Complex & Intelligent Systems (2022) 8:4251–4263
https://doi.org/10.1007/s40747-022-00712-x

ORIGINAL ARTICLE

Interactive spatio-temporal feature learning network for video foreground detection

Hongrui Zhang[1] · Huan Li[2]

Received: 24 August 2021 / Accepted: 5 March 2022 / Published online: 25 March 2022

Abstract

Video foreground detection (VFD), as one of the basic pre-processing tasks, is very essential for subsequent target tracking and recognition. However, due to the interference of shadow, dynamic background, and camera jitter, constructing a suitable detection network is still challenging. Recently, convolution neural networks have proved its reliability in many fields with their powerful feature extraction ability. Therefore, an interactive spatio-temporal feature learning network (ISFLN) for VFD is proposed in this paper. First, we obtain the deep and shallow spatio-temporal information of two paths with multi-level and multi-scale. The deep feature is conducive to enhancing feature identification capabilities, while the shallow feature is dedicated to fine boundary segmentation. Specifically, an interactive multi-scale feature extraction module (IMFEM) is designed to facilitate the information transmission between different types of features. Then, a multi-level feature enhancement module (MFEM), which provides precise object knowledge for decoder, is proposed to guide the coding information of each layer by the fusion spatio-temporal difference characteristic. Experimental results on LASIESTA, CDnet2014, INO, and AICD datasets demonstrate that the proposed ISFLN is more effective than the existing advanced methods.

Keywords Video foreground detection · Feature enhancement · Interactive multi-scale feature · Deep learning

Introduction

Video foreground detection (VFD), which aims to identify the changing targets in a video sequence, has become a popular research topic in computer vision. Many applications adopt this technique, including autonomous driving [1], remote sensing [2], action recognition [3–5], and video surveillance [6, 7]. As an important pre-processing component, its detection accuracy directly impacts the quality of subsequent work. However, illumination change, dynamic background, shadows, and camera jitter make the process challenging.

Over the past few decades, a wide variety of techniques have been proposed for VFD [8–11]. In general, existing methods can be approximately divided into two broad classes, traditional machine learning-based and deep learning-based approaches. As the classical pixel-based techniques in conventional methods (e.g., GMM [12], KDE [13]), the detection accuracy suffers from a great negative impact in the face of illumination changes and camera movement. Furthermore, region-based approaches [14, 15] lack motion information, and multiple block-level computations increase the complexity of the algorithm. Overall, traditional methods rely on low-level manual characteristics such as color features, texture features, and spatial distribution. All of these features lack high-level semantic information, which will lead to serious target missing and detection errors, as well as a weak response to complicated environments. In recent years, convolutional neural networks (CNNs) have significantly improved the quality of many image processing tasks by virtue of their powerful feature extraction capabilities. Using this technique, high-level semantic cues can be gleaned that might not be obtainable using traditional methods. Although numerous deep learning-based approaches have shown promising results in VFD, there are still some issues as follows.

Firstly, several existing methods [11, 16–18] only performed analysis spatial clues without considering temporal characteristics, resulting in the isolation of information. Secondly, different types of features have variations between

✉ Huan Li
lihuan2216@163.com

[1] Department of Electronics and Information Engineering, Central China Normal University, Wuhan 430079, China

[2] Department of Special Technology, Special Operations College of the PLA, Guilin 541000, China

different levels. In video detection, spatio-temporal difference can provide more accurate target information, but many approaches [19, 20] mixed them together for training. Additionally, some scholars directly employed skip connections in the encoder-decoder structure to enhance feature expression [20, 21], however, this will result in noise and unnecessary information flow to the decoder and affect the performance.

Based on the limitations of existing methods discussed above, the motivation of our approach is to construct a model that can make full use of spatio-temporal characteristics in the coding phase. Moreover, valuable target cues are also crucial for the decoder. To realize these objectives, we propose an interactive spatio-temporal feature learning network for video foreground detection. Our thought is to mine multi-level and multi-scale spatio-temporal features and to encourage different types of knowledge to communicate with each other. For this purpose, we design a two-path spatio-temporal information extraction module (TSIEM) to obtain rich spatio-temporal features while strengthening the intrinsic connection between features. Besides, a vital challenge is how to cope with the nuisance caused by the loss of some details after information pass through deeper layers in an encoder-decoder network. Our solution to this concern is to propose, rather than having a simple skip connection between encoder and decoder, a multi-level feature enhancement module (MFEM) that can share powerful target information with the decoder.

In brief, the contributions of this paper are summarized as follows.

1. We propose a novel end-to-end interactive spatio-temporal feature learning network for video foreground detection. Compared with the existing advanced methods, our model is fast in speed (24 fps) while having a higher detection accuracy.
2. We design two-path spatio-temporal information extraction module (TSIEM) to obtain multi-level and multi-scale spatio-temporal difference information. In particular, the proposed IMFEM promotes the learning among low-level, intermediate-level and high-level features.
3. We construct multi-level feature enhancement module (MFEM) to deliver fine coding features to the decoder, which can provide an effective way to solve the problem of blurred boundaries and ambiguous pixels caused by rough features.

Related work

As a hot topic in the field of artificial intelligence, various techniques for video foreground detection are constantly being proposed. We organize and analyze these approaches from the perspective of traditional methods and deep learning methods.

Traditional method

Initially, the popular traditional method was kicked off by Gaussian mixture model (GMM) [12], which is a background representation model based on the statistical information of pixel samples. Specifically, a background model is gained in GMM by counting the pixel values of each point in a video image, followed by a process of background subtraction that extracts the moving object. Nevertheless, this method will cause misdetection because of the following factors: (i) The scene changes substantially, such as sudden changes in light or camera jitter; (ii) The colors of the foreground and background are similar. Subsequently, Barnich et al. [22] proposed a non-parametric method called Vibe. Unlike GMM, Vibe adopts a random background update strategy. Due to the pixel changes are uncertain, it is difficult to use a fixed model to describe them. Hence, Vibe algorithm assumed that a random model is to some extent more suitable for simulating the uncertainty of pixel change when the model of pixel change cannot be determined. Additionally, the main disadvantage of this method is that noise and static targets are blended into the background, which brings interference to the foreground detection. To deal with dynamic background problem, Zhao et al. [23] first applied an adaptive threshold segmentation approach to segment the input frame into multiple binary images. Second, the foreground detection was performed by lateral suppression and an improved template matching method. Sajid et al. [24] proposed multimode background subtraction (MBS) to overcome multiple challenges. Here, binary masks of RGB and YCbCr color spaces were created by denoising the merged image pixels, thus separating foreground pixels from background. Roy et al. [25] constructed 3 pixel-based background models to deal with complex and changing real-world scenarios. Tom et al. [26] employed the spatio-temporal dependency between background and foreground to build a video foreground detection algorithm in a tensor framework.

Deep learning method

With the continuous development of deep learning, scholars have also introduced this technology to VFD [27–31]. Akula et al. [32] employed LeNet-5 structure for infrared target recognition. Patil et al. [33] first employed the temporal histogram to estimate the background, and then sent two saliency maps with different resolutions to the CNN to obtain segmentation results. On the basis of fully convolutional network (FCN), Yang et al. [34] introduced dilated convolution to expand the receptive field. Furthermore, to prevent long-term stationary objects from blending into the background, a strategy of increasing the interval of multi-frame video sequence images is proposed. Guerra et al. [35] utilized a U-Net-based background subtraction method to extract the target after acquiring the background using a set of video frames.

Recently, attention mechanisms have been proven effective in image processing [36–38]. Using this mechanism not only highlights important knowledge, but also establishes contextual relevance. To obtain location information, Minematsu et al. [39] added an attention module to the proposed weakly supervised frame-level labeling network. Chen et al. [19] introduced attention mechanism and residual block into ConvLSTM to extract temporal context cues. Additionally, the STN model and CRF layer are added to the end of network for feature refinement. After that, Qu et al. [11] designed a symmetrical pyramid attention model in CNN to get close contextual connections.

