MOHAMED KHIDER UNIVERSITY - BISKRA	جامعة محمد خيضر - بسكرة
FACULTY OF SCIENCE AND TECHNOLOGY	كلية العلوم والتكنولوجيا
DEPT. ELECTRICAL ENGINEERING	قسم الهندسة الكهربائية
DIVISION OF ELECTRONICS	شعبة الإلكترونيك
Ref:	المرجع: ....

Thesis title:

Moving Object Detection based on RGBD Information

By

IHSSANE HOUHOU

A thesis submitted to the Department of Electrical Engineering in candidacy for the Degree of **Doctorate (3rd Cycle)** in **Electronics (Biometrics and Surveillance)**.

Members of the jury:

President:	Pr. Salim Sbaa	Prof	University of Biskra
Supervisor:	Pr. Athmane Zitouni	Prof	University of Biskra
Co-Supervisor:	Pr. Yassine Ruichek	Prof	University of Technology of Belfort-Montbéliard
Examiner:	Dr. Soraya Zehani	MCA	University of Biskra
Examiner:	Dr. Messaoud Hettiri	MCA	University of Eloued

2022/2023

Dedication

This work is dedicated
to

My beloved parents.
My dear brothers and sisters.
My teachers.
All the ones I care about.

Acknowledgements

First, I thank Allah the Almighty for guiding me through this path. Nothing I made by myself but with HIS Almighty blessing.

I gratefully acknowledge my supervisor, professor Athmane Zitouni, and my co-advisors, professor Yassine Ruichek and professor Abdelmalik Taleb-Ahmed. This project would be nothing without their help, support, and companionship throughout these years.

I would like to thank the defense committee, who dedicated their time to reading and examining this thesis and making it more valuable to the scientific community.

I would thank professor Salah Eddine Bekhouche and professor Mohamed Kas for their collaboration and contribution to this study. My thanks also go to our lab mates from LESIA (University of Biskra) and the EPAN team from CIAD (University of Technology Belfort-Montbéliard). Especially Fahad Lateef, Abderrazak Chahi, and Ibrahim Kajo.

I also thank the Algerian Ministry of Higher Education and Scientific Research for funding the scholarship PNE (Programme National Exceptionnel). Also, the Faculty of Science and Technology, University of Biskra, for providing the internship programs.

My special thanks to my family for their support and prayers. To my father Abdelkarim, who taught me how to be a real man. To my dear mother Naima, who has the Paradise beneath her feet. To my dear brothers and sisters whom I care about, especially my precious twin brother Mohcine.

Abstract

This thesis is targeting the Moving Object Detection topic, more specifically, the Background Subtraction. In this study, we proposed two approaches using color and depth information to solve the background subtraction. The following two paragraphs will give a brief abstract for each approach.

In this research study, we propose a framework for improving traditional Background Subtraction techniques. This framework is based on two data types: color and depth; it stands for obtaining preliminary results of the background segmentation using Depth and RGB channels independently, then using an algorithm to fuse them to create the final results. The experiments on the SBM-RGBD dataset using four methods: ViBe, LOBSTER, SuBSENSE, and PAWCS, proved that the proposed framework achieves an impressive performance compared to the original RGB-based techniques from the state-of-the-art.

This dissertation also proposes a novel deep learning model called Deep Multi-Scale Network (DMSN) for Background Subtraction. This convolutional neural network is built to use RGB color channels and Depth maps as inputs with which it can fuse semantic and spatial information. Compared with previous Deep Learning Background Subtraction techniques that lack information due to their use of only RGB channels, our RGBD version can overcome most of the drawbacks, especially in some particular challenges. Further, this study introduces a new protocol for the SBM-RGBD dataset regarding scene-independent evaluation, dedicated to Deep Learning methods to set up a competitive platform that includes more challenging situations. The proposed method proved its efficiency in solving the background subtraction in complex problems at different levels. The experimental results verify that the proposed work outperforms the state-of-the-art on SBM-RGBD and GSM datasets.

Keywords: Computer vision, Moving object detection, Background subtraction, Traditional approaches, Deep learning, DMSN, Scene-independent evaluation.

الملخص:

تستهدف هذه الأطروحة موضوع "الكشف عن الأجسام المتحركة" ، وبشكل أكثر تحديدًا "طرح الخلفية". في هذه الدراسة ، اقترحنا طريقتين تستخدمان معلومات اللون والعمق لحل مشاكل طرح الخلفية. الفقرتان التاليتان تمثلان ملخصًا موجزًا لكل نهج.

في هذه الدراسة البحثية ، نقترح نهجا لتحسين التقنيات التقليدية في مجال طرح الخلفية. يعتمد هذا النهج على نوعين من البيانات، هي اللون والعمق. ترتكز دراستنا على استخراج النتائج الأولية لطرح الخلفية باستخدام معلومات العمق و اللون بشكل مستقل ، ثم استخدام خوارزمية لدمجها وإنشاء النتائج النهائية. أثبتت التجارب على مجموعة بيانات SBM-RGBD باستخدام أربع طرق هي: ViBe، LOBSTER، SuBSENSE و PAWCS، أن الطريقة المقترحه تحقق أداءً رائعًا مقارنة بالتقنيات الأصلية المستندة اساسا على المعلومات اللون فقط.

تقترح هذه الرسالة أيضًا نموذجًا جديدًا للتعلم العميق يسمى Deep Multi-Scale Network (DMSN) لطرح الخلفية. تم تصميم هذه الشبكة العصبية خصيصا من اجل استخدام قنوات ألوان الـRGB وخرائط العمق كمدخلات، بحيث يمكن لهذه التقنية دمج المعلومات الدلالية والمكانية للجسم المراد تحديده في المشهد. بالمقارنة مع تقنيات الطرح الخلفية السابقة للتعلم العميق التي تفتقر إلى المعلومات بسبب اعتمادها على المعلومات اللونية فقط ، فإن إصدار هذه التقنية خاصتنا التي تعتمد على معلومات الـRGBD قادرة على التغلب على معظم العيوب التي تواجه التقنيات السابقة، خاصة في أنواع معينة من التحديات. علاوة على ذلك ، تقدم هذه الدراسة بروتوكولًا جديدًا لمجموعة بيانات الـSBM-RGBD، فيما يتعلق بـ"التقييم المستقل عن المشهد"، والمخصص لأساليب التعلم العميق، وذلك لإنشاء قاعدة تنافسية تتضمن مواقف وتحديات أكثر واقعية. أثبتت الطريقة المقترحة (DMSN) كفاءتها في مجال طرح الخلفية في المواقف المعقدة وعلى مستويات مختلفة. تأكد النتائج التجريبية من أن العمل المقترح يتفوق على أحدث ما توصل إليه من تقنيات في اثنين من بين اشهر المجموعات البيانية في مجال طرح الخلفية، SBM-RGBD و GSM.

الكلمات المفتاحية: الرؤية الحاسوبية، كشف الأجسام المتحركة، طرح الخلفيه، المناهج التقليدية، التعلم العميق، DMSN، التقييم المستقل عن المشهد.

Scientific Productions

Publications in journals

- **I. Houhou**, A. Zitouni, Y. Ruichek, SE. Bekhouche, M. Kas, A. Taleb-Ahmed, "RGBD deep multi-scale network for background subtraction," International Journal of Multimedia Information Retrieval, vol.11, pp.395–407, May. 2022. (Q1, A class, IF:3.2)

Publications in international conferences

- **I. Houhou**, A. Zitouni, Y. Ruichek, A. Taleb-Ahmed, "Detection of Moving Objects Using Codebook with Image Pyramid," Image and Signal Processing and their Applications (ISPA'17), Mostaganem, Dec 2017.

- **I. Houhou**, A. Zitouni, Y. Ruichek, F. Gouizi, SE. Bekhouche, A. Taleb-Ahmed, "Background Subtraction using a scaled SuBSENSE," International Conference on Electrical Engineering (ICEEB'18), Biskra, Dec 2018.

- **I. Houhou**, A. Zitouni, Y. Ruichek, SE. Bekhouche, A. Taleb-Ahmed, "Improving ViBe-based Background Subtraction Techniques Using RGBD Information," Image and Signal Processing and their Applications (ISPA'22), Mostaganem, pp.1–6, May 2022.

Table of Contents

Page

List of Tables viii

List of Figures ix

List of Acronyms xiii

1 INTRODUCTION 1

1.1 Overview . . . 2

1.1.1 Background Subtraction . . . 2

1.1.2 Depth extraction . . . 4

1.2 Purpose of study and research objectives . . . 7

1.3 Problematic . . . 7

1.4 Contributions . . . 8

1.5 Thesis structure . . . 8

2 LITERATURE REVIEW 10

2.1 Introduction . . . 11

2.2 Datasets . . . 11

2.2.1 RGB-based datasets . . . 11

2.2.2 RGBD-based datasets . . . 17

2.3 Evaluation metrics . . . 20

2.4 Existing works . . . 20

2.4.1 Traditional approaches . . . 21

2.4.2 CNN-based approaches . . . 34

2.5 Limitations and general thoughts . . . 48

2.6 Conclusion . . . 49

3 TRADITIONAL APPROACHES: MULTI-SCALE RESOLUTIONS & RGBD-BASED TECHNIQUES 50

3.1 Introduction . . . 51

3.2 Multi-Scale Resolution for background subtraction . . . 51

3.2.1 The SuBSENSE Approach . . . 52

3.2.2 Multi-Scale Resolution (Image Pyramid) . . . 52

3.2.3 Performance Evaluation . . . 54

3.2.4 Conclusions . . . 56

3.3 RGBD ViBe-based background subtraction . . . 56

3.3.1 The proposed framework . . . 56

3.3.2 Experiments and discussions . . . 59

3.3.3 Conclusion . . . 66

3.4 Conclusion . . . 66

4 DEEP-LEARNING APPROACHES: DEEP MULTI-SCALE NETWORK (DMSN) 67

4.1 Introduction . . . 68

4.2 The proposed method: Deep Multi-Scale Network (DMSN) . . . 68

4.2.1 Introduction . . . 68

4.2.2 The proposed framework . . . 68

4.2.3 The proposed architecture . . . 69

4.3 The proposed protocols: Scene independent protocols . . . 74

4.3.1 Introduction . . . 74

4.3.2 Scene Independent Evaluation protocol for SBM-RGBD . . . 74

4.3.3 Leave One Video Out protocol for GSM . . . 76

4.4 Experiments and performance evaluation . . . 77

4.4.1 Results and discussion . . . 77

4.5 Conclusion . . . 88

5 CONCLUSIONS AND FUTURE WORKS 90

5.1 Conclusions . . . 91

5.2 Future perspectives . . . 93

Bibliography 95

List of Tables

Table Page

2.1 Change detection dataset 12
2.2 SBM-RGBD dataset 18

3.1 *F-score* calculation 54
3.2 Possibilities of foreground-foreground and background-background effects. 61
3.3 The *F-Score* Calculation for the four *c/d* possible combinations (tested on ViBe) 61
3.4 The *F-Score* Calculation for ViBe-based techniques on SBM-RGBD Dataset 61
3.5 The video-based *F-Score* calculation for ViBe-based techniques on SBM-RGBD Dataset 65

4.1 DMSN Network Details 73
4.2 Scene Independent Evaluation (SIE) protocol for SBM-RGBD 75
4.3 Leave One Video Out (LOVO) protocol for GSM 76
4.4 *F-score* on SBM-RGBD dataset (original protocol). 78
4.5 *F-score* on SBM-RGBD dataset. 79
4.6 *F-score* on GSM dataset. 87
4.7 *F-score* from experiment 1 and experiment 2. 88

List of Figures

Figure **Page**

1.1 Background subtraction mechanism. 2

1.2 The basic system of background subtraction. 3

1.3 The difference between RGB images and Depth maps 4

1.4 Stereo Vision setup example. 5

1.5 Time-of-Flight setup example. 6

1.6 Structured Light setup. 6

2.1 Frame samples from LASIESTA dataset [7]. 13

2.2 Color (RGB) frame samples and their corresponding groundtruth from the MarDCT dataset. 14

2.3 Samples from the SBI dataset. 15

2.4 Samples from the SBMnet dataset. 16

2.5 SBM-RGBD groundtruth description. 17

2.6 Frame samples from GSM dataset, from up to bottom, Cespatx_ds, Sleeping_ds, Shadows_ds, BootStraping_ds, Despatx_ds, LS_ds, TimeOfDay_ds, respectively. Color frame, Depth map, and the groundtruth, from left to right respectively. 19

2.7 Classical Gaussian Mixture Models (GMM), Type-2 Fuzzy Gaussian Mixture Model (T2-FGMM), and IV-Fuzzy Gaussian Mixture Model (IV-FGMM) algorithms, from left to right respectivelly [23]. 22

2.8 The framework presented by Hanzi and David[25]. 23

2.9 The Auto-Adaptive Parallel SOM Architecture (AAPSA) framework presented by Graciela and Mario I.[27]. 24

2.10 An example of the CodeBook background model[29]. 25

2.11 The difference between the comparison process of VIsual Background Extractor (ViBe) (left) and Vein-ViBe (V-ViBe) (right)[36]. 25

2.12 The proposed framework by Wang *et al.*[36]. 26

2.13 Local Binary Patterns (LBP) calculation[37]. 27

2.14 SeungJong and Moongu framework[31]. 28

2.15 An example to understand the Local Binary Similarity Patterns (LBSP) process[38]. 29

2.16 Self-Balanced SENsitivity SEgmenter (SuBSENSE) framework [33]. . . . 29

2.17 Pixel-based Adaptive Word Consensus Segmenter (PAWCS) framework [35]. 30

2.18 The framework used in Bohyung and Larry [40] paper. 30

2.19 An example to go from 2x3 image to 6x9 neuronal map structure[41]. . . . 31

2.20 The proposed framework from Leens *et al.* [44]. Where, G.R. stands for Geodesic Reconstruction and T&I stands for Transformation and Interpolation. 33

2.21 The proposed framework by Ottonelli *et al.* [45]. 33

2.22 The Compensation Factor (CF) explanation by Ottonelli *et al.*[45]. Where, S represents the “Subtraction mask”, AF stands for “Averaging filter”, D stands for “Depth enhanced mask” and TH is the decision threshold in pixel unit. 33

2.23 The proposed framework by Huang *et al.* [46]. 34

2.24 The model of ConvNet as represented in [48]. 35

2.25 The Multiscale Fully Convolutional Network (MFCN) model [50]. 35

2.26 The MFCN framework [50]. 36

2.27 The Random Permutation of Temporal Pixels (RPoTP) framework [52]. . . 36

2.28 The Dynamic Deep Pixel Distribution Learning (D-DPDL) pipeline [53]. . 37

2.29 The difference between Foreground Segmentation Network (FgSegNet) and FgSegNet V2 architectures [54, 55]. 38

2.30 Multispectral background subtraction architecture[56]. 38

2.31 The Background Subtraction of Unseen Videos Network (BSUV-Net) architecture [57]. 39

2.32 The Background Subtraction conditional Generative Adversarial Networks (BScGAN) proposed framework and architecture [60]. 40

2.33 The Deep Context Prediction (DCP) framework[61]. 41

2.34 The Conditional Generative Adversarial Networks Background Subtraction (CGAN-BS) framework[62]. 42

2.35 The generator architecture presented by Patil *et al.* [63]. 43

2.36 The generator architecture presented by Patil *et al.* [64]. 43

2.37 The Moving Objects Segmentation Generative Adversarial Network (MOS-GAN) as presented by Sultana *et al.* [65]. 44

2.38 The Representation Learning Networks Architecture (RLNA) architecture [65]. 45
2.39 The M-cGAN framework as presented by Sultana *et al.* [65]. 46
2.40 The BackGround Subtraction neural Networks for Depth videos (BGSNet-D) framework as presented by Wang *et al.* [66]. 46
2.41 The Foreground detection GAN RGBD (ForeGAN-RGBD) framework as presented by Sultana *et al.* [67]. 47
2.42 The Conditional Generative Adversarial Networks RGBD (CGAN-RGBD) generator as presented by Sultana *et al.* [68]. 47
2.43 The CGAN-RGBD framework as presented by Sultana *et al.* [68]. 48

