IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 25, 2019, accepted July 24, 2019, date of publication July 30, 2019, date of current version August 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2931922

Moving Object Detection Method via ResNet-18 With Encoder–Decoder Structure in Complex Scenes

XIANFENG OU, PENGCHENG YAN, YIMING ZHANG, BING TU, GUOYUN ZHANG, JIANHUI WU, AND WUJING LI

School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, China
Machine Vision and Artificial Intelligence Research Center, Hunan Institute of Science and Technology, Yueyang 414006, China

Corresponding authors: Bing Tu (tu_bing@163.com) and Guoyun Zhang (gyzhang@hnist.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 51704115, in part by the Science and Technology Program of Hunan Province under Grant 2016TP1021, in part by the Hunan Provincial Innovation Foundation for Postgraduate under Grant CX2018B776, Grant CX2018B779, and Grant YCX2019A14, and in part by the Hunan Provincial Natural Science Foundation of China under Grant 2019JJ40104.

ABSTRACT In complex scenes, dynamic background, illumination variation, and shadow are important factors, which make conventional moving object detection algorithms suffer from poor performance. To solve this problem, a moving object detection method via ResNet-18 with encoder–decoder structure is proposed to segment moving objects from complex scenes. ResNet-18 with encoder–decoder structure possesses pixel-level classification capability to divide pixels into foreground and background, and it performs well in feature extraction because of its layers are so shallow that many more low-scale features will be retained. First, the object frames and their corresponding artificial labels are input to the network. Then, feature vectors will be generated by the encoder, and they are converted into segmentation maps by the decoder through deconvolution processing. Third, a rough matching of the moving object regions will be obtained, and finally, the Euclidean distance is used to match the moving object regions accurately. The proposed method is suitable for the scenes where dynamic background, illumination variation, and shadow exist, and experimental results on the public standard CDnet2014 and I2R datasets, from both qualitative and quantitative comparison aspects, demonstrate that the proposed method outperforms state-of-the-art algorithms significantly, and its mean *F-measure* increased by 1.99%∼29.17%.

INDEX TERMS Complex scenes, moving object detection, ResNet-18, encoder-decoder network, background subtraction.

I. INTRODUCTION

Moving object detection is one of the most extensively studied topics in computer vision and the digital image processing [1]–[3], which is usually used as a preprocessing step in numerous vision applications including object tracking [4], object detection [5], behavior analysis [6] and so on. The purpose of moving object detection is to extract the motion regions (foreground objects) in the image sequences from the backgrounds. However, moving object detection is still a challenging problem, since some background regions are contained in complex scenes, where water surface, shaking leaves, light changing and moving cloud exist. In addition, the moving object detection algorithms should also adapt to some factors, such as illumination variation and the shadow.

Background subtraction is one of the most popular methods to detect moving objects. Background subtraction divides the foreground and background by building the background model and calculating the difference between the current frame and the background model. In recent years, researchers have done a lot of optimizations on background subtraction [7]–[9]. Roy and Ghosh [10] has proposed an efficient real-time background subtraction algorithm, this algorithm was characterized by using a single sliding window to update the model in adaptive, which could overcome sudden and/or gradual lighting changes in scenes. Chen *et al.* [11] has proposed a background subtraction model based on hierarchical super-pixel segmentation and robust estimator, which

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

improved the robustness of the system in dynamic background. However, these aforementioned methods are less robust to frequent appearance changes of scenes, and there are multiple variables in some scenes (such as light changes through the shaking leaves and the brightness changes by water waves, *etc*.) that will lead to the background model cannot be updated accurately. So, if segmentation of the possible moving objects from the background is available and then the detection of their motions can be obtained, the effect of the background on moving object detection will be greatly weakened.

Recently, the Convolution Neural Network (CNN) has been successfully employed in many research fields, such as computer vision [12], [13] and nonlinear system [14]. Long *et al*. [15] has proposed the fully convolutional neural network (FCN). FCN uses 1×1 convolution to replace the full connection layer, then up-samples the last convolution layer's feature map by deconvolution and restores to the size of the input image, these characteristics make FCN with the ability to predict each pixel in the image. Badrinarayanan *et al*. [16] has proposed the SegNet for image segmentation, in which SegNet's decoder used pooling indices to compute and to perform non-linear upsampling in the max-pooling step of the corresponding encoder. Bian *et al*. [17] has proposed a network that was a composition of n FCNs, the network operated at different scales, which means this network could use multi-scale networks to make use of their merits of multiple networks, and then the network merged the predictions to produce a single output. All of these networks had encoder-decoder architecture, in which the encoder extracted features from an input image, and then the decoder converted them into a specific prediction. In many cases, image classification networks were fine-tuned and employed as the encoders [18]–[20], while the decoders were designed in various ways according to the purposes.

Conventional background subtraction algorithms focus on how to build and update a background model or how to compare an object frame with the background model. In this paper, we develop a moving object method with the encoder-decoder architecture CNN network, in which the ResNet-18 is fine-tuned and employed as the encoder. The network's input contains object frames and corresponding artificial labels. Segmentation maps and a rough matching of the moving object regions will be obtained, then the Euclidean distance is used to match the moving object regions accurately. By using the prior information of the foreground object, our proposed method does not need to build the background model. Therefore, the background no longer affects the segmentation of foreground and background, and the update of the background model is not needed. Our proposed method can be applied to complex scenes where dynamic background, illumination variation and shadow exist.

