Deep Video Foreground Target Extraction With Complex Scenes

Die Li
School of Information Science and Engineering,
Yunnan University
Kun Ming, China
e-mail: Annie5242@163.com

Murong Jiang
School of Information Science andEngineering,
Yunnan University
Kun Ming, China
e-mail: jangmr@ynu.edu.cn

Yuan Fang
School of Information Science and Engineering,
Yunnan University
Kun Ming, China
e-mail: fangyhag@gmail.com

Yaqun Huang
School of Information Science andEngineering,
Yunnan University
Kun Ming, China
e-mail: huangyq@ynu.edu.cn

Chunna Zhao
School of Information Science andEngineering,
Yunnan University
Kun Ming, China
e-mail: chunnazhao@163.com

Abstract—**The foreground extraction of video is derived from frame difference and background subtraction. These methods have always relied on the temporal consistency of successive video frames. If the video consistency is suddenly destroyed, such as an object being occluded, some frames are lost, or the camera is shaken, etc., the results of foreground extraction may be significantly degraded. In this paper, our extends One-Shot Video Object Segmentation by adding two branchs for Positioning foreground targets and propagating foreground semantic information. These two branches are parallel to the existing foreground segmentation branches, which facilitates efficient iterative and fine-tuning of the foreground segmentation branches. This foreground extraction method is different from the previous background subtraction and frame difference methods. Our method does not rely on the time information and only needs to give the manual annotation of one frame, which can extract all the specific foreground in the video sequence. Then, the coloring can get complete foreground information for the separated target. Experiments show that the foreground extraction algorithm applied in this paper has better robustness in dynamic background, camera shake, intermittent motion of objects, low contrast, and so on.**

***Keywords*-foreground extraction; background subtraction; dynamic background; intermittent motion**

I. Introduction

In video sequences, algorithms separate the mutative foreground from a complex background is a major problem in computer vision, and it is often used as a cornerstone in advanced applications[1-5]. At present, various methods were proposed for this problem and the most widely used one was background subtraction. The basic idea of background subtraction is similar to the frame difference method. They both use the difference operation for different images to extract the target area. However, the background subtraction and the interframe difference method are different. The frame difference method subtracts the current frame from the adjacent frame, while background subtraction subtracts the current frame from a constantly updated background model, and then it extracts the moving targets in the difference images. With automatic processing techology for surveillance video gotten more and more attention in computer vision, the number of researchers whose study based on background subtraction were also increasing. Vibe[1], PBAS[2], SUBSENSE[3] were the most widely used background subtraction. Nowadays, with the development of deep learning in computer vision, the background modeling method based on deep neural network has appeared[4-6]. Background modeling methods are very effective for static background processing. But these methods are not ideal for dealing with dynamic background, static or intermittent foreground, and low frame rate video and so on. Based on these problems, this paper applies One-Shot Video Object Segmentation[7](OSVOS) to the foreground extraction. This method does not need to model the background. It only needs to provide the manual annotation of one frame in the video, which combines the frame image to select a specific foreground target. Then, the foreground features are automatically propagated backwards, and the foreground in subsequent frames is separated separately. The overview of our foreground extraction methods is shown in Figure 1.

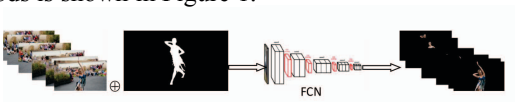

Figure 1. Overview of the method

The foreground extraction method studied in this paper is a semi-supervised learning algorithm based on the complete convolutional neural network architecture. The model is derived from Ref[7], and it can be subdivided into three networks. They are basic network, main network and finetuning network. The idea of the model is as follows. First, the basic network was learned semantic information on ImageNet in advance, and the objects of the image were marked. Then, the DAVIS dataset were trained on the main network. And all the objects of the video frame were splited instead of specific object. Finally, further finetuning was

978-1-5386-7413-0/18/$31.00 ©2018 IEEE
DOI 10.1109/SNSP.2018.00089

performed. This paper improved the basis of this model and applied it to the foreground extraction of complex background video. The main improvements are as follows:

1. Adding foreground contour branch and instance segmentation branch.

In order to the foreground to be more accurately positioned, and the semantic information of the foreground is automatically propagated backwards, we proposed to add two parallel network to the main network. One was trained to detect the foreground target outline. The other was trained to get the foreground mask and category label of each frame.