3.1 The proposed framework based on the Multi-Scale Resolution (MSR) . . . 51
3.2 MSR explanation (in general): (a) Reducing the size. (b) Expanding the size. 53
3.3 The visual comparison between the proposed method and the original SuBSENSE. Rows from top to bottom ranged to contain three frame examples from three categories, *dynamicBackground*, *lowFramerate*, and *turbulence* respectively. Columns from left to right ranged to contain input frame, groundtruth, SuBSENSE result (without MSR), first resolution level, second resolution level, and third resolution level result, respectively. 55
3.4 The proposed framework for background subtraction based on RGBD information. 57
3.5 Some frame samples from SBM-RGBD Dataset, to illustrate the challenging categories. From left to right, *ColorCamouflage*, *DepthCamouflage*, and *Shadows* 58
3.6 The visual comparison between the proposed method and the original ViBe approach. Rows from top to bottom represent three frame samples from the categories, *ColorCamouflage*, *DepthCamouflage*, and *Shadows*, respectively. Columns from left to right correspond to the input RGB frame, input Depth frame, groundtruth, RGB ViBe result, and the proposed RGBD framework result, respectively. 63
3.7 The visual comparison between the proposed method and the original PAWCS approach. Rows from top to bottom represent three frame samples from the categories, *IlluminationChanges*, *ColorCamouflage*, and *Shadows*, respectively. Columns from left to right correspond to the input RGB frame, input Depth frame, groundtruth, RGB PAWCS result, and The proposed RGBD framework result, respectively. 64

4.1 The proposed Framework 69

4.2 DMSN Architecture for Background Subtraction. 70

4.3 Average *F-score* at different thresholds on SBM-RGBD. 80

4.4 The visual comparison. Rows from top to bottom show four video examples from the SBM-RGBD dataset: *ChairBox*, *Fall01cam0*, *Sleeping_ds*, and *TimeOfDay*, respectively. Columns from left to right represent the RGB input, Depth, groundtruth, and example frames from the results of, *FgSegNet_S*[54], *FgSegNet_M*[54], *FgSegNet_v2*[55], *MFCN*[50], and the proposed method *DMSN*, respectively. 81

4.5 *F-score* calculation at different thresholds on SBM-RGBD for each split (S1 to S6). 83

4.6 Average *F-score* at different thresholds on GSM dataset. 85

4.7 The visual comparison. Rows from top to bottom show four examples from the GSM dataset: *Bootstrapping_ds*, *Sleeping_ds*, *TimeOfDay*, and *LS_ds*, respectively. Columns from left to right represent the RGB input, Depth, groundtruth, and example frames from the results of, *FgSegNet_S*[54], *FgSegNet_M*[54], *FgSegNet_v2*[55], *MFCN*[50], and the proposed method *DMSN*, respectively. 86

List of Acronyms

AAPSA	Auto-Adaptive Parallel SOM Architecture
AI	Artificial Intelligence
ANN	Artificial Neural Network
ASM	Analysis of Specialized Modules
BErR	Background Error Ratio
BGSNet-D	BackGround Subtraction neural Networks for Depth videos
BS	Background Subtraction
BScGAN	Background Subtraction conditional Generative Adversarial Networks
BSIF	Binarized Statistical Image Features
BSUV-Net	Background Subtraction of Unseen Videos Network
BWs	Background Words
CEL	Cross-Entropy Loss
CF	Compensation Factor
CFE	Contrast Feature Extractor
CGAN	Conditional Generative Adversarial Networks
CGAN-RGBD	Conditional Generative Adversarial Networks RGBD
CGAN-BS	Conditional Generative Adversarial Networks Background Subtraction
CL	Content Loss
CNN	Convolutional Neural Networks
CV	Computer Vision
DANN	Domain-Adversarial Neural Networks
DCP	Deep Context Prediction

DL	Deep-Learning
DMSN	Deep Multi-Scale Network
DPDL	Deep Pixel Distribution Learning
DSLR	Digital Single-Lens Reflex
D-DPDL	Dynamic Deep Pixel Distribution Learning
FB	False Background
FCN	Fully Convolutional Networks
FF	False Foreground
FgSegNet	Foreground Segmentation Network
FErR	Foreground Error Ratio
ForeGAN-RGBD	Foreground detection GAN RGBD
FPM	Feature Pooling Module
GANs	Generative Adversarial Networks
GMM	Gaussian Mixture Models
HSV	Hue-Saturation-Value
IR	Infrared
IV-FGMM	IV-Fuzzy Gaussian Mixture Model
KDE	Kernel Density Estimation
KDA	Kernel Density Approximation
LBP	Local Binary Patterns
LBSP	Local Binary Similarity Patterns
LOVO	Leave One Video Out
LPQ	Local Phase Quantization
LSL	Least Squares Loss
MFCN	Multiscale Fully Convolutional Network
MNRL	Maximum Negative Run-Length
MoG	Mixture of Gaussians
MPB	Modified Poisson Blending

MRF	Markov Random Field
MSR	Multi-Scale Resolution
MTPA	Multi-scale Temporal Edge Aggregation
MOS-GAN	Moving Objects Segmentation Generative Adversarial Network
OF	Optical Flow
PAWCS	Pixel-based Adaptive Word Consensus Segmenter
PMD	Photonic Mixer Device
Relu	Rectified Linear Unit
ResNet	RESidual neural NETwork
RGB	Red Green Blue
RGBD	Red Green Blue Depth
RL	Reconstruction Loss
RLNA	Representation Learning Networks Architecture
RPoTP	Random Permutation of Temporal Pixels
SALBP	Scene Adaptive Local Binary Patterns
SDE	Scene Dependent Evaluation
SIE	Scene Independent Evaluation
SOBS	Self-Organizing Background Subtraction
SuBSENSE	Self-Balanced SENsitivity SEgmenter
SVM	Support Vector Machine
TB	Total Background
TBMOD	Texture-Based MODeling
TF	Total Foreground
ToF	Time-of-Flight
T2-FGMM	Type-2 Fuzzy Gaussian Mixture Model
VHR	Very High Resolution
ViBe	VIsual Background Extractor
V-ViBe	Vein-ViBe

CHAPTER 1

INTRODUCTION

Contents

Page

1.1 Overview 2

1.1.1 Background Subtraction 2

1.1.2 Depth extraction 4

1.2 Purpose of study and research objectives 7

1.3 Problematic 7

1.4 Contributions 8

1.5 Thesis structure 8

1.1 Overview

1.1.1 Background Subtraction

Our human body is designed in such a manner that it collects information from the environment to cope with several challenges facing us in daily life. This information is solicited using the five senses (vision, hearing, smell, taste, and touch). The visual sense is the one that captures the most significant amount of information which can exceed 10 Mb of data per second[1]. Indeed, using cameras to mimic how the human eyes work and capture such amounts of data is helpful for many tasks nowadays using technology based on Artificial Intelligence (AI), which we simply call Computer Vision (CV).

One of the well-known applications in Computer Vision is Background Subtraction (BS). Background subtraction is a research topic where the main goal is to separate the moving objects from the static ones in a video frame sequence. It is an essential step in most computer vision systems that are based on object detection.

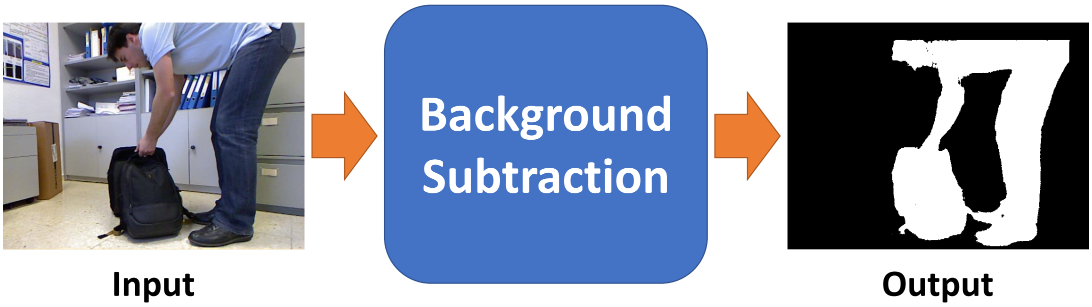


Figure 1.1: Background subtraction mechanism.

Background subtraction has been one of the prevalent topics for the last decades, dedicated to the evolution of applied intelligence technology. Since that, this topic has seen progress thanks to many types of research for better performance and has contributed to many applications related to security, safety, autonomous vehicles, and more. Background subtraction is essential for such applications; it simplifies the treated scene from complex, colorful images full of unnecessary data to a sequence of binary-labeled frames that illustrates the moving objects and eliminates the static objects from the scene, i.e., each pixel of these frames either labeled with a value of "1" as a part of the Foreground objects or with a value of "0" a part of the Background objects. These results obtained from the background subtraction operation will give the complete system (autonomous vehicle, for example) the ability to perform better and cope with Real-Time applications. Figure 1.1 illustrates the main objective of a background subtraction system.

We believe that the use of background subtraction goes way back several decades to the author Potter [2], the first who mentioned the two main terms related to background subtraction: Segmentation and Motion. His study is based on detecting moving objects using two consecutive frames, which is actually inspired by previous research claiming that many animals do not see the static objects, but only the moving ones [3]. Where the fundamental approach of a background subtraction system can be explained as follows:

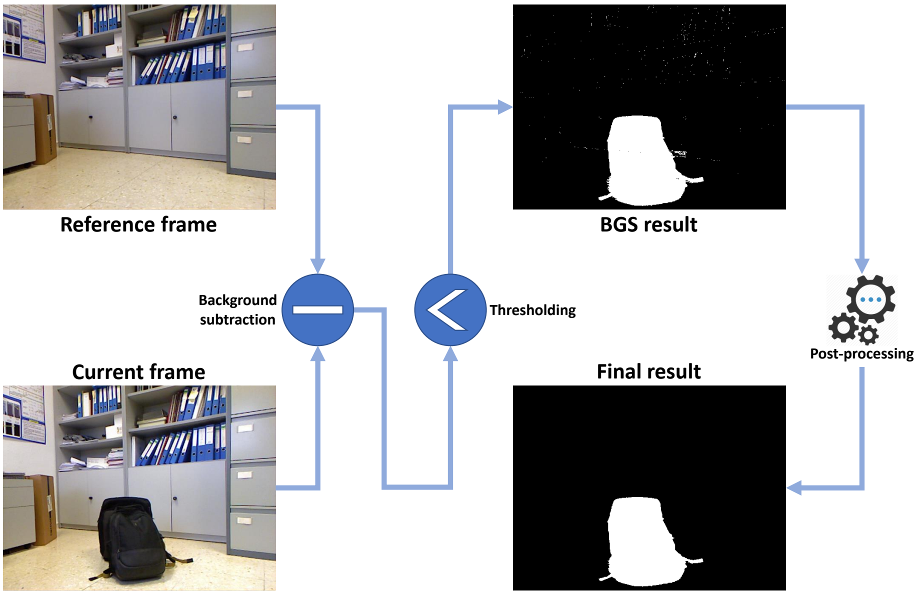


Figure 1.2: The basic system of background subtraction.

Let's suppose f_{t1} and f_{t2} are two images taken from the same scene in two different moments, $t1$ and $t2$. Assuming f_{t1} as a reference frame containing only static objects. Applying a pixel-wise comparison between this frame and another frame from the same sequence that includes new object(s) results in canceling the stationary elements, keeping only nonzero values corresponding to the non-stationary components.

This explained operation can be defined as:

$$p_{bg_{t2}}(x,y) = \begin{cases} 1 & if \quad \left|p_{f_{t2}}(x,y) - p_{f_{t1}}(x,y)\right| > T \\ 0 & otherwise \end{cases} \tag{1.1}$$

Where $p_{bg_{t2}}(x, y)$ is the pixel value at the moment $t2$ and the position (x, y) from the background subtraction result frame. $p_{f_{t1}}(x, y)$ and $p_{f_{t2}}(x, y)$ are the pixel value in the two different moments $t1$ and $t2$, at the same position (x, y) from the reference frame and the current frame, respectively. Note that T is the threshold, where: $0 < T < 255$.

Figure 1.2 illustrates, visually, the basic system of background subtraction. Note that some noise will occur in the resulting frame after the thresholding. Using post-processing operations such as connecting contours, filling holes, and/or other known filters will remove the salt noise and unwanted parts.

Although the approach just described is essential, it is commonly employed as the backbone of imaging systems meant to identify patterns in controlled environments, including parking spaces, buildings, as well as other fixed locales. Therefore, this thesis will go through very advanced techniques that solve more complex scenarios.

1.1.2 Depth extraction

As standard users, everyone is familiar with digital camera devices such as Digital Single-Lens Reflex (DSLR), Mirrorless, and Phone cameras. This ordinary tool provides gray-scaled or colored images. These kinds of cameras are often used for research and applied technologies (Surveillance, Autonomous cars... etc.), and they are called Red Green Blue (RGB) cameras, referring to their functioning mechanism. On the other hand, there are different kinds of cameras that are generally not used for the same purpose as RGB cameras. These ones are mainly dedicated to research and technology applications. In this study, we will build our proposed methods to use the Depth information, which will be extracted using unique cameras or frameworks.

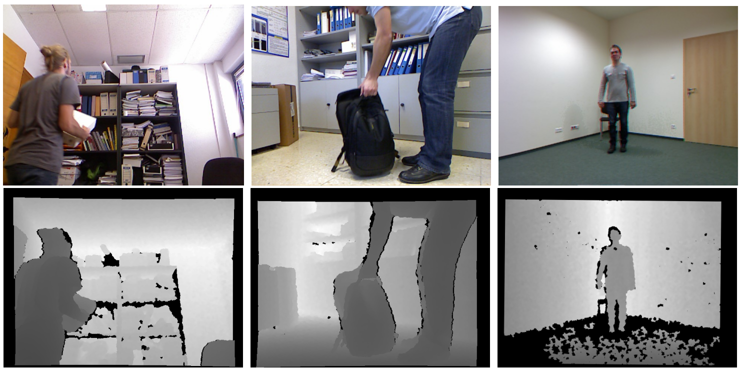


Figure 1.3: The difference between RGB images and Depth maps

Depth cameras are equipped with sensors to give some geometrical information about the scene, which can help solve some issues with systems that rely only on RGB data. A depth image is constructed of pixel values for depth (different than the RGB values), which are proportional to the estimated distance between the depth camera and the corresponding point in the real scene (see Figure 1.3). Here, we explain some different approaches to extracting Depth information to give the reader an overview of this subject:

Stereo vision: Which is actually based on the approach of the human eyes. This method requires two ordinary cameras placed and calibrated to spot the scene from two different angles simultaneously. The two corresponding frames will be processed using a stereo-to-depth algorithm based on the calibration of the setup of the equipment. Figure 1.4 gives an example of how the setup should be.

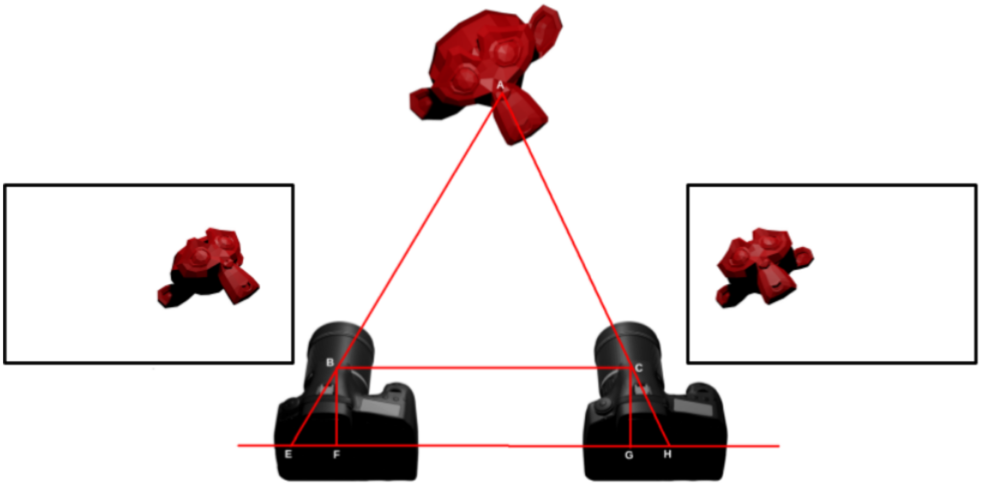


Figure 1.4: Stereo Vision setup example.

Time-of-Flight (ToF): This model of depth cameras is based on transmitter-receiver Infrared (IR). Its approach is to calculate the IR time of flight Camera-Object-Camera, then estimate the distance between the camera and the targeted objects. Figure 1.5 explains more about the ToF method.