The rest of this paper is organized as follows: Section II briefly describes related work. Section III presents the proposed method in detail, and Section IV shows the experiment results and assesses the performance from both qualitative and quantitative aspects in comparison with the state-of-the-art algorithms. Finally, Section V concludes the whole work.

II. RELATED WORK

In consideration of the advantages of the encoder-decoder CNN network, which is able to segment the foreground and background of the image, and then the classification of foreground and background pixels are also be realized. Generally, the CNN network should be trained in supervised, and artificial labels are used in many image processing tasks [21], [22].

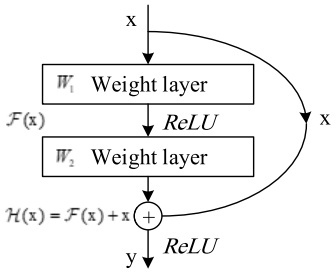


FIGURE 1. The residual block.

The deep residual network (ResNet) is one of the most commonly convolution neural network (CNN). The residual block is shown in Fig. 1, in which the curved arrows represent shortcut connection.

The residual block is defined as:

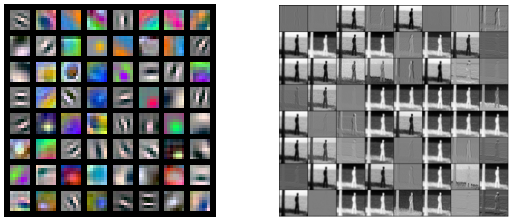
$$y = \mathcal{F}(x, \{W_i\}) + x \tag{1}$$

where x, y are the input and output of the layers considered. The function $\mathcal{F}(x, \{W_i\})$ represents the residual mapping to be learned. The residual block in Fig. 1 has two weight layers, W_1 and W_2 represent the first layer and the second layer respectively. As for $\mathcal{F} = W_2\sigma(W_1 x)$, σ denotes *ReLU* and the biases are omitted for simplifying notations. Formally, denote the desired underlying mapping as $\mathcal{H}(x)$. Let the stacked nonlinear layers fit another mapping of $\mathcal{F}(x) = \mathcal{H}(x) - x$, the original mapping is recast into $\mathcal{F}(x)+x$. Hypothesize that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. To the extreme, if an identity mapping were optimal, it would be easier to push the residual to zero than to fit an identity mapping by a stack of nonlinear layers [23]. In the meantime, shortcut connection retains additional information of the previous layer.

Due to the introduction of residual block, the degradation caused by the increase in the number of layers in the network is well resolved by ResNets. He *et al*. [23] showed that ResNets (including ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152, whose main difference lies in the number of network layers) perform better in image classification than other CNN models in ImageNet dataset, which indicated that image features could be well extracted by ResNets. Therefore, after excluding the fully connected layers, we can use ResNets as the encoder, and the pre-trained

model that has been trained on the ImageNet dataset is used for fine-tuning.

Another research [24], [25] showed that too much deep layers were redundant for dense prediction of the image. For CNN, the shallower layers tend to learn low-scale features (object edges feature, structures feature and textures feature, *etc*.), while the deeper layers learn higher-scale features (spatial context, global semantic and the local features of the objects), and as the number of network layers deepens, low-scale features will be lost. As shown in Fig. 2, filters and outputs of ResNet-18's second convolution layer are visualization. However, our work needs to segment the background by making an intensive prediction of the image, and loss of low-scale features will lead to image segmentation inaccurate.



(a) filters visualization (b) outputs visualization

FIGURE 2. The second convolutional layer of ResNet-18 is visualized, where (a) represents the 64 filters (convolution kernels), and the outputs in (b) correspond to each convolution kernel in (a) respectively. Outputs in (b) consist of rich low-scale features.

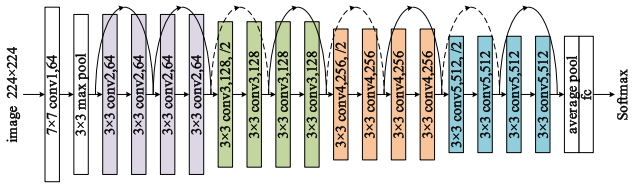


FIGURE 3. The Network structure of ResNet-18.

The performance of ResNet-18 is similar to other ResNets, which can retain more of the low-scale features due to the reason that it is shallow. Therefore, we use ResNet-18 pre-trained model as feature extractor (encoder) for our network model, the Network structure of ResNet-18 is described in Fig.3. ResNet-18 consists of 16 convolution layers, 2 downsampling layers and some fully connected layers(fc). The input image size of ResNet is 224×224, in addition to the first convolution layer, the convolution kernel size is 7×7, and the other layers are 3×3. After average pooling the feature map of the last convolution layer, an eigenvector is obtained by full connection, then the classification probability is obtained by normalization with Softmax. The convolution layer that outputs the same size feature map has the same number of filters, as shown in Fig.3, two convolution layers of the same color form a residual block. Shortcut connections are those skipping two layers (curved arrows in Fig. 3), the dotted shortcuts increase dimensions.

ResNet-18 will obtain an eigenvector that contains multiple probabilities, which are used to indicate that the input image belongs to a certain class, and the class with the highest probability will be the output finally. The number of input channels of the fully connected layer must be fixed, so the input image of ResNet-18 needs to be a fixed size.

