided by time-varying information. However, in moving object detection, leveraging time-varying information is crucial for accurately locating the target position. Unlike existing methods [20,21,23], we leverage the network design to effectively capture change information and intelligently utilize it to enhance network performance. Figure 1 illustrates the overall pipeline of the DAN. Briefly, a Siamese convolutional network is devised to extract different levels of encoded features and exploit them to obtain information about changes at different scales. It should be noted that the single-branch encoder consists of 5 convolutional blocks with the number of channels 32, 64, 128, 256, and 256, respectively. Further, we design a Change-Aware Module (CAM) for mining semantic information of deep-level features. In the decoding stage, we propose a Motion-Attentive Selection Module (MASM) that utilizes change information, reference frame, and current frame information to autonomously optimize the target features and generate high-quality prediction results. CAM is dedicated to mining the value of depth change information, while MASM optimizes the motion cues at each stage from the features of different states. The joint use of the above two designs can provide complementary and comprehensive object information for the decoder through mutual learning between features and mutual influence.

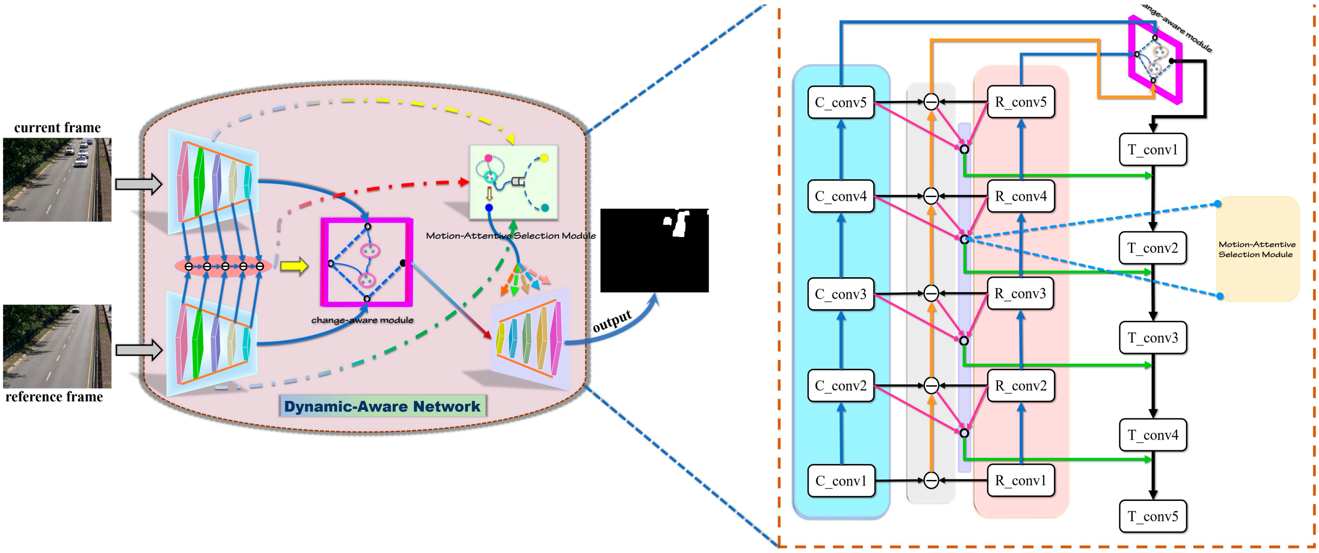

Figure 1. The network architecture of DAN.

3.2. Change-Aware Module

The identification of information changes is crucial for moving object detection as it allows for the quick detection of discrepancies in the scene. This operation plays a pivotal role in the accurate detection of moving objects, which is conducive to improving the efficiency of the detection process. However, many existing methods fail to consider the target cues from time-varying information, resulting in the inability to accurately perceive scene elements. Based on the above analysis, we design a symmetric Siamese Convolutional network (SCN) to acquire change features. First, we use five convolutional layers to extract coding features from both the current frame and reference frame. Next, the change information is captured hierarchically by pixel-wise subtraction, as illustrated in Figure 2.

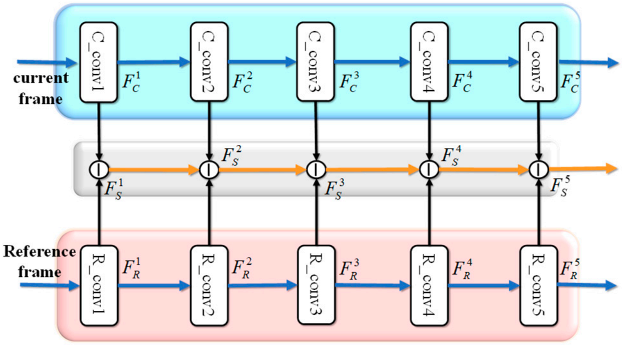

Figure 2. Illustration of the architecture of Siamese convolutional network (SCN).

After obtaining the target change information from various levels through the aforementioned process, we proceed to merge the change information from the five levels. Typically, in the subsequent step, the highest-level encoded features and change information are combined and sent to the decoder. However, deep-level encoded features possess rich semantic cues and strong feature discrimination capabilities. Simply stitching them together with previously acquired change information or adding them up by element may result in a lack of context awareness in the network. Therefore, to effectively exploit the benefits of deep features, we aim to maximize the perception of object change information by extracting multiscale deep change characteristics. Based on this, we designed a module called the Change-Aware Module (CAM), as shown in Figure 3. The following section outlines the detailed implementation steps for this approach.

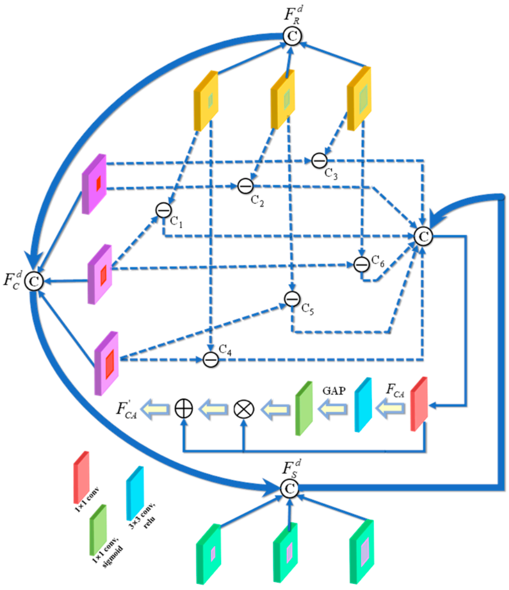

Figure 3. The structure of change-aware module.

First, we utilize SCN to acquire different levels of encoded features F_C^i F_R^i. Also, change information F_S^i is obtained through the above features. F_S^i can be represented by Equation (1).

$$F_S^i = \left| F_C^i - F_R^i \right| \tag{1}$$

where $i \in \{1,2,3,4,5\}$. F_C^i and F_R^i denote information of the current frame and reference frame at different levels, respectively.

