Moving Object Segmentation using GAN

A dissertation report submitted in fulfillment of the requirements for the degree of Bachelor of Technology

by

ASHISH GOYAL (2016UCP1100)
SWARAJ THAKRE (2016UCP1663)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MALAVIYA NATIONAL INSTITUTE OF TECHNOLOGY

MAY 2020

Certificate

We,

ASHISH GOYAL (2016UCP1100)
SWARAJ THAKRE (2016UCP1663)

Declare that this thesis titled, “Moving Object Segmentation using GAN” and the work presented in it is our own. I confirm that:

- This project work was done wholly or mainly while in candidature for a B.Tech. degree in the department of computer science and engineering at Malaviya National Institute of Technology Jaipur (MNIT).

- Where any part of this thesis has previously been submitted for a degree or any other qualification at MNIT or any other institution, this has been clearly stated. Where we have consulted the published work of others, this is always clearly attributed. Where we have quoted from the work of others, the source is always given. With the exception of such quotations, this Dissertation is entirely our own work.

- We have acknowledged all main sources of help.

Signed : Ashish Goyal and Swaraj Thakre

Date : 20 May 2020

Dr. Namita Mittal
Associate Professor
Department of Computer Science and Engineering
Malaviya National Institute of Technology Jaipur

May 2020

Abstract

The moving object segmentation (MOS) in videos with bad climate unpredictable movement of articles, dynamic background, jittering of camera, thermal videos etc. situations is still an open issue for computer vision applications.

In various computer vision tasks, for example video synopsis, vehicle navigation, objects tracking, video surveillance, person re-identification, magnetic particle imaging, traffic monitoring etc. moving object segmentation is used

The moving object segmentation task is comprehensively classified as background subtraction, saliency estimation, frame difference, pixel-level, optical flow, region level, and deep learning based techniques. In literature, the most of work accomplished for MOS is with background subtraction technique.

The dynamic background in videos (falling of snow, waving of tree and sporadic movement of item) influence the exactness of background subtraction based procedures. In reasonable situations (like waving of tree, waving of water, bad climate, low contrast, sporadic movement of item, shadow effect and unexpected light changes), the classification of pixels into foreground or background is a difficult task for MOS. Therefore to address these challenges, we propose a novel deep learning based approach wherein a background is evaluated for the video utilizing few initial frames of the video and the work of foreground segmentation is finished utilizing GAN's.

Acknowledgements

We want to give our genuine thanks and appreciation to our esteemed supervisor Dr. Namita Mittal**,** Associate Professor, Dept. of Computer Science & Engineering, Malaviya National Institute of Technology (MNIT) Jaipur, for providing her valuable direction & encouraged us to learn new things and helped in making this project a wonderful learning experience.

We want to express our profound thankfulness towards Dr. Pilli Emmanuel Shubhakar, HOD of Computer Science & Engineering for the guidance, support and for permitting us to utilize all the necessary facilities of the institution.

Finally we offer our true thanks to all the staff individuals from Computer Science and Engineering Department who assist and guide us during this course of project work.

ASHISH GOYAL (2016UCP1100)
SWARAJ THAKRE (2016UCP1663)

Table of Contents

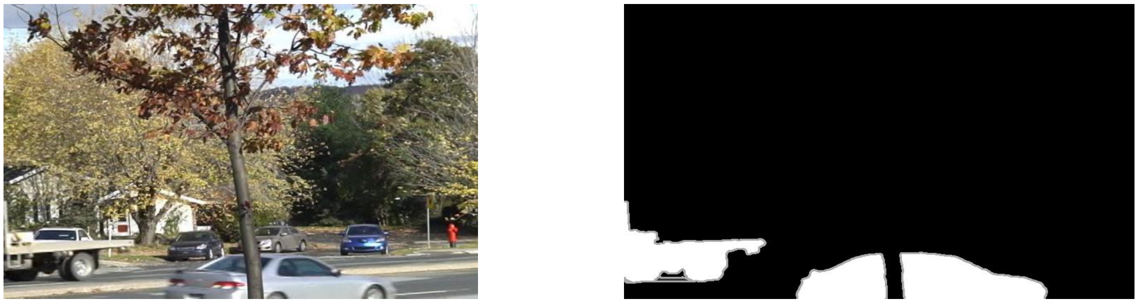
List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation

The MOS (moving object segmentation) is still an open issue for computer vision applications due to videos with awful climate, sporadic movement of items, dynamic background, jittering of camera, and shadow scenarios. Video frame pixels segmentation into either background or foreground is pixel-level classification task. The MOS task is extensively divided as background subtraction, region level, pixel-level, optical flow, frame difference, saliency estimation and deep learning based technique. The majority of work accomplished for moving object segmentation is with background subtraction method. In this methodology, at first clean background image is assessed using few input video frames and afterwards the pixel-wise segmentation is done between input video frames & evaluated background. Videos in the dynamic background (waving of tree, sporadic movement of items, falling of snow etc.) influence the exactness of background subtraction based procedures. In reasonable situations (like bad climate, waving of tree, waving water, low contrast, irregular movement of items, impact of shadow and abrupt brightening changes), the classification of pixels into background or foreground is a difficult assignment for MOS. Thus to address these matter, we suggest a novel deep learning based methodology wherein background is estimated for the video using few initial frames of the video and the task of foreground segmentation is done utilizing GAN's.



a) Frame from dynamic background category

b) Segmented ground truth

Fig 1.1 Example of MOS from CDnet2014

1.2 Application of MOS (Moving Object Segmentation)

In various computer vision tasks, for example video synopsis, vehicle navigation, objects tracking, video surveillance, person re-identification, magnetic particle imaging, traffic monitoring etc. MOS is used.

1.3 Thesis organization

The rest of the report is designed as shown below:
In chapter 2, we will discuss about important terms & concepts like Generative Adversarial Networks (GANs), Variants of GANs, Motion Saliency, F-score, PWC etc. In chapter 3, we will discuss about the related work in the field of Moving Object Segmentation. In chapter 4, we will discuss about our proposed approach. In chapter 5, we will discuss the experimental result, analysis and ablation studies. In chapter 6, we gave the conclusion.