III. PROPOSED METHOD

In general, moving object detection needs to process each pixel in the image to get the foreground object, we take advantage of encoder-decoder network's pixel-level classification capability, and try to divide the pixels in the image into foreground pixels and background pixels. Through supervised learning, the model acquires the features of the foreground object, and then segments the pixels belonging to the foreground object. In the proposed method, after the network is trained, the pre-trained model is obtained. Image input pre-training model will output a rough matching image. Then, the Euclidean distance is used to further refine the motion pixels to obtain fine matching region.

Combined with the related work, we propose an encoder-decoder network based on ResNet-18. We elaborate our improved network structure in detail in the following part, of which the main network structure of our improved method is described in Fig. 4.

A. THE STRUCTURE OF THE NETWORK

Since moving object detection is the binary classification task (the foreground and the background), the over-deep layers in the network are redundant, so we design our feature extractor (encoder) similarly, Fig. 4 (the part in the dotted box) shows our encoder part.

Since the fully connected layer is replaced by a 1×1 convolution, without the limit of the number of input channels in the fully connected layer, our network can process images with any size. The hidden layer extracts abstract features by deepening convolution layer by layer, the feature maps of the previous layer of the 1×1 convolution layer contain a large amount of semantic information, this information is used in CNN to classify the object, while in our network it can be used to classify the pixels. In our network, 1×1 convolution reduces dimensions and predicts image pixels, for example, in ResNet-18, the dimension of the feature map exported by conv5 is $7 \times 7 \times 512$, by convolving with kernel of 1×1, a heatmap with a dimension of $7 \times 7 \times 1$ can be generated, which contains predicted values for all pixels in the input image. Finally, through a properly transposed convolution, the heatmap is restored to the size of the input image and the result is the output.

1) NETWORK INPUT

Since our network is an encoder-decoder structure, it doesn't need a fixed-size input. To make our network robust to dynamic appearance variations during the learning process, we choose frames randomly to generate labels in order to avoid successive frames (with high similarity). To obtain the training dataset, for each video, we randomly select 20% of the frames for labeling. And for each frame in the training

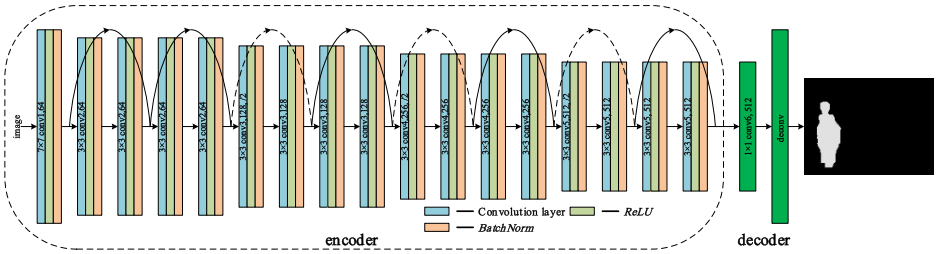


FIGURE 4. The structure of the proposed network. As shown in the figure, the convolution layer before the 1 × 1 convolution layer (the part in the dotted box) constitutes a feature extractor (encoder). In the figure, the blue part represents the convolution layer, the cyan part represents the *ReLU*, and the orange part represents the Norm.

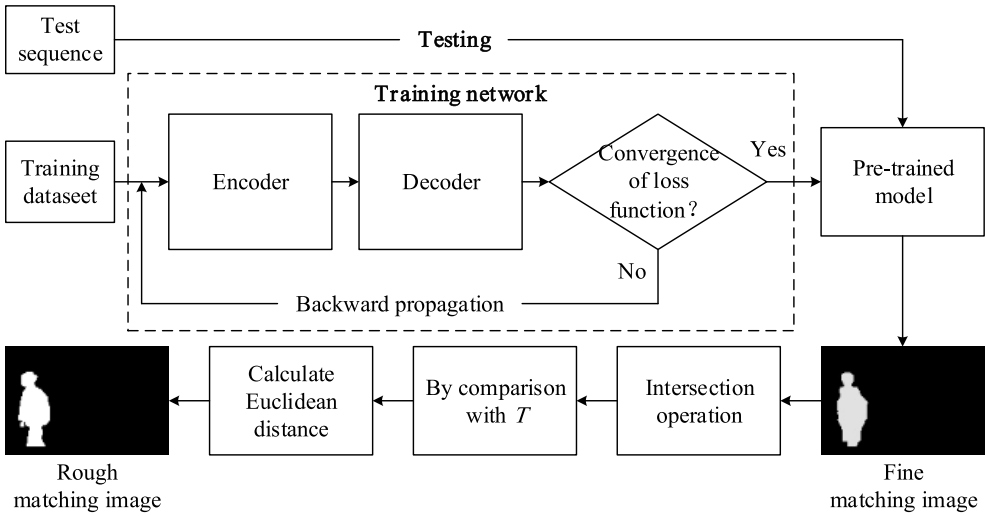


FIGURE 5. The proposed method flow chart.

dataset, we label the bounding box containing the object leaving margins around it.