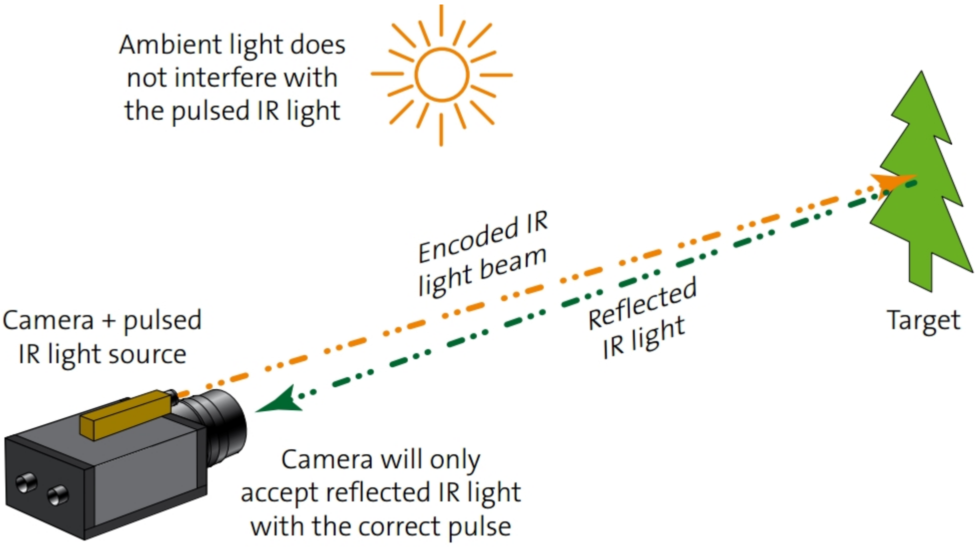

Ambient light does not interfere with the pulsed IR light
Encoded IR light beam
Reflected IR light
Camera + pulsed IR light source
Target
Camera will only accept reflected IR light with the correct pulse

Figure 1.5: Time-of-Flight setup example.

Structured Light: This approach is basically built with a projector that sends an invisible structured light on an object. These structured lights will be captured by a special sensor/camera. After that, the algorithm will create a depth map based on the curves that appear on the targeted objects. Figure 1.6 illustrates more about how this technique works. It is worth mentioning that Microsoft used this approach to build their famous XBOX camera, the Microsoft Kinect.

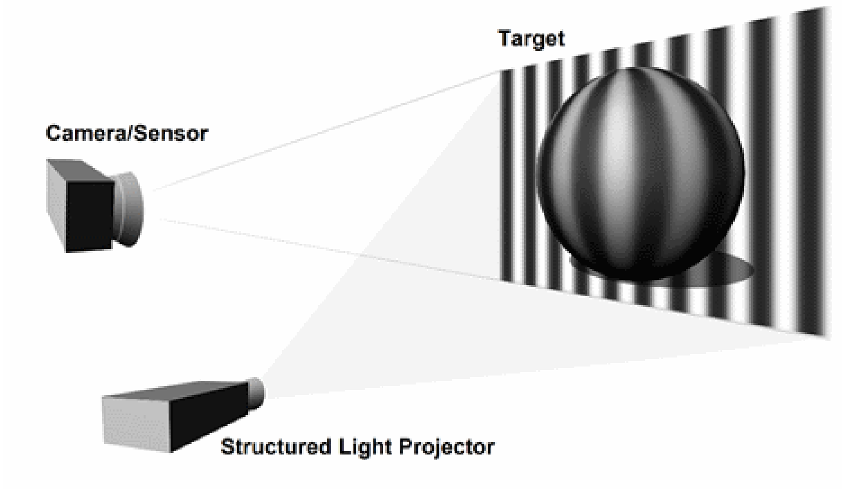


Figure 1.6: Structured Light setup.

1.2 Purpose of study and research objectives

This thesis is dedicated to creating a novel method to improve the background subtraction performance for better Computer Vision systems in the future. Our study should rely on the two main research branches in the field: *(i)* Traditional and *(ii)* Deep-Learning. Further, the data in the background subtraction is essential to have an efficient model; using multiple data sources will help our proposed system reach the desired performance. However, the evaluation protocol has to be reasonable to ensure the effectiveness of the proposed approach.

1.3 Problematic

Background subtraction is a domain that became more attractive in the late nineties when Stauffer and Grimson proposed the Mixture of Gaussians [4]. Since then, several datasets and competitions have been created in this field to offer new challenges. Although most of these datasets are overrated by advanced new techniques, some categories are still unbeatable even by recent Deep Learning approaches. These challenging categories, such as illumination changes, depth camouflage, intermittent motion, and more, give competitive scenes that cannot be easily handled. We believe these scenarios require an advanced technique that extracts more features and eliminates unnecessary information from the scene. Hence, utilizing more than one source of data will provide much recognition for our system rather than using only color images like most state-of-the-art approaches.

The second major problem in background subtraction is related to the protocol followed to evaluate the approaches. Using a scene-dependent protocol will lead to an unfair evaluation. Scene-dependent protocol stands for training (or creating) a model from a set of frames chosen from one video and then evaluating that model using the remaining frames from the same video. This causes the Training/Testing overlapping, mainly in the Deep Learning techniques.

The state-of-the-art methods are based on building a model from a frame reference or temporal history (multiple frames). Only a few techniques can provide background subtraction without using any references; nevertheless, they use scene-dependent evaluation. This approach will lead to a model that can not be effective for real scenarios.

This study aims to create a method to solve these major problems based on the existing state-of-the-art techniques and the latest approaches from traditional and Deep

Learning research studies.

1.4 Contributions

The aim of this thesis is the investigation and development of background subtraction. Our main contributions are listed as follows:

- An in-depth survey about background subtraction.
- Proving the importance of using Depth information for background subtraction.
- Providing a new framework for background subtraction using RGBD information based on the traditional approaches.
- Providing a novel model of Deep Learning for background subtraction using RGBD information.
- Proposing a new protocol for the SBM-RGBD dataset that guarantees a fair evaluation process for Deep Learning methods.

1.5 Thesis structure

The rest of this thesis is organized as follows:

In Chapter 2, we provide an in-depth literature review of background subtraction. This chapter will start by summarizing the most-used datasets in the literature. Then a brief description of the evaluation metrics used to evaluate the background subtraction methods. A detailed presentation of the current work, the latest updates in the field, and a comparison between the different existing approaches are given. A brief conclusion will summarise the chapter at the end. Chapter 3 is devoted to our contribution related to traditional approaches. This chapter starts with explaining our very first contribution, the use of the Multi-Scale Resolution (MSR) in background subtraction. Later in the same chapter, we present the color-depth information fusion framework, applied to the ViBe-based approaches. We organized this chapter by giving each contribution three sections, a description of the approach, a performance evaluation, then conclusions. In the end, a general conclusion of the Chapter is given. Chapter 4 is reserved for our contribution related to deep learning-based approaches. It starts with introducing a novel CNN-based architecture that fuses color and depth information to apply the background subtraction. Then, a detailed description of the proposed scene-independent protocol is provided. Experiments and evaluations of the proposed method are analyzed. Later, the chapter conclusion is presented at the end. The last chapter is a general conclusion of our work and an envisioning for some future works.

CHAPTER 2

LITERATURE REVIEW

Contents

Page

2.1 Introduction 11

2.2 Datasets 11

2.2.1 RGB-based datasets 11

2.2.2 RGBD-based datasets 17

2.3 Evaluation metrics 20

2.4 Existing works 20

2.4.1 Traditional approaches 21

2.4.2 CNN-based approaches 34

2.5 Limitations and general thoughts 48

2.6 Conclusion 49

2.1 Introduction

In the world of technology, where computer vision has become a giant industry, the moving object detection and background subtraction fields took the researchers' attention. The last decades have witnessed a lot of research and studies dedicated to these domains. Indeed, datasets, challenges, and compositions are continuously rising to seek the ideal approach that solves most of these problems and can be applied in real scenarios.

We present an in-depth literature overview on background subtraction in this chapter. The first section will summarize the most often used datasets in this field. The type of data used by the owners and the represented categories for each dataset are provided. The metrics used to evaluate the background subtraction techniques are briefly described. We then give a complete description of the previous work, as well as the most recent developments in the area and a comparison of the many available approaches. Later, some limitations and general thoughts about the field are given. At the end of the chapter, there will be conclusions and a quick summary.

2.2 Datasets

Several datasets have been created in the past to evaluate background subtraction methods. In this chapter, we attend to cite the frequently used ones. We have chosen this list of datasets based on the variety of challenges, the type of data provided, and the amount of data offered by the dataset founders.

2.2.1 RGB-based datasets

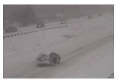
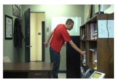
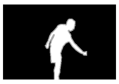
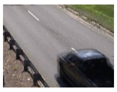
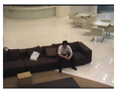
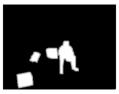
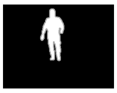
2.2.1.1 CDNET

The Change Detection NET benchmark dataset[1]. To the best of our knowledge, it is the most significant benchmark that targets background subtraction. Two successful projects, Goyette *et al.* [5] (2012) and Wang *et al.* [6] (2014), have been launched to create a solid base for researchers to work on motion detection with an enormous amount of data and various real scenarios.

In total, *CDnet 2012* and its extended version *CDnet2014*, built by 13 experts from 7 different universities, captured and then manually segmented and labeled more than 70

[1] http://changedetection.net/

Table 2.1: Change detection dataset

Category	Challenges and properties	Number of videos	Input example	Groundtruth example
badWeather	• Outdoor • Winter storm conditions • Traffic & pedestrians	4		
baseline	• Indoor/outdoor • Subtle background motion • Traffic & pedestrians	4		
cameraJitter	• Indoor/outdoor • Unstable cameras • Traffic & pedestrians	4		
dynamicBackground	• Outdoor • Strong background motion • Traffic, boats, & pedestrians	6		
intermittentObjectMotion	• Indoor/outdoor • Parking & abandoned objects • Cars & pedestrians	6		
lowFramerate	• Outdoor • Low Frame-Rate (0.17 to 1 fps) • Traffic, boats, & pedestrians	4		
nightVideos	• Outdoor • Night videos • Traffic	6		
PTZ	• Outdoor • Background motion • Traffic & pedestrians	4		
shadow	• Indoor/outdoor • Strong and soft moving shadows • Traffic, bikes & pedestrians	6		
thermal	• Indoor/outdoor • Far-infrared cameras • Pedestrians & kayaks	5		
turbulence	• Indoor/outdoor • Turbulence background & small moving objects • Traffic	4		

000 frames. These frames are from 53 videos, divided into 11 categories; each category contains from 4 to 6 videos, each video folder contains two sub-folders, input and groundtruth, and two files named ’ROI.bmp’ and ’ROI.jpg’ illustrating the spatial region of interest. Table 2.1 summarizes the dataset content.

2.2.1.2 LASIESTA

LASIESTA (Labeled and Annotated Sequences for Integral Evaluation of SegmenTation Algorithms) [7][2] is a large dataset that combines pixel-level and object-level annotation for moving object detection. It has six-pixel labels: Black for Background, Red for the first moving object, Green for the second moving object, Yellow for the third moving object, White for the moving objects remaining static, and Gray for the uncertainty pixels. The LASIESTA dataset consists of 48 sequences covering several scenarios from indoor and outdoor conditions. The majority of the videos are dedicated to pedestrians, and only four sequences are related to car parking situations. Figure. 2.1 shows some frame samples.

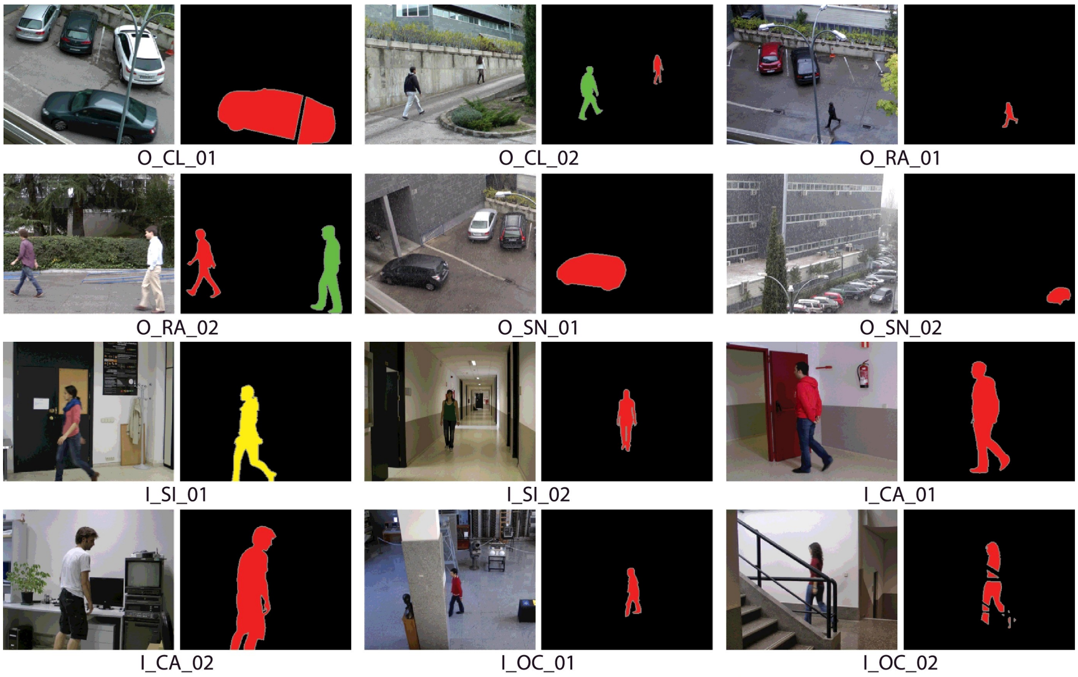


Figure 2.1: Frame samples from LASIESTA dataset [7].

[2] https://www.gti.ssr.upm.es/data/lasiesta_database

2.2.1.3 MarDCT

Maritime Detection, Classification, and Tracking (MarDCT)[8][3] is a dataset made by a group of researchers from Italy. This dataset targets three main tasks of computer vision: Detection, Classification, and Tracking. The detection dataset is dedicated to background subtraction; it contains ten videos (RGB data) of small boats from the city of Venice and six other videos of ships from the Mediterranean coast (Italy), plus three videos (IR data) of other ships from the Northern Europe coast. The provided data quality is mainly low; it includes vibrations and In/Out zooming of the scene. The dataset contains several challenges, such as dynamic background, bootstrapping, reflections, night/day scenes, and shadows, which are recommended to be used by researchers seeking these kinds of environments to solve the background subtraction. In Figure. 2.2 we represent some samples from the MarDCT dataset.

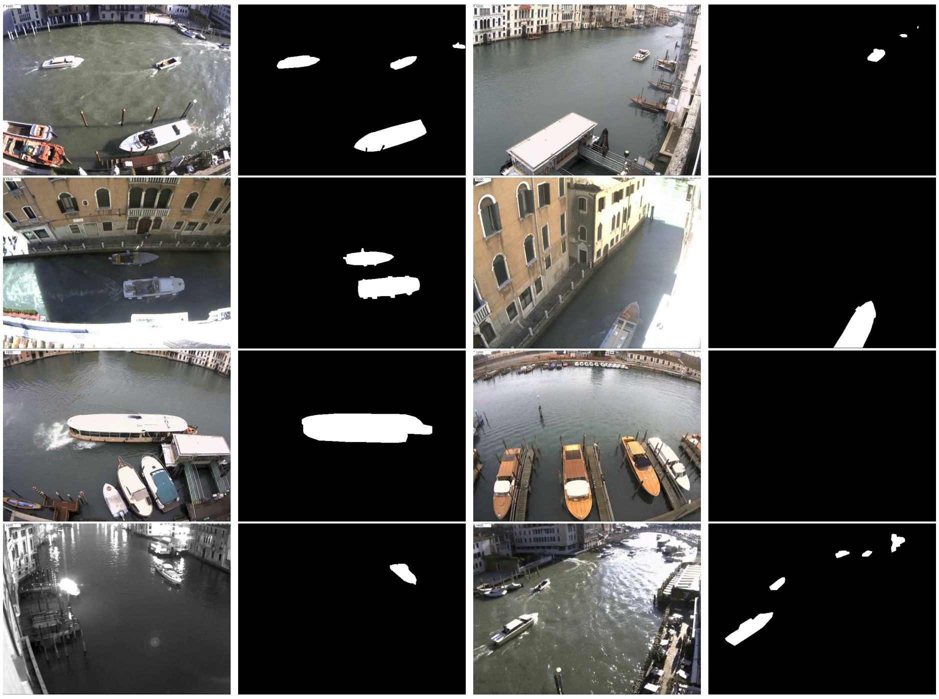


Figure 2.2: Color (RGB) frame samples and their corresponding groundtruth from the MarDCT dataset.