Chapter 2

Important Terms and Concepts

2.1 Generative Adversarial Network

2.1.1 Introduction

Generative Adversarial Networks [1] were introduced in the year 2014 by Ian Goodfellow. They are extensively used in many applications because of its ability to mimic any kind of distribution of data. It comprises of 2 networks, a discriminator model and a generator model competing against one another (that’s why it is called “adversarial” learning). Generator generates the data from scratch, while the discriminator distinguishes the real features from fake features. Let us understand from the example of forger and a policeman. The forger plays the role of the generator by producing currency notes that are similar to the original currency notes and thus trying to fool the policeman. While the policeman plays the role of the discriminator and its goal is to distinguish between the forged currency notes and real currency notes. Initially, the policeman would be able to easily distinguish between the forged and the original notes as the forger lacks the experience, but with time and learning from the mistakes made the forger will learn to create currency notes that are more and more similar to the original currency notes. In summary, GAN’s are neural systems that are trained in an adversarial manner to create information mirroring a similar appropriation/distribution.

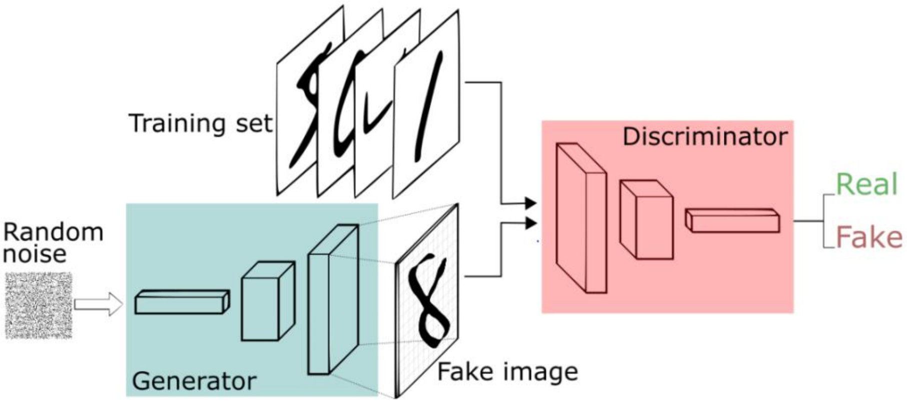
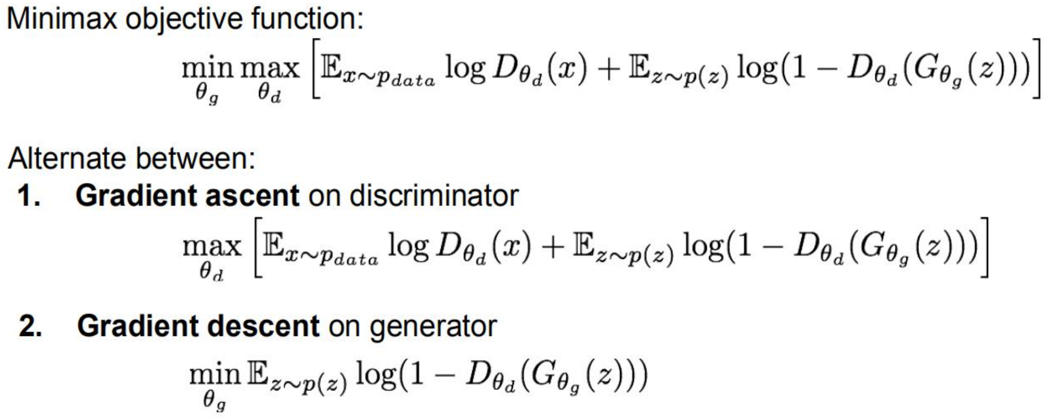


Fig 2.1 Discriminator and Generator model [2]

2.1.2 Objective function of the GAN.



Minimax objective function:

$$\min_{\theta_g} \max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

Alternate between:

1. **Gradient ascent** on discriminator

$$\max_{\theta_d} \left[\mathbb{E}_{x \sim p_{data}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z))) \right]$$

2. **Gradient descent** on generator

$$\min_{\theta_g} \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

Fig 2.2 Objective function of GAN [1]

Objective function of the GAN [1]:

- Discriminator yields the likelihood between 0 and 1 of real image.

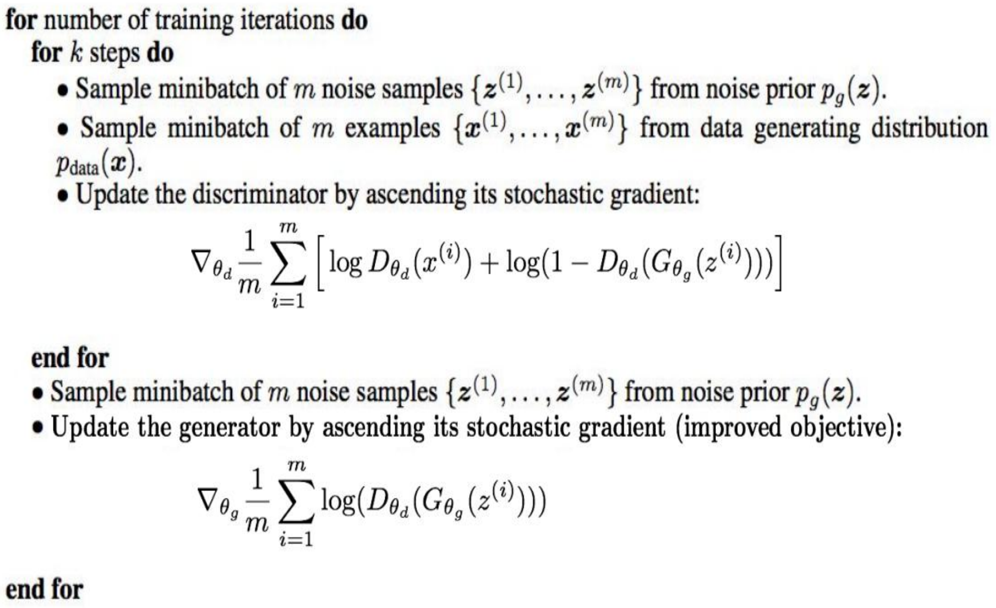
- P(z) is the random distribution from which our input samples z (input to the generator) are sampled.

- In the training process, the probability distribution of the real samples is denoted by P_{data}.