2) NETWORK STRUCTURE

In the moving object detection processing, it is required to segment foreground object. Considering the diversity of the foreground, the network should be strong during classification. Since moving object detection is a binary classification task (foreground or background), too many deep layers will lead to the loss of structural information in the image, and information loss will also be caused by pooling. Therefore, our network encoder adopts the structure similar to ResNet-18, which shows high performance in terms of accuracy and processing time. The encoder extracts the features from inputs through several combinations of convolution, *Rectified Linear Units* (*ReLU*) [26]. Moreover, pooling layers are not included in the network. The network similar to ResNet-18, in addition to the first convolution layer, the filter size is 7 × 7, and the other layers are 3 × 3. The decoder (the green part of Fig. 4) is composed of 1 × 1 convolution layer and deconv layer. In order to speed up the network training, we execute the Batch Normalization (*BatchNorm*) [27] for each convolution layer in the encoder.

3) LOSS FUNCTION

Since the network outputs are binary values (foreground: 1, background: 0), we use element-wise Euclidean distance as the loss function $\mathcal{L}$. Let P the probability of output and $\mathcal{L}$ the value of the groundtruth label. The $\mathcal{L}$ score is then estimated by

$$\mathcal{L} = \frac{1}{NM} \sum_{x=1}^{N} \sum_{y=1}^{M} \|P(x, y) - L(x, y)\|_2 \tag{2}$$

where $L \in \{0,1\}$, N and M are the output sizes, which are the same size as the input image and vary with the input size. (x, y) is the pixel location in the probability map. In the network, the normalization in each layer can effective preadjusts the feature scale to [0, 1], resulting in stable loss convergence with the L_2 norm.

B. MOVING OBJECT DETECTION

1) ROUGH MATCHING

The overall descriptions of our proposed method are described in Fig. 5. The dotted box shows training network, in which training data contains image sequences and corresponding artificial labels. Firstly, they are input into the

network for training until the loss function converged. And then pre-trained model will be generated when network training is completed. Finally, finish the prediction of the classification of pixel points and output rough matching images by importing the image sequence into the pre-trained model.

We use the MXNet library [28] to train and test the network. To initialize the parameters of the convolutional layers in the encoder, we fine-tune the ResNet-18 parameters, which pre-trained on the large image dataset (ImageNet) for an image classification task. We train network via the SGD and set the initial learning rate to 10^{-2}, the decay factor of learning rate is 0.1, every 500 iterations, the learning rate decays once and finally decays to 10^{-6}, network is trained by 5000 iterations. We set the batahsize to 8.

2) FINE MATCHING

In the process of feature extraction by the encoder, the size of the feature map decreases gradually, which results in the loss of some structure and edge information. Therefore, the edge structure of the foreground mask in the predicted image is relatively rough. Moreover, since the network may segment similar objects that have not been moved, the Euclidean distance is used to further refine the motion pixels in the region.

Motion pixels in an image can be quickly detected by calculating the Euclidean distance between adjacent frames. We use three frames of the interval for calculation (frame F_i, frame F_{i-1} and frame F_{i-2}), which is inspired by [4], to eliminate global moving and reduce the error caused by slow moving. Since the network has segmented the foreground rough matching region and the background, the global moving of the background has been eliminated, it only needs to eliminate the error caused by the slow moving of the object. The encoder-decoder network output segmentation graph is binary, we first map it to the real image. Let F be the image sequences, F_i and F_{i-1} are adjacent frames, here i is the index of the image sequence. Firstly, the Euclidean distance between two adjacent images is calculated by

$$F_{i,i-1}(x, y) = |F_i(x, y) - F_{i-1}(x, y)| \tag{3}$$

where (x, y) is the pixel location in the image. Then, $F_{i,i-1}(x, y)$ and $F_{i-1,i-2}(x, y)$ are converted into binary $B_{i,i-1}(x, y)$ and $B_{i-1,i-2}(x, y)$ by comparison with the threshold T, here the value of T in this paper is adjusted according to different test image sequences.

$$B_{i,i-1}(x, y) = \begin{cases} 1 & if\ F_{i,i-1}(x, y) \geq T \\ 0 & otherwise \end{cases} \tag{4}$$

Finally, the intersection operation is performed on $B_{i,i-1}(x, y)$ and $B_{i-1,i-2}(x, y)$

$$B_{i-1}(x, y) = \begin{cases} 1 & if\ B_{i,i-1}(x, y) \bigcap B_{i-1,i-2}(x, y) = 1 \\ 0 & otherwise \end{cases} \tag{5}$$

IV. EXPERIMENTS RESULTS

In this section, we describe implementation in detail, and comparative *F-measure* show the effectiveness of the proposed method. The proposed method is compared with other 11 moving object detection algorithms (the improved Gaussian Mixture Model (GMM) [29], PBAS [30], ViBe [31], SOBS+ [32], 3dSOBS+ [33], LSD [34], LBP-P [35], SCS-LBP [36], HCS-LBP [37], VKS [38], DFB [39]).

We conduct a comparative experiment on 15 publicly available standard video datasets from I2R [40] and CDnet2014 [41], including AirportHall, Bootstrap, Curtain, Escalator, Fountain, ShoppingMall, Lobby, Campus, WaterSurface, Boats, Canoe, Fall, Fountain01, Founatin02, and Overpass. I2R dataset provides 20 frames for each video as groundtruth, and the CDnet2014 dataset provides the groundtruth for each frame in video sequences to evaluate the performance. These videos contain various difficult challenges, such as busy human flows (AirportHall, Bootstrap), multiple types of moving objects (Campus, Boats, Fall), moving cast shadows (AirportHall, Bootstrap, ShoppingMall), sudden illumination changes (Lobby), and dynamic background (Curtain, Escalator, Fountain, Campus, WaterSurface, Boats, Canoe, Fall, Fountain01, Fountain02, Overpass).