Next, the deep features are processed using hierarchical dilation convolution to obtain sufficient target information, which is calculated as follows.

$$d_C^l = D_{r=j}(f^{3\times3}(F_C^5)) \tag{2}$$

$$d_R^l = D_{r=j}(f^{3\times3}(F_R^5)) \tag{3}$$

$$d_S^l = D_{r=j}(f^{3\times3}(F_S')) \tag{4}$$

where $F_S' = \sum_{i=1}^{5} F_S^i$, $D_{r=j}(\cdot)$ is dilated convolution operation, j denotes the dilation rate, and $j \in \{1,2,5\}, l \in \{1,2,3\}$.

Then, pixel-level subtraction is performed on features at different scales to obtain complementary change information (i.e., C_i $i \in \{1,2,3,4,5,6\}$), which can be written as follows.

$$C_1 = \left|d_C^1 - d_R^2\right|, C_2 = \left|d_C^2 - d_R^1\right|, C_3 = \left|d_C^3 - d_R^1\right| \tag{5}$$

$$C_4 = \left|d_C^1 - d_R^3\right|, C_5 = \left|d_C^2 - d_R^3\right|, C_6 = \left|d_C^3 - d_R^2\right| \tag{6}$$

Finally, feature fusion at different levels contributes to improving network performance. Thus, we aggregate complementary information (F_m and F_m') and obtain global information by global average pooling (GAP). The above operations help the network to select useful channel features while reducing redundant connections. Meanwhile, their unique information is retained by element-wise addition. The whole process can be formulated as:

$$F_m = cat(C_1, C_2, C_3, C_4, C_5, C_6) \tag{7}$$

$$F_m' = cat(F_R^d, F_C^d, F_S^d) \tag{8}$$

$$F_{CA} = f^{1\times1}(cat(F_m, F_m')) \tag{9}$$

$$F_{CA}' = F_{CA} \oplus [\delta(f^{1\times1}(GAP(f^{3\times3,relu}(F_{CA})))) \otimes F_{CA}] \tag{10}$$

where $cat(\cdot)$ is a concatenation operation.

Overall, the change-aware module acquires differential features at different levels to help the network quickly locate target locations. Moreover, important target features are emphasized by extracting complementary change information. This method maximizes the acquisition and utilization of differential features in an effective way to improve detection efficiency.

3.3. Motion-Attentive Selection Module

When the network is extended to deeper layers, some details are lost during the process of feature extraction, resulting in issues like indistinct target outlines and incomplete targets in the detection results [54]. In the face of the above problems, a conventional method is to use a skip connection to directly transfer the information from the encoding stage to the decoder. However, this technique will introduce irrelevant information, such as noise. To alleviate this problem, we design a Motion-Attentive Selection Module (MASM) to enhance the effective transfer between features. Figure 4 presents the details of MASM.

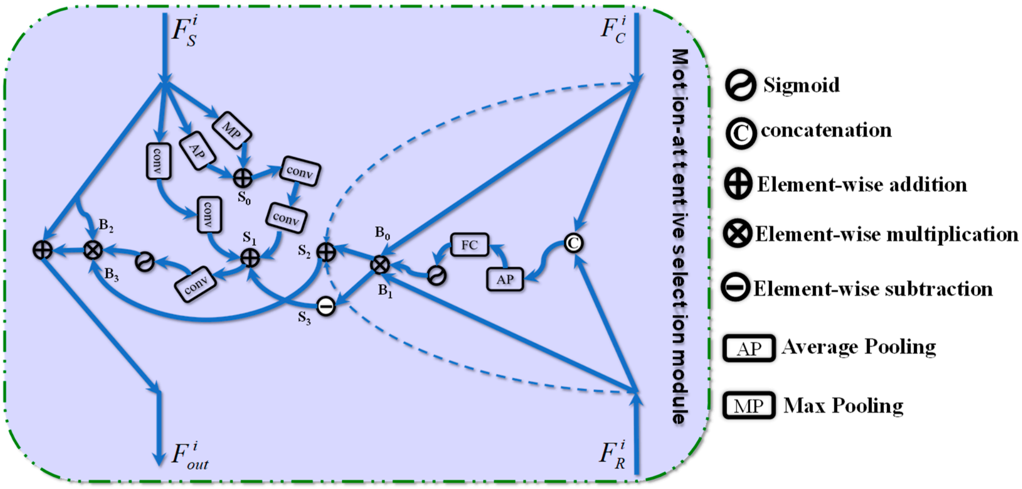

Figure 4. Detailed configuration of motion-attentive selection module.

The motion-attentive selection module selects and enhances motion information automatically using change information, current frame, and reference frame features to provide powerful target features for the decoder. Specifically, we discard the lower-level change information due to its excessive coarse details and high background noise. The proposed MASM consists of 3 parts. First, global and local operations are applied to the change information to aggregate the target cues. In the global channel, the dependencies between channels are integrated using max-pooling (*MP*) and average pooling (*AP*). In the local channel, a 3×3 convolution is employed to capture the local context. The above process can be written as follows.

$$S_0 = MP(F_S^i) \oplus AP(F_S^i) \tag{11}$$

$$S_1 = [f^{1\times1}(f^{1\times1,relu}(S_0))] \oplus [f^{3\times3}(f^{3\times3,relu}(F_S^i))] \tag{12}$$

where $f^{1\times1}(\cdot)$ is 1×1 convolution, $f^{3\times3}(\cdot)$ denotes 3×3 convolution, and *relu* is the activation function.

Meanwhile, we notice that salient object characteristics can be obtained from both the current frame and reference frame while also optimizing motion information. To obtain more accurate cues, we employ the current frame and the reference frame as the input of MASM to learn their correlation. Concretely, we aggregate these two types of features (F_C^i, F_R^i), then acquire a scale factor using average pooling and a fully connected layer to adaptively adjust the fusion information. Last, the reinforced information is employed to obtain the change information S_3 and salient target features S_2 further, which can be formulated as follows.

$$B_0 = F_C^i \otimes \delta(f_c(AP(cat(F_C^i, F_R^i)))) \tag{13}$$

$$B_1 = F_R^i \otimes \delta(f_c(AP(cat(F_C^i, F_R^i)))) \tag{14}$$

$$S_2 = [F_C^i \oplus B_0] \oplus [F_R^i \oplus B_1] \tag{15}$$

$$S_3 = Sub[B_0, B_1] \tag{16}$$

where $f_c(\cdot)$ denotes fully connected layer, $Sub(\cdot)$ represents pixel-wise subtraction, and δ is sigmoid function.