- Discriminator (θ_d) need to maximize objective such that for fake images/frames D(G(z)) is near to 0 and for real images/frames D(x) is near to 1

- Generator (θ_g) need to minimize objective such that D(G(z)) is near to 1.

2.1.3 Training the GAN.



```
for number of training iterations do
  for $k$ steps do
    • Sample minibatch of $m$ noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
    • Sample minibatch of $m$ examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
    • Update the discriminator by ascending its stochastic gradient:
        $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[ \log D_{\theta_d}(x^{(i)}) + \log(1 - D_{\theta_d}(G_{\theta_g}(z^{(i)}))) \right]$
  end for
  • Sample minibatch of $m$ noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
  • Update the generator by ascending its stochastic gradient (improved objective):
      $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(D_{\theta_d}(G_{\theta_g}(z^{(i)})))$
end for
```

Fig 2.3 Algorithm for training the GAN [1]

2.1.4 Problems with GANs

These are some problems with GANs [1]:

1. **Mode collapse:** During the training the generator may collapse to a setting where it always produces same output. This problem occurs due to the multi-modal distribution of real world data.

2. **Vanishing gradients:** If during initial phase of training if the discriminator learns well, the gradients vanishes for the generator.

3. Hard to reach the Nash equilibrium.

2.1.5 Notable researches involving GAN

As improvement to the GAN proposed in [1] researchers made many modifications to its objective function, architectures of generator/discriminator leading to the discovery of GAN's potential in many applications. Currently there are many variants of GAN, few of the notable variants include DCGAN, Conditional GAN, Pix2Pix, CycleGAN, etc.

a) Deep convolutional GAN (DCGAN). [3]

This is one of the most successful and popular design for GAN. A Deep Convolution GAN (DCGAN) does something very similar to that of traditional GAN [1], but it specifically focuses on using Deep Convolutional networks in place of those fully connected networks. Generator uses strided convolutions for down sampling and transposed convolutions for upsampling.

b)Conditional GAN (DCGAN). [4]

In GAN, there is no power over methods of the information to be produced. The conditional GAN changes that by including the label y as an extra parameter to the generator and expectations that the corresponding pictures are created. In CGAN, labels act as an augmentation to the latent space **z** to create and discriminate pictures better.

c) CycleGAN. [5]

CycleGAN introduced the concept of image to image translation via unpaired GAN based learning. Unpaired GAN based learning involves training your generator and discriminator network based of unpaired training data. i.e. input and output pairs of the training data are not mapped to each other. They accomplished unpaired based GAN learning by using a cycle consistency loss in addition to the adversarial loss for GAN training.

d) Pix2Pix. [6]

Pix2Pix is a GAN model intended with the end goal of picture to picture interpretation. Input picture is used as the label in Pix2Pix (which is a Conditional GAN). The generator model is provided with a picture as input and generates a translated form of the picture. The discriminator model is given an input picture and a genuine (real) or generated paired picture and should decide whether the paired picture is genuine or counterfeit. At last, the generator is trained to minimize the loss between generated picture & the expected target picture and fool the discriminator model. The Pix2Pix architectural details are shown below:

1) Generator model (U-Net architecture): For its generator model Pix2Pix utilizes U-Net architecture, rather than a typical encoder decoder model. The encoder-decoder generator architecture includes accepting a picture as input and downsampling it over a couple of layers until a bottleneck layer, where the portrayal is then upsampled again over a couple of layers before yielding the final image with the desired size. The U-Net model design is fundamentally the same as in that it includes downsampling to a bottleneck and upsampling again to an output image, but skip-connections are there between the layers of the similar size in the decoder and the encoder permitting the bottleneck to go around..

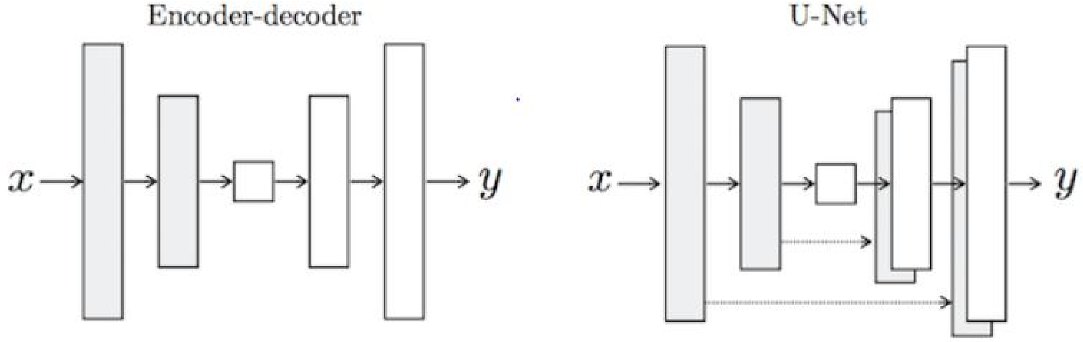


Fig 2.4 Encoder-decoder and U-Net architecture [6]

2) Discriminator model (PatchGAN): Not at all like the conventional GAN model that utilizes a deep convolutional neural system to classify pictures, the Pix2Pix model uses a PatchGAN. This is a deep convolutional neural network intended to classify patches of an input image as genuine or fake, as opposed to the whole picture. The PatchGAN discriminator model is actualized as a deep convolutional neural network, however the no. of layers is configured such that the effective receptive field of every output of the network maps to a specific size in the input picture. The yield of the network is a single feature map of genuine/counterfeit forecasts that can be averaged to give a single score.

3) Composite adversarial + L1 loss: The generator model is prepared utilizing the L1 or mean absolute pixel contrast between the created translation of the source pictures & the expected target pictures and the adversarial loss for the discriminator model. The L1 loss regularizes the generator model to yield pictures that are a conceivable interpretation of source pictures, whereas the adversarial loss impacts whether the generator model can yield pictures that are conceivable into the target area.