[3] http://labrococo.dis.uniroma1.it/MAR/index.htm

2.2.1.4 SBI & SBMnet

These two datasets are devoted to the background initialization and background modeling methods, where the primary purpose of these methods is to extract the background scene from the image sequence. It stands for removing all the moving objects and keeping only the static ones. The result will be an image representing the background scene, which will be used later in further applications, including background subtraction. These datasets are represented as follows:

Scene Background Initialization (SBI) [9][4]: It is a dataset that was created for the Scene Background Modeling and Initialization (SBMI2015) workshop[10]. It contains fourteen different scenarios that differ between Outdoor-Indoor scenes and Pedestrians-Vehicles as foreground objects. These sequences are extracted from seven various publicly frequently used datasets. Figure. 2.3 shows some samples from the SBI dataset. Note that the groundtruth of background initialization is different than the background subtraction; they are represented as an RGB frame that does not include any foreground objects.

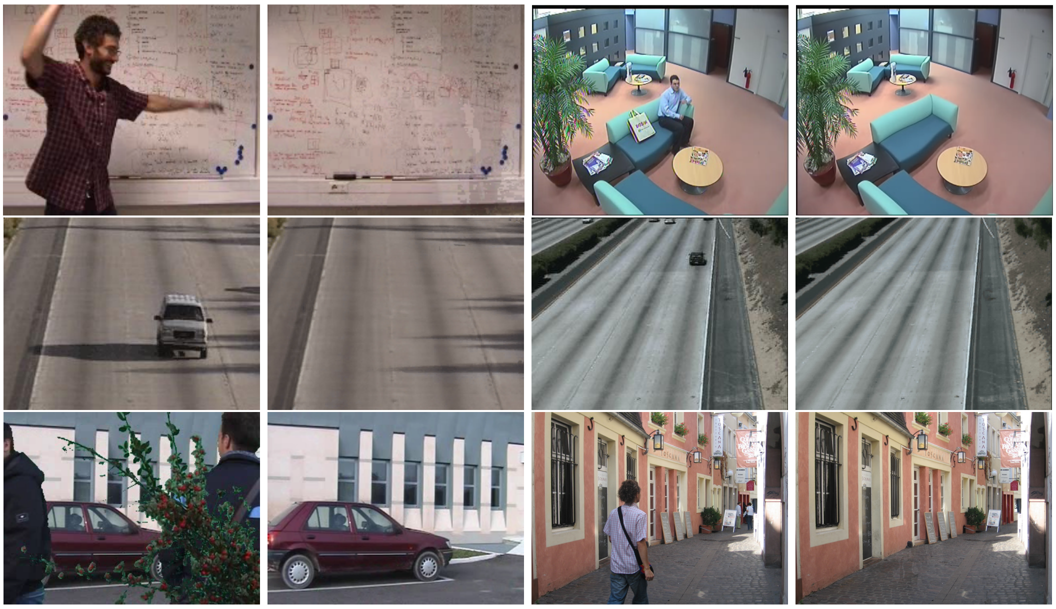


Figure 2.3: Samples from the SBI dataset.

[4] https://sbmi2015.na.icar.cnr.it/SBIdataset.html

Scene Background Modeling (SBMnet) [11][5]: It is a dataset that collects various scenarios from sixteen existing datasets to set up a new benchmark devoted to the background initialization task. This dataset consists of seventy-nine videos, divided into eight categories; each category represents a kind of challenge. This dataset also provides videos with different light settings, resolutions, lengths, and frame rates, allowing researchers to have a wide range of challenges. Figure. 2.4 shows some samples from the SBMnet dataset.

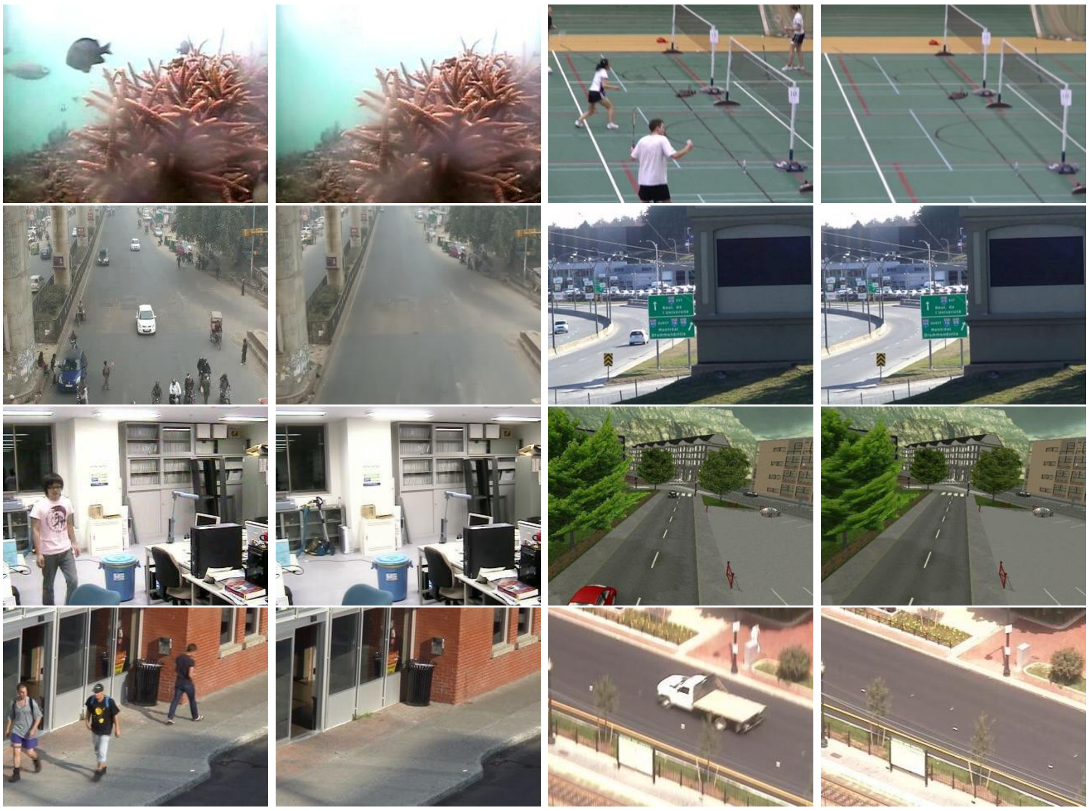


Figure 2.4: Samples from the SBMnet dataset.

[5] http://scenebackgroundmodeling.net/

2.2.2 RGBD-based datasets

2.2.2.1 SBM-RGBD

SBM-RGBD[12][6], is a dataset made for the SBM-RGBD Challenge in 2017. This dataset is a combination of five different small datasets [13–17], making it one of the most used and variegated ones, especially when it comes to background subtraction based on using Depth information. The SBM-RGBD dataset targets indoor environments where many challenges are related to critical situations such as overlapping, illumination, shadows, and others. The presence of Depth maps along with the RGB information in this dataset allows researchers to study the fusion of these two pieces of information and solve many scenarios that are considered difficult to handle using only one source of data. The depth maps provided by the dataset founders are extracted from the Microsoft Kinect camera.

This dataset contains seven categories; each category has from four to six video sequences (33 videos in total), and each sequence has two kinds of input information (RGB and Depth), including the corresponding groundtruth for results evaluation. The length of these sequences varies between 70 and 1400 frames per video. The seven categories provided in this dataset represent seven different challenges: Illumination Changes, Color Camouflage, Depth Camouflage, Intermittent Motion, Out of Sensor Range, Shadows, and Bootstrapping (See Table. 2.2). The groundtruth is made based on the concept of the ChangeDetection dataset. It has four pixel-wise value labels: 0 for Background, 85 for Outside Region of Interest, 170 for Unknown, and 255 for Foreground. (See Figure. 2.5)

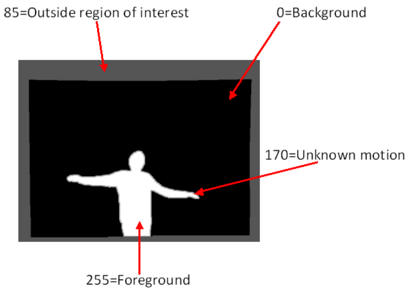
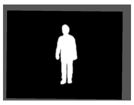
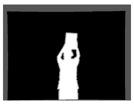
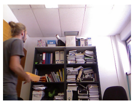
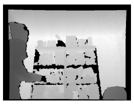
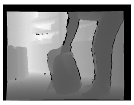
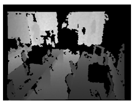
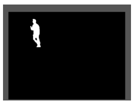
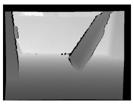


Figure 2.5: SBM-RGBD groundtruth description.

[6] https://rgbd2017.na.icar.cnr.it/SBM-RGBDdataset.html

Table 2.2: SBM-RGBD dataset

Category	Challenges and properties	Number of videos	RGB example	Depth example	Groundtruth example
Bootstrapping	• Foreground objects in all frames • Low light conditions • Some background regions stay hidden	5			
Color Camouflage	• Background/foreground color similarity • Some scenes include lighting contrast • Some scenes include intermittent motion	4			
Depth Camouflage	• Background/foreground distance similarity • Some scenes include lighting contrast • Some scenes include intermittent motion	4			
Illumination Changes	• Difficult lighting conditions & sudden illumination changes • Some scenes include intermittent motion • Some scenes does not include foreground objects	4			
Intermittent Motion	• Abandoned objects • Background objects start moving • Some scenes include color and distance similarity	6			
Out of Sensor Range	• Failure to measure depth in some regions • Some scenes include reflective surfaces • Some scenes include color and distance similarity	5			
Shadows	• Foreground object shadows • Lighting contrast • Some scenes include color and distance similarity	5			

2.2.2.2 GSM

GSM[13][7] is a small dataset that provides seven videos. Each video presents a particular category: BootStraping, Color camouflage, Depth camouflage, Sleeping, Shadows, TimeOfDay, and LightSwitch; these categories are well described in the previous Section 2.2.2.1.

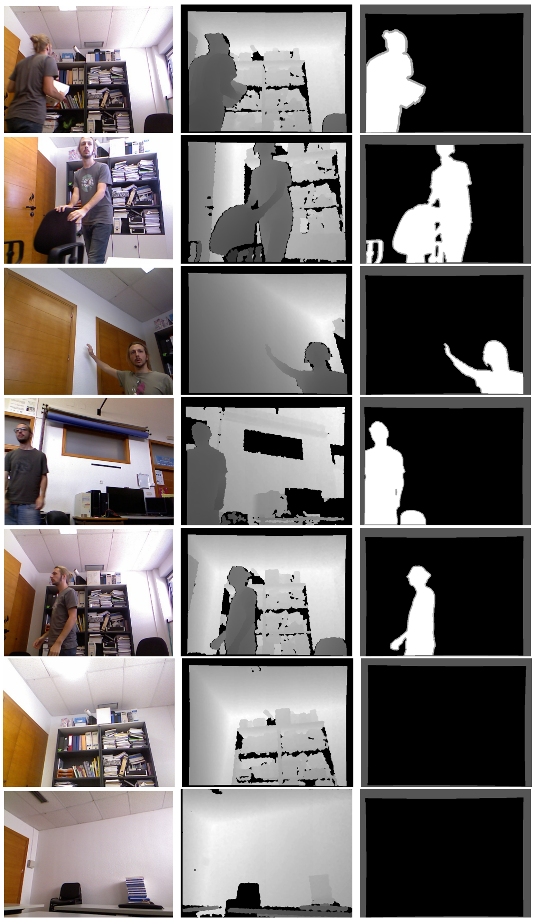


Figure 2.6: Frame samples from GSM dataset, from up to bottom, Cespatx_ds, Sleeping_ds, Shadows_ds, BootStraping_ds, Despatx_ds, LS_ds, TimeOfDay_ds, respectively. Color frame, Depth map, and the groundtruth, from left to right respectively.

In fact, the GSM dataset videos are included in the SBM-RGBD dataset. The length of the sequences varies between 200 and 1231 frames per each. The frame size is fixed to 640 X 480 for all sequences. This dataset is a better choice to challenge deep learning networks using a small amount of data. The founder of GSM provides two

[7] http://gsm.uib.es

types of data, color and depth information. The groundtruth is labeled following the ChangeDetection dataset. Frame samples of color, depth and groundtruth from each sequence are presented in Figure 2.6.

2.3 Evaluation metrics

The *F-score*, or the *F-measure*, is the most frequently used metric for background subtraction in the state-of-the-art. It is based on several indicators such as True Positives (*TP*), True Negatives (*TN*), False Positives (*FP*), and False Negatives (*FN*). These indicators are computed from the comparison between the estimated results and the provided Groundtruth, as follows:

- *TP*: The pixels which are foreground, and classified as foreground.
- *TN*: The pixels which are background, and classified as background.
- *FP*: The pixels which are background, but classified as foreground.
- *FN*: The pixels which are foreground, but classified as background.

$$F_{score} = 2.\frac{Pr.Re}{Pr + Re} \tag{2.1}$$

Where Recall (*Re*) and Precision (*Pr*) are computed as follows:

$$Re = \frac{TP}{TP + FN} \tag{2.2}$$

$$Pr = \frac{TP}{TP + FP} \tag{2.3}$$

2.4 Existing works

In recent decades, computer vision has been an active research area with vast and quick development in several applications such as robotics, autonomous cars, surveillance, and so on. This magnificent evolution results from many factors that have an impact on improving the performance of the new approaches and overcoming their drawbacks in this field. Background subtraction has become fertile ground in computer vision due to the limitless challenges that provoke researchers to seek and propose new solutions. This section will summarize the most well-known and very recent methods that may be interesting for the new researchers in this domain. These methods can be categorized into many kinds of classes, but we preferred to split them into two kinds of approaches: (i) Traditional-based techniques (Modeling and clustering); (ii) CNN-based techniques.

2.4.1 Traditional approaches

As we mentioned at the beginning of the previous chapter, the initial idea of background subtraction goes back to 1975 of the last century [2]; from then until now, any unsupervised method based on subtraction, modeling, statistics, and/or clustering is considered a traditional (or classical) approach.

Mixture of Gaussians (MoG) [4] is one of the most well-known classical approaches for background subtraction. This method is based on statistics; it consists of modeling each individual pixel into color intensities using a Mixture of Gaussian probability density functions. Due to its background model update system, this method is one of the early approaches that handle multiple challenges in background subtraction, such as light changing, shadows, objects overlapping, etc. It has been improved many times; one of the most riveting studies is presented by Pakorn and Richard [18]; it is an improved version dedicated to better detection in the shadow category to prevent labeling the shadow as a moving object. At the same time, the authors consider the processing speed to make this method run for real-time detection. Later, Zoran [19] added a new important feature for the GMM since it is considered a parametric approach; this new version includes a constant update for the parameters and improves selecting the proper amount of components for each pixel. The next version of GMM was presented by Lee[20], who presented an efficient technique to balance the model convergence speed and stability. In a new paper, again, Zoran and Ferdinand [21] presented a new improvement developing their previous work [19] by creating a new formula to automatically update the GMM parameters. Some interesting studies also included Fuzzy Logic for improving the MoG approach. Zhao *et al.*[22] used T2-FGMM supported by Markov Random Field (MRF) to solve dynamic backgrounds such as waving trees and water rippling. T2-FGMM also has been improved by Darwich *et al.*[23], involving more extra steps to make accurate decisions named IV-FGMM. Figure 2.7 tells the differences between the classical GMM, T2-FGMM, and IV-FGMM algorithms.