2. Adding training set.

In order to test the foreground extraction effect of the model in dynamic background, jitter, and intermittent monitoring video, we proposed that the main network should be trained in a surveillance video dataset with a complex background. So the main network was trained simultaneously in the DAVIS data set and the 2014 dataset of CDNet.

3. Coloring the segmented foreground.

Combined with the color channel of each pixel, the segmented foreground is colored.

The rest of the paper is organized as follows. Section 2 gives a brief review of related works. The Improved algorithm is elaborated in Section 3. In section 4, the experimental and comparison results are presented. Section 5 concludes the paper.

II. Related Content

Deep learning is a new field in machine learning research. In recent years, it is obvious that deep learning has been applied in images, sounds and texts. Deep neural networks, especially convolutional neural networks (CNNs), have made a qualitative leap in the development of image classification [8,9] and target detection [10-12]. Recently, many researchers have applied CNN to image semantic segmentation [13,14]and instance segmentation [15-20]. The essence of semantic segmentation is that an image is segmented according to its semantics, and usually each pixel is classified. Semantic segmentation does not distinguish between target instances. If a target has two or more similar targets in an image, those targets will only be assigned the same category tag. However, a particular instance of each type of target in the image can be assigned a category tag through an instance segmentation algorithm. In the foreground extraction algorithm of this paper, we want to extract not only a type of target, but a single target with specific information. Therefore, an instance segmentation network branch is added to our algorithm.

The essence of instance segmentation is the combination of instance perception and semantic segmentation. It mainly includes two parts: target detection and semantic segmentation. In detail, different instances are framed from the image by the target detection method. Then, within different instance regions, each instance is tagged pixel by pixel through a semantic segmentation method and given its category tag. Ref.[17] uses FPN for target detection and semantic segmentation by adding mask branches to obtain three output vectors, categories, bounding boxes and binary mask Mask for each ROI. This method has a small overhead and can be used to detect various poses specific to the instance. In this paper, the semantic segmentation is applied to the foreground extraction of video. The segmented frame with the mask and category is matched with the first frame of the manual annotation frame, and the particular instance that we needed is selected and its semantic signal is automatically propagated to subsequent frames. The instance segmentation network branch is combined with the foreground information extracted in the first round, and the accuracy of the extracted foreground object can be improved.

Image edges often carry a lot of information about an image. The edges exist in the irregular structure and unevenness of the image. That is to say, at the abrupt points of the signal, these points give the location of the image outline. When performing foreground extraction, it is important to be able to pinpoint the contour of the target. In order to align the foreground of the segmentation with the objects in the original image frame, we have added a contour branch network. This network is different from the general edge detection method. We only extract the outline of the target instance and do not deal with the background and internal texture details.

III. Improved Method

The foreground extraction algorithm in this paper applies the same end-to-end complete convolutional neural network architecture as OSVOS algorithm [7]. As the CNN network is gradually fine-tuned, a powerful appearance model is constructed for video object segmentation. The network model mainly includes three networks. They are base network, main network, and test network. Our improvement mainly includes three aspects: adding the data set, adding the positioning branch network and the semantic selection and propagation branch network, And coloring the segmented foreground. The network diagram of this paper is shown in Figure 2.

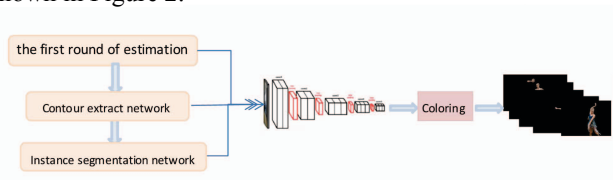

Figure 2. Network diagram

A. *Training and testing details*

In the pre-training phase, the base CNN of our architecture is pre-trained on ImageNet for image labeling. Then the main network is further trained on the binary masks of the training set of DAVIS2016 and the CDNet. At the same time, the outer contour extraction branch network and the instance segmentation branch network are trained in both data sets. At this stage, our main network is able to extract all instances in the video frame instead of a single instance. In order to obtain the foreground goal of a specific instance, the segmentation result is combined with two

branch networks for fine tuning. And the training before fine tuning is called the first round of foreground estimation.

In the testing phase, segmenting a particular entity in a video, given the image and ground-truth of the first frame. We proceed by further fine-tuning the main network for the particular image and ground-truth pair, and then testing on the entire sequence, using the new weights. At last, the foreground of the extraction is colored. Finally, the foreground is exactly the same as the target in the input video frame.