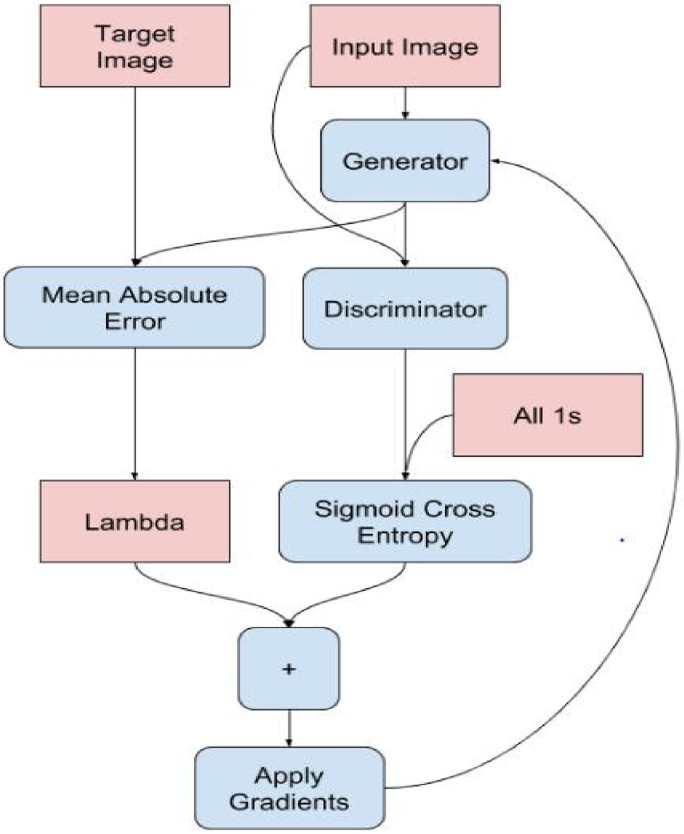


Fig 2.5 Generator loss calculation

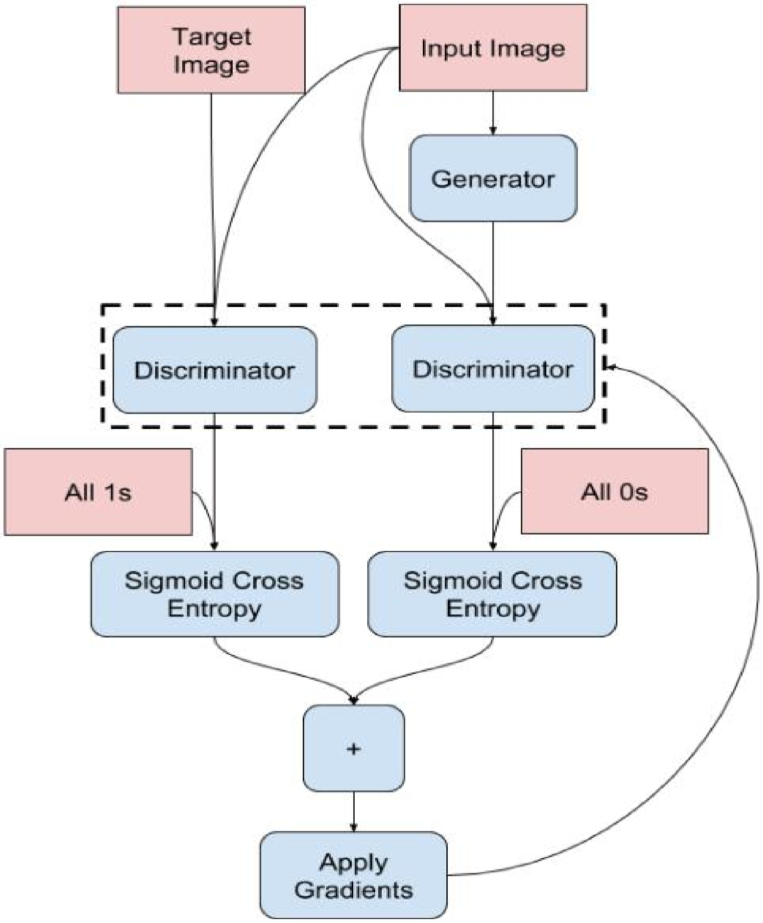


Fig 2.6 Discriminator loss calculation

2.2 Motion Saliency

It is the method of applying computer vision concepts & image processing concepts to naturally find the most “notable” areas of an image. In this way, motion saliency detection includes keeping track or finding the regions of the items that move.

2.3 F1 score

The F score [16] is a measure of a test’s accuracy. It is additionally called the F measure or F1 score. The F score is characterized as the weighted harmonic mean of the test’s recall and precision. This is determined by:

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precisior}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precisior} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Fig 2.7 F1-score formula. [16]

with the precision and recall of a test considered. Precision, additionally known as the positive predictive value, is the extent of positive outcomes that really are positive. Recall, additionally known as sensitivity, is the capacity of a test to effectively recognize positive outcomes to get the true positive rate. The F score arrives at the best worth, means perfect recall & precision, at an estimation of 1. The worst F score, which implies least recall & least precision, would be an estimation of 0.

The F score is utilized to gauge a test’s accuracy, and it adjusts the utilization of recall & precision to do it. The F score can give an increasingly sensible proportion of a test’s performance by utilizing recall & precision.

2.4 PWC

PWC [18] is an acronym for percentage of wrong classification. It is a classification metric used for qualitative analysis of the model. (Here TN refers to true negative, FN refers to false negative, FP refers to false positive and TP refers to true positive)

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

Fig 2.8 PWC calculation formula. [18]

Chapter 3

Related Work

The MOS task is comprehensively classified as background subtraction, optical flow, frame difference, pixel-level, saliency estimation, region level and deep learning based techniques.

3.1 Traditional methods.

Classification of every pixel of video frame into background & foreground is done in the customary MOS approaches. Yeh et al. [7] gave frame difference based methodology with movement remuneration & hysteresis thresholding to extract the temporal & spatial features for MOS. Processing time to estimate total foreground for single input frame isn't appropriate for real world applications.

The main errand for any foreground extraction calculation is to recognize the spotless forefront. To identify the spotless forefront, the movement data of articles is most critical component. Lin et al. [8] proposed completely unsupervised method to beat the issue of unpredictable movement utilizing region of difference method. They expected that color distribution of foreground object is not same as background. As color data is very sensitive to enlightenment varieties, this technique neglects to address the issue happening because of light varieties.

Liao et al. [9] proposed a background subtraction technique which defeat the issue activated by illumination varieties and dynamic backgrounds utilizing the idea of kernel density estimation procedure and scale invariant local ternary pattern operator.