In order to further illustrate the advantages of the proposed method, the quantitative comparison is made by calculating the *F-measure* of 15 videos. The *F-measure* measures the weighted average of the Precision and Recall.

$$F - measure = \frac{2 \cdot recall \cdot precision}{recall + precision} \tag{6}$$

where *recall* measures the percentage of all pixels belonging to the moving object which is correctly detected, and *precision* measures the percentage of all detected pixels which belongs to moving object. They are defined as

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$precision = \frac{TP}{TP + FP} \tag{8}$$

where TP is the number of moving objects detected pixels corresponding to detected pixels in the groundtruth. FN is the number of moving object non-detected pixels corresponding to detected pixels in the groundtruth. FP is the number of moving objects detected pixels corresponding to nondetected pixels in the groundtruth.

A. COMPARATIVE EVALUATION

In this section, we test 15 videos where dynamic background, illumination variation and shadow exist. We compare the performance from both qualitative and quantitative aspects. Fig. 6 shows the qualitative comparison of 9 videos in the I2R dataset, and Fig. 7 shows the quantitative comparison of 6 videos in the CDnet2014 dataset. Table 1 presents all of the 15 videos' quantitative comparisons results.

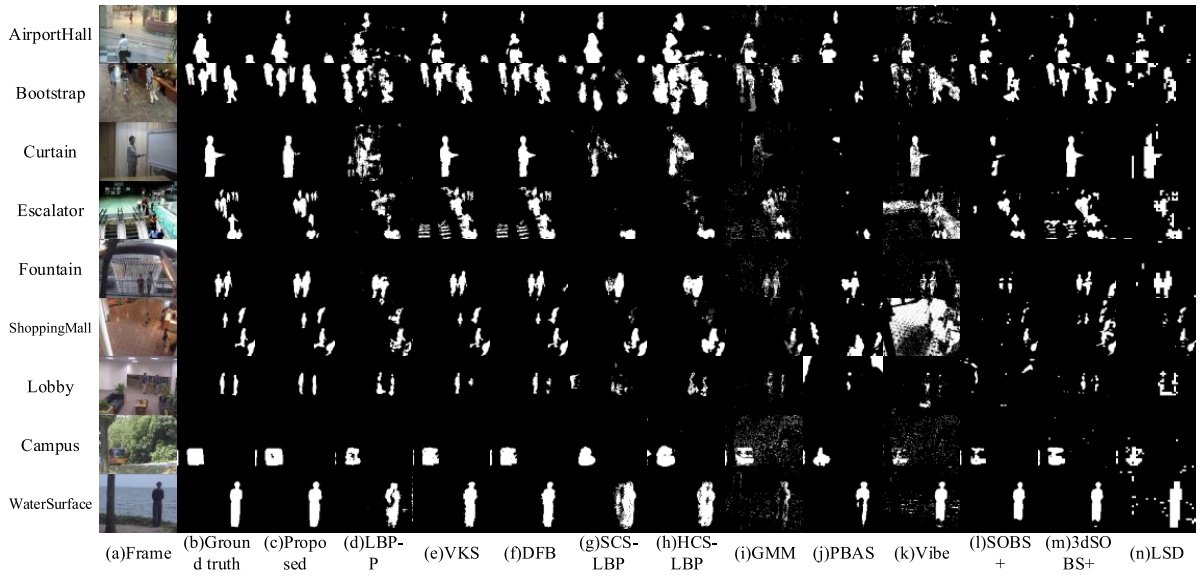


FIGURE 6. The results of 9 videos from the I2R dataset. Column (a) is the truth, (b) is groundtruth, (c) is the proposed method, and others are traditional algorithms.

FIGURE 7. The results of 6 videos from the CDnet2014 dataset. Column (a) is the truth, (b) is groundtruth, (c) is the proposed method, and others are traditional algorithms.

1) QUALITATIVE COMPARISON

Fig. 6 shows 9 videos from the I2R dataset. As for the ''AirportHall'' and the ''Bootstrap'', there are multiple moving objects and soft shadows. The proposed method, VKS, DFB and 3DSOBS+ have better detected the foreground, and the foreground of the proposed method is more complete. In the ''Curtain'', a person wearing a bright coat that resembles the curtain's color as it swings in the background. The proposed method, VKS, DFB and 3DSOBS+ have good detection effectiveness. In the ''Escalator'', the foreground is crowded and has a contrasting background, compared with other methods, the proposed method has better detection performance. In the ''Fountain'', the proposed method, VKS and DFB are all effective, but the proposed method loses some details. In the ''ShoppingMall'', some objects are standing still all the time, while others are moving all the time. The proposed method, VKS and DFB all detect foreground objects, but the proposed method loses some details. In the ''Lobby'', the illumination is suddenly changed, and only the proposed method yields good results, but the outline of the foreground object is incomplete. In the ''Campus'', the shaking of leaves results in continuous movement in the background. The proposed method, VKS, DFB and GMM all detect the foreground objects, but GMM has a large number of noise points, and VKS, DFB lose more foreground. In the ''WaterSurface'', water waves cause the background keep moving, and a person wearing flat texture of rippling water surface. The results show that there is ''ghost'' in the results of SCS-LBP and HCS-LBP, and there are noise points in Vibe and LSD. Most of the algorithms have effectively detected the foreground, and the proposed methods, VKS and DFB have similar detection performance.