Moreover, there are many other types of deep learning techniques. In 2017, Sakkos et al. [40] used 3D convolution to obtain spatial and temporal changes of the target simultaneously. Likewise, Akilan et al. [41] proposed a 3D CNN-LSTM network consists of dual coding and slow decoding. Further, in the D-DPDL model proposed by Zhao et al. [42], the convolutional neural network received random temporal pixel arrangement of features as input, and a Bayesian refinement module was constructed to suppress random noise. In addition, Bakkay et al. [43] adopted conditional generative adversarial network for foreground object detection. Patil et al. [44] fed the features gained by optical flow encoder and edge extraction mechanism into a bridge network composed of dense residual blocks, and propagated the predicted mask of previous frame to decoder to get exact motion targets. Li et al. [45] improved the detection performance by acquiring and adjusting multi-scale complementary knowledge of the change map in three steps (i.e., feature extraction, feature fusion, and feature refining).

Proposed method

The proposed Interactive spatiotemporal feature learning network (ISFLN) mainly composed of two components, namely, two-path spatio-temporal information extraction module (TSIEM) and multi-level feature enhancement module (MFEM). In the following subsections, we will provide a detailed analysis of the designed modules.

Overview

The overall structure of our method is given in Fig. 1. Specifically, TSIEM is conducted in two stages. For the first stage, we employ a Siamese Convolutional Network to obtain multi-level features of the current frame and reference frame. Then, multi-level spatio-temporal difference information is derived via element-wise subtraction. For the second stage, we analyze the different scale spatio-temporal context cues of the object using an interactive multi-scale feature extraction module. Two advantages exist in the design above. On the one hand, it emphasizes the change information of object. On the other hand, it strengthens the learning of multi-type features from different perspectives. Next, we guide and enhance the original coding features with two-paths spatio-temporal information. Finally, this knowledge is shared with the decoder to improve the expression ability of features.

Two-path spatio-temporal information extraction module (TSIEM)

Most models directly concatenated the current frame and reference frame for feature extraction, which ignores the differences between different features. Hence, to capture detailed spatio-temporal difference characteristics, we constructed a two-path feature extraction module at different levels.

For the first path, we send an input $F \in \mathbb{R}^{H\times W\times C_i}$ into the siamese network to get the multi-resolution feature maps $F_{\frac{1}{2}}$, $F_{\frac{1}{4}}$, $F_{\frac{1}{8}}$, $F_{\frac{1}{16}}$, and $F_{\frac{1}{16}}$. Specifically, the number of filters in five convolution operations Conv_ 1 → Conv_ 2 → Conv_ 3 → Conv_ 4 → Conv_ 5 are 32, 64, 128, 256, 256 (that is, C_i), respectively. All convolution blocks except Conv_5 are accompanied by BN, ReLu, and Max-Pooling, whereas Conv_5 only contains BN and ReLu. After that, perform the absolute difference operation on the corresponding block to obtain the first path spatio-temporal difference features.

Although the above process has acquired multi-level features, objects have different scales in various scenarios. Thus, we propose an interactive multi-scale feature extraction module (IMFEM) to fully capture information at multiple scales, as shown in Fig. 2.

The IMFEM divides the features into low-level, intermediate-level, and high-level for processing, and the entire learning process involves four steps. First of all, perform multi-scale feature extraction operations on low-level features. Usually, two branches of 3 × 3 and 5 × 5 convolution are used, but a large convolution kernel will cause expensive calculations, so we replace 5 × 5 convolution with two 3 × 3 convolution. Additionally, to reduce the number of channels, 1 × 1 convolution is added before 3 × 3 convolution. Then, the 1 × 1 convolution → 3 × 3 convolution branches are cross-merged to promote the exchange of characteristics on the same level. Next, low-level features fused at the far end are sent to the near end

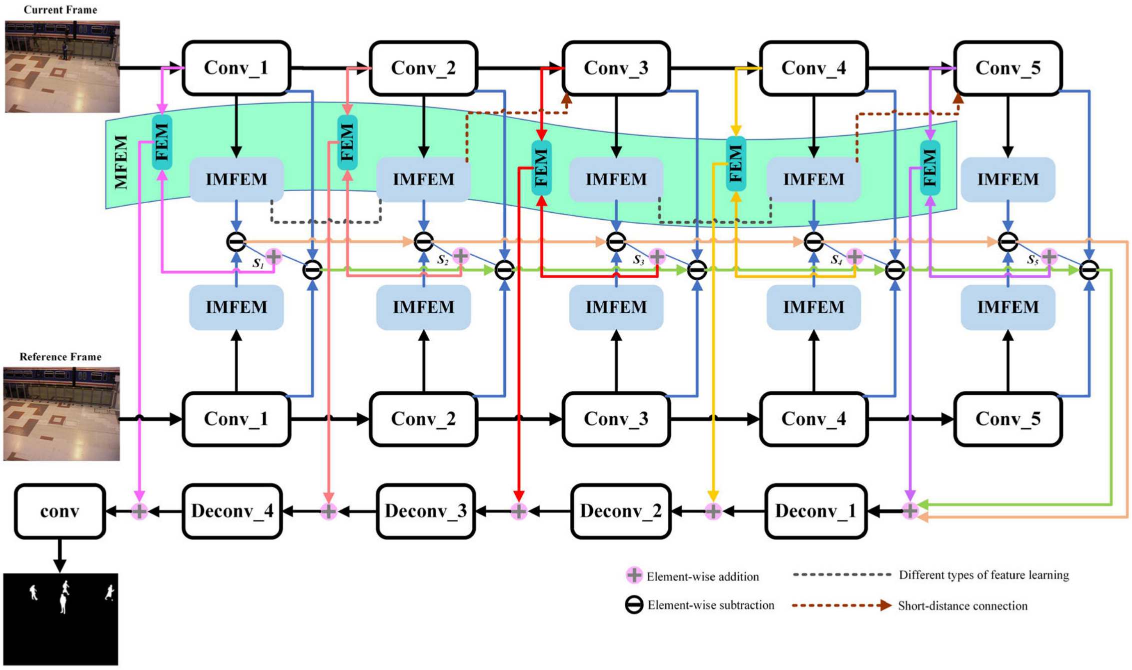


Fig. 1 An overview of the proposed model. *IMFEM* interactive multi-scale feature extraction module, *MFEM* multi-level feature enhancement module

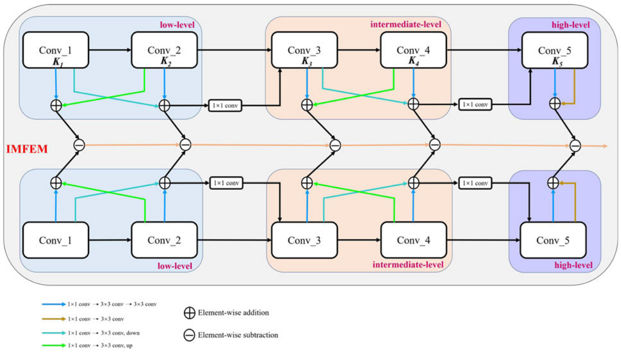


Fig. 2 The interactive multi-scale feature extraction module (IMFEM). Different types of multi-scale information learn from each other

of intermediate-level features via a short-distance connection. Here, the fused features first undergo 1 × 1 convolution before connection, which reduces the number of channels. Finally, spatio-temporal difference information of the corresponding locations is extracted. Likewise, intermediate-level and high-level features also follow the above steps.