One of the critical limitations of the traditional background subtraction approaches back then was the presence of the parameters adjustment process. Some studies presented automatic approaches to handle this issue; on the other hand, several studies

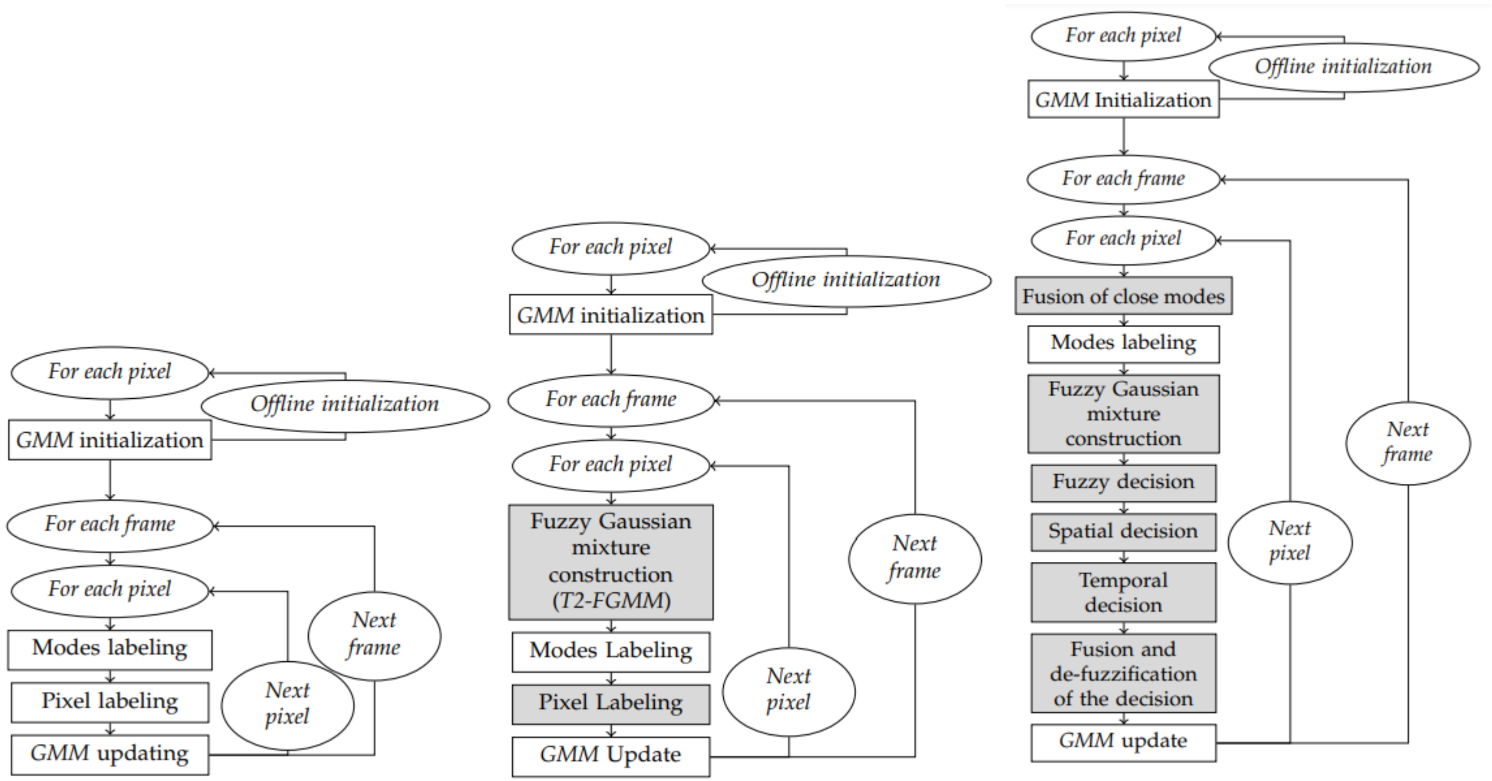


Figure 2.7: Classical GMM, T2-FGMM, and IV-FGMM algorithms, from left to right respectivelly [23].

appear to introduce a new concept: non-parametric methods. Researchers were developing these approaches to generate a new scheme that can help the user interact less with the systems in multiple scenarios. Using statistical estimation based on local intensity observations, Elgammal *et al.*[24] proposed Kernel Density Estimation (KDE) Background Subtraction as a non-parametric approach. This method is based on local intensity observations and estimates the background probability statistically at the pixel-level. This method was an inspiration for the following background subtraction approaches. Hanzi and David [25] proposed a non-parametric method that provides three tasks: Background subtraction, Segmentation, and Tracking objects. Their targets focused on handling multiple conditions at a time and solving the problem of human segmentation occlusion. Figure 2.8 shows the main steps of their proposed method.

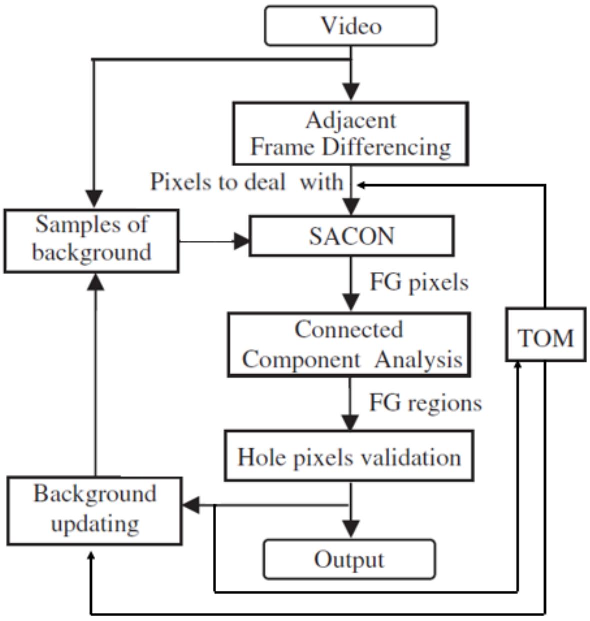


Figure 2.8: The framework presented by Hanzi and David[25].

Most methods in the last decade are considered non-parametric because of the evolutionary ideas presented in the previous two studies. Bin and Piotr [26] presented a new method that contains twenty-one parameters. These parameters are fixed to cope with several situations to solve the background subtraction. This study aims to produce a method that proportionally provides real-time performance with high accuracy. Later, Graciela and Mario I. proposed AAPSA [27], a complex framework (represented in Figure 2.9) that contains several steps, including two background models with three inputs, Analysis of Specialized Modules (ASM), then an output that feeds the system again to update the background model. This approach performed well in two public datasets at that time using the same parameter values.

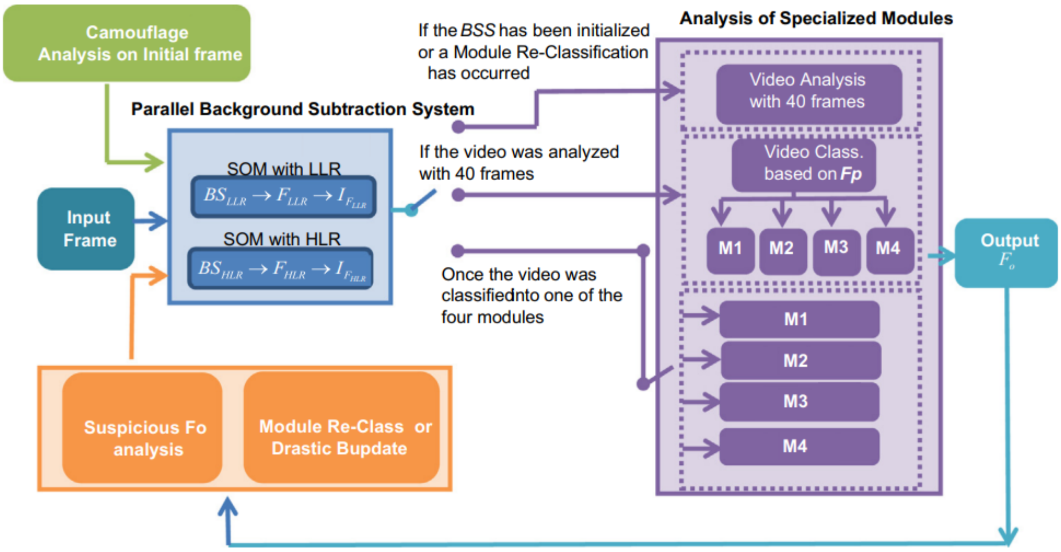

Camouflage Analysis on Initial frame
Parallel Background Subtraction System
If the *BSS* has been initialized or a Module Re-Classification has occurred
Analysis of Specialized Modules
Video Analysis with 40 frames
Input Frame
SOM with LLR
$BS_{LLR} \rightarrow F_{LLR} \rightarrow I_{F_{LLR}}$
SOM with HLR
$BS_{HLR} \rightarrow F_{HLR} \rightarrow I_{F_{HLR}}$
If the video was analyzed with 40 frames
Video Class. based on ***Fp***
M1
M2
M3
M4
Output F_o
Once the video was classifiednto one of the four modules
M1
M2
M3
M4
Suspicious Fo analysis
Module Re-Class or Drastic Bupdate

Figure 2.9: The AAPSA framework presented by Graciela and Mario I.[27].

On the other hand, clustering was an effective tool to improve background subtraction performance. CodeBook by Kim *et al.*[28] was one of the first and most well-known methods based on creating the background model from a set of frames in the learning phase. We can simply explain the way of creating a CodeBook background model. The model has the exact size of the video frames, and each pixel will be represented separately in this model by a victor named: CodeBook. Each CodeBook may contain between 1 to N CodeWords (N is the number of the learning frames). Figure 2.10 explains how the CodeBook model should be constructed. Later, during the testing phase, the algorithm will compare the built model to the current frame and estimate the background. At the same time, an updating process is applied to re-adapt the current model with the new situations appearing in the scene. The background model is frequently updated based on the number of appearances and duration of the objects from the previous testing frames.

Olivier and Marc [30] presented an approach called ViBe, which is considered one of the simplest methods in the background subtraction field. It is a pixel-based method that creates a background model containing the intensity information for each pixel position. Compared to other methods, this was the first approach to creating the background model from the first frame using a random strategy that uses the pixel neighbors. This

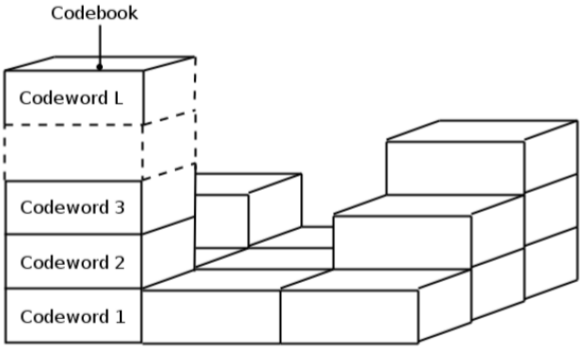


Figure 2.10: An example of the CodeBook background model[29].

background model also is frequently and randomly updated after each newly treated frame during the testing phase. This method has many chances to be improved, which makes it a fascinating technique for researchers. Many other approaches were inspired by CodeBook and ViBe such as [31–35]. One of the most recent studies that were inspired by the ViBe approach is the one proposed by Wang *et al.*[36]. Their approach, V-ViBe, is dedicated to detecting veins of palms, back of hands, and fingers using background subtraction. The main improvement is represented in creating two kinds of background sample sets, one is static, and the second is dynamic. Then in the step of pixel values comparison, they used two radios, R_1 and R_2, and the Euclidean distance to determine the matching points for the output results and update the background model. Figure 2.11 explains the differences between the original method ViBe and the proposed version V-ViBe, specifically in the level of comparison and updating process. Figure 2.12 summarizes the proposed framework of this method.

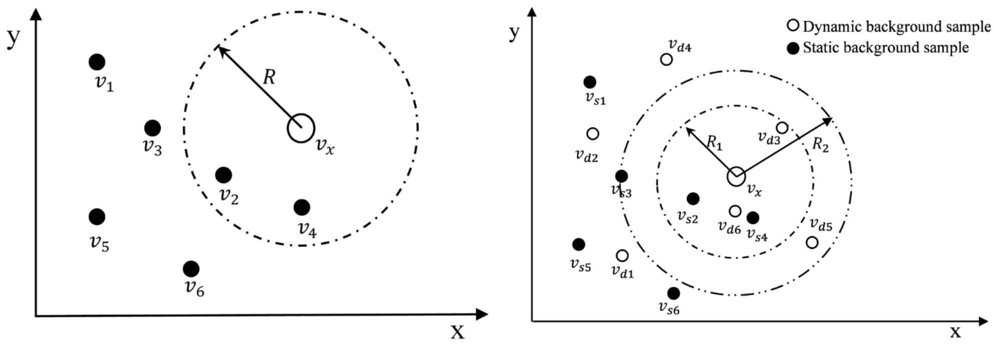


Figure 2.11: The difference between the comparison process of ViBe (left) and V-ViBe (right)[36].

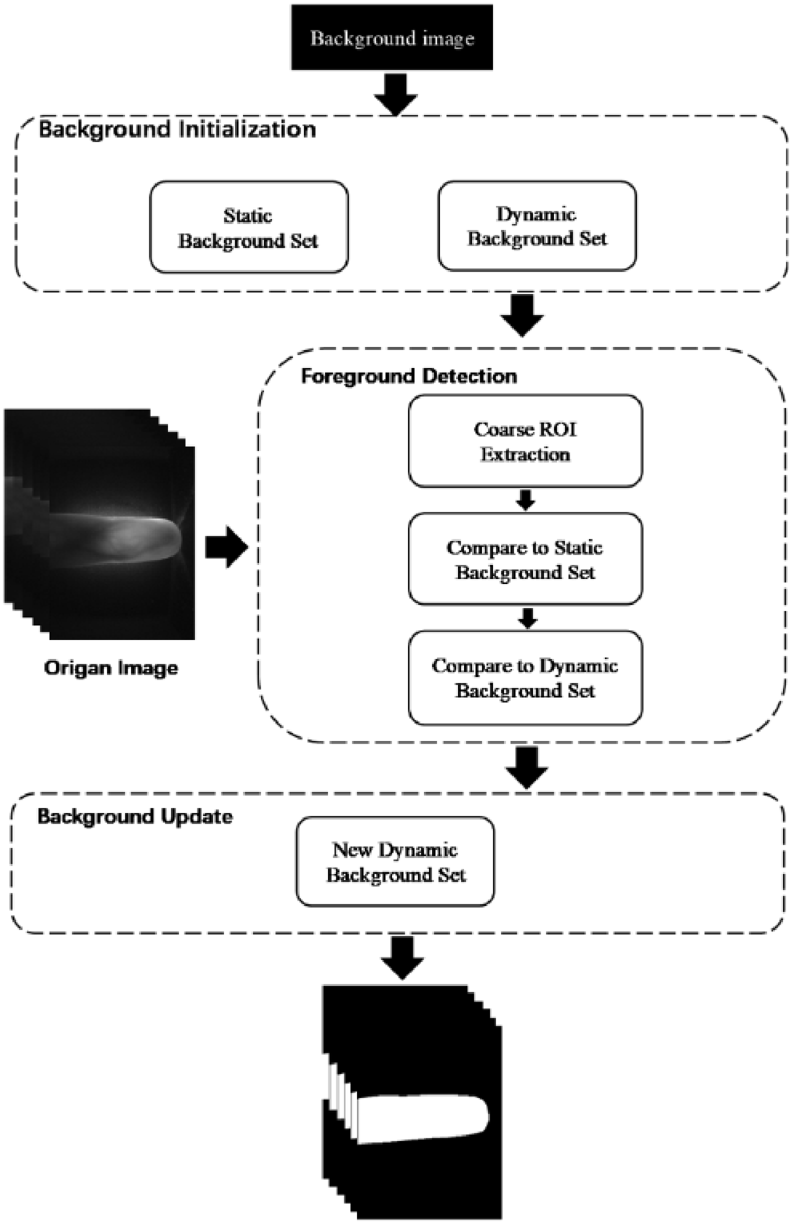


Figure 2.12: The proposed framework by Wang *et al.*[36].

However, most of the techniques were based on RGB channels or Gray-scale data. Some studies appeared to point out the ability to use other kinds of data that may provide useful information about the scene. Heikkila and Pietikainen [37] used the LBP calculation of the RGB frame as an input for their system. The background model in this method, Texture-Based MODeling (TBMOD), is built on a group of adaptive LBP histograms for each pixel (see Figure 2.13). Next, in the background detection process, each pixel from the current frame will then be compared to the corresponding group of LBP histograms from the background model using the histogram intersection.

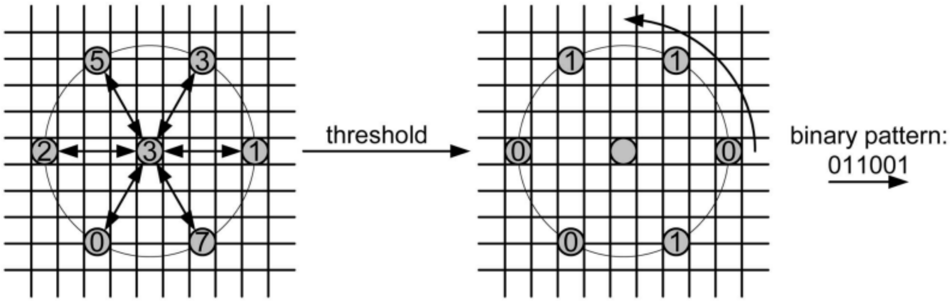


Figure 2.13: LBP calculation[37].

The TBMOD method has been an inspiration for utilizing other helpful descriptors in the background modeling field. SeungJong and Moongu [31] proposed a new framework based on three primary information: pixel texture Scene Adaptive Local Binary Patterns (SALBP), pixel color RGB and region appearance. This framework is represented in Figure2.14, which shows the four main steps of this scheme to reach the final results (Preprocessing, Background modeling, Background subtraction, and Model maintenance). This framework also involves some updated features from the CodeBook method [28].