Next, we integrate change information (S_1 and S_3) in different states by element-wise addition. Further, the integrated information is used to refine the motion features (B_2 and B_3). Ultimately, the global and local contextual features are aggregated to obtain fine-grained motion information F_{out}. The whole process is implemented as follows.

$$B_2 = F_S^i \otimes (\delta(f^{3\times3}(S_1 \oplus S_3))) \tag{17}$$

$$B_3 = S_2 \otimes (\delta(f^{3\times3}(S_1 \oplus S_3))) \tag{18}$$

$$F_{out} = F_S^i \oplus B_2 \oplus B_3 \tag{19}$$

The designed CAM and MASM help to improve the accuracy of moving object detection. However, CAM has a slightly simpler fusion mechanism for the same type of features, and the extraction and fusion of key information can be enhanced by the attention mechanism in subsequent research. MASM adopts the same processing strategy for different levels of change information and salient information; it can be considered to construct a sub-strategy processing mechanism in MASM for the characteristics of different levels of features.

4. Experiments

4.1. Datasets and Evaluation Metrics

(1) Datasets: To verify the validity of our devised DAN, we conduct experimental comparisons on three commonly used benchmark datasets, including LASIESTA [55], CDnet2014 [56], and INO [57]. The LASIESTA dataset contains 48 videos acquired from indoor and outdoor scenes with a size of 352×288 pixels. CDnet2014 is a large-scale moving object detection dataset that includes 11 categories of video scenes. The INO dataset contains a wealth of videos of outdoor scenes captured by the VIRxCam platform installed outdoors.

(2) Evaluation metrics: F1 is one of the most commonly used comprehensive evaluation metrics in MOD, which is the reconciled average of precision and recall. Moreover, we used seven other metrics to analyze the performance of different models, including accuracy (Acc), FPR, FNR, Sp, AUC, mIoU, and PWC. Detailed information about the above metrics can be found in [18,56,58].

4.2. Implementation Details

We performed experimental deployments on the TensorFlow framework. The training process is performed on an NVIDIA RTX 3060 GPU. We optimized the proposed DAN using Adam and set the initial learning rate to 0.0001. The loss function adopts binary cross-entropy. Additionally, we set the epoch and batch size to 50 and 2, respectively.

4.3. Ablation Study

We provide a series of ablation analyses on the LASIESTA, CDnet2014, and INO datasets to validate each component in the DAN. Table 2 presents the quantitative comparison of ablation analysis. Moreover, Figure 5 shows the qualitative results of different combinations for a more intuitive comparison. In particular, the red rectangular boxes show where there are large differences in the results obtained from different combinations of modules.

Change-Aware Module (CAM). We employ CAM to quickly capture differences in the scene, which is a crucial step to improve the detection efficiency. As shown in combination ③ in Table 2, when we remove the CAM under the proposed framework, the F1 decreases from 87.9% to 85.69%, which is a 2.21% performance reduction. Figure 5 gives the visualization results of the ablation experiment. As can be seen in column 5 of Figure 5, interference information appears in the detected objects after the removal of CAM. The above results indicate that the design of CAM is an important part of the overall framework.

Table 2. Effectiveness of each module in the designed model. (w/o: without. ETDD: The encoder transmits information directly to the decoder. ↑ means the higher the better, ↓ means the lower the better).

Modules	**Metrics**								
	Acc↑	**Precision↑**	**Recall↑**	**F1↑**	**PWC↓**	**FPR↓**	**FNR↓**	**Sp↑**	**AUC↑**
① Ours	0.9709	0.882	0.8903	0.879	0.8906	0.0061	0.1097	0.9939	0.9842
② w/o MASM	0.9692	0.8702	0.8443	0.846	1.0411	0.0061	0.1557	0.9939	0.9695
③ w/o CAM	0.9678	0.8532	0.8828	0.8569	1.218	0.0095	0.1172	0.9905	0.9813
④ w/o MASM + CAM	0.9696	0.8541	0.8309	0.8263	1.0899	0.0066	0.1691	0.9934	0.9564
⑤ w/o MASM + CAM + ETDD	0.9649	0.7693	0.7791	0.7481	1.6923	0.011	0.2209	0.989	0.9479

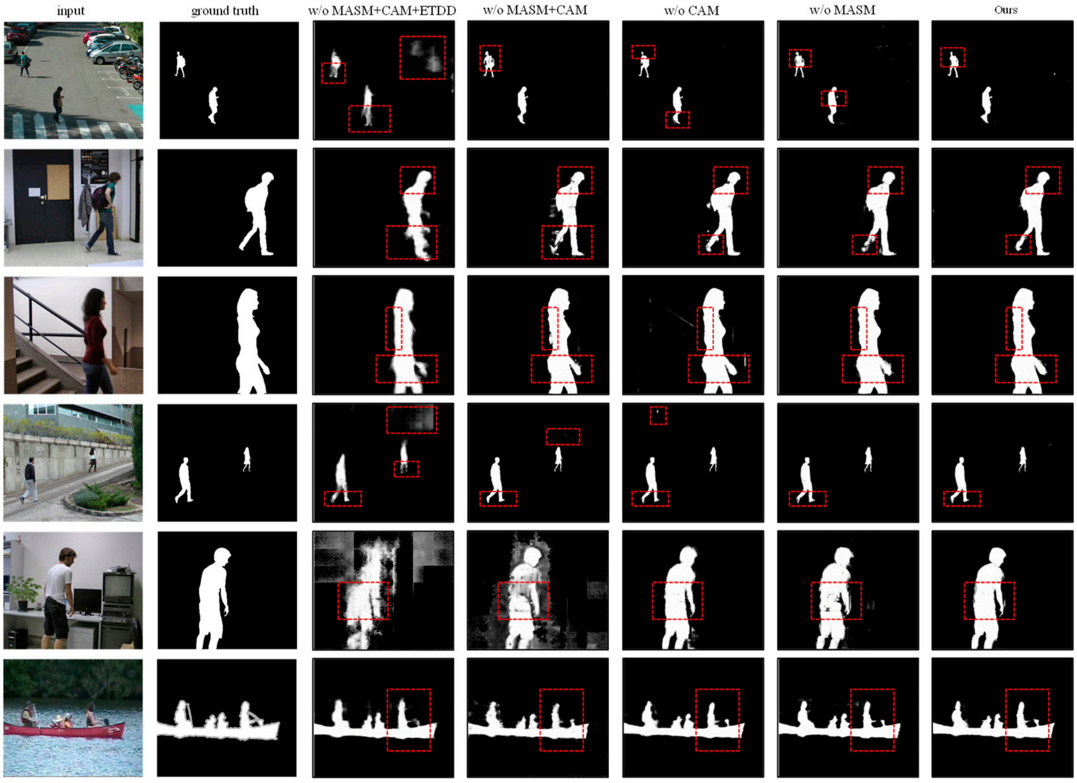

Figure 5. Visual results of ablation analysis on LASIESTA and CDnet2014 datasets.

Motion-Attentive Selection Module (MASM). As described in Section 3.3, we design MASM to enhance the expressiveness of features. In Table 2, combination ② displays the quantitative results obtained by removing MASM from combination ①. As can be seen from the results, the performance of F1 is reduced by 3.3% (from 87.9% to 84.6%). Furthermore, the qualitative results are presented in column 6 of Figure 5. From the figure, it can be noticed that there are voids in the captured moving objects after removing the MASM. Both quantitative and qualitative results reflect the rationality of the proposed MASM.