Technically, one path in TSIEM is used to get multi-level features, and the other path is employed to get multi-scale context features, which can provide relatively sufficient target information for the network. Moreover, the design of IMFEM also strengthens the learning between the same type and different types of features. By doing this, the flow of information across levels is promoted, thereby enhancing the effectiveness of detection.

Multi-level feature enhancement module (MFEM)

When features pass through a deeper convolution layer, some knowledge and details are lost [46]. Numerous studies [20, 47–49] take a skip connection approach to fixing this issue, which adds the encoding features directly to the decoder. Unfortunately, this will introduce noise and rough information existing in the encoding stage, which is not conducive to accurate segmentation of foreground objects. Consequently, we employ spatio-temporal difference information obtained in the previous section to design a multi-level feature enhancement model (MFEM) to enhance the sharing of encoding features and decoder, as shown in Fig. 3.

The core of MFEM is to use fused spatio-temporal difference information S_i $\{i = 1, 2, 3, 4, 5\}$ to guide and refine coding feature K_m $\{m = 1, 2, 3, 4, 5\}$. We take the first feature enhancement module (FEM) as an example for detailed introduction. In the first step, perform a set of average-pooling and max-pooling on S_1 and K_1 to aggregate spatial information, respectively. An element-wise addition is adopted to fuse the two-path aggregation features, and the output is sent to a sigmoid activation function to adjust the hybrid features, denoted as M:

$$M = \sigma\{[\mathrm{AP}(K_1) + \mathrm{MP}(K_1)] + [\mathrm{AP}(S_1) + \mathrm{MP}(S_1)]\}, \quad (1)$$

where $\mathrm{AP}(\cdot)$ and $\mathrm{MP}(\cdot)$ denote the average-pooling and max-pooling, respectively.σ is sigmoid function.

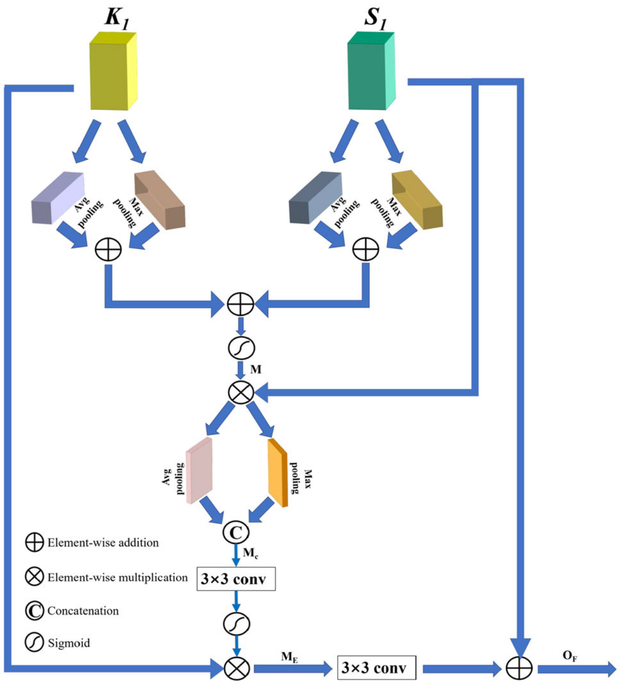


Fig. 3 Illustration of the feature enhancement module (FEM)

After that, the new features are used as weights and multiplied by S_I. Further, the weighted features are respectively performed average-pooling and max-pooling along the channel axis. The above process can be expressed by Eq. 2.

$$M_c = \mathrm{Cat}[\mathrm{AP}(S_1 \otimes M), \mathrm{MP}(S_1 \otimes M)], \tag{2}$$

where Cat(·) denotes concatenate operation.⊗ refers to element-wise multiplication.

Then, the concatenated feature map passes through 3 × 3 convolution and sigmoid activation function, and the enhanced coding features are obtained by element-wise multiplication with K_I.

Finally, the enhanced features are fed to 3 × 3 convolution and combined with S_I to gain output $\mathrm{O_F}$, which is sent to the corresponding decoder. Similarly, other levels of coding features also perform the above operations. The $\mathrm{O_F}$ can be formulated from Eqs. 3 and 4.

$$M_E = \sigma(f^{3\times3}(M_c)) \otimes K_1, \tag{3}$$

$$O_F = f^{3\times3}(M_E) + S_1, \tag{4}$$

where $f^{3\times3}(\cdot)$ represents a 3 × 3 convolutional layer.

In short, MFEM utilizes the fused spatio-temporal difference information to guide multi-level coding features, telling them which information is important and where the information is located, thereby improving the expressive ability of coding features. Meanwhile, it also provides more valuable clues for the decoder and a strong guarantee for higher accuracy of detection.

Experiments

Datasets and parameter settings

Video sequences employed in experiments come from the LASIESTA [50], CDnet2014 [51], INO [52], and AICD [53] datasets, including indoor and outdoor scenes. In the training process, 70% of the samples are employed as a training set and the rest as a testing set.

We perform experiments with Tensorflow in Python 3.7 and run the proposed model on workstation with processor GeForce RTX 3060 Laptop GPU and i7-10870H CPU. The input frame size is adjusted to 224 × 224, and the network adopts 50 epochs with a batch size 5 for training. Adam as the optimizer has an initial learning rate of 0.001. In addition, the loss function of our network utilizes binary cross-entropy.

In experimental analysis, the evaluation indicators [45, 51, 54] used include accuracy (Acc), precision (Pre), recall (Rec), F1, percentage of wrong classifications (PWC), false positive rate (FPR), false negative rate (FNR), Specificity (Sp), area under curve (AUC), and mean intersection over union (mIoU).

Ablation study

To verify the effectiveness of the proposed modules, we conduct a comprehensive analysis on three datasets (i.e., LASIESTA, CDnet2014, INO). Here, nine indicators are used to observe the performance of the designed module.

As shown in Table 1, we compare the performance of one-path spatio-temporal difference module (here as the baseline), interactive multi-scale feature extraction module (IMFEM), and multi-level feature enhancement module (MFEM). Specifically, F1 is the weighted average of precision and recall, and AUC represents the area under ROC curve, which is the comprehensive result of False Positive Rate (FPR) and True Positive Rate (TPR, that is, recall). From the viewpoint of these two compositive indicators, our proposed modules are effective. Additionally, Fig. 4 gives some visual results of different modules, as seen, baseline (BL), TSIEM, and encoder and decoder are directly connected (TSIEM + EDDC) have different degrees of boundary blur and error detection. Especially in baseline, due to the lack of enough target information, foreground pixels cannot be judged correctly, resulting in problems such as unclear target contours and missing targets. When IMFEM and MFEM modules are added sequentially, the above phenomenon is better alleviated.