Guillaume-Alexandre *et al.*[38] proposed another interesting descriptor named LBSP. Figure 2.15 perfectly explains the thorough process of change detection using LBSP. It shows an example of the outcomes of four distinct binary operations on image regions (B and NF). Bc and NFc represent the center pixels of sections B and NF, respectively. The pixels surrounding the center of the areas B and NF are denoted by Bn and NFn, respectively. The first and second columns at the bottom of the figure employ comparisons to produce binary representations of the regions using intra-region operations in one case and both intra-region and inter-region operations in the other. The third and fourth columns employ absolute difference thresholding to construct binary representations utilizing intra-region operations in one case and both intra-region and inter-region operations in the other. A threshold of 5 is utilized in all cases. Because of its pixel-based design, LBSP may be computed both inside an area in an image and across regions between different images or two regions within the same image to capture variations in intensity and texture. These features made the LBSP valuable to be adapted in future background subtraction approaches.

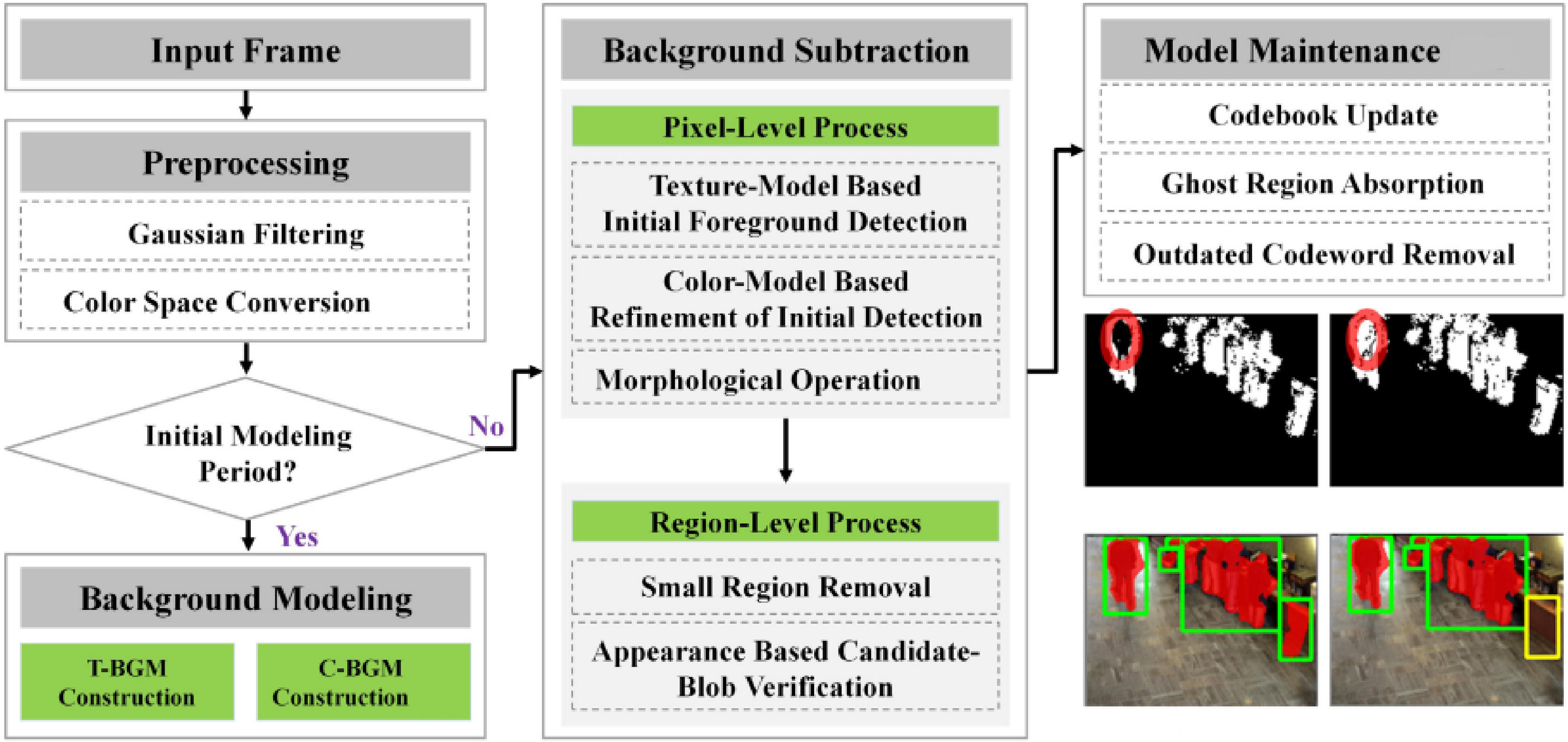


Figure 2.14: SeungJong and Moongu framework[31].

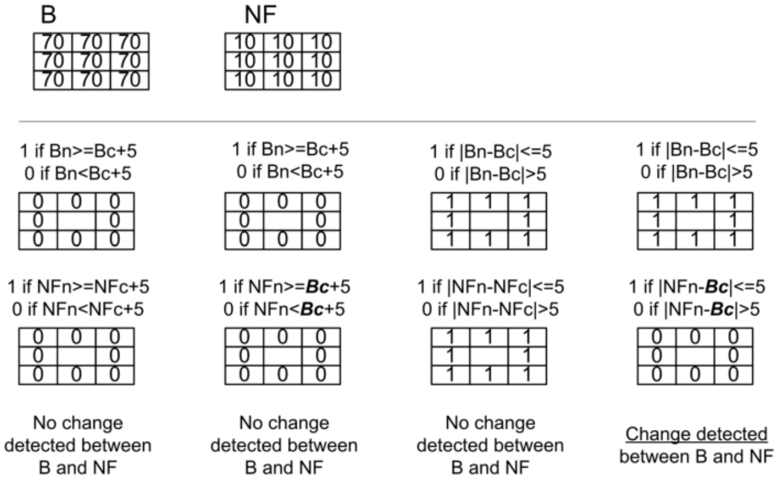


Figure 2.15: An example to understand the LBSP process[38].

SuBSENSE[33, 34] is one of the well-known approaches in the background subtraction domain. This method is characterized by the feature of automatic adjustment of the parameters, which makes it a non-parametric approach that solves multiple scenarios, but mostly it was made to target *Camouflage* and *IlluminationChange* categories. The main framework of this method is to build a background model that consists of two pieces of information, 8-bit RGB intensities, and 16-bit LBSP binary strings, then using Hamming distance to determine if the corresponding pixel from the current frame is considered as a background or foreground pixel. More details about the framework of this approach are presented in Figure2.16.

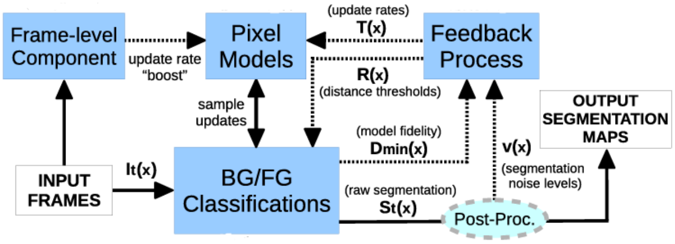


Figure 2.16: SuBSENSE framework [33].

The same authors proposed a developed version of SuBSENSE which was then named PAWCS[35, 39]. This new version is based on building the model from a triplet of Color-LBSP-Persistence for each pixel. This triplet is defined as Background Words (BWs). The Persistence is inspired by the Maximum Negative Run-Length (MNRL) from the CodeBook method, which determines the less important BWs to be eliminated from the background model and keeps only the most important ones. This adapted feature plus the internal parameters update caused more complexity compared to SuBSENSE but definitely assured better precision in different stages. Figure 2.17 shows the PAWCS framework in a simplified diagram.

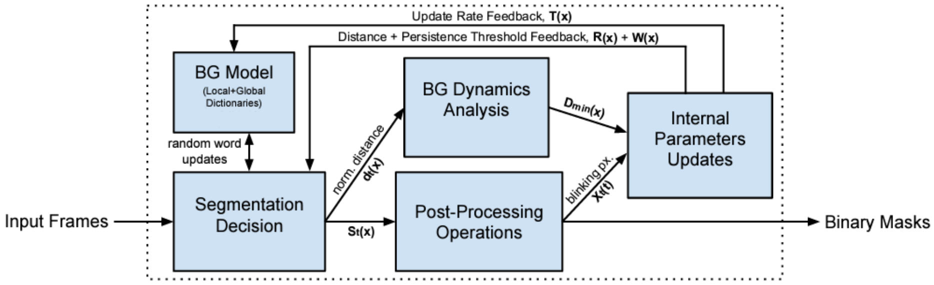

Update Rate Feedback, T(x)
Distance + Persistence Threshold Feedback, R(x) + W(x)
BG Model
(Local+Global Dictionaries)
random word updates
BG Dynamics Analysis
$D_{min}(x)$
Internal Parameters Updates
norm. distance
$d_t(x)$
blinking px.
$X_t(t)$
Input Frames
Segmentation Decision
$S_t(x)$
Post-Processing Operations
Binary Masks

Figure 2.17: PAWCS framework [35].

Machine learning has also made significant contributions to the background subtraction area in the last few years. By including Kernel Density Approximation (KDA) and Support Vector Machine (SVM) into their model, Bohyung and Larry[40] proposed a technique that uses color, gradient, and Haar-like characteristics, solving three main challenges in the field: Spatio-temporal variations, shadow, and illumination changes. The authors simplify the training approach used in this method in Figure 2.18, where the background modeling is achieved using KDA, then the background/foreground classifier is built based on SVM. Despite using a one-dimensional KDA to overcome the slow performance, the complexity of this approach keeps it far from real-time performance.

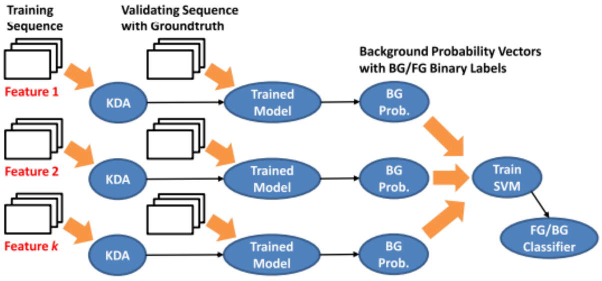
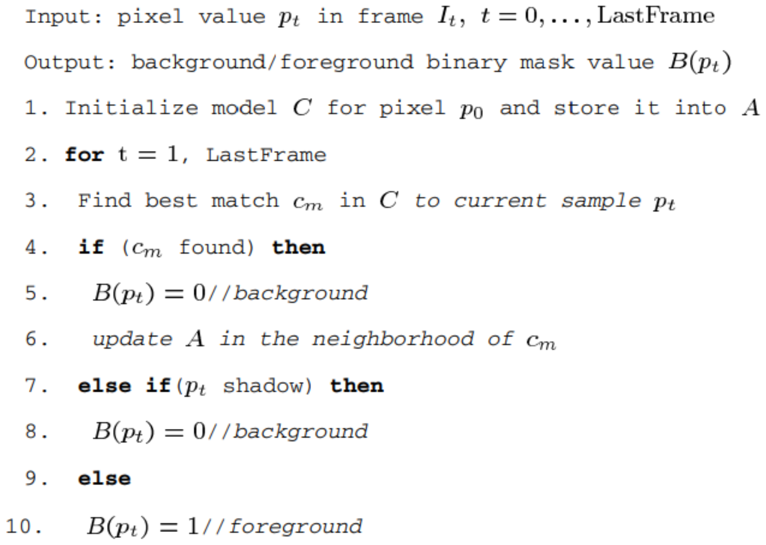


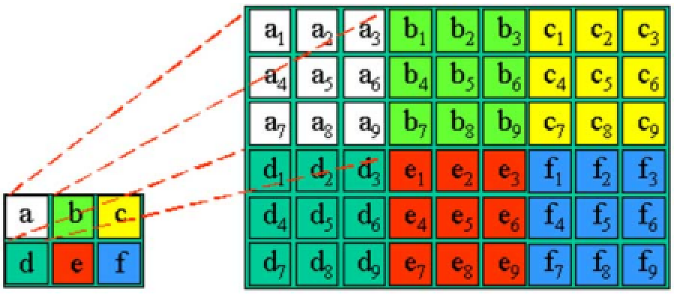
Figure 2.18: The framework used in Bohyung and Larry [40] paper.

Maddalena and Petrosino [41] used Artificial Neural Network (ANN) to build the background model on their proposed method, Self-Organizing Background Subtraction (SOBS). This method stands for representing each pixel by multiple neurons to store as much information about the scene; an example of the background model initialization is represented in Figure 2.19. The authors consider this method a self-organizing approach, referring to its ability to adapt to dynamic scenarios automatically. This technique uses a different color space instead of RGB to build the background model. They consider the Hue-Saturation-Value (HSV) because they mention that it allows for the expression of colors as a human eye experience. This approach process is simplified in the following algorithm (presented in the original paper):

Algorithm 1: SOBS (Self-Organizing Background Subtraction)[41]



```
Input: pixel value $p_t$ in frame $I_t$, $t = 0, \ldots, \mathrm{LastFrame}$
Output: background/foreground binary mask value $B(p_t)$
Initialize model $C$ for pixel $p_0$ and store it into $A$
for t = 1, LastFrame
 Find best match $c_m$ in $C$ to current sample $p_t$
 if ($c_m$ found) then
  $B(p_t) = 0$//background
  update $A$ in the neighborhood of $c_m$
 else if($p_t$ shadow) then
  $B(p_t) = 0$//background
 else
  $B(p_t) = 1$//foreground
```



a	b	c
d	e	f

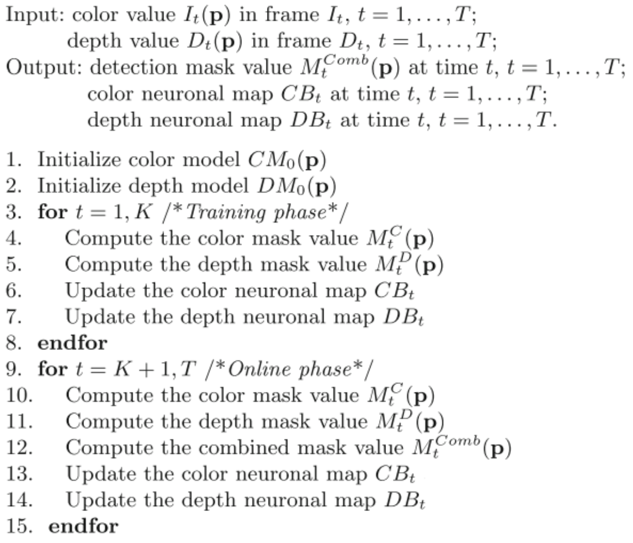
a_1	a_2	a_3	b_1	b_2	b_3	c_1	c_2	c_3
a_4	a_5	a_6	b_4	b_5	b_6	c_4	c_5	c_6
a_7	a_8	a_9	b_7	b_8	b_9	c_7	c_8	c_9
d_1	d_2	d_3	e_1	e_2	e_3	f_1	f_2	f_3
d_4	d_5	d_6	e_4	e_5	e_6	f_4	f_5	f_6
d_7	d_8	d_9	e_7	e_8	e_9	f_7	f_8	f_9

Figure 2.19: An example to go from 2x3 image to 6x9 neuronal map structure[41].

The same authors upgraded their approach, called SOBS_CF[42]. This new version has two additional features compared to the original approach. One is related to incorporating spatial coherence, exploiting the contiguous pixel intensity difference. The second is to automatically update the background model using a novel Fuzzy approach proposed by the authors.

Maddalena and Petrosino again proposed a new version of SOBS under the SBM-RGBD challenge in 2017[12], named RGBD-SOBS[43]. This last version proves the additional value that can be provided by the depth maps along with the RGB color information to deal with some specific challenges such as illumination changes and shadows. These challenges are hard to handle when it comes to an approach that relies only on RGB channels. The new proposed algorithm is presented as follows:

Algorithm 2: RGBD-SOBS[43]



```
Input: color value $I_t(\mathbf{p})$ in frame $I_t$, $t = 1, \ldots, T$;
       depth value $D_t(\mathbf{p})$ in frame $D_t$, $t = 1, \ldots, T$;
Output: detection mask value $M_t^{Comb}(\mathbf{p})$ at time $t$, $t = 1, \ldots, T$;
        color neuronal map $CB_t$ at time $t$, $t = 1, \ldots, T$;
        depth neuronal map $DB_t$ at time $t$, $t = 1, \ldots, T$.

Initialize color model $CM_0(\mathbf{p})$
Initialize depth model $DM_0(\mathbf{p})$
for $t = 1, K$ /*Training phase*/
    Compute the color mask value $M_t^C(\mathbf{p})$
    Compute the depth mask value $M_t^D(\mathbf{p})$
    Update the color neuronal map $CB_t$
    Update the depth neuronal map $DB_t$
endfor
for $t = K + 1, T$ /*Online phase*/
    Compute the color mask value $M_t^C(\mathbf{p})$
    Compute the depth mask value $M_t^D(\mathbf{p})$
    Compute the combined mask value $M_t^{Comb}(\mathbf{p})$
    Update the color neuronal map $CB_t$
    Update the depth neuronal map $DB_t$
endfor
```

More related works on traditional RGBD-based approaches for background subtraction connect directly with our main study in this thesis. The authors in [44] used a ToF camera called Photonic Mixer Device (PMD), which provides three kinds of information in three parallel matrices (distance, amplitude, and intensity), plus the Grayscale information provided by an RGB camera. The principle of their method is to extract the background subtraction from each matrix individually and then apply specific logical operations between them to reach the final results. Figure2.20 illustrates the followed framework proposed by the authors.