Fig. 7 shows 6 videos from the CDnet2014 dataset, all of which contain a complex background. Both the ''Boats'' and

TABLE 1. Performance of F-measure (%) on dataset I2R and CDnet2014(1st: Bold, 2nd: Underline, 3rd: Incline). The scores on the videos in the AirportHall, Bootstrap, Curtain, Escalator, Fountain, ShoppingMall, Lobby, Campus, WaterSurface, Boats, Canoe, Fall, Fountain01 and Fountain02 are separately listed.

Video	Proposed method	LBP-P	VKS	DFB	SCS-LBP	HCS-LBP	GMM	PBAS	Vibe	SOBS+	3DSOBS+	LSD
AirportHall	72.42	50.15	71.34	70.77	46.65	42.21	53.14	58.43	61.16	62.84	**76.44**	*72.22*
Bootstrap	*70.70*	61.70	76.99	**77.87**	46.44	50.69	61.51	52.80	63.30	64.71	70.19	58.42
Curtain	*92.12*	57.26	**94.20**	93.16	39.66	62.07	40.29	57.25	82.49	40.84	91.98	85.57
Escalator	*64.62*	67.51	51.18	50.64	24.70	39.27	49.14	26.49	57.04	62.59	47.17	**72.14**
Fountain	86.64	79.09	**87.11**	85.28	73.33	70.54	51.64	72.60	55.80	47.75	58.55	*83.71*
ShoppingMall	*78.42*	60.11	**83.19**	80.75	61.51	53.61	64.13	71.53	68.54	65.58	65.48	73.62
Lobby	*76.92*	60.98	76.93	**77.85**	32.50	40.54	33.23	19.34	26.41	7.59	23.72	73.13
Campus	**94.40**	70.35	88.31	*87.56*	68.50	66.30	32.11	77.93	36.17	72.53	82.73	76.13
WaterSurface	**95.82**	81.52	*92.63*	93.86	59.60	66.83	36.75	74.79	86.02	85.00	88.67	90.5
Boats	**90.74**	56.98	76.58	79.52	44.92	60.60	31.61	21.24	63.30	58.01	87.62	*79.71*
Canoe	65.10	66.77	64.43	76.81	68.84	77.30	53.92	39.94	68.44	92.84	**94.83**	*83.51*
Fall	76.61	71.70	56.42	60.11	92.95	*80.30*	52.62	**95.02**	32.77	68.33	51.64	74.27
Fountain01	*38.76*	15.64	17.51	19.84	64.96	33.89	17.41	**81.96**	6.05	38.07	30.00	33.15
Fountain02	55.76	36.94	**92.49**	91.97	68.80	58.11	62.42	*90.60*	63.38	87.96	89.73	85.36
Overpass	*82.74*	78.43	59.55	66.02	80.92	85.77	64.27	66.31	70.51	82.76	**92.51**	69.51
mean	**76.12**	61.01	72.59	74.13	58.28	59.20	46.95	60.42	56.09	62.49	70.08	*74.06*

the ''Canoe'' represent ''the ship on the water surface'', water waves cause their background keep moving. The ''Canoe'' with several people boating, so it has a lot of detail on the outline. The results show that the proposed approach, LBP-P, VKS and DFB are superior to the other methods, but the results of the ''Canoe'' show that the proposed method is a little inferior to VKS and DFB in handling the details of the outline. The ''Fall'' describes a variety of moving objects, including people, cars, and trucks. The flickering leaves cause the background keep moving. The results show that the proposed method, SCS-LBP, HCS-LBP and SOBS+ all inhibited background motion. There are smaller moving objects at the ''Fountain01'' and the ''Fountain02'', and they all pass behind the fountain. The results show that the proposed method is inferior to VKS and DFB in dealing with small objects. As shown in the ''Overpass'', the proposed method and 3dSOBS+ are obviously superior to other algorithms.

2) QUANTITATIVE COMPARISON

As shown in Table 1, the performance of our proposed method is compared with 11 other methods. The proposed method shows the best performance in mean value, its *F-measure* mean is 76.12%. And it is worth noting that the performance of this method in most videos is ranked within top 3.

On the one hand, the data in table 1 shows that the proposed method can adapt to multiple variables in complex scenes. The ''Airport'' and the ''Bootstrap'' both have ''busy human flows'' and ''moving cast shadows'', the proposed method is 2nd and 3rd respectively in them. In the ''Lobby'', which have sudden illumination changes (it also affects the foreground and the background), the proposed method is 3rd, its *F-measure* just lower 0.93% than DFB (the 1st). In the ''Campus'', the scene contains a variety of variables (''multiple types of moving objects'' and ''dynamic background''), but the proposed method's *F-measure* is higher than the 2nd (VKS) by 6.09%, it is a significant increase. The proposed method reduces the influence of background on foreground segmentation through CNN autonomous learning of object features, so it has good performance in the dynamic background. Even in the ''Canoe'' and the ''Fall'', our results are acceptable. On the other hand, except for the proposed method, other algorithms commonly perform well just in some type of scenes, and they have very poor performance in the other scenes. Such as, VKS's *F-measure* is 94.20% in the ''Curtain'' but only 17.51% in the ''Fountain01'', SCS-LBP's *F-measure* is 92.95% in the ''Fall'' but 24.70% in the ''Escalator'', and 3DSOBS+'s *F-measure* is 94.83% in the ''Canoe'' but 23.72% in the ''Lobby''. However, compared with the aforementioned methods, the proposed method's *F-measure* indicates much stable in different scenes, and its mean *F-measure* increased by 1.99%∼29.17%. In summary, the proposed method can be applied to the scenes where dynamic background, illumination variation and shadow exist.