B. Contour branch

The extracted foreground target through the first round of foreground estimation is not able to be accurately positioned. Contour detection can locate the exact position of the outer contour of the foreground. Therefore, the outer contour extraction branch (referred to as a contour branch) network is added. The contour branch proposed in this paper is the same network architecture as our foreground extraction network. Selection of foreground in contour branch network, we applied interactive segmentation algorithm[21] to select foreground targets. The binary contour mask is a self-built label data set by applying the Canny operator[22] and the Ground-Truth of foreground. Our contour branch also require only one frame of binary contour mask to extract the outer contour of a particular instance of all frames. The resulting outline does not contain background edges and foreground texture details, only included the outline of the specific instance. This outline contains the edge information of the foreground target that can be pinpointed to the specific location of the foreground. The contour branch network architecture is shown in Figure 3.

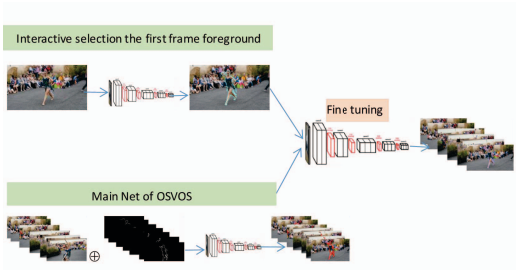

Figure 3. Contour branch network architecture

The extracted foreground target through the first round of foreground estimation is continuously iterated and fine-tuned by the contour branch network, and it can be obtained that the foreground target with accurate positioning. In this paper, the contour extraction branch is processed in parallel with the main network branch. The accuracy is greatly improved on the basis of less reduction in operational efficiency.

C. Instance segmentation branch

We applies the Mask-RCNN algorithm[17] as our example segmentation algorithm. The output of an instance segmentation network is a set of video frames containing a binary mask and its category labels. We match the output of instance segmentation network with the first frame manual annotation. The objects we are interested in will be automatically selected and their characteristics passed to subsequent frames. This process is divided into two parts, semantic selection and semantic propagation[19]. The first frame of the segmentation frame is overlapped with the first frame manual annotation, and the object that we are interested is selected. This step is called semantic selection. The number of instances and categories in this process are determined by the manual annotation frame. Once the semantic selection is done on the first frame, the information is propagated throughout the video. This step is occurred in subsequent frames and called semantic propagation.

D. Coloring

In this paper, the final step in the foreground extraction is to color the extracted binary mask. The method of coloring has a fast bilateral solution [24] and color padding. In this paper, we applied color padding method. Its process is as follows. Firstly, the pixel position of the extracted binary mask is matched with the input video frame position. And then the extracted binary mask is filled with colors according to the gray values of the RGB channels of the foreground pixels. With this simple method, a good coloring effect is obtained.

IV. EXPERIMENTS

A. Our Results

Our algorithm was tested on the DAVIS test set and the 2014 test dataset, and we achieved the ideal foreground extraction effect in various types of video. Our performance is better than most classical and representative background subtraction models. Some experimental results are shown in Figure 4.

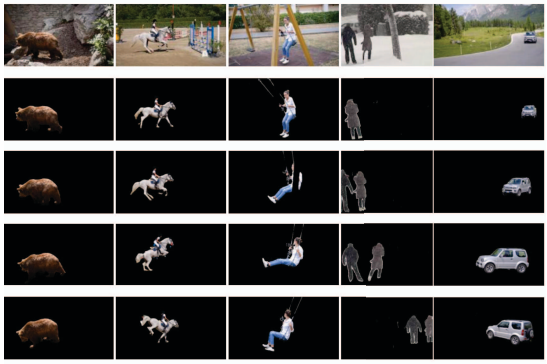

Figure 4. The foreground extraction effect of our algorithm

Figure 4 shows the foreground extraction results of our algorithm in different scenarios (people, vehicles, animals, etc.). The first line is the first frame of each video, and the following is the extraction effect of subsequent frames. Although the foreground form in some videos has changed a lot, the application of this algorithm has achieved considerable results.

B. Comparison results

In order to evaluate the performance of our algorithm, We compared our algorithm with several representative background subtractions. The background modeling methods are Type-2FuzzyGMM-UV with MRF algorithm[24] (Referred to as T2FMRF-UV), PBAS algorithm[2], SuBSENSEBGS algorithm[3]. Results are shown in Figures 5, 6, and 7.