Table 1 Ablation study for different modules

	Metrics								
Modules	Acc ↑	Pre ↑	Rec ↑	F1 ↑	PWC ↓	FPR ↓	FNR ↓	Sp ↑	AUC ↑
Baseline (BL)	0.9725	0.7848	0.7515	0.7603	1.3545	0.0082	0.2485	0.9918	0.9476
BL + IMFEM (TSIEM)	0.9735	0.8324	0.8099	0.8125	1.1616	0.0077	0.1901	0.9923	0.9703
TSIEM + EDDC	0.9749	0.8562	0.8257	0.8338	0.8721	0.0053	0.1743	0.9949	0.9724
TSIEM + MFEM (our)	0.9773	0.8786	0.8547	0.8609	0.7173	0.0042	0.1454	0.9960	0.9739

TSIEM + EDDC: encoder and decoder are directly connected. ↑indicates higher value is better, ↓ indicates lower value is better

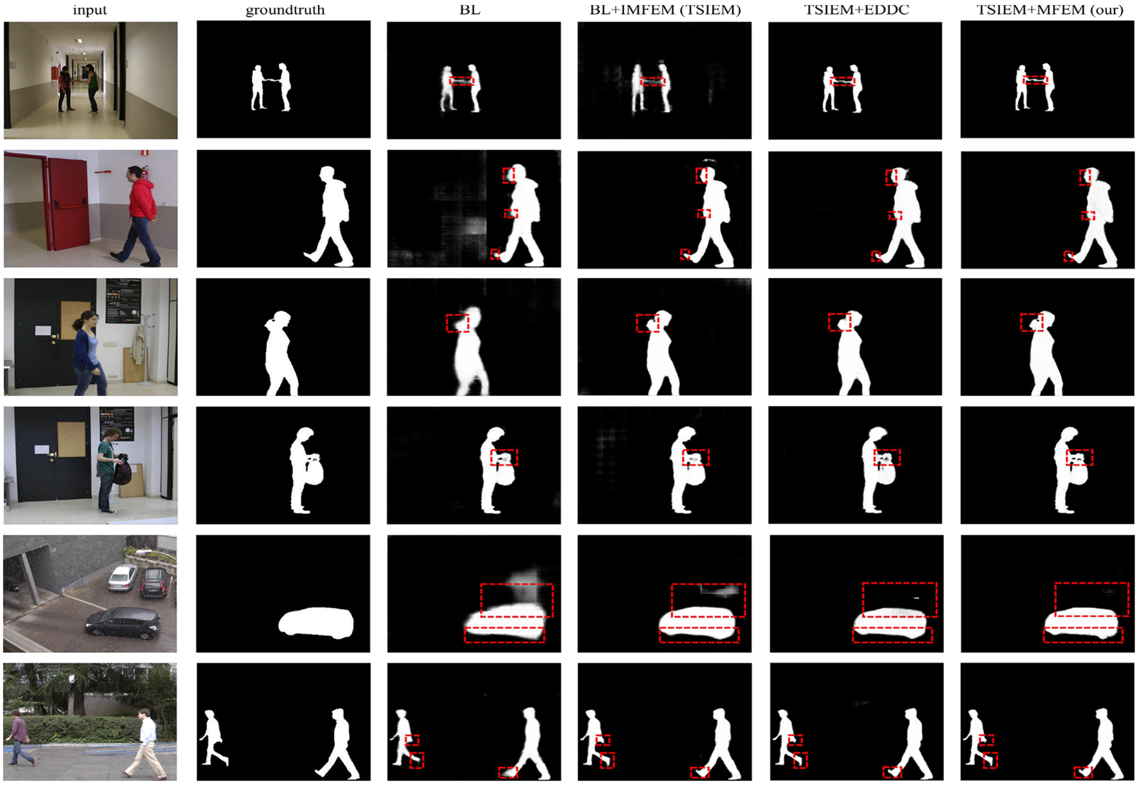


Fig. 4 Visual comparison results of different modules on LASIESTA dataset. (Columns 1 and 2 are the input frame and ground truth, respectively. Columns 3 to 6 indicate the detection results of different modules, respectively.)

Comparison with state-of-the-arts

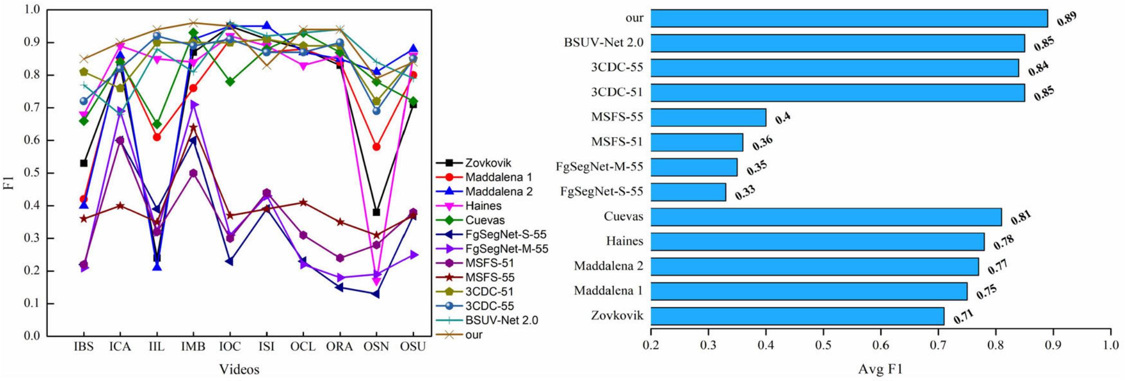
We compare the ISFLN with the existing traditional techniques and deep learning methods on LASIESTA, CDnet2014, INO, and AICD datasets.

1. LASIESTA dataset: Table 2 presents the F1 value of different approaches on LASIESTA [16, 55–62]. The last column indicates the average F1 value of these approaches across all videos, which is 89% in the proposed network. Compared with the same type of deep learning methods FgSegNet-S[16], FgSegNet-M [16], MSFS-51 [59], MSFS-55 [59], 3CDC-51 [60], 3CDC-55 [60], and BSUV-Net 2.0 [61], our technology has improved by 56%, 54%, 53%, 49%, 4%, 5%, and 4%, respectively. Figure 5 illustrates the graphical comparison of F1 and average F1 for different methods. As observed, the proposed framework performs stably on different videos, without rapid rise or fall.
Moreover, to validate the performance of the proposed method on unseen videos, we conduct experiments on LASIESTA dataset, as shown in Table 3 and Fig. 6. Specifically, training videos and test videos are different. It should be noted here that these videos used for experimental work also contain different challenges (see Table 3). Quantitative and qualitative analyses show that our approach can obtain comparable results for unseen videos.

Table 2 F1 score of different methods on LASIESTA dataset

Methods	Videos										
	IBS	ICA	IIL	IMB	IOC	ISI	OCL	ORA	OSN	OSU	Avg
Zovkovik [62]-2004	0.53	0.83	0.24	0.87	*0.95*	0.91	0.88	0.83	0.38	0.71	0.71
Maddalena 1[55]-2008	0.42	0.85	0.61	0.76	0.91	0.87	0.88	0.84	0.58	0.80	0.75
Maddalena 2[56]-2012	0.40	0.86	0.21	0.91	*0.95*	**0.95**	0.87	0.85	*0.81*	**0.88**	0.77
Haines [57]-2014	0.68	*0.89*	0.85	0.84	0.92	0.89	0.83	0.86	0.17	*0.86*	0.78
Cuevas [58]-2018	0.66	0.84	0.65	*0.93*	0.78	0.88	*0.93*	0.87	0.78	0.72	0.81
FgSegNet-S-55 [16]-2018	0.22	0.60	0.39	0.60	0.23	0.39	0.23	0.15	0.13	0.37	0.33
FgSegNet-M-55 [16]-2018	0.21	0.69	0.32	0.71	0.31	0.43	0.22	0.18	0.19	0.25	0.35
MSFS-51 [59]-2020	0.22	0.60	0.32	0.50	0.30	0.44	0.31	0.24	0.28	0.38	0.36
MSFS-55 [59]-2020	0.36	0.40	0.35	0.64	0.37	0.39	0.41	0.35	0.31	0.37	0.40
3CDC-51 [60]-2021	*0.81*	0.76	0.90	0.90	0.90	0.91	0.89	0.89	0.72	0.85	*0.85*
3CDC-55 [60]-2021	0.72	0.82	*0.92*	0.89	0.91	0.87	0.87	*0.90*	0.69	0.85	0.84
BSUV-Net 2.0 [61]-2021	0.77	0.68	0.88	0.81	**0.96**	*0.92*	*0.93*	**0.94**	**0.84**	0.79	*0.85*
Our	**0.85**	**0.90**	**0.94**	**0.96**	*0.95*	0.83	**0.94**	**0.94**	0.79	0.84	**0.89**