Ottonelli *et al.* proposed in their paper[45] another schema to fuse color and depth information. It stands for considering both background subtraction results of RGB and

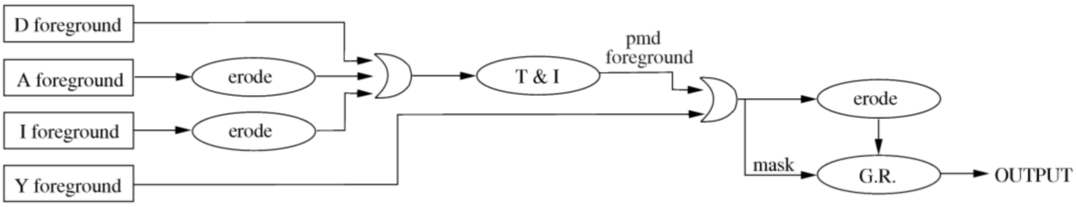


Figure 2.20: The proposed framework from Leens *et al.* [44]. Where, G.R. stands for Geodesic Reconstruction and T&I stands for Transformation and Interpolation.

Depth independently, plus a grayscale version of the RGB frame, then computing a Depth-based CF, and finally applying an "OR" logical operation between the CF and the RGB background subtraction result followed by a noise removal. The authors of this paper used the ViBe technique to extract the preliminary background subtraction results, as we can notice in Figure 2.21, which explains the global framework. Figure 2.22 illustrates the CF utility in this framework.

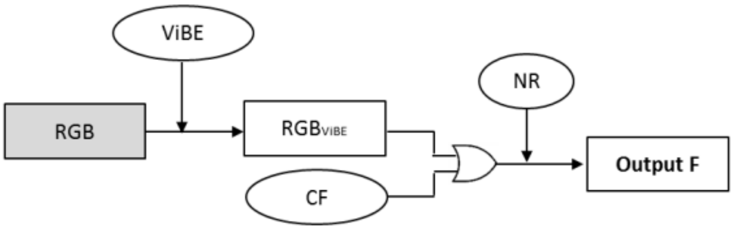


Figure 2.21: The proposed framework by Ottonelli *et al.* [45].

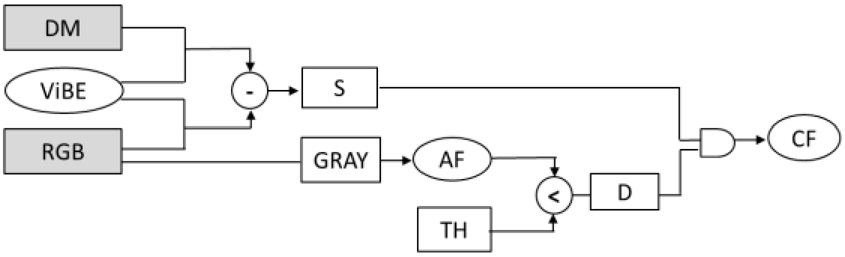


Figure 2.22: The CF explanation by Ottonelli *et al.*[45]. Where, S represents the "Subtraction mask", AF stands for "Averaging filter", D stands for "Depth enhanced mask" and TH is the decision threshold in pixel unit.

Huang *et al.*[46], proposed a framework that also uses the ViBe approach to extract the background separately from RGB and Depth information, then apply a weighted formula to fuse these results, followed by an adaptive refinement with Spatio-temporal consistency based on edge detection, and finally update the background model with the original ViBe updating technique. The proposed framework is represented in Figure2.23.

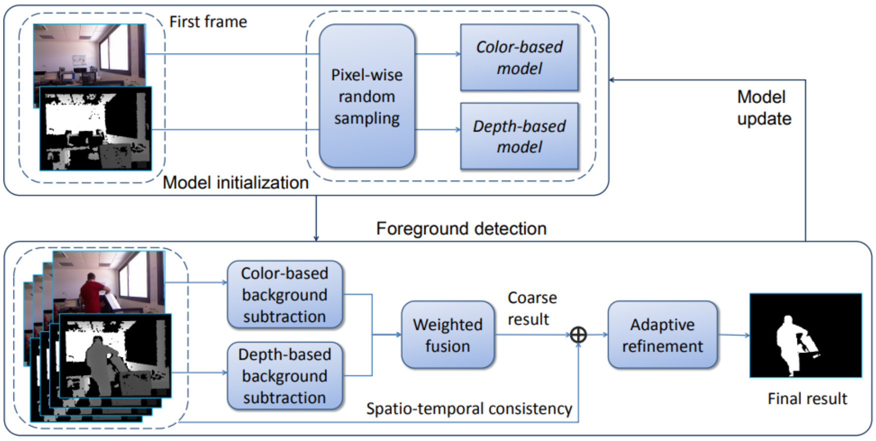


Figure 2.23: The proposed framework by Huang *et al.* [46].

2.4.2 CNN-based approaches

Deep Learning has recently proven to be one of the top solutions for various tasks. Background subtraction is one of the fields where Deep Learning succeeded in the majority of datasets and competitions [6, 9, 12, 47]. Braham and Van Droogenbroeck [48] pioneered the use of Convolutional Neural Networks (CNN) in background subtraction. The ConvNet architecture is very similar to the handwriting digit classification LeNet-5[49]. This network was appropriate for the Background Subtraction task. Its main framework is to choose a background reference from the tested video and a set of frames from the same video that include moving objects plus their corresponding groundtruth. These three inputs are given to the CNN network to create a background model that will be used later for the testing phase. An illustration of the ConvNet model is presented in Figure 2.24.

Zeng and Zhu [50] used a U-Net-kind architecture, named MFCN, inspired by the Fully Convolutional Networks (FCN)[51]. As illustrated in Figure 2.25, the skipping convolutional layers and the additional operations linking the encoder-decoder on different stages make the feature extraction more efficient to build an effective background

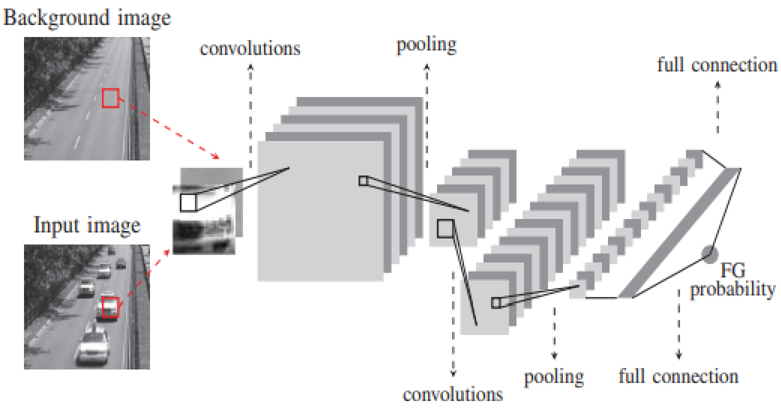


Figure 2.24: The model of ConvNet as represented in [48].

model. The interesting feature of this architecture is that, it does not require a reference frame to detect the background. However, the followed training protocol causes Training/Testing overlapping that affects the final results. This flop can be easily detected in the framework illustrated by the authors in the original paper (See Figure 2.26), where they split the frames of the same video for training and testing.

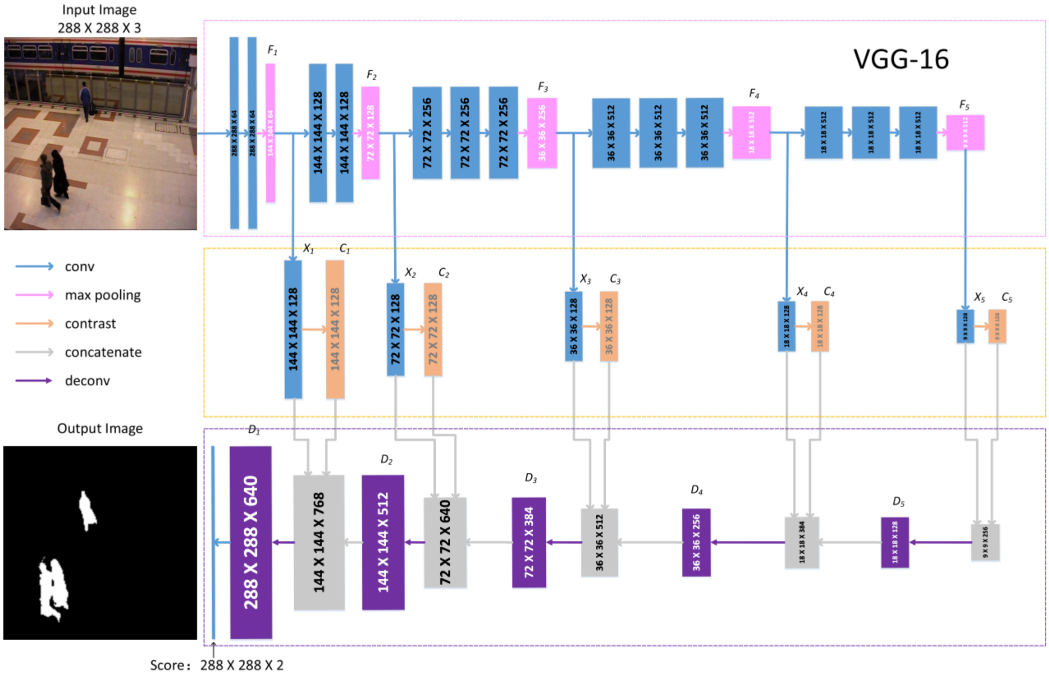


Figure 2.25: The MFCN model [50].

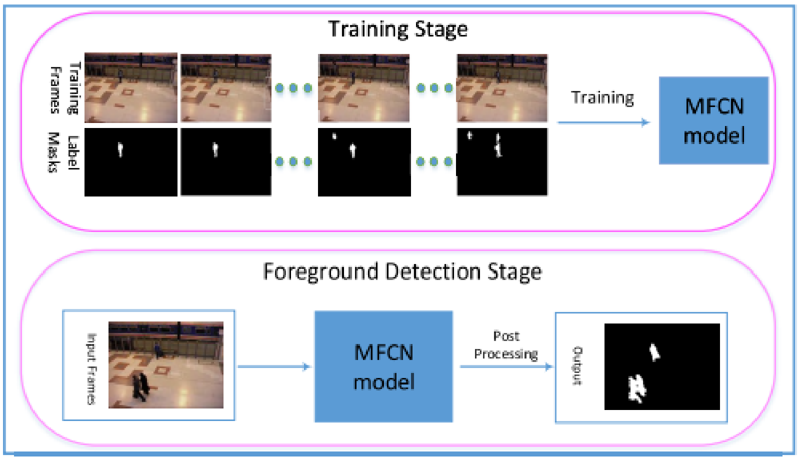


Figure 2.26: The MFCN framework [50].

Zhao and Basu[52] created a new approach named Deep Pixel Distribution Learning (DPDL). This approach uses the model of ConvNet[48] in the deep learning modeling stage. This method contains two blocks: The first one is the features generation pixel distribution, using a technique named RPoTP. The second block is to feed the network model with these features and build a model that will be used later for solving the background subtraction. An illustration of this approach is presented in Figure 2.27; where the input to arithmetic distribution layers for learning distributions is a histogram of subtractions between a pixel's current observation and its past counterparts. A convolutional operation is used to integrate the output histograms of the arithmetic distribution layers, which are then fed into a classification architecture including a fully connected layer for classification, a Rectified Linear Unit (Relu) layer, and a convolution layer.

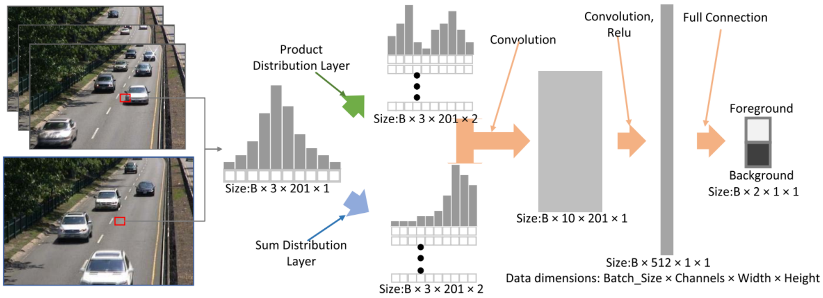


Figure 2.27: The RPoTP framework [52].

This approach (RPoTP) has been developed by the same authors in their next study[53]. This improved version, D-DPDL, has a more advanced RPoTP technique for feature generation where the authors focused on making the features dynamically generated during the training to prevent the network from over-fitting the pattern implied in random permutations. Also, to reduce the noise, they proposed a Bayesian refinement model for post-processing. Figure 2.28 summarizes the pipeline proposed by the authors in the original paper.

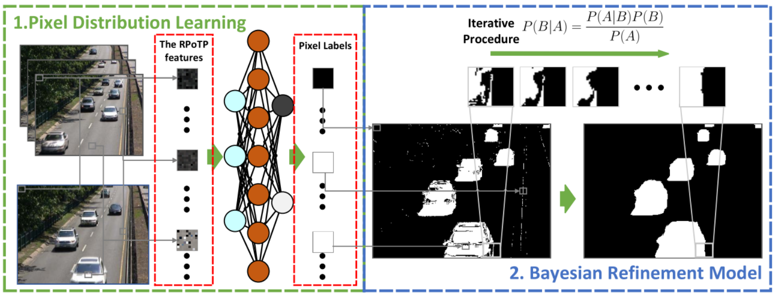

1.Pixel Distribution Learning
The RPoTP features
Pixel Labels
Iterative Procedure
$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$
2. Bayesian Refinement Model

Figure 2.28: The D-DPDL pipeline [53].

Lim and Keles [54] proposed the FgSegNet, in which they use different scales of the same frame input to obtain information diversity. This architecture assures the parallel convolutional filters to scan the entire foreground object rather than just a tiny part, which makes the model absorb more features of the scene. An upgraded version, named FgSegNet V2 [55], was made by the same authors, where they improved the Feature Pooling Module (FPM) with two additional skip connections containing convolutional and global average pooling layers. The authors also avoided the 3-scaled feature to gain less complexity and speed up the process. These modifications make significant improvements and put the FgSegNet V2, again, in the top-ranked methods among all previous methods on the Change Detection challenge[6]. The difference between FgSegNet and FgSegNet V2 architectures is illustrated in Figure 2.29 from the original papers.

The same model (FgSegNet V2) inspired Liu *et al.* [56] to propose a background subtraction method based on multispectral images and deep learning. Simply, the authors extracted three channels of the multispectral images provided by the chosen dataset, then created a model (inspired by FgSegNet V2) that can adapt with this kind of information to be used later for the background subtraction operation. Figure 2.30 shows the proposed model mentioned in the authors' paper.

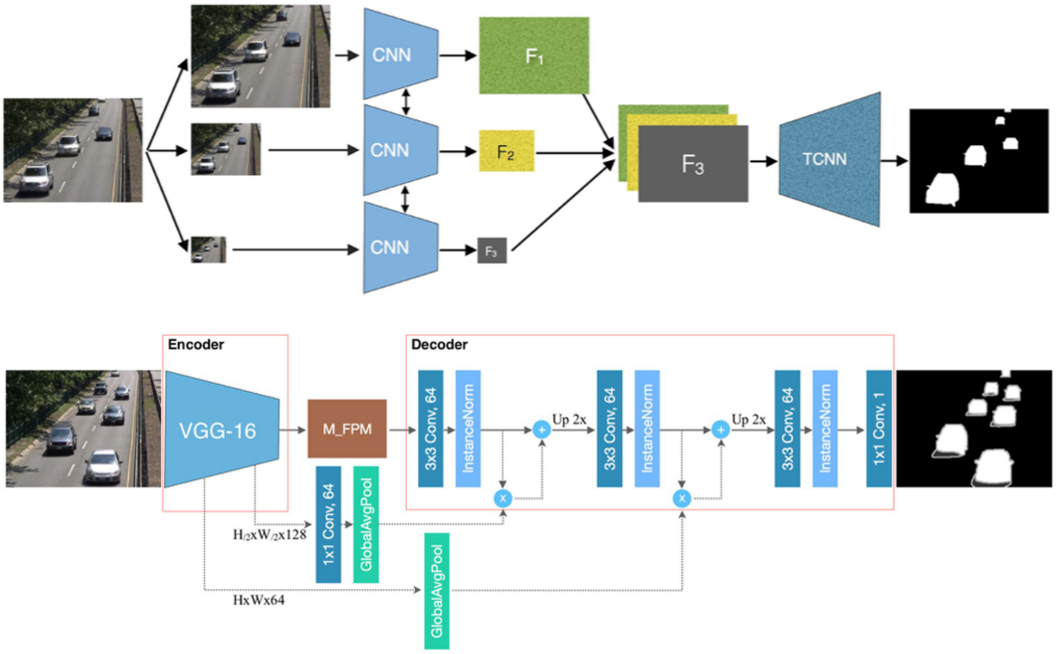


Figure 2.29: The difference between FgSegNet and FgSegNet V2 architectures [54, 55].