Figure 5. Comparison of bad weather experiments

Figure 5 is a surveillance video scene on a snowy street at night. We extracted the foreground of the target vehicle on the lane. From top to bottom, the extraction effect of T2-FMRF-UV, PBAS, SuBSENSEBGS, our algorithm before coloring and our algorithm after coloring. As can be seen from the experimental results, there are a lot of holes in the foreground target extracted by PBAS algorithm and T2FM RF-UV algorithm. The PBAS algorithm is better than the T2FMRF-UV algorithm, because the PBAS algorithm can basically extract the overall structure of the vehicle at the appropriate lens. The SuBSENSEBGS algorithm works better than the above two background subtraction algorithms. It basically extracts foreground targets from various angles, but there are still holes. Our algorithm is able to extract the complete contour of the vehicle without voiding. We conducted a comparative evaluation of the results as shown in Table 1.

Table 1. comparative evaluation of results

Method	advantage	disadvantage
T2FMRF-UV	Good noise immunity	Low accuracy
PBAS	Good noise immunity	Low accuracy
SuBSENSEBGS	good noise immunity and high accuracy	Structural incompleteness
Our algorithm	good noise immunity , high accuracy and good integrity	GPU environment is needed

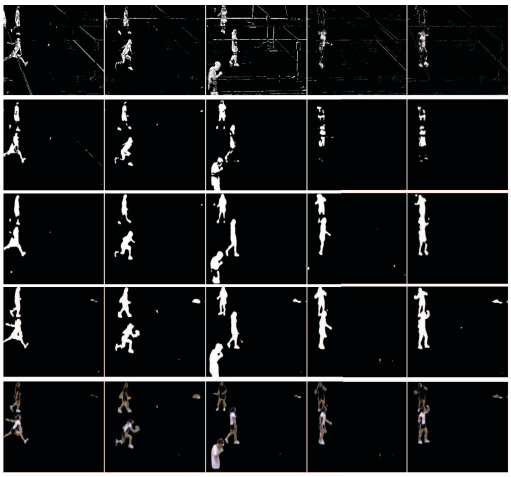

Figure 6. Comparison of jitter video experiments

Figure 6 is a video of a multiplayer scene taken while the camera is shaking. From top to bottom, the extraction effect of T2FMRF-UV, PBAS, SuBS-ENSEBGS, our algorithm before coloring and our algorithm after coloring. It shows the results of T2FMRF-UV algorithm, the PBAS algorithm, the SuBSENSEBGS algorithm, and our algorithm before and after coloring. From figure 6, we know that PBAS has serious hollow phenomena, and some foregrounds are missing, such as the fourth and fifth columns of the second row. The T2FMRF-UV algorithm extracts the edges of the scene but extracts a large number of unnecessary background edges. The SuBSENSEBGS algorithm has achieved good results in this motion scene, but there are still some foregrounds were built into the background to make the foreground appear hollow. In the sixth line, some backgrounds are also extracted, but they have little effect on the foreground characters. And the moving characters are structurally intact in different postures, and there is almost no void phenomenon. We conducted a comparative evaluation of the results as shown in Table 2.

Table 2. comparative evaluation in jitter video

Method	Experimental effect
T2FMRF-UV	Low noise immunity, Low accuracy and high error classification ratio
PBAS	Low accuracy , low foreground integrity
SuBSENSEBGS	Low accuracy and high error classification ratio
Our algorithm	good noise immunity , high accuracy and good integrity