Moreover, the proposed method is more sensitive to the bigger foreground. Such as bigger foreground objects for the ''Campus'', the ''Watersurface'' and the ''Boats'', the proposed method's *F-measure* is 1st in these scenes. But, due to

the loss of information during network coding, the proposed method is now cannot show satisfactory performance when dealing with small objects and complex contour details (such as the ''Fountain01'' and the ''Fountain02''), and this issue will be improved in our following research work.

V. CONCLUSION

In this paper, we propose a moving object detection method via ResNet-18with encoder-decoder structure in complex scenes. The proposed method requires supervised training, and the network has the encoder-decoder structure. As for the input, the network accepts the object frame and corresponding artificial labels. Then, the feature vectors generated by the encoder are converted into segmentation maps by the decoder. Furthermore, the foreground mask is further precisely positioned by Euclidean distance. Experimental results show that the proposed method provides better performance than the conventional algorithms on the I2R and the CDnet2014 dataset. However, there is still some challenges exist in smaller objects detection, and we are trying to improve the performance in our following research work.

REFERENCES

[1] K. A. Joshi and D. G. Thakore, ''A survey on moving object detection and tracking in video surveillance system,'' *Int. J. Soft Comput. Eng.*, vol. 2, no. 3, pp. 44–48, 2012.

[2] M. Yazdi and T. Bouwmans, ''New trends on moving object detection in video images captured by a moving camera: A survey,'' *Comput. Sci. Rev.*, vol. 28, pp. 157–177, May 2018.

[3] J. S. Kulchandani and K. J. Dangarwala, ''Moving object detection: Review of recent research trends,'' in *Proc. IEEE Int. Conf. Pervasive Comput.*, Jan. 2015, pp. 1–5.

[4] H. Yin, Y. Chai, S. X. Yang, and X. Yang, ''Fast-moving target tracking based on mean shift and frame-difference methods,'' *J. Syst. Eng. Electron.*, vol. 22, no. 4, pp. 587–592, Aug. 2011.

[5] X. Zhou, C. Yang, and W. Yu, ''Moving object detection by detecting contiguous outliers in the low-rank representation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 597–610, Mar. 2013.

[6] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, ''Crowded scene analysis: A survey,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[7] T. Bouwmans, ''Traditional and recent approaches in background modeling for foreground detection: An overview,'' *Comput. Sci. Rev.*, vol. 11, pp. 31–66, May 2014.

[8] T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah, ''Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset,'' *Comput. Sci. Rev.*, vol. 23, pp. 1–71, Feb. 2016.

[9] A. Sobral and A. Vacavant, ''A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos,'' *Comput. Vis. Image Understand.*, vol. 122, pp. 4–21, May 2014.

[10] S. M. Roy and A. Ghosh, ''Real-time adaptive histogram min-max bucket (HMMB) model for background subtraction,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1513–1525, Jul. 2017.

[11] M. Chen, X. Wei, Q. Yang, Q. Li, G. Wang, and M.-H. Yang, ''Spatiotemporal GMM for background subtraction with superpixel hierarchy,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1518–1525, Jun. 2018.

[12] C. Han, Y. Duan, X. Tao, and J. Lu, ''Dense convolutional networks for semantic segmentation,'' *IEEE Access*, vol. 7, pp. 43369–43382, 2019.

[13] L. Fan, H. Kong, W.-C. Wang, and J. Yan, ''Semantic segmentation with global encoding and dilated decoder in street scenes,'' *IEEE Access*, vol. 6, pp. 50333–50343, 2018.

[14] Q. Zhou, P. Shi, H. Liu, and S. Xu, ''Neural-network-based decentralized adaptive output-feedback control for large-scale stochastic nonlinear systems,'' *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1608–1619, Dec. 2012.

[15] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 2015, pp. 3431–3440.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, ''SegNet: A deep convolutional encoder-decoder architecture for image segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[17] X. Bian, S. Lim, and N. Zhou, ''Multiscale fully convolutional network with application to industrial inspection,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–8.

[18] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon, ''Pixel-level matching for video object segmentation using convolutional neural networks,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2167–2176.

[19] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, ''Detect globally, refine locally: A novel approach to saliency detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3127–3135.

[20] F. Qi, C. Lin, G. Shi, and H. Li, ''A convolutional encoder-decoder network with skip connections for saliency prediction,'' *IEEE Access*, vol. 7, pp. 60428–60438, 2019.

[21] J. Dolz, C. Desrosiers, and I. B. Ayed, ''3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study,'' *NeuroImage*, vol. 170, pp. 456–470, Apr. 2018.

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

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

[24] C. Xiang, L. Zhang, Y. Tang, W. Zou, and C. Xu, ''MS-CapsNet: A novel multi-scale capsule network,'' *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1850–1854, Dec. 2018.