Effectiveness of our designed structure. From the previous analysis, we verify the efficacy of CAM and MASM, respectively. In this part, we validate whether the combination of these two modules improves the network performance. We also analyze the way information is transmitted between the encoder and decoder. Combination ④ in Table 2 gives the performance after removing MASM and CAM in DAN. Specifically, F1 is 82.63%, compared with the combination ①, ②, and ③, the performance is reduced by 5.27%, 1.97%, and 3.06%, respectively. Based on the combination ④, the combination ⑤ is obtained by removing the way that the encoder transmits information directly to the decoder (ETDD). In combination ⑤, the decrease in F1 is more obvious. Compared with combinations ①, ②, ③, and ④, the performance decreased by 13.09%, 9.79%, 10.88%, and 7.82%, respectively. Besides, the visual results shown in Figure 5 indicate that there are problems, such as incomplete objects and wrong object judgments in the results obtained by combinations ④ and ⑤. The above analysis indicates that our designed structure can effectively improve the accuracy of moving object detection. Also, we test the real-time speed on the employed platform, with the proposed model taking approximately 0.056 s to process one frame.

4.4. Comparisons to the State-of-the-Arts

To further validate the validity of our method, we compare it with state-of-the-art algorithms on LASIESTA, CDnet2014, and INO datasets.

(1) **LASIESTA dataset**: In Table 3, we report the quantitative performance of nine techniques on the LASIESTA dataset. Figure 6 illustrates the performance trends in different approaches on the LASIESTA dataset. It can be seen that our designed network is competitive compared with others. The last row of Table 3 presents the average F1 obtained by the different algorithms, where our method achieves 89%. The performance is improved by 8%, 54%, 49%, 5%, 5%, 3%, 4% and 2% compared to Cuevas [59], FgSegNet-M-55 [25], MSFS-55 [21], Fast-D [60], 3DCD-55 [61], Pardas [62], DFC-D [63], and CUAN [64], respectively. Besides, our method also presents a superior performance on single-type videos.

Table 3. Performance comparison of different approaches in terms of F1 on the LASIESTA dataset. (Bold indicates the best result).

Videos	Methods								
	Cuevas [59]	**FgSegNet-M-55 [25]**	**MSFS-55 [21]**	**Fast-D [60]**	**3DCD-55 [61]**	**Pardas [62]**	**DFC-D [63]**	**CUAN [64]**	**DAN (Ours)**
O_SN	0.78	0.19	0.31	0.88	0.69	0.86	**0.90**	0.89	0.85
O_SU	0.72	0.25	0.37	**0.87**	0.85	**0.87**	0.82	**0.87**	**0.87**
O_RA	0.87	0.18	0.35	**0.94**	0.90	0.90	0.93	0.90	**0.94**
O_CL	0.93	0.22	0.41	**0.94**	0.87	0.80	**0.94**	0.84	0.92
I_SI	0.88	0.43	0.39	**0.93**	0.87	0.88	**0.93**	0.84	0.85
I_OC	0.78	0.31	0.37	0.92	0.91	0.90	0.92	0.83	**0.93**
I_MB	0.94	0.71	0.64	0.94	0.89	0.78	0.94	0.93	**0.95**
I_IL	0.65	0.32	0.35	0.50	**0.92**	0.82	0.51	0.82	0.83
I_CA	0.84	0.69	0.40	0.89	0.82	0.89	**0.94**	0.89	0.93
I_BS	0.66	0.21	0.36	0.62	0.72	0.85	0.63	**0.86**	**0.86**
Average	0.81	0.35	0.40	0.84	0.84	0.86	0.85	0.87	0.89

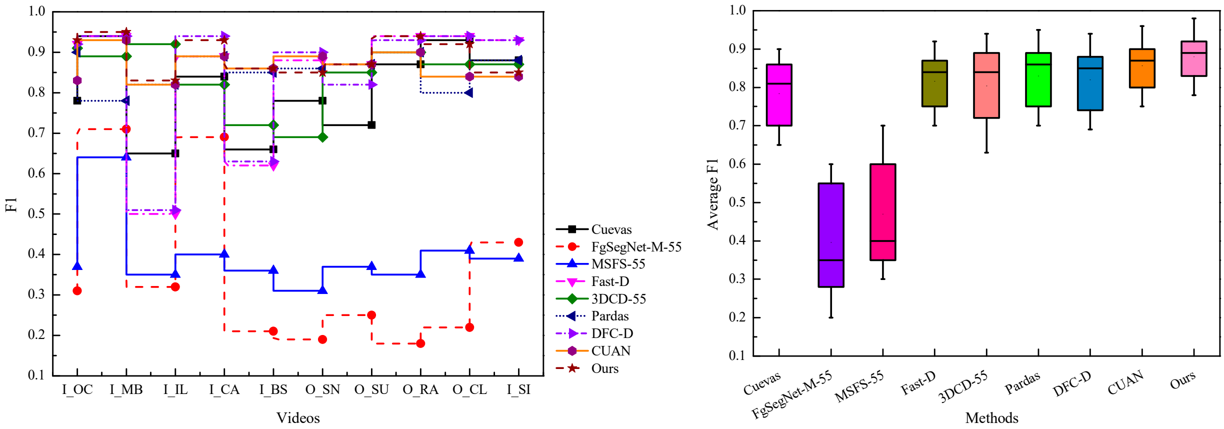

Figure 6. Analysis of the performance of various approaches on the LASIESTA dataset (Metrics are F1 and average F1).

(2) **CDnet2014 dataset**: Table 4 presents the quantitative results of different techni ques [23,24,26,65–69] on the CDnet2014 dataset. Specifically, the proposed DAN achieves 89% on the average F1. Although DAN does not outperform advanced methods in overall performance, our method demonstrates relative stability when facing different types of challenges. For example, in video *turbulence0*, the performance of approaches BMN-BSN [23] and BSUV-Net [26] fluctuates significantly, with F1 of only 2% and 44%. In the low frame rate video *turnpike_0_5fps*, the F1 value obtained by Deepbs [24] is only 49%. In short, the designed network is more suitable for scenes with variability. Furthermore, Figure 7 shows the performance trends in different techniques on the CDnet2014 dataset.

Table 4. Performance comparison of different methods in terms of F1 on the CDnet2014 dataset. (Bold indicates the best result).