The best and second results are highlighted in bold and italics, respectively
Avg means average F1



(a) (b)

Fig. 5 Comparison of F1 and Avg F1 of different methods on LASIESTA dataset

Table 3 Quantitative analysis for unseen videos on LASIESTA dataset

Training		Testing		Metrics		
Video	Challenge	Video	Challenge	Pre ↑	Rec ↑	F1 ↑
I_BS_01	Bootstrap	I_CA_01	Camouflage	0.6035	0.7268	0.6417
I_IL_01	Illumination change	I_SI_01	Simple sequence	0.9010	0.8754	0.8805
O_CL_01	Cloudy	O_SN_01	Snowy	0.7926	0.7049	0.7356
I_CA_02	Camouflage	I_MB_01	Modified background	0.5637	0.6113	0.5763
O_CL_01	Cloudy	O_RA_01	Rainy	0.8574	0.8265	0.8375
I_CA_01	Camouflage	I_OC_02	Occlusion	0.7366	0.5031	0.5842
O_SN_01	Snowy	O_RA_01	Rainy	0.8572	0.7438	0.7851
I_IL_01	Illumination change	I_CA_01	Camouflage	0.8914	0.5985	0.7069
I_IL_02	Illumination change	I_MB_01	Modified background	0.8483	0.9256	0.8746
O_CL_02	Cloudy	O_SU_02	Sunny	0.6109	0.5264	0.5514
Average				0.7663	0.7042	0.7174

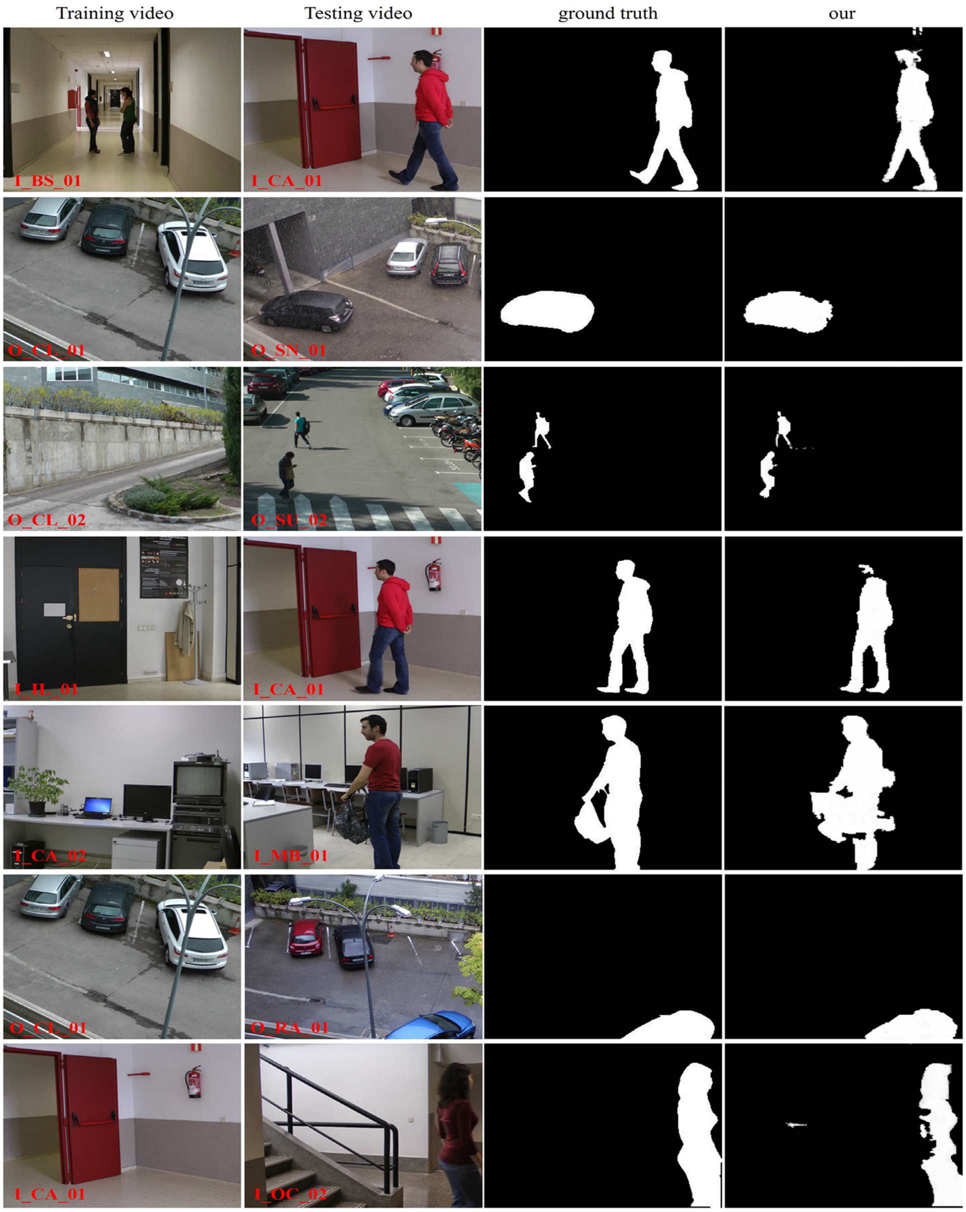


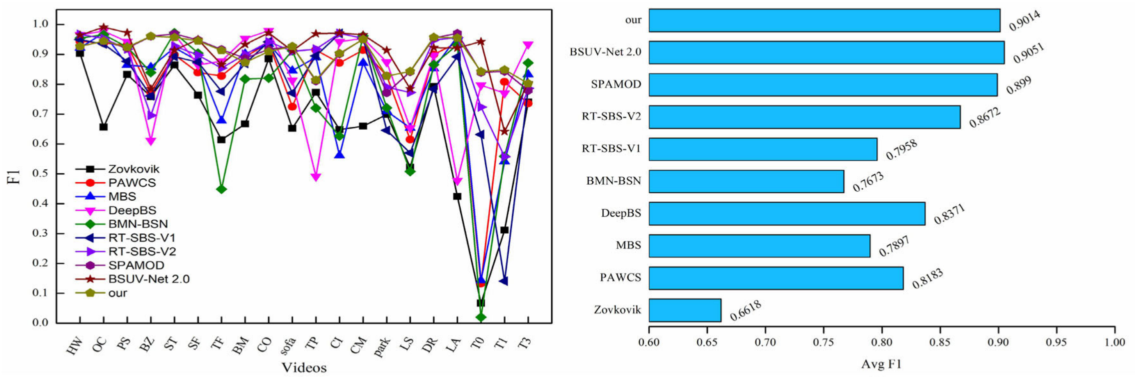
Fig. 6 Visual results of scene independence analysis on LASIESTA dataset

2. CDnet2014 dataset: Table 4 gives the performance of different methods on CDnet2014 [11, 24, 35, 61–65], in which the F1 value of the proposed network reaches 90.14%. The video for the experimental analysis contains many challenges, including dynamic background, camera jitter, shadow, and bad weather. As can be seen from the Table 4, no single method is omnipotent. Clearly, from the overall performance, our model is more stable than other methods, which is also verified in Fig. 7. In Fig. 7a, the trend of the curve indicates that our proposed approach is more capable of dealing with various challenges.
3. INO dataset: As reported in Table 5, we use accuracy (Acc), recall (Rec), specificity (Sp), and area under curve (AUC) on INO dataset to qualitatively analyze the existing methods [11, 32, 54, 66–68]. According to various indicators, the proposed method has a competitive edge. In terms of AUC, the proposed ISFLN is 96%, which is 23% and 6% higher than the deep learning techniques Akula-CNN [32] and SPAMOD [11], respectively.
4. AICD dataset: We utilize this dataset to examine how well the proposed technique performs when faced with small targets and large changes in perspective. Table 6 gives a quantitative comparison with the existing advanced methods (the metrics are F1 and mIoU). As seen in Table 6, the proposed approach outperforms the competition in both metrics. Specifically, compared with CNN-feat [69], WS-Net [70], DeconvNet [71], Mask-CDNet [72], and SASCNet [45], the F1 of proposed