The most generally utilized conventional techniques for MOS is background subtraction in which the background is evaluated utilizing numerous video frames by joining the local and global information of pixel intensities. Straightforward pixel-level subtraction among the determined background frame (after evaluating the clean background) and input frames is performed to calculate the foreground.

However, the customary strategies fail to address the difficulties, for example jittering of camera, dynamic background, and so forth wherein the background isn't static. This happens as it is difficult to obtain a good background estimate in such cases using the traditional techniques.

3.2 Deep learning based methods.

Recently, For MOS, researchers & scientists have focused on saliency estimation & deep learning based techniques. As they accomplish the huge improvement in segmentation accuracy when contrasted with background subtraction method. Additionally, they conquer the practical situations like unpredictable movement of items, shadow, dynamic background (waving of tree, sporadic movement of items, falling of snow etc.), terrible climate, enlightenment changes, etc. for some degree.

Wenguan et al. [12] proposed learning based static and dynamic saliency based methodology for predominant item location. The learning based system to extricate the pixel-level semantic highlights. Yang et al. [11] proposed block based deep learning technique to assess the background. As learning saliency based strategy accomplished noteworthy improvement in segmentation accuracy. Chen et al. [13] presented the idea of deep convolutional encoder-decoder network with the assistance of pre-trained VGG-16 design.

These deep learning based strategies have accomplished critical improvement in performing MOS task when contrasted with the conventional methodologies, but still their segmentation accuracy isn't sufficient for utilizing these techniques in real world scenarios.

Chapter 4

Proposed Method

4.1 Limitations of previous approaches and motivation for using GAN.

The most generally utilized conventional strategies for Moving object segmenation is background subtraction in which the background is assessed utilizing different video frames by joining the local and global data of pixel intensities. Straightforward pixel-level subtraction among the determined background frame and input video frames is gauge to determine the foreground object. Anyway these techniques fail in performing MOS for cases wherein the background isn't static for example camera jitter, dynamic background etc. As background estimation in such cases is a troublesome task.

Deep learning based strategy have accomplished noteworthy improvement in performing MOS task when contrasted with the customary methodologies, but still their segmentation accuracy is not up to the mark for utilizing these methods in real world scenarios.

Recently, for semantic division in medical as well as natural images e.g. [14], [15], adversarial training methods are being utilized These GAN-based methodologies have accomplished promising outcome for semantic division.

Inspired by them, GAN's based approaches have been used for MOS too. These GAN's based approaches are broadly divided into 1) calculating a background estimate using GAN and performing background subtraction for foreground estimation 2) calculating a motion saliency estimate for each frame and enhancing it using GAN's for foreground segmentation. However, instead of performing background subtraction or using motion saliency for foreground estimation we use GAN and let it implicitly learn a function for foreground estimation using the raw input image conditioned on its background.

We have determined both quantitative & qualitative outcomes utilizing our proposed technique. From these outcomes, it is apparent that our proposed technique beats the state-of-the-art strategies for MOS and show significant improvement in F-measure.

4.2 CDnet 2014 dataset

CDnet 2014 dataset is utilized by us in this project [10] (it is the extended version of dataset CDnet 2012). It consists of 11 categories wherein every category holds different challenges to be addressed. These categories incorporate thermal, bad weather, shadow, camera jitter etc. The 2014 CDnet dataset gives sensible, camera caught (without CGI), differing set of indoor and outdoor videos. These videos have been recorded utilizing cameras ranging from higher resolution consumer grade camcorders, business PTZ cameras to near-infrared cameras and low-resolution IP cameras. As an outcome, spatial resolutions differ from 320×240 to 720×486 of the videos in the 2014 CDnet

Details about ground truth in Dataset (CDnet2014):

- Manual annotation at pixel level is done for each frame.
- Gray scale value ‘0’ is alloted to **Static pixels** .
- Gray scale value ‘255’ is allotted to **Moving pixels**.
- Gray scale value ‘85’ is allotted to **Non-ROI pixels**.
- Gray scale value ‘50’ is allotted to **Shadow pixels**.
- Gray scale value “170” allotted to pixels that are corrupted by motion blur or half-occluded
- Below figure showing the 5 labels:

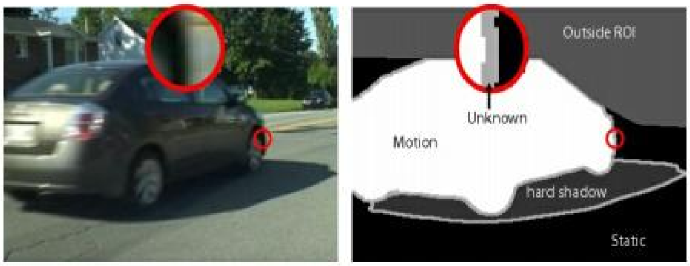


Fig 4.1 Ground Truth Labels [10]