Videos	Methods								
	IUTIS-5 [65]	SemanticBGS [66]	Deepbs [24]	BMN-BSN [23]	RT-SBS-V1 [67]	BSUV-Net [26]	GSTO [68]	ISFLN [69]	Ours
highway	0.95	0.96	0.97	0.95	0.95	0.98	0.88	0.93	0.95
office	0.97	0.96	0.98	0.97	0.93	0.97	0.84	0.94	0.93
PETS2006	0.94	0.94	0.94	0.92	0.88	0.95	0.83	0.92	0.92
canoe	0.95	0.95	0.98	0.82	0.94	0.91	0.84	0.91	0.91
turbulence1	0.65	0.30	0.77	0.56	0.14	0.66	0.32	0.85	0.84
sofa	0.79	0.84	0.81	0.91	0.77	0.89	0.73	0.93	0.93
turnpike_0_5fps	0.88	0.88	0.49	0.72	0.90	0.91	0.79	0.81	0.80
peopleInShade	0.91	0.92	0.92	0.89	0.92	0.90	0.97	0.89	0.87
lakeSide	0.60	0.66	0.65	0.51	0.57	0.76	NA	0.84	0.79
cubicle	0.92	0.98	0.94	0.63	0.97	0.92	0.78	0.90	0.90
turbulence0	0.89	0.89	0.80	0.02	0.63	0.44	0.46	0.84	0.84
diningRoom	0.92	0.93	0.90	0.87	0.78	0.91	NA	0.96	0.95
copyMachine	0.93	0.96	0.95	0.96	0.95	0.84	0.84	0.95	0.95
Average	0.87	0.86	0.85	0.75	0.79	0.85	0.75	**0.90**	0.89

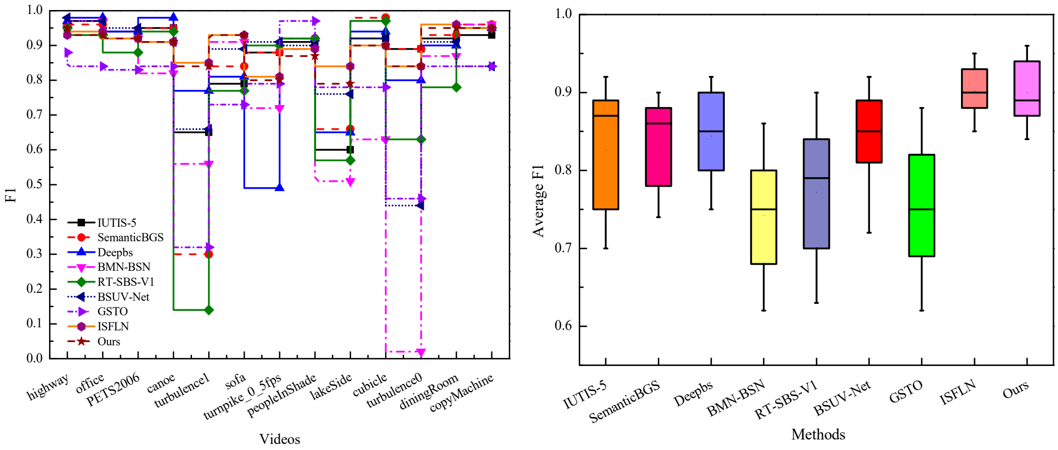

Figure 7. Analysis of the performance of various approaches on the CDnet2014 dataset (Metrics are F1 and average F1).

(3) **INO dataset**: In Table 5, we utilize four metrics to compare the performance of different approaches [20,58,69–73] on the INO dataset. The data presented in the table indicates that our method performs well overall and has advantages in several metrics. In particular, the proposed model obtains 98% on AUC, which improves the performance by 8%, 17%, and 2% compared to the recent advanced techniques SPAMOD [20], Qiu [58], and ISFLN [69], respectively.

Table 5. Performance comparison of different methods on the INO dataset. (Bold indicates the best result. ↑ means the higher the better, ↓ means the lower the better).

Metrics	Methods							
	Li [70]	Akula-CNN [71]	DL [72]	MRF [73]	SPAMOD [20]	Qiu [58]	ISFLN [69]	Ours
Acc↑	0.75	0.79	0.80	0.81	0.98	0.83	**0.98**	**0.98**
recall↑	0.70	0.73	0.75	**0.79**	0.62	0.80	0.77	0.78
Sp↑	0.28	0.26	0.20	0.19	0.90	0.16	**0.99**	**0.99**
AUC↑	0.70	0.73	0.74	0.78	0.90	0.81	0.96	**0.98**

(4) **Visual analysis**: Figures 8 and 9 illustrate the qualitative comparison of different methods and our approach [23,24,26,65,67,69]. These examples involve many challenging and complex scenarios, such as shadows, lighting variations, small-sized objects, atmospheric turbulence, and background disturbances. Clearly, the proposed network is able to correctly localize the object position and acquire moving objects with clear contours. The qualitative results highlight the effectiveness of our method in suppressing background interference and accurately distinguishing the object area. Moreover, the designed DAN exhibits the capability to detect objects at different scales.

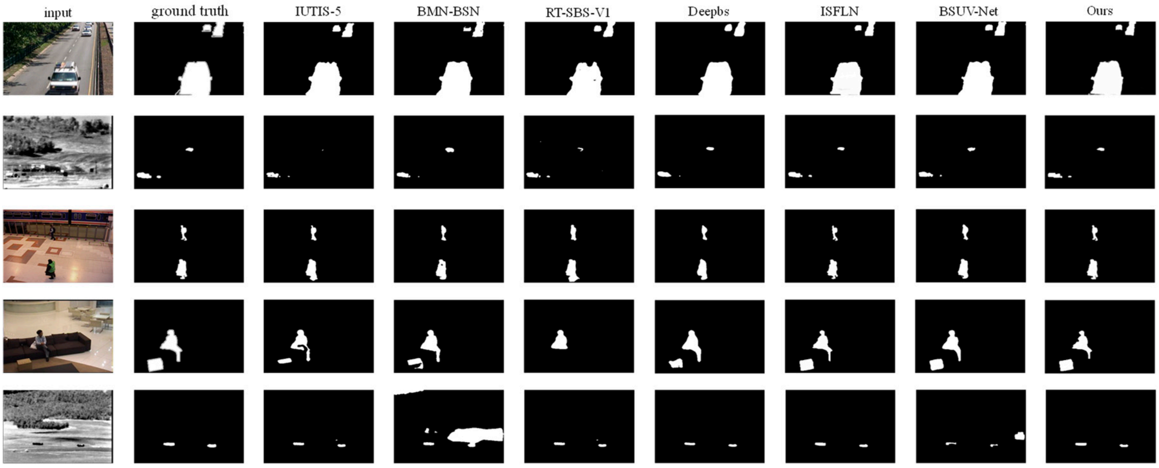

Figure 8. Visual results on the CDnet2014 dataset.