Table 4 F1 score of different methods on CDnet2014 dataset

	Methods									
Videos	Zovkovik [62] 2004	PAWCS [63] 2015	MBS [24] 2017	DeepBS [64] 2018	BMN-BSN [35] 2019	RT-SBS-V1 [65] 2020	RT-SBS-V2 [65] 2020	SPAMOD [11] 2021	BSUV-Net 2.0 [61] 2021	Our
Highway (HW)	0.9038	0.9436	0.9217	*0.9655*	0.9542	0.9493	**0.9672**	0.9265	0.9648	0.9268
Office (OC)	0.6564	0.9375	0.9719	*0.9780*	0.9666	0.9347	0.9534	0.9445	**0.9913**	0.9448
PETS2006 (PS)	0.8327	0.9315	0.8648	*0.9425*	0.9244	0.8769	0.9178	0.9211	**0.9727**	0.9242
Blizzard (BZ)	0.7585	0.7737	0.8572	0.6115	0.8395	0.7607	0.6958	*0.9603*	0.7850	**0.9609**
Skating (ST)	0.8644	0.8984	0.9223	0.9669	**0.9722**	0.8928	0.9279	*0.9689*	0.9143	0.9570
Snowfall (SF)	0.7631	0.8393	0.8782	0.8648	0.9041	0.8738	0.8974	*0.9481*	**0.9501**	0.9452
Traffic (TF)	0.6137	0.8278	0.6781	0.8776	0.4488	0.7761	0.8543	**0.9165**	0.8655	*0.9133*
Badminton (BM)	0.6669	0.8920	0.9021	**0.9527**	0.8176	0.8695	0.9001	0.8870	*0.9334*	0.8729
Canoe (CO)	0.8851	0.9379	0.9345	**0.9794**	0.8206	0.9420	0.9422	0.9171	*0.9717*	0.9095
Sofa	0.6524	0.7247	0.8455	0.8134	0.9122	0.7706	0.9104	*0.9266*	0.9112	**0.9270**
Turnpike_0_5fps (TP)	0.7729	0.9146	0.8901	0.4917	0.7203	0.8952	*0.9176*	0.8108	**0.9691**	0.8140
Cubicle (CI)	0.6480	0.8713	0.5613	0.9427	0.6264	*0.9702*	*0.9702*	0.9016	**0.9715**	0.9034
Copymachine (CM)	0.6597	0.9143	0.8711	0.9534	*0.9620*	0.9541	0.9541	0.9513	**0.9657**	0.9506
Park	0.6989	0.8286	0.7099	*0.8741*	0.7210	0.6454	0.7903	0.7712	**0.9136**	0.8274
Lakeside (LS)	0.5221	0.6147	0.6541	0.6535	0.5082	0.5699	0.7725	*0.8419*	0.7849	**0.8434**
Dingroom (DR)	0.7925	0.8997	0.8531	0.8970	0.8659	0.7838	0.9482	*0.9548*	0.9226	**0.9576**
Library (LA)	0.4247	0.9390	*0.9594*	0.4773	0.9320	0.8920	0.9564	**0.9706**	0.9215	0.9554
Turbulence0 (T0)	0.0673	0.1326	0.1435	0.7971	0.0200	0.6319	0.7237	0.8390	**0.9429**	*0.8428*
Turbulence1 (T1)	0.3118	0.8083	0.5413	0.7698	0.5588	0.1411	0.5573	*0.8431*	0.6416	**0.8485**
Turbulence3 (T3)	0.7405	0.7356	0.8329	**0.9340**	*0.8714*	0.7865	0.7865	0.7791	0.8053	0.8024
Avg	0.6618	0.8183	0.7897	0.8371	0.7673	0.7958	0.8672	0.8990	**0.9051**	*0.9014*

The best and second results are highlighted in bold and italics, respectively
Avg means average F1



(a) (b)

Fig. 7 Comparison of F1 and Avg F1 of different methods on CDnet2014 dataset

method increases by 60.2%, 74.4%, 54.8%, 51.9%, and 2.2%, respectively, while the mIoU increases by 26.6%, 41.2%, 17.6%, 15.2%, and 2.1%, respectively.

5. Visual results: Fig. 8 provides a qualitative analysis through visual comparison. There are a variety of challenges in the videos shown, such as dynamic background, camera jitter, shadows, turbulence, and bad weather. The visual results given comprise two traditional methods, Zovkovik [62] and PAWCS [63], as well as four deep learning methods, DeepBS [64], RT-SBS-V2 [65], SPAMOD [11], and BSUV-Net 2.0 [61]. It can be seen from Fig. 8 that our proposed network can obtain stable output in the face of various challenges. As an example, in rows 4, 8, 9, 10, and 12, there are problems with missed detection, holes, and false detection in targets obtained by other approaches. Figure 9 shows visualization results of the proposed method on AICD dataset. Particularly, the first row gives input frame, the second row is ground truth, and the third row pictures detection results of the proposed approach. It is obvious that the proposed model is able to obtain relatively clear objects under small targets and large view variations.

   The proposed method relies on the reference frame and the current frame to obtain spatio-temporal difference information for foreground detection. It is mainly applied to video surveillance with fixed cameras. In experiments, we observe that the detection performance of the proposed framework is limited when scene changes greatly and in a turbulent situation. In view of the above limitations, we plan to guide the network to identify features by propagating the prediction mask of the previous frame in future work.

6) Real-time: Time complexity is also one of the factors to investigate model performance. In Table 7, we present a comparison of the FPS and the number of trainable parameters for some recent methods. Specifically, the reasoning speed of the proposed network is 24 frames per second with about 5.27 M trainable parameters. Overall, our method is competitive in terms of real-time performance.

Table 5 Acc, Rec, Sp, and AUC scores of different methods on INO dataset

	Metrics			
Methods	Acc ↑	Rec ↑	Sp ↑	AUC ↑
STSM [66]-2015	0.75	0.70	0.28	0.70
Akula-CNN [32]-2016	0.79	0.73	0.26	0.73
DL [67]-2017	0.80	0.75	0.20	0.74
MRF [68]-2018	0.81	*0.79*	0.19	0.78
Qiu [54]-2019	*0.83*	**0.80**	0.16	0.81
SPAMOD [11]-2021	**0.98**	0.62	*0.90*	*0.90*
Our	**0.98**	0.77	**0.99**	**0.96**

The best and second results are highlighted in bold and italics, respectively

Avg means average F1

Table 6 F1 and mIoU scores of different models on AICD dataset

	Metrics	
Methods	F1 ↑	mIoU ↑
CNN-feat [69]-2015	0.287	0.535
WS-Net [70]-2017	0.145	0.389
DeconvNet [71]-2018	0.341	0.625
Mask-CDNet [72]-2020	0.370	0.649
SASCNet [45]-2020	*0.867*	*0.780*
Our	**0.889**	**0.801**