4.3 Data Preprocessing

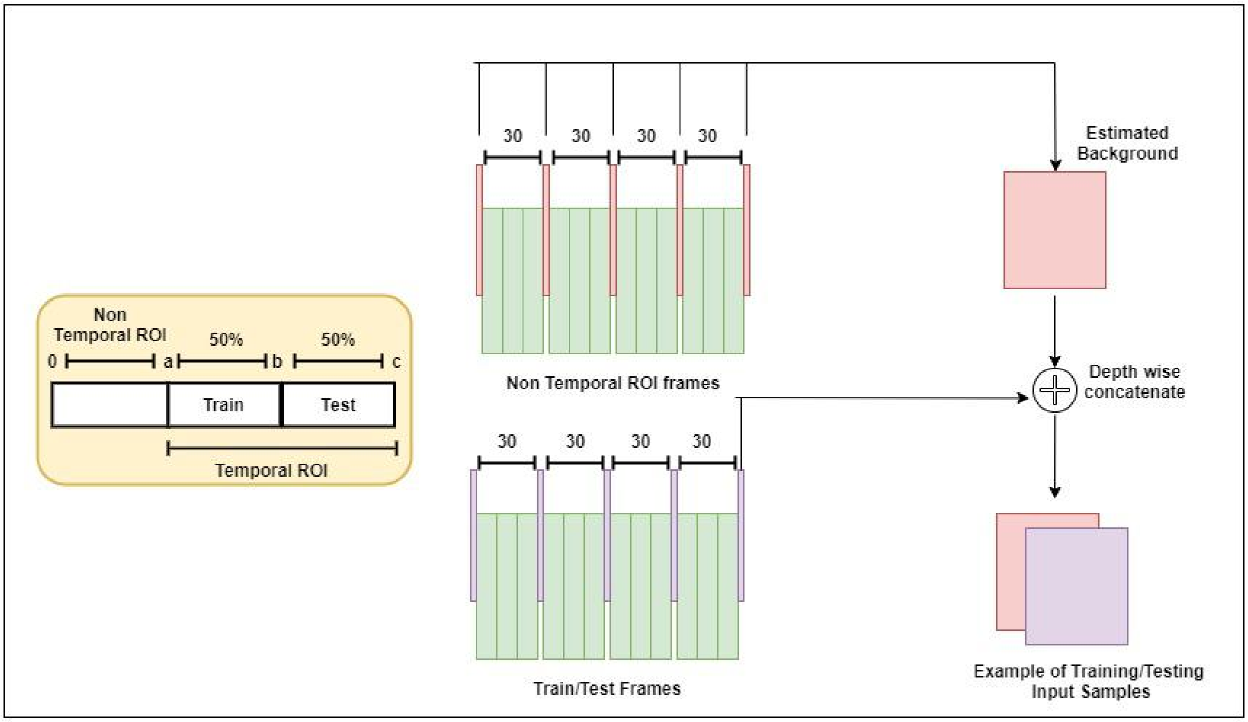


Fig 4.2 Data preprocessing step of our proposed framework

This is the first step of our proposed framework. In our proposed framework result for each category of the CDnet 2014 dataset is obtained separately. For a given category first few frames of videos in the category are not in our temporal region of interest (ROI) [10]. We call these frames as Non temporal ROI frames as shown in Fig 4.2. These first few frames from each video of a given category are thus used to calculate background estimation for the respective videos of given category. Background is estimated by taking the depth wise median of these Non temporal ROI frames. However, not all the Non temporal ROI frames are used for background estimation. We select few frames from these Non temporal ROI frames for background estimation by picking them at the gap of 30 frames as outlined in Fig 4.2.

Frames in the temporal region of interest are break into training and testing set. The first 50% of the frames (in the temporal region of interest) from each video of a given category are utilized for training purpose, while the remaining 50% are utilized for the testing purpose as outlined in Fig 4.2. Thus now we have a train/test split for a given category, also we have estimated background for each video of a given category (Fig 4.3)