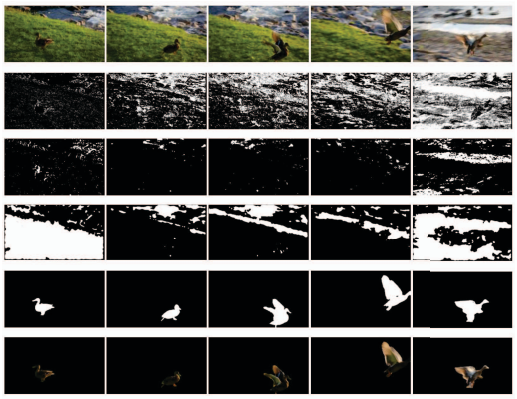

Figure 7. Dynamic background experiment comparison

Figure 7 shows the process of wild ducks moving from rest, walking, to take off in a natural scene with dynamic background and strong illumination. The first line is the original video frame, and the other lines are the T2FM RF-UV, PBAS, SuBSENSEBGS, our algorithm before coloring and our algorithm after coloring. The results of the three background subtraction methods shown are not ideal. The SuBSENSEBGS algorithm misinterprets the foreground as the background (the wild duck in the picture above) when the first few frames contain the foreground and the foreground moves slowly. In the case of a dynamic background, the dynamic background is extracted as the foreground (The water flow in the above picture). The same PBAS and T2FMRF-UV algorithms make it difficult to accurately extract foreground targets for dynamic backgrounds and fast frame rate scenarios. The algorithm of this paper has strong robustness to high frequency and dynamic background. Even in the case of motion blur, the target foreground can be extracted as shown in the first column of the fifth and sixth rows on the right side of Figure 7. We conducted a comparative evaluation of the results as shown in Table 3.

Table 3. comparative evaluation in dynamic background video

Method	Experimental effect
T2FMRF-UV	Low noise immunity, low accuracy positioning and high error classification ratio
PBAS	Low noise immunity, low accuracy positioning and high error classification ratio
SuBSENSEBGS	Low noise immunity, low accurate positioning and high error classification ratio
Our algorithm	good noise immunity , precise positioning, high accuracy and good integrity

In Tables 1 to 3, we can see from the performance comparison that our algorithm can accurately locate and error-free classification and immune noise in the background of bad weather, jitter camera and dynamic strong illumination. However, the other three algorithms have lower accuracy and higher misclassification.

C. *Quantitative assessment*

In order to quantitatively evaluate our algorithm and other three foreground extraction algorithms, we conducted a large number of comparative experiments on the DAVIS2016 dataset and the 2014 dataset. In this paper, we extracts 9 sets of video from DAVIS2016 data set (first nine in Figure 8) and 4 sets of video from the 2014 dataset (the last four in Figure 8). The performances of the four algorithms are evaluated by calculating the average dice coefficients of the foreground extraction results in the video. The evaluation effect is shown in Figure 8.

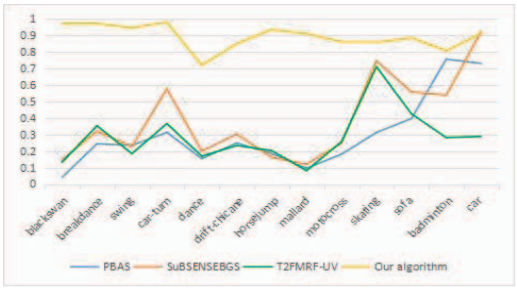

Figure 8. Comparison of average dice coefficients

In Figure 8, The T2FMRF-UV, PBAS, and SuBSENSEBGS algorithms are extracted better in surveillance video than in natural scene video. But their performance is not as good as our algorithm. Especially in the natural scene, the average dice coefficient of our algorithm is up to 90%, while the other three algorithms are almost less than 40%. The SuBSENSEBGS algorithm is more prominent in carturn and car video, but it is still lower than our algorithm. In addition, we classify the videos in the two databases by dynamic background, camera jitter, interm object motion, shadow, and more. After extracting the foreground of each type of video by using four algorithms, the average value of each type of dice coefficient is obtained. The results are shown in Table 4.

Table 4. Average dice coefficients in each type of scenario

Method	Camera jitter	Dynamic background	Interm.object motion	shadow
T2FMRF-UV	0.32	0.26	0.40	0.31
PBAS	0.75	0.37	0.43	0.45
SuBSENSEBGS	0.58	0.46	0.56	0.49
Our algorithm	0.81	0.91	0.88	0.90

Through experimental verification, we know that the T2FMRF-UV algorithm can detect the target foreground in a typical dynamic background video. But the general dynamic scene video extraction effect is not ideal. The PBAS algorithm can eliminate a certain degree of shadow and adapt to illumination changes, but the accuracy is low, and it is easy to update the slowly moving foreground to the background in interm object motion video. The SUBSENSE algorithm has improved in anti-jitter and shadow elimination but not as good as our algorithm. Our algorithm has achieved very good results in dynamic background video, camera jitter video, shadow, intermittent and so on.