The best and second results are highlighted in bold and italics, respectively

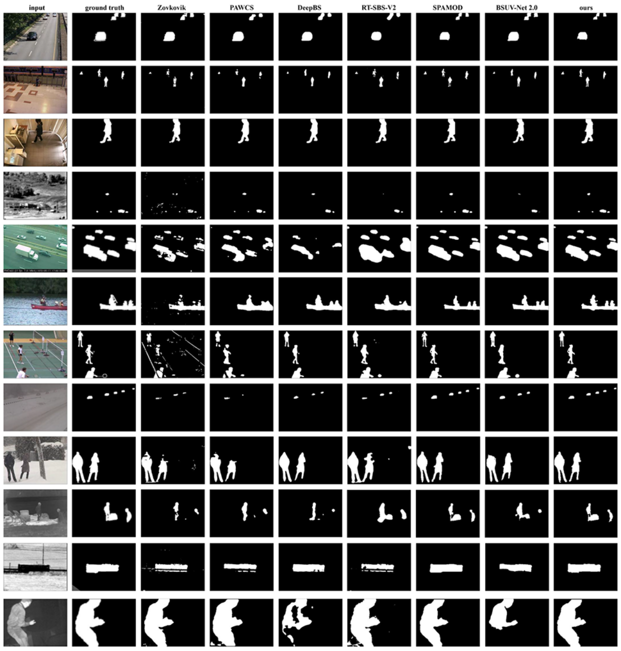


Fig. 8 Qualitative comparison of proposed method with other existing approaches on CDnet2014 dataset. (Columns 3–9 display the results of Zovkovik [62], PAWCS [63], DeepBS [64], RT-SBS-V2 [65], SPAMOD [11], BSUV-Net 2.0 [61] and our method, respectively.)

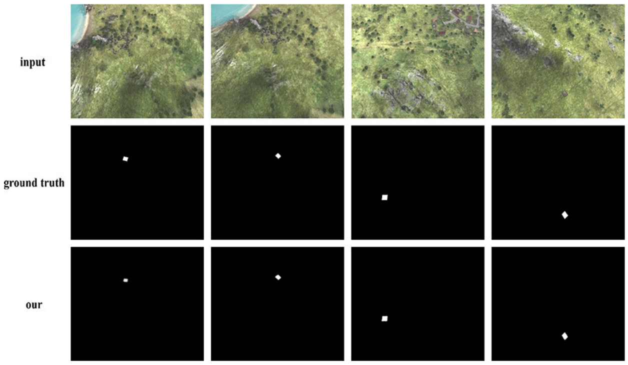


Fig. 9 Visual results on AICD dataset

Table 7 Speed comparison of different methods

Method	FPS	#Param
FgSegNet [16]-2018	18	2.60 M
3CDC [60]-2021	25	0.13 M
BSUV-Net 2.0 [61]-2021	6	NA
MFCN [17]-2018	27	20.83 M
RT-SBS [65]-2020	25	NA
SASCNet [45]-2020	10	NA
SPAMOD [11]-2021	13	35.1 M
our	24	5.27 M

Conclusion

In this paper, we propose a novel end-to-end video foreground detection approach called Interactive Spatio-temporal Feature Learning Network (ISFLN). Our innovations are the design of a two-path spatio-temporal information extraction module (TSIEM), and a multi-level feature enhancement module (MFEM). The design of TSIEM strengthens the learning between different types of information, and obtains sufficient spatio-temporal difference knowledge from multi-level and multi-scale aspects, which is extremely key for video object detection tasks. Further, fine coding features are captured by MFEM and shared with the decoder to enhance feature expression ability to get more accurate detection results. We demonstrate the effectiveness of designed modules by ablation experiments. Moreover, compared with the existing advanced approaches in LASIESTA, CDnet2014, INO, and AICD datasets, ISFLN can achieve better detection results, and also achieves a competitive processing speed (24 fps). In future work, we will try to enhance the learning of spatial features by incorporating edge label training, and perform in-depth research on scene independence analysis.

Declarations

Conflict of interest There is no conflict of interest from the authors.

References

1. Tang X, Tu W, Li K, Cheng J (2021) DFFNet: an IoT-perceptive dual feature fusion network for general real-time semantic segmentation. Inf Sci 565:326–343
2. Cheng G, Si Y, Hong H, Yao X, Guo L (2021) Cross-scale feature fusion for object detection in optical remote sensing images. IEEE Geosci Remote Sens 18(3):431–435
3. Zhang M, Yang Y, Ji Y, Xie N, Shen F (2018) Recurrent attention network using spatial-temporal relations for action recognition. Signal Process 145:137–145
4. Li F, Zhu A, Liu Z, Huo Y, Xu Y, Hua G (2021) Pyramidal graph convolutional network for skeleton-based human action recognition. IEEE Sens J 21(14):16183–16191