[25] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, ''A review on deep learning techniques applied to semantic segmentation,'' 2017, *arXiv:1704.06857*. [Online]. Available: https://arxiv.org/abs/1704.06857

[26] V. Nair and G. E. Hinton, ''Rectified linear units improve restricted Boltzmann machines,'' in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[27] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167

[28] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, ''MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems,'' 2015, *arXiv:1512.01274*. [Online]. Available: https://arxiv.org/abs/1512.01274

[29] Z. Zivkovic, ''Improved adaptive Gaussian mixture model for background subtraction,'' in *Proc. 17th Int. Conf. Pattern Recognit.*, Aug. 2004, pp. 28–31.

[30] M. Hofmann, P. Tiefenbacher, and G. Rigoll, ''Background segmentation with feedback: The pixel-based adaptive segmenter,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 38–43.

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

[32] L. Maddalena and A. Petrosino, ''The SOBS algorithm: What are the limits?'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 21–26.

[33] L. Maddalena and A. Petrosino, ''The 3dSOBS+ algorithm for moving object detection,'' *Comput. Vis. Image Understand.*, vol. 122, pp. 65–73, May 2014.

[34] X. Liu, G. Zhao, J. Yao, and C. Qi, ''Background subtraction based on low-rank and structured sparse decomposition,'' *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.

[35] M. Heikkila and M. Pietikäinen, ''A texture-based method for modeling the background and detecting moving objects,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, Apr. 2006.

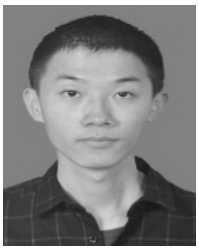
[36] G. Xue, J. Sun, and L. Song, ''Dynamic background subtraction based on spatial extended center-symmetric local binary pattern,'' in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 1050–1054.

[37] G. Xue, L. Song, J. Sun, and M. Wu, ''Hybrid center-symmetric local pattern for dynamic background subtraction,'' in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.

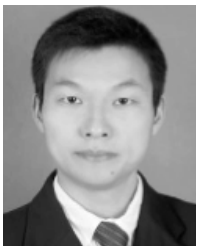
[38] M. Narayana, A. Hanson, and E. Learned-Miller, "Background modeling using adaptive pixelwise kernel variances in a hybrid feature space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2104–2111.
[39] M. Narayana, A. R. Hanson, and E. G. Learned-Miller, "Improvements in joint domain-range modeling for background subtraction," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–11.
[40] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.
[41] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 393–400.

XIANFENG OU received the B.S. degree in electronic information science and technology and the M.S. degree in communication and information system from Xinjiang University, Urumchi, China, in 2006 and 2009, respectively, and the Ph.D. degree in communication and information system from Sichuan University, Chengdu, China, in 2015. He was a Visiting Researcher with the Internet Media Group, Polytechnic di Torino, Turin, Italy, from January to April 2014, working on distributed video coding and transmission. He is currently an Associate Professor with the School of Information Science and Engineering, Hunan Institute of Science and Technology. His main research interests include machine vision and artificial intelligence, object detection, and image and video coding process technologies.

PENGCHENG YAN received the B.S. degree in electronic information engineering, in 2017. He is currently a Postgraduate Student with the Machine Vision and Artificial Intelligence Research Center, Hunan Institute of Science and Technology. His main research interests mainly include face detection, and moving object detection and recognition.



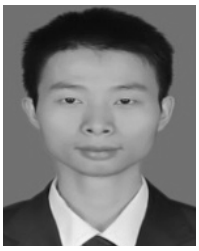
YIMING ZHANG received the B.S. degree in electronic information engineering, in 2018. He is currently a Postgraduate Student with the Machine Vision and Artificial Intelligence Research Center, Hunan Institute of Science and Technology. His main research interests mainly include hyper spectral image object detection, and moving object detection and recognition.



BING TU received the B.S. degree from the Guilin University of Technology, Guilin, China, in 2006, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China. He is currently pursuing the Ph.D. degree in electrical engineering with Hunan University, Changsha, China. In 2013, he joined the College of Information and Communication Engineering, Hunan Institute of Science and Technology. Since September 2013, he has been a Lecturer with the Department of Computer Science. He is engaged in image fusion, pattern recognition, hyper spectral image classification, and image and video coding technologies.

GUOYUN ZHANG received the B.S. degree in automation from Xiangtan University, in 1993, and the M.S. and Ph.D. degrees in control theory and control engineering from Hunan University, Changsha, China, in 2000 and 2003, respectively. He was a Visiting Researcher with George Fox University, from January to June 2014. He is currently a Professor with the Hunan Institute of Science and Technology. His research interests include image processing, computer vision, and pattern recognition.

JIANHUI WU received the M.S. and Ph.D. degrees in optical information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2006 and 2009, respectively. He was a Visiting Researcher with the Computer Vision Group, University of Nevada, Reno, USA, from August 2015 to August 2016, working on object detection and recognition. His main research interests include object detection and object recognition.



WUJING LI received the Ph.D. degree in computer science and technology from Sichuan University, Chengdu, China, in 2012. He joined the School of Information Science and Engineering, Hunan Institute of Science and Technology, in 2017. His main research interests include image enhancement and restoration, machine vision, and image recognition.