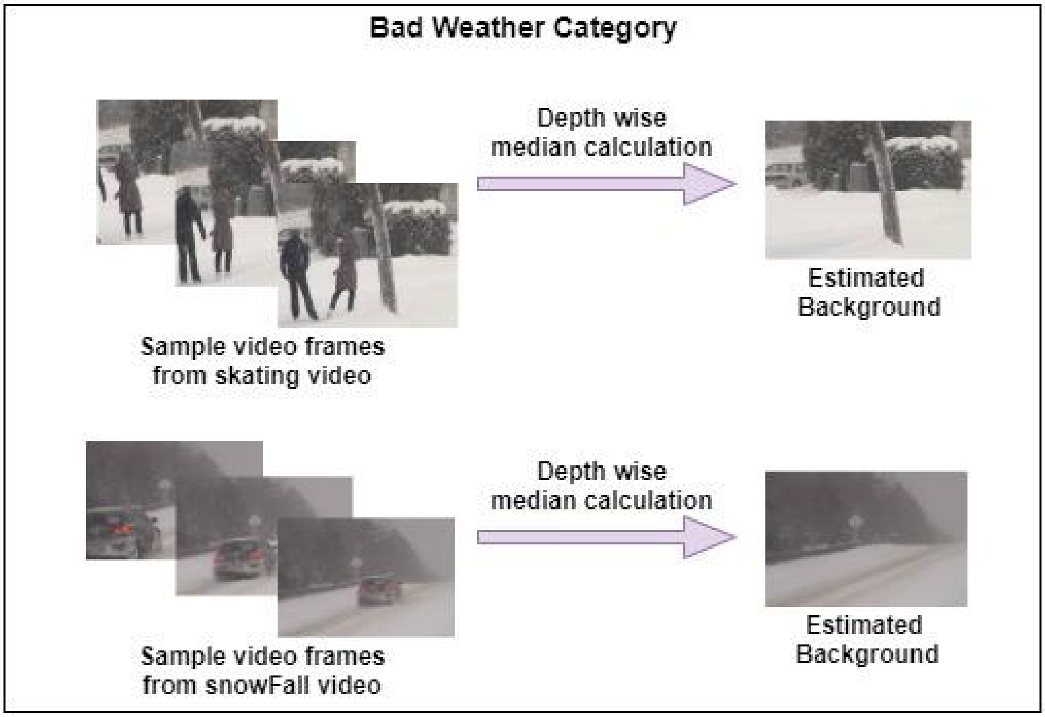


Fig 4.3 Example of estimated background for videos in CDnet2014 dataset

4.4 Training the GAN using background conditioned input

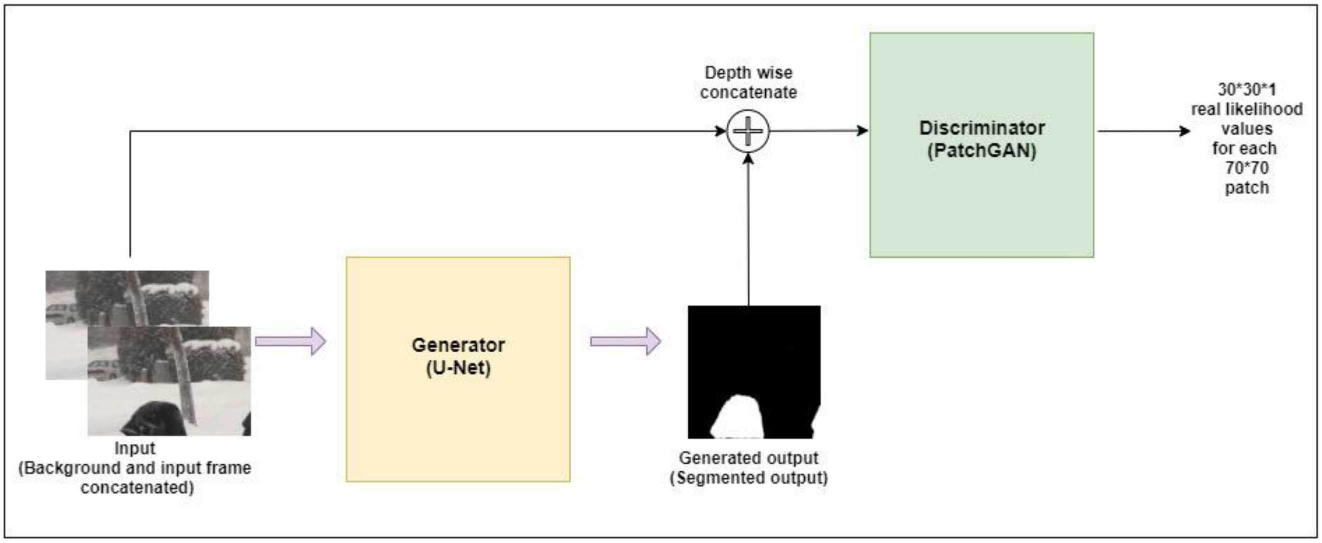


Fig 4.4 Training the pix2pix GAN on background conditioned input frames

After the data preprocessing step we have estimated background, training frames and testing frames for each video in all the categories of the CDnet2014 dataset. For the training purpose we do not use all the frames in the training set of the particular video. However, we sample few frames from the training set by picking them up at a gap of 30 frames and use them for the training purpose. These sampled frames from a particular video are than depth wise concatenated with the estimated background of that video. These depth wise concatenated frames form our input for training GAN for foreground segmentation. We use pix2pix GAN for foreground segmentation purpose as shown in Fig 4.4

4.5 Testing approach

While testing we use the trained generator and the input to the generator are the background conditioned test frames. As MOS is a pixel level segmentation task we calculate the F1-score and PWC for a particular category by calculating the confusion matrix at pixel level. While calculating the values for confusion matrix thresholding is done to maximize the F1-score. This threshold is further used to calculate PWC.

Chapter 5

Experimental Results

5.1 Experimental Setup.

For all our training and testing purposes we have used Google Colab and used tensorflow for the coding purpose. We have used pix2pix GAN as our GAN model. The input layer of the generator (U-net) of pix2pix GAN model was modified to accept input samples of size 256*256*6. The input layer of the discriminator (PatchGAN) of the pix2pix GAN model was modified to accept input samples of size 256*256*9. Hyperparameter ‘LAMBDA’ in the generator loss of pix2pix GAN was set to 200. Training was performed for about 200-250 epochs.

We have done our experiment on 7 categories from the Dataset (CDnet2014) that are Intermittent Object Motion, Shadow, Bad Weather, Thermal, Baseline, Dynamic Background and Camera Jitter.

5.2 Quantitative and Qualitative analysis

Category	(Chen et al.) [13]	PWS [19]	DBS [20]	FgGAN[18]	Proposed Method
Bad Weather	0.8949	0.8152	0.8301	0.9734	0.9711
Baseline	0.9594	0.9397	0.9530	0.9740	0.9849
Camera Jitter	0.9422	0.8137	0.8990	0.9727	0.9173
Dynamic Background	0.7356	0.9109	0.8906	0.9746	0.8190
Intermittent Object Motion	0.7538	0.8169	0.6718	0.9658	0.9097
Shadow	0.9084	0.8913	0.9304	0.9663	0.9359
Thermal	0.8546	0.8324	0.7946	0.9612	0.9678
Average	0.8641	0.8600	0.8528	0.9697	0.9294

Table 5.1 Comparison of various categories between our proposed method & other methods in terms of F-score

- As seen in table 5.1 our proposed methods beats Chen et al. [13] by approximately 6 %, PWS [19] by approximately 7% and DBS [20] by approximately 8% in terms of F-measure.
- Our trained model is light weight as compared to FgGAN still it achieves comparable result with FgGAN [18].
- In our proposed method there is no need of post processing.
- From table 5.2 we can clearly see that our proposed approach has the lowest PWC of about 12.75% which is best among all.

Category	PWS [19]	DBS [20]	FgGAN [18]	Proposed Method
Bad Weather	0.5319	0.3784	0.0619	0.0561
Baseline	0.4491	0.2424	0.1807	0.0296
Camera Jitter	1.4220	0.8994	0.3806	0.1527
Dynamic Background	0.2723	0.2933	0.0656	0.3063
Intermittent Object Motion	2.8371	4.4473	0.4250	0.1656
Shadow	1.0230	0.7403	0.2476	0.1203
Thermal	1.4018	3.5773	0.4956	0.0623
Average	1.1339	1.5112	0.2653	0.1275

Table 5.2 Comparison of various categories between our proposed method & other methods in terms of PWC

Figure shows estimated foreground by our proposed method for some category:

(i) Bad weather category of CDnet2014 Dataset

(ii) Dynamic Background category of CDnet2014 Dataset

(iii) Baseline category of CDnet2014 Dataset

(iv) Thermal category of CDnet2014 Dataset

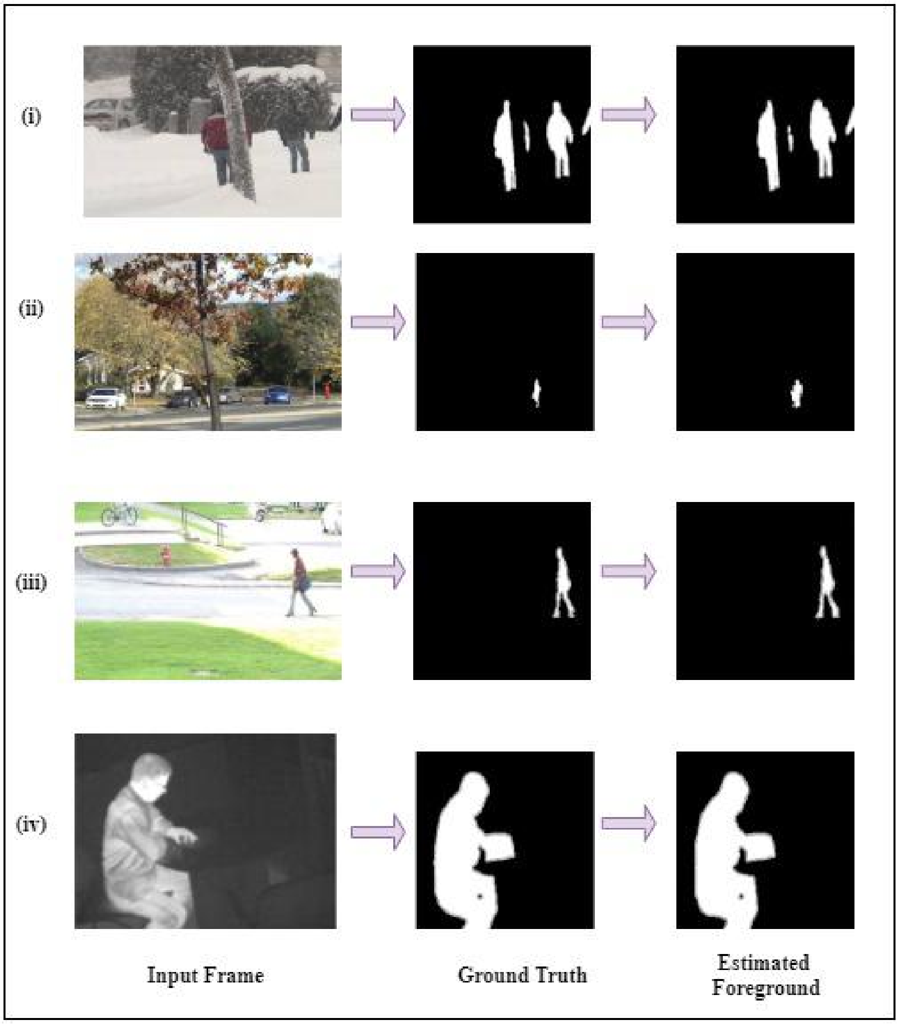


Fig 5.1 Output illustration

5.3 Ablation studies

For deciding the value of hyperparameter 'LAMBDA' in the generator loss, we varied its value from 10 to 300 at certain intervals and calculated the F1-score for dynamic background category. As you can see in table 5.3 the F1-score increases 'LAMBDA' value increases from 10 to 200, however the F1-score decreases when ‘LAMBDA’ reaches 300. So, we fixed our hyperparameter ‘LAMBDA’ value to 200.

LAMBDA	F1-score
10	0.758
100	0.790
200	0.819
300	0.795

Table 5.3 Lambda value and F1-score for Dynamic background category

For testing the effectiveness of background conditioned input we also performed an experiment (saliency input experiment) wherein we fed background subtracted frames to the pix2pix GAN keeping other conditions the same as our proposed methods and as seen in table 5.4 our method beats the saliency input experiment.

Category	Saliency input experiment	Proposed method
Dynamic background	0.7170	0.8190

Table 5.4 F1-Score by Saliency input experiment and Proposed method on Dynamic background category.

Chapter 6

Conclusion

The moving object segmentation (MOS), in the field of computer vision is still a challenge with varying background conditions. To deal with this problem, a novel approach is proposed in this report. The greater part of the work accomplished for MOS is with background subtraction approaches. Although, GAN based approaches have already been used for MOS. These GAN based approaches either calculates a background estimate using GAN and performing background subtraction for foreground estimation or calculates a motion saliency estimation for each frame and enhancing it for foreground segmentation. The proposed method however uses GAN to implicitly learn a function for foreground estimation, which is inspired by the idea to learn a complex function which can utilize both the raw video frame and its background information. In the presented method, background is estimated for a particular video and the current frame of the video is conditioned over this estimated background and used for foreground segmentation. This decreases the number of parameters involved in the model along with improving the accuracy of the model. Calculation of both qualitative and quantitative results is done using the proposed method using CDnet2014 dataset. From these results, it is evident, that our proposed approach shows notable improvements compared to the various state of the art strategies for MOS. It also shows major improvement in F-score because it learnt a complex function for segmentation instead of just subtraction and also the features of the input frame are preserved in our proposed approach which might further help in segmentation, instead of feeding motion saliency estimate where some important information might be lost even before performing segmentation. There is also no need for post-processing in our approach.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In *Advances in neural information processing systems*(pp.2672-2680).

[2] https://pathmind.com/wiki/generative-adversarial-network-gan

[3] Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[4] Mirza, M. and Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2242–2251. IEEE, 2017.

[6] isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[7] C.-H. Yeh, C.-Y. Lin, K. Muchtar, H.-E. Lai, and M.-T. Sun. Three-pronged compensation and hysteresis thresholding for moving object detection in real-time video surveillance. *IEEE Transactions on Industrial Electronics*, 64(6):4945– 4955, 2017.

[8] Y. Lin, Y. Tong, Y. Cao, Y. Zhou, and S. Wang. Visualattention-based background modeling for detecting infrequently moving objects. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6):1208–1221, 2017.

[9] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. ¨ Li.Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1301–1306. IEEE,2010.

[10] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 387–394, 2014.

[11] L. Yang, J. Li, Y. Luo, Y. Zhao, H. Cheng, and J. Li. Deep background modeling using fully convolutional network. IEEE Transactions on Intelligent Transportation Systems, 19(1):254–262, 2018.

[12] W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing, 27(1):38–49, 2018.

[13] Y. Chen, J. Wang, B. Zhu, M. Tang, and H. Lu. Pixel-wise deep sequence learning for moving object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[14] A. Lahiri, V. Jain, A. Mondal, and P. K. Biswas. Retinal vessel segmentation under extreme low annotation: A generative adversarial network approach. arXiv preprintarXiv:1809.01348, 2018.

[15] Y. Li and L. Shen. cc-gan: A robust transfer-learning framework for hep-2 specimen image segmentation. IEEE Access, 6:14048–14058, 2018.

[16] https://deepai.org/machine-learning-glossary-and-terms/f-score

[17] Sultana, Maryam, et al. "Unsupervised deep context prediction for background estimation and foreground segmentation." Machine Vision and Applications 30.3 (2019): 375-395.

[18] Patil, Prashant, and Subrahmanyam Murala. " Fggan: A cascaded unpaired learning for background estimation and foreground segmentation." 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019.

[19] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. A selfadjusting approach to change detection based on background word consensus. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 990–997. IEEE, 2015

[20] M. Babaee, D. T. Dinh, and G. Rigoll. A deep convolutional neural network for video sequence background subtraction.*Pattern Recognition*, 76:635–649, 2018.