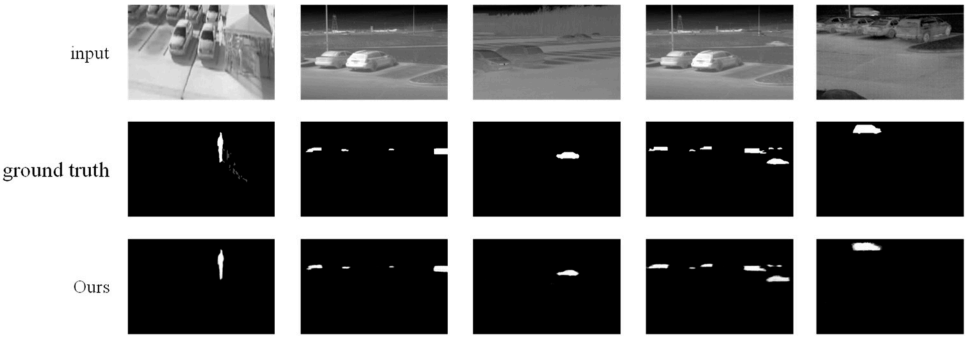

Figure 9. Visual results on the INO dataset.

(5) **Complexity analysis:** The main constraints for model application are the number of FLOPs and parameters. Table 6 illustrates a comparison of the model complexity of some advanced techniques [21,24,25,61,69,74,75]. Notably, the number of parameters and FLOPs of our model are 4.64 M and 6.87 G, respectively. Collectively, the presented model exhibits impressive performance compared to other approaches.

Table 6. Comparison of Parameters and FLOPs for different approaches.

Methods	DeepBS [24]	FgSegNet-M-55 [25]	MSFS [21]	3DCD [61]	ISFLN [69]	BSUV-Net 2.0 [74]	MAAN [75]	Ours
#Params	3.15 M	15.83 M	7.49 M	0.13 M	5.27 M	15.9 M	2.97 M	4.64 M
FLOPs	1750 G	220 G	181 G	NA	19.49 G	540 G	12.3 G	6.87 G

4.5. Limitations and Future Work

The designed dynamic-aware network performs well in most situations, but when the scene changes considerably, the object capture ability decreases obviously. The main reasons for the above problems can be summarized as the following two points: (i) the amount of information provided on the input side is not sufficient; (ii) the types of cues that can be captured in the network are single. To address these issues, the following two aspects will be investigated in the next work: (i) provide additional reference frame information for the network, for example, adding the averaging result of the frame before the current frame to the input; (ii) construct an edge information supervision mechanism to guide the network to extract more complete object features.

5. Conclusions

In this paper, we propose a moving object detection model named Dynamic-Aware Network (DAN). Our core idea is to fully utilize time-varying information and complementary features to enhance the model's reasoning ability. To this end, we first build a Siamese convolutional network to extract time-varying information. Then, we design CAM to learn the intrinsic connection between depth-varying features and time-varying information, which enhances the context-awareness of the proposed model. Further, we construct the MASM to guide the transfer of high-quality information between the encoder and decoder. The whole design concept improves the feature representation of the model and reduces the interference of background information. Experimental results on three datasets exhibit the capability of the proposed approach to achieve competitive performance. In the future, we will adequately exploit the target position relationship for moving object detection.

Author Contributions: Conceptualization, H.Z. and L.Y.; methodology, H.Z. and L.Y.; software, H.Z.; validation, H.Z. and X.D.; formal analysis, H.Z. and X.D.; writing—original draft preparation, H.Z.; writing—review and editing, H.Z. and X.D.; supervision, L.Y.; project administration, L.Y.; funding acquisition, H.Z. and L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research and Development Program of ShanXi Province (No. 202102010101008) and the Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No. 2024L295).

Data Availability Statement: The original contribution of this research is included in the paper. For further inquiries, please contact the first author.

Conflicts of Interest: The author declares no conflicts of interest.

Nomenclature

$D_{r=j}$	dilated convolution operation	δ	sigmoid function
cat	concatenation operation	Sub	pixel-wise subtraction
GAP	global average pooling	$f^{1\times1}(\cdot)$	1×1 convolution
$f^{3\times3}(\cdot)$	3×3 convolution	AP	average pooling
MP	max-pooling	$\otimes$	element-wise multiplication
$\oplus$	element-wise addition	f_c	fully connected layer