5. Hua S, Wang C, Xie Z, Wu X (2020) A force levels and gestures integrated multi-task strategy for neural decoding. Complex Intell Syst 6(3):469–478
6. Zhang H, Qu S, Li H, Luo J, Xu W (2020) A moving shadow elimination method based on fusion of multi-feature. IEEE Access 8:63971–63982
7. Wang Z, Ma Y (2021) Detection and recognition of stationary vehicles and seat belts in intelligent Internet of Things traffic management system. Neural Comput Appl. https://doi.org/10.1007/s00521-021-05870-6
8. Chiu C, Ku M, Liang L (2010) A robust object segmentation system using a probability-based background extraction algorithm. IEEE Trans Circ Syst Video Technol 20(4):518–528
9. Zhao C, Sain A, Qu Y, Ge Y, Hu H (2019) Background subtraction based on integration of alternative cues in freely moving camera. IEEE Trans Circ Syst Video Technol 29(7):1933–1945
10. Xu Y, Ji H, Zhang W (2020) Coarse-to-fine sample-based background subtraction for moving object detection. Optik 207:164195
11. Qu S, Zhang H, Wu W, Xu W, Li Y (2021) Symmetric pyramid attention convolutional neural network for moving object detection. Signal Image Video Process 15:1747–1755
12. Stauffer C, Grimson WEL (1999) Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 246–252
13. Elgammal A, Harwood D (2000) Non-parametric Model for Background Subtraction. In: Proceedings of the European conference on computer vision, pp 751–767
14. Varadarajan S, Miller P, Zhou H (2015) Region-based mixture of Gaussians modelling for foreground detection in dynamic scenes. Pattern Recogn 48(11):3488–3503
15. Minaee S, Wang Y (2019) An ADMM approach to masked signal decomposition using subspace representation. IEEE Trans Image Process 28(7):3192–3204
16. Lim LA, Keles HY (2018) Foreground segmentation using convolutional neural networks for multiscale feature encoding. Pattern Recogn Lett 112:256–262
17. Zeng D, Zhu M (2018) Multiscale fully convolutional network for foreground object detection in infrared videos. IEEE Geosci Remote S 15(4):617–621
18. Wang Y, Luo Z, Jodoin P (2017) Interactive deep learning method for segmenting moving objects. Pattern Recogn Lett 96:66–75
19. Chen Y, Wang J, Zhu B, Tang M, Lu H (2017) Pixelwise deep sequence learning for moving object detection. IEEE Trans Circ Syst Video Technol 29(9):2567–2579
20. Tezcan MO, Ishwar P, Konrad J (2020) 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), pp 2774–2783
21. Akilan T, Wu QMJ (2020) sEnDec: an improved image to image CNN for foreground localization. IEEE Trans Intell Transp 21(10):4435–4443
22. Barnich O, Droogenbroeck M (2011) ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans Image Process 20(6):1709–1724
23. Zhao Y (2018) ALI-TM: a moving objects detection algorithm for infrared images with dynamic background. Infrared Phys Techn 93:205–212
24. Sajid H, Cheung S (2017) Universal multimode background subtraction. IEEE Trans Image Process 26(7):3249–3260
25. Roy SM, Ghosh A (2020) Foreground segmentation using adaptive 3 phase background model. IEEE Trans Intell Transp 21(6):2287–2296
26. Tom AJ, George SN (2021) A three-way optimization technique for noise robust moving object detection using tensor low-rank approximation, $l_{1/2}$, and TTV regularizations. IEEE Trans Cybern 51(2):1004–1014
27. Zhou T, Li J, Wang S, Tao R, Shen J (2020) MATNet: motion-attentive transition network for zero-shot video object segmentation. IEEE Trans Image Process 29:8326–8338
28. Wang B, Liu W, Han G, He S (2020) Learning long-term structural dependencies for video salient object detection. IEEE Trans Image Process 29:9017–9031
29. Xu K, Wen L, Li G, Huang Q (2020) Self-supervised deep TripleNet for video object segmentation. IEEE Trans Multimed 23:3530–3539
30. Lu X, Wang W, Shen J, Tai Y, Crandall D, Hoi S (2020) Learning video object segmentation from unlabeled videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 8957–8967
31. Patil PW, Dudhane A, Murala S (2020) End-to-end recurrent generative adversarial network for traffic and surveillance applications. IEEE Trans Veh Technol 69(12):14550–14562
32. Akula A, Singh A, Ghosh R, Kumar S, Sardana HK (2016) Target recognition in infrared imagery using convolutional neural network. In: Proceedings of the international conference on computer vision and image processing, pp 25–34
33. Patil PW, Murala S, Dhall A, Chaudhary S (2018) MsEDNet: multi-scale deep saliency learning for moving object detection. In: Proceedings of the IEEE international conference on systems, man, and cybernetics, pp 1670–1675
34. Yang L, Li J, Luo Y, Zhao Y, Cheng H, Li J (2018) Deep background modeling using fully convolutional network. IEEE Trans Intell Transp 19(1):254–262
35. Guerra VM, Rouco J, Novo J (2019) An end-to-end deep learning approach for simultaneous background modeling and subtraction. In: Proceedings of the British machine vision conference, pp 1–12
36. Huang Z, Li W, Li J, Zhou D (2021) Dual-path attention network for single image super-resolution. Expert Syst Appl 169:114450
37. Fu J, Liu J, Tian H (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3141–3149
38. Xiao B, Xu B, Bi X, Li W (2021) Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion. IEEE Trans Image Process 30:163–175
39. Minematsu T, Shimada A, Taniguchi R (2019) Simple background subtraction constraint for weakly supervised background subtraction network. In: Proceedings of the ieee international conference on advanced video and signal based surveillance (AVSS), pp 1–8
40. Sakkos D, Liu H, Han J, Shao L (2018) End-to-end video background subtraction with 3d convolutional neural networks. Multimed Tools Appl 77(17):23023–23041
41. Akilan T, Wu QJ, Safaei A, Huo J, Yang Y (2020) A 3D CNN-LSTM-based image-to-image foreground segmentation. IEEE Trans Intell Transp 21(3):959–971
42. Zhao C, Basu A (2020) Dynamic deep pixel distribution learning for background subtraction. IEEE Trans Circ Syst Vid 30(11):4192–4206
43. Bakkay MC, Rashwan HA, Salmane H, Khoudour L, Puig D, Ruichek Y (2018), BScGAN: deep background subtraction with conditional generative adversarial networks. In: Proceedings of the IEEE international conference on image processing (ICIP), pp 4018–4022
44. Patil PW, Biradar K, Dudhane A, Murala S (2020) An end-to-end edge aggregation network for moving object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 8146–8155
45. Li S (2020) Change detection in images using shape-aware siamese convolutional network. Eng Appl Artif Intel 94:103819
46. Dosovitskiy A, Brox T (2016) Inverting visual representations with convolutional networks. In: Proceedings of the IEEE computer

society conference on computer vision and pattern recognition, pp 4829–4837
47. Zhang K, Zhang Y, Cheng H (2020) Self-supervised structure learning for crack detection based on cycle-consistent generative adversarial networks. J Comput Civil Eng 3(34):4020004
48. Han T, Ivo RF, Rodrigues D, Peixoto SA, Albuquerque V, Filho P (2020) Cascaded volumetric fully convolutional networks for whole-heart and great vessel 3D segmentation. Future Gener Comput Syst 108:198–209
49. Gao K (2021) Dual-branch combination network (DCN): towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. Med Image Anal 67:101836
50. Carlos C, Eva M, Narciso G (2016) Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA. Comput Vis Image Underst 152:103–117
51. Wang Y, Jodoin P, Porikli F, Konrad J, Benezeth Y, Ishwar P (2014) CDnet 2014: an expanded change detection benchmark dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 393–400
52. https://www.ino.ca/en/technologies/video-analytics-dataset/
53. Bourdis N, Marraud D, Sahbi H (2011) Constrained optical flow for aerial image change detection. In: Proceedings of the IEEE international geoscience and remote sensing symposium, pp 4176–4179
54. Qiu S, Luo J, Yang S, Zhang M, Zhang W (2019) A moving target extraction algorithm based on the fusion of infrared and visible images. Infrared Phys Technol 98:285–291
55. Maddalena L, Petrosino A (2008) A self-organizing approach to background subtraction for visual surveillance applications. IEEE Trans Image Process 17(7):1168–1177
56. Maddalena L, Petrosino A (2012) The SOBS algorithm: what are the limits? In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops, pp 21–26
57. Haines TSF, Xiang T (2014) Background subtraction with Dirichlet process mixture models. IEEE Trans Pattern Anal 36(4):670–683
58. Berjón D, Cuevas C, Morán F, García N (2018) Real-time nonparametric background subtraction with tracking-based foreground update. Pattern Recogn 74:156–170
59. Lim LA, Keles HY (2020) Learning multi-scale features for foreground segmentation. Pattern anal appl 23(3):1369–1380
60. Mandal M, Dhar V, Mishra A, Vipparthi SK, Mottaleb M (2021) 3DCD: scene independent end-to-end spatiotemporal feature learning framework for change detection in unseen videos. IEEE Trans Image Process 30:546–558
61. Tezcan MO, Ishwar P, Konrad J, Janusz K (2021) BSUV-Net 2.0: spatio-temporal data augmentations for video-agnostic supervised background subtraction. IEEE Access 9:53849–53860
62. Zivkovic Z (2004) Improved adaptive Gaussian mixture model for background subtraction. In: Proceedings of the IEEE conference pattern recognit (ICPR), pp 28–31
63. Charles P, Bilodeau G, Bergevin R (2015) A self-adjusting approach to change detection based on background word consensus. In: Proceedings of the IEEE Winter conference on applications of computer vision, pp 990–997
64. Babaee M, Dinh DT, Rigoll G (2018) A deep convolutional neural network for video sequence background subtraction. Pattern Recogn 76:635–649
65. Cioppa A, Droogenbroeck M, Braham M (2020) Real-time semantic background subtraction. http://arxiv.org/abs/2002.04993v3
66. Li Z (2015) Infrared small moving target detection algorithm based on joint spatio-temporal sparse recovery. Infrared Phys Technol 69:44–52
67. Bhattacharjee SD, Talukder A, Alam MS (2017) 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, SPIE, pp 102030O
68. Sun B, Li Y, Guo G (2018) Moving target segmentation using Markov random field-based evaluation metric in infrared videos. Opt Eng 1(57):13106
69. Sakurada K, Okatani T (2015) Change Detection from a street image pair using CNN features and superpixel segmentation. In: Proceedings of the British machine vision conference, pp 1–12
70. Khan S, He X, Porikli F, Bennamoun M, Sohel F, Togneri R (2017) Learning deep structured network for weakly supervised change detection. In: Proceedings of the international joint conference on artificial intelligence, pp 2008–2015
71. Alcantarilla P (2018) Street-view change detection with deconvolutional networks. Auton Robot 42(7):1301–1322
72. Bu S, Li Q, Han P, Leng P, Li K (2020) Mask-CDNet: a mask based pixel change detection network. Neurocomputing 378:166–178

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