V. Conclusion

In this paper, the OSVOS video segmentation algorithm was applied to the video foreground extraction of natural scenes and surveillance video. The foreground extraction algorithm is robust to dynamic background, intermittent objects, high frame rate, occlusion and camera shooting scenes. In this paper, the algorithm was compared with the existing advanced background subtraction algorithms through experiments. The experimental results showed that our algorithm have good noise immunity, high accuracy and good integrity. It has better performance than other algorithms in dynamic background and jitter video and so on. Our algorithms can be used not only for real-time monitoring video target extraction, but also for natural images. When this technology is mature enough, we believed that it can apply one photo of the suspect to automatically find and extract the video frame containing the suspect from the monitoring video of the incident location. Because of its good performance in target intermittent video, it is even used for tracking and identification of aerospace equipment.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61862062) and Yunnan University Level Project(Y2000211). We gratefully acknowledge all supporters for this research.

References

[1] O.Barnich and M.Van Droogenbroeck.ViBe: A Universal Background Subtraction Algorithm for Video Sequences. Image Processing, IEEE Transactions on, 20(6):1709-1724,june 2011

[2] M. Hofmann, P. Tiefenbacher, G. Rigoll, "Background segmentation with feedbacks: The pixel-based adaptive segmenter", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, pp. 38-43, Jun. 2012.

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

[4] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puigtt, Y. Ruichek, "BSCGAN: Deep Background Subtraction with Conditional Generative Adversarial Networks", Image Processing (ICIP) 2018 25th IEEE International Conference on, pp. 4018-4022, 2018.

[5] D. Zeng, M. Zhu, "Correction to "Background Subtraction Using Multiscale Fully Convolutional Network'"", Access IEEE, vol. 6, pp. 32225-32225, 2018.

[6] S.Jain,B.Xiong,K.Grauman. FusionSeg Learning to combine motion and appearance for fully automatic segmention of generic objects in videos,2017

[7] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taix´e, D. Cremers, and L. Van Gool. One-shot video object segmentation. In CVPR, 2017.

[8] L. Shen, S. Jia, "Three-Dimensional Gabor Wavelets for Pixel-Based Hyperspectral Imagery Classification", IEEE Trans. on Geoscience and Remote Sensing, vol. 49, no. 12, pp. 5039-5046, 2011.

[9] J. Li, X. Huang, P. Gamba et al., "Multiple feature learning for hyperspectral image classification", IEEE Trans. on Geoscience and Remote Sensing, vol. 53, no. 3, pp. 1592-1606, 2015.

[10] R. Girshick. Fast R-CNN. In ICCV, 2015. 1, 3

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy and S. Reed. SSD: Single shot multibox detector. In ECCV, 2016.

[12] J.Dai,Y.Li,K.He,andJ.Sun.R-FCN:Object detection via region-based fully convolutional networks. InNIPS,2016.

[13] L.C.Chen,G.Papandreou,L.Kokkinos and K Murphy. DeepLab:Semantic Image Segmentation with Deep Convolutional Nets,Atrous Convolution,and Fully Connected CRFs.Computer Vision and Pattern Recognition.2017

[14] J. Long,E.Shellhamer and T. Darrel. Fully Convolutional Networks for Semantic Segmentation.IEEE conference on Representation Learning,2016

[15] J. Dai, K. He, J. Sun, "Instance-aware semantic segmentation via multi-task network cascades", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3150-3158, Jun. 2016.

[16] J. Dai, K. He, J. Sun, "Convolutional feature masking for joint object and stuff segmentation", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3992-4000, Jun. 2015.

[17] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask R-CNN. In ICCV, 2017.

[18] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, "Simultaneous detection and segmentation", Proc. Eur. Conf. Comput. Vis., pp. 297-312, 2014.

[19] K.-K. Maninis, S. Caelles,Y.Chen, J. Pont-Tuset, L. Leal-Taix´e, D. Cremers, and L. Van Gool.Video Object Segmentation Without Temporal Information. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018.

[20] T.Bouwmans, F. E.Baf, B.Vachon. Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. Recent Patents on Computer Science, Bentham Science Publishers, 2008, 1 (3), pp.219-237

[21]N.Xu, B. Price,and T.Huang.Deep Interactive Object Selection.CVPR.2016.pp374-381

[22] Canny J. A computational approach to edge detection[J].IEEE Trans. Pattern Anal. Mach. Intell, 1986,8 (6):679-698.

[23] J.T.Barron,B.Poole. The fast bilateral solver. InECCV, 2016

[24] Z.Zhao,T.Bouwmans, X.Zhang,Y.C.Fang. A Fuzzy Background Modeling Approach for Motion Detection in Dynamic Backgrounds. In CMSP,2012.