References

1. Wang, Y.; Zhang, W.; Lai, C.; Wang, J. Adaptive temporal feature modeling for visual tracking via cross-channel learning. *Knowl. Based Syst.* **2023**, *265*, 110380. [CrossRef]
2. Gong, F.; Gao, Y.; Yuan, X.; Liu, X.; Li, Y.; Ji, X. Crude Oil Leakage Detection Based on DA-SR Framework. *Adv. Theory Simul.* **2022**, *5*, 2200273. [CrossRef]
3. Latif, G.; Alghmgham, D.A.; Maheswar, R.; Alghazo, J.; Sibai, F.; Aly, M.H. Deep learning in Transportation: Optimized driven deep residual networks for Arabic traffic sign recognition. *Alex. Eng. J.* **2023**, *80*, 134–143. [CrossRef]
4. Jegham, I.; Alouani, I.; Ben Khalifa, A.; Mahjoub, M.A. Deep learning-based hard spatial attention for driver in-vehicle action monitoring. *Expert Syst. Appl.* **2023**, *219*, 119629. [CrossRef]
5. Hussain, M.I.; Rafique, M.A.; Kim, J.; Jeon, M.; Pedrycz, W. Artificial Proprioceptive Reflex Warning Using EMG in Advanced Driving Assistance System. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 1635–1644. [CrossRef]
6. Munir, F.; Azam, S.; Rafique, M.A.; Sheri, A.M.; Jeon, M.; Pedrycz, W. Exploring thermal images for object detection in underexposure regions for autonomous driving. *Appl. Soft Comput.* **2022**, *121*, 108793. [CrossRef]
7. Sofuoglu, S.E.; Aviyente, S. GLOSS: Tensor-based anomaly detection in spatiotemporal urban traffic data. *Signal Process.* **2022**, *192*, 108370. [CrossRef]
8. Zhang, L.; Xie, X.; Xiao, K.; Bai, W.; Liu, K.; Dong, P. MANomaly: Mutual adversarial networks for semi-supervised anomaly detection. *Inf. Sci.* **2022**, *611*, 65–80. [CrossRef]
9. López-Rubio, E.; Molina-Cabello, M.A.; Castro, F.M.; Luque-Baena, R.M.; Marín-Jiménez, M.J.; Guil, N. Anomalous object detection by active search with PTZ cameras. *Expert Syst. Appl.* **2021**, *181*, 115150. [CrossRef]
10. Herrmann, M.; Pfisterer, F.; Scheipl, F. A geometric framework for outlier detection in high-dimensional data. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1491. [CrossRef]
11. Shao, M.; Sun, Y.; Liu, Z.; Peng, Z.; Li, S.; Li, C. GPNet: Key Point Generation Auxiliary Network for Object Detection. *Adv. Theory Simul.* **2023**, *6*, 2200894. [CrossRef]
12. Kourbane, I.; Genc, Y. A graph-based approach for absolute 3D hand pose estimation using a single RGB image. *Appl. Intell.* **2022**, *52*, 16667–16682. [CrossRef]
13. Wu, T.; Peng, J.; Zhang, W.; Zhang, H.; Tan, S.; Yi, F.; Ma, C.; Huang, Y. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowl. Based Syst.* **2022**, *235*, 107676. [CrossRef]
14. Yu, J.-M.; Ham, G.; Lee, C.; Lee, J.-H.; Han, J.-K.; Kim, J.-K.; Jang, D.; Kim, N.; Kim, M.-S.; Im, S.G.; et al. A Multiple-State Ion Synaptic Transistor Applicable to Abnormal Car Detection with Transfer Learning. *Adv. Intell. Syst.* **2022**, *4*, 2100231. [CrossRef]
15. Wang, T.; Hou, B.; Li, J.; Shi, P.; Zhang, B.; Snoussi, H. TASTA: Text-Assisted Spatial and Temporal Attention Network for Video Question Answering. *Adv. Intell. Syst.* **2023**, *5*, 2200131. [CrossRef]
16. Goh, G.L.; Goh, G.D.; Pan, J.W.; Teng, P.S.P.; Kong, P.W. Automated Service Height Fault Detection Using Computer Vision and Machine Learning for Badminton Matches. *Sensors* **2023**, *23*, 9759. [CrossRef]
17. Naik, B.T.; Hashmi, M.F. YOLOv3-SORT: Detection and tracking player/ball in soccer sport. *J. Electron. Imaging* **2023**, *32*, 011003. [CrossRef]
18. Li, S.; Han, P.; Bu, S.; Tong, P.; Li, Q.; Li, K.; Wan, G. Change detection in images using shape-aware siamese convolutional network. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103819. [CrossRef]
19. Zhang, H.; Qu, S.; Li, H. Dual-Branch Enhanced Network for Change Detection. *Arab. J. Sci. Eng.* **2022**, *47*, 3459–3471. [CrossRef]
20. Qu, S.; Zhang, H.; Wu, W.; Xu, W.; Li, Y. Symmetric pyramid attention convolutional neural network for moving object detection. *Signal Image Video Process.* **2021**, *15*, 1747–1755. [CrossRef]
21. Lim, L.A.; Keles, H.Y. Learning multi-scale features for foreground segmentation. *Pattern Anal. Appl.* **2020**, *23*, 1369–1380. [CrossRef]
22. Yang, L.; Li, J.; Luo, Y.; Zhao, Y.; Cheng, H.; Li, J. Deep Background Modeling Using Fully Convolutional Network. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 254–262. [CrossRef]
23. Mondéjar-Guerra, V.; Rouco, J.; Novo, J.; Ortega, M. An end-to-end deep learning approach for simultaneous background modeling and subtraction. In Proceedings of the 30th British Machine Vision Conference, Cardiff, UK, 9–12 September 2019; pp. 1–12.
24. Babaee, M.; Dinh, D.T.; Rigoll, G. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognit.* **2018**, *76*, 635–649. [CrossRef]
25. Lim, L.A.; Yalim Keles, H. Foreground segmentation using convolutional neural networks for multiscale feature encoding. *Pattern Recognit. Lett.* **2018**, *112*, 256–262. [CrossRef]
26. Tezcan, M.O.; Ishwar, P.; Konrad, J. BSUV-Net: A Fully-Convolutional Neural Network for Background Subtraction of Unseen Videos. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 2763–2772.
27. Zhu, M.; Wang, H. Fast detection of moving object based on improved frame-difference method. In Proceedings of the 6th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 21–22 October 2017; pp. 299–303.
28. Kang, Y.; Huang, W.; Zheng, S. An improved frame difference method for moving target detection. In Proceedings of the Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 1537–1541.

29. Luo, X.; Jia, K.; Liu, P. Improved Three-Frame-Difference Algorithm for Infrared Moving Target. In Proceedings of the 5th International Conference on Image, Vision and Computing (ICIVC), Beijing, China, 10–12 July 2020; pp. 108–112.
30. Sengar, S.S.; Mukhopadhyay, S. A novel method for moving object detection based on block based frame differencing. In Proceedings of the 3rd International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, India, 3–5 March 2016; pp. 467–472.
31. Sengar, S.S.; Mukhopadhyay, S. Moving object detection based on frame difference and W4. *Signal Image Video Process.* **2017**, *11*, 1357–1364. [CrossRef]
32. Boufares, O.; Boussif, M.; Aloui, N. Moving Object Detection System Based on the Modified Temporal Difference and OTSU algorithm. In Proceedings of the 18th International Multi-Conference on Systems, Signals & Devices (SSD), Monastir, Tunisia, 22–25 March 2021; pp. 1378–1382.
33. Zeng, W.; Xie, C.; Yang, Z.; Lu, X. A universal sample-based background subtraction method for traffic surveillance videos. *Multimed. Tools Appl.* **2020**, *79*, 22211–22234. [CrossRef]
34. Pan, H.; Zhu, G.; Peng, C.; Xiao, Q. Background subtraction for night videos. *PeerJ Comput. Sci.* **2021**, *7*, e592. [CrossRef]
35. Cioppa, A.; Braham, M.; Van Droogenbroeck, M. Asynchronous Semantic Background Subtraction. *J. Imaging* **2020**, *6*, 50. [CrossRef]
36. Kalli, S.; Suresh, T.; Prasanth, A.; Muthumanickam, T.; Mohanram, K. An effective motion object detection using adaptive background modeling mechanism in video surveillance system. *J. Intell. Fuzzy Syst.* **2021**, *41*, 1777–1789. [CrossRef]
37. Braham, M.; Droogenbroeck, M.V. Deep background subtraction with scene-specific convolutional neural networks. In Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP), Bratislava, Slovakia, 23–25 May 2016; pp. 1–4.
38. Wang, Y.; Luo, Z.; Jodoin, P.-M. Interactive deep learning method for segmenting moving objects. *Pattern Recognit. Lett.* **2017**, *96*, 66–75. [CrossRef]
39. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
40. Vijayan, M.; Raguraman, P.; Mohan, R. A Fully Residual Convolutional Neural Network for Background Subtraction. *Pattern Recognit. Lett.* **2021**, *146*, 63–69. [CrossRef]
41. Lin, C.; Yan, B.; Tan, W. Foreground Detection in Surveillance Video with Fully Convolutional Semantic Network. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4118–4122.
42. St-Charles, P.-L.; Bilodeau, G.-A.; Bergevin, R. SuBSENSE: A Universal Change Detection Method with Local Adaptive Sensitivity. *IEEE Trans. Image Process.* **2015**, *24*, 359–373. [CrossRef]
43. Qiu, M.; Li, X. A Fully Convolutional Encoder–Decoder Spatial–Temporal Network for Real-Time Background Subtraction. *IEEE Access* **2019**, *7*, 85949–85958. [CrossRef]
44. Li, Y.; Zhang, Y.; Liu, J.Y.; Wang, K.; Zhang, K.; Zhang, G.S.; Liao, X.F.; Yang, G. Global Transformer and Dual Local Attention Network via Deep-Shallow Hierarchical Feature Fusion for Retinal Vessel Segmentation. *IEEE Trans. Cybern.* **2023**, *53*, 5826–5839. [CrossRef] [PubMed]
45. Chen, S.-B.; Ji, Y.-X.; Tang, J.; Luo, B.; Wang, W.-Q.; Lv, K. DBRANet: Road Extraction by Dual-Branch Encoder and Regional Attention Decoder. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
46. Gaudio, A.; Smailagic, A.; Faloutsos, C.; Mohan, S.; Johnson, E.; Liu, Y.; Costa, P.; Campilho, A. DeepFixCX: Explainable privacy-preserving image compression for medical image analysis. *WIREs Data Min. Knowl. Discov.* **2023**, *13*, e1495. [CrossRef]
47. Minematsu, T.; Shimada, A.; Taniguchi, R.-i. Simple background subtraction constraint for weakly supervised background subtraction network. In Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
48. Zhang, L.; Hu, X.; Zhang, M.; Shu, Z.; Zhou, H. Object-level change detection with a dual correlation attention-guided detector. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 147–160. [CrossRef]
49. Sakkos, D.; Liu, H.; Han, J.; Shao, L. End-to-end video background subtraction with 3d convolutional neural networks. *Multimed. Tools Appl.* **2018**, *77*, 23023–23041. [CrossRef]
50. Gao, Y.; Cai, H.; Zhang, X.; Lan, L.; Luo, Z. Background Subtraction via 3D Convolutional Neural Networks. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1271–1276.
51. Yu, R.; Wang, H.; Davis, L.S. ReMotENet: Efficient Relevant Motion Event Detection for Large-Scale Home Surveillance Videos. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1642–1651.
52. Zheng, W.; Wang, K.; Wang, F.-Y. A novel background subtraction algorithm based on parallel vision and Bayesian GANs. *Neurocomputing* **2020**, *394*, 178–200. [CrossRef]
53. Bahri, F.; Shakeri, M.; Ray, N. Online Illumination Invariant Moving Object Detection by Generative Neural Network. In Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing, Hyderabad, India, 18–22 December 2018; pp. 1–8.
54. Dosovitskiy, A.; Brox, T. Inverting Visual Representations with Convolutional Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4829–4837.

55. Carlos, C.; Maria Yanez, E.; Narciso, G. Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA. *Comput. Vis. Image Underst.* **2016**, *152*, 103–117.
56. Wang, Y.; Jodoin, P.; Porikli, F.; Konrad, J.; Benezeth, Y.; Ishwar, P. CDnet 2014: An Expanded Change Detection Benchmark Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 393–400.
57. Video Analytics Dataset [DS]. Available online: http://www.ino.ca/en/video-analytics-dataset/ (accessed on 1 March 2022).
58. Qiu, S.; Luo, J.; Yang, S.; Zhang, M.; Zhang, W. A moving target extraction algorithm based on the fusion of infrared and visible images. *Infrared Phys. Technol.* **2019**, *98*, 285–291. [CrossRef]
59. Berjón, D.; Cuevas, C.; Morán, F.; García, N. Real-time nonparametric background subtraction with tracking-based foreground update. *Pattern Recognit.* **2018**, *74*, 156–170. [CrossRef]
60. Hossain, M.A.; Hossain, M.I.; Hossain, M.D.; Thu, N.T.; Huh, E.-N. Fast-D: When Non-Smoothing Color Feature Meets Moving Object Detection in Real-Time. *IEEE Access* **2020**, *8*, 186756–186772. [CrossRef]
61. Mandal, M.; Dhar, V.; Mishra, A.; Vipparthi, S.K.; Abdel-Mottaleb, M. 3DCD: Scene Independent End-to-End Spatiotemporal Feature Learning Framework for Change Detection in Unseen Videos. *IEEE Trans. Image Process.* **2021**, *30*, 546–558. [CrossRef] [PubMed]
62. Pardàs, M.; Canet, G. Refinement Network for unsupervised on the scene Foreground Segmentation. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–21 January 2021; pp. 705–709.
63. Hossain, M.A.; Hossain, M.I.; Hossain, M.D.; Huh, E.-N. DFC-D: A dynamic weight-based multiple features combination for real-time moving object detection. *Multimed. Tools Appl.* **2022**, *81*, 32549–32580. [CrossRef]
64. Canet Tarrés, G.; Pardàs, M. Context-Unsupervised Adversarial Network for Video Sensors. *Sensors* **2022**, *22*, 3171. [CrossRef] [PubMed]
65. Bianco, S.; Ciocca, G.; Schettini, R. Combination of Video Change Detection Algorithms by Genetic Programming. *IEEE Trans. Evol. Comput.* **2017**, *21*, 914–928. [CrossRef]
66. Braham, M.; Piérard, S.; Droogenbroeck, M.V. Semantic background subtraction. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 4552–4556.
67. Anthony, C.; Marc Van, D.; Braham, M. Real-Time Semantic Background Subtraction. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 3214–3218.
68. Li, L.; Wang, Z.; Hu, Q.; Dong, Y. Adaptive Nonconvex Sparsity Based Background Subtraction for Intelligent Video Surveillance. *IEEE Trans. Ind. Inform.* **2021**, *17*, 4168–4178. [CrossRef]
69. Zhang, H.; Li, H. Interactive spatio-temporal feature learning network for video foreground detection. *Complex Intell. Syst.* **2022**, *8*, 4251–4263. [CrossRef]
70. Li, Z.; Hou, Q.; Fu, H.; Dai, Z.; Yang, L.; Jin, G.; Li, R. Infrared small moving target detection algorithm based on joint spatio-temporal sparse recovery. *Infrared Phys. Technol.* **2015**, *69*, 44–52. [CrossRef]
71. Akula, A.; Singh, A.; Ghosh, R.; Kumar, S.; Sardana, H.K. Target Recognition in Infrared Imagery Using Convolutional Neural Network. In *Proceedings of International Conference on Computer Vision and Image Processing*; Springer: Singapore, 2016; pp. 25–34.
72. Bhattacharjee, S.D.; Talukder, A.; Alam, M.S. Graph clustering for weapon discharge event detection and tracking in infrared imagery using deep features. In Proceedings of the Conference on Pattern Recognition and Tracking XXVII, Anaheim, CA, USA, 1 May 2017; p. 102030O.
73. Sun, B.; Li, Y.; Guosheng, G. Moving target segmentation using Markov random field-based evaluation metric in infrared videos. *Opt. Eng.* **2018**, *1*, 013106. [CrossRef]
74. Ozan, T.M.; Prakash, I.; Konrad, J.; And Janusz Konrad, F.I. BSUV-Net 2.0: Spatio-Temporal Data Augmentations for Video-Agnostic Supervised Background Subtraction. *IEEE Access* **2021**, *9*, 53849–53860. [CrossRef]
75. Zhang, H.; Qu, S.; Li, H.; Xu, W.; Du, X. A motion-appearance-aware network for object change detection. *Knowl.-Based Syst.* **2022**, *255*, 109612. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.