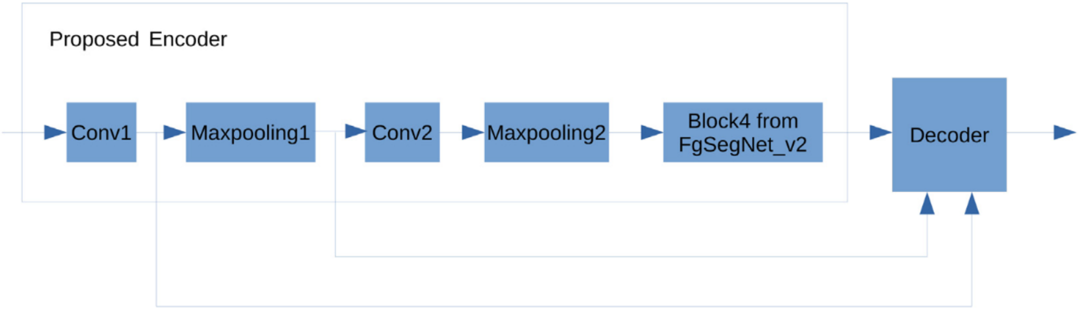


Figure 2.30: Multispectral background subtraction architecture[56].

As one of the first studies to apply Deep-Learning (DL)-based background subtraction on unseen videos, Tezcan *et al.* [57, 58] used a new fully-convolutional neural network named BSUV-Net. In this study, they used as inputs: the current frame and two other frames considered as background samples (the recent frame and another one with no foreground objects) plus a foreground probability map for each frame obtained from applying semantic segmentation (See Figure 2.31). However, the inputs used for this model in order to perform the background subtraction required a reference frame. We can argue that using a reference frame will give the system a preview of the targeted scenario because this frame has to be chosen by the user to be a reference of how the background should be. This approach may cause low performance in many scenarios, especially Bootstrapping category, where there will be moving objects in all the sequence's frames.

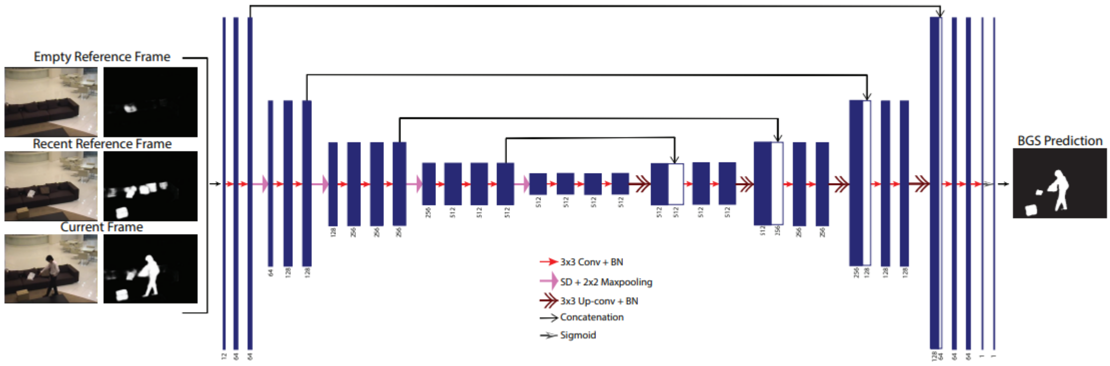


Figure 2.31: The BSUV-Net architecture [57].

Generative Adversarial Networks (GANs)[59] have also been applied in background modeling. GANs are built of two main parts; both of them are CNN-based models. The first model is named the generator, and the second one is the discriminator. The generator is made to generate new examples; these examples (named the fake data) will then go through the discriminator along with the dataset examples (named the real data) to define if the generator is making more similar examples compared to the dataset examples or not depending on the critical setup level made by the user.

Bakkay *et al.* [60] proposed a new approach for background subtraction using GANs. In the generator encoder, they used eight convolutional layers followed by six RESidual neural NETwork (ResNet) blocks. On the other hand, the decoder consists of eight deconvolutional layers. Figure 2.32 shows an overview of the proposed framework and the proposed architecture of the BScGAN approach. This method, BScGAN, also

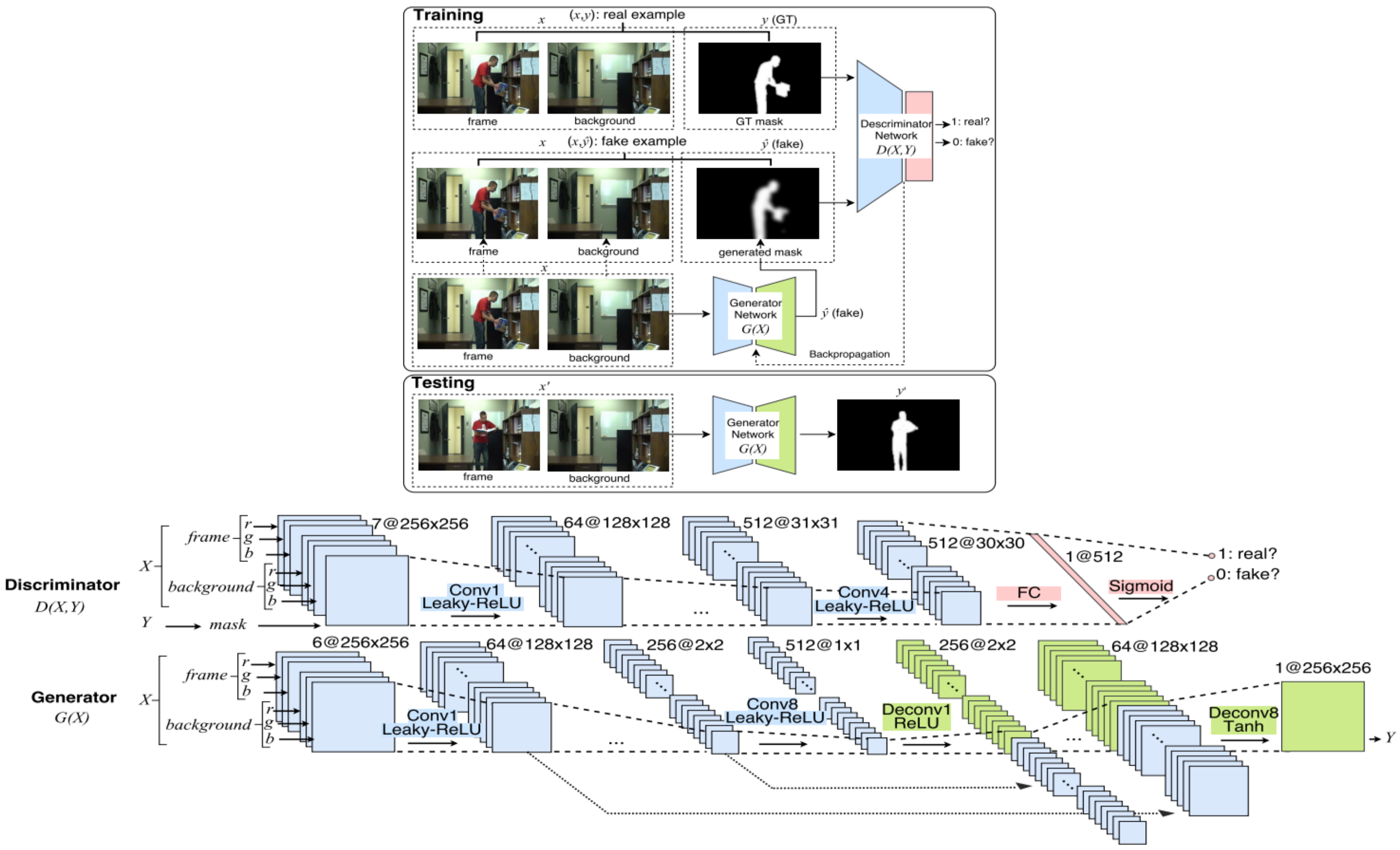


Figure 2.32: The BScGAN proposed framework and architecture [60].

takes into consideration some reference frames to help the model solve the background subtraction, which is one of the cons that we mentioned before.

Sultana *et al.*[61] proposed a framework named DCP. It is based on two models: context prediction initialization and fine texture optimization. They also used the Modified Poisson Blending (MPB) technique for better context prediction. This study is a combination of many techniques; we consider it one of the most complex approaches: Pre-processing using Optical Flow (OF), context estimation using GANs based on AlexNet, Texture optimization using VGG-19, post-processing using MPB, then a classical foreground detection boosted with some morphological operations to optimize the results. A simplified presentation from the original paper in Figure 2.33 is presented to explain the proposed framework.

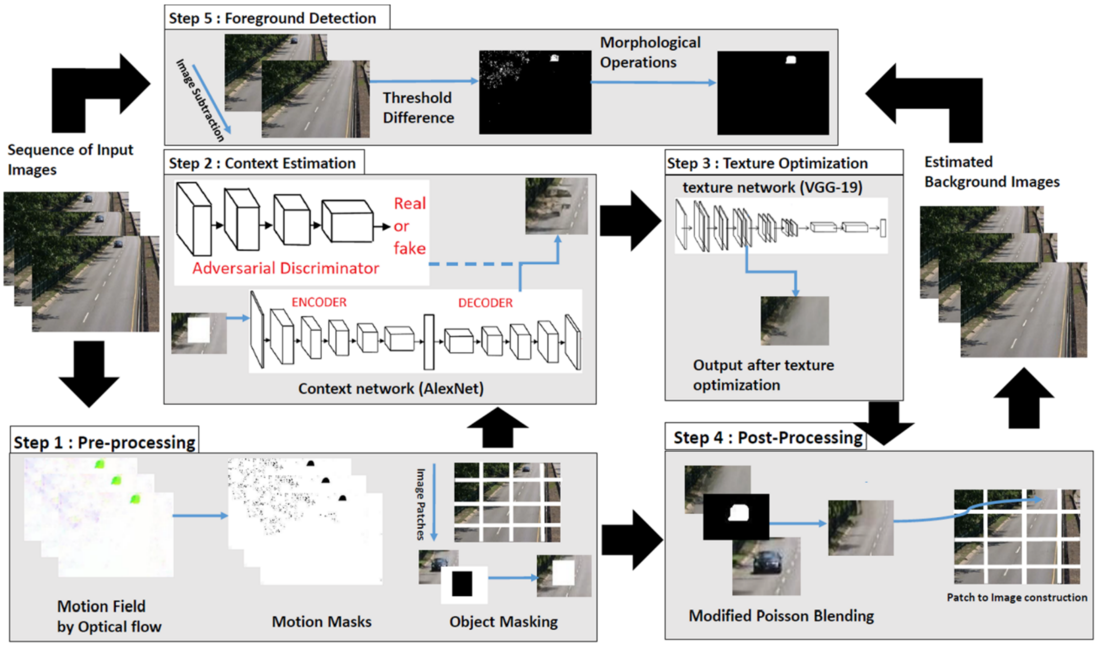
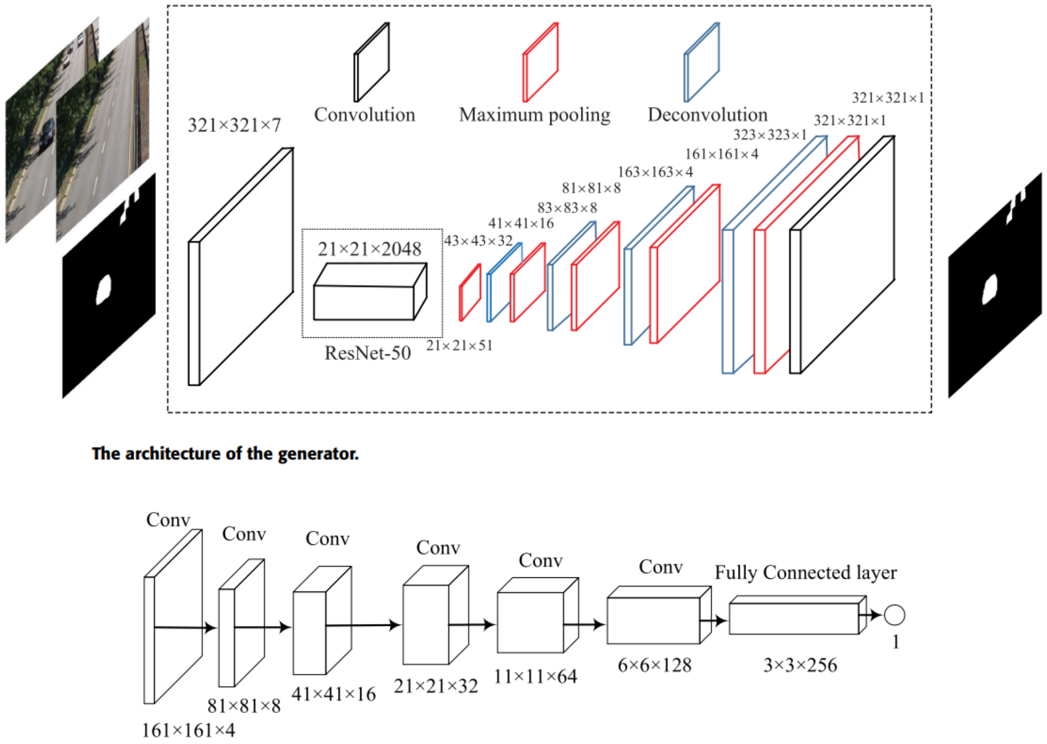

Step 5 : Foreground Detection
Image Subtraction
Threshold Difference
Morphological Operations
Sequence of Input Images
Step 2 : Context Estimation
Real or fake
Adversarial Discriminator
ENCODER
DECODER
Context network (AlexNet)
Step 3 : Texture Optimization
texture network (VGG-19)
Output after texture optimization
Estimated Background Images
Step 1 : Pre-processing
Motion Field by Optical flow
Motion Masks
Image Patches
Object Masking
Step 4 : Post-Processing
Modified Poisson Blending
Patch to Image construction

Figure 2.33: The DCP framework[61].

Moreover, most background subtraction studies focus on natural videos. Yu *et al.* [62] went through both natural videos (obtained by regular cameras) and Very High Resolution (VHR) optical remote sensing videos. Their proposed approach is based on Conditional Generative Adversarial Networks (CGAN) and Domain-Adversarial Neural Networks (DANN), and it has two main sections as any GANs model. The first is the generator based on ResNet-50 with additional layers such as deconvolution (transposed convolution). The inputs used to the generator are three, one is the current frame, the second is a reference frame, and the third is a SuBSENSE[33] foreground mask of the current frame as leverage for the background generating. The discriminator section consists of six successive convolutional layers with a stride of two, ending with a fully connected layer. Figure 2.34 shows the generator and the discriminator architectures.



The architecture of the generator.

The architecture of the discriminator.

Figure 2.34: The CGAN-BS framework[62].

Patil *et al.* [63] proposed a novel GANs-based model that uses temporally sampled multiple frames and spatial features at multiple scales to predict the foreground segmentation from unseen videos. This approach shows an innovative feature extraction method, as illustrated in Figure 2.35. The authors used multiple encoder-decoder (Three) associated with skip connections, and each encoder-decoder produces a background subtraction frame directed to the discriminator (only the result frame) and the next encoder-decoder (associated with three RGB frames as reference frames). The same authors proposed a similar method which is also a GANs-based approach. This time, Patil *et al.*[64] succeeded in creating a framework with less complexity where they used only one frame instead of multiple frame references. This method considers the first frame as a reference frame, which is a disadvantage that commonly affects the background subtraction process in the Bootstrapping category. Another contribution added to this study is the proposed Multi-scale Temporal Edge Aggregation (MTPA) network. They used three MTPAs to process the features sent through the skip connections from the encoder heading to the decoder; These MTPAs are also connected to each other, as illustrated in Figure 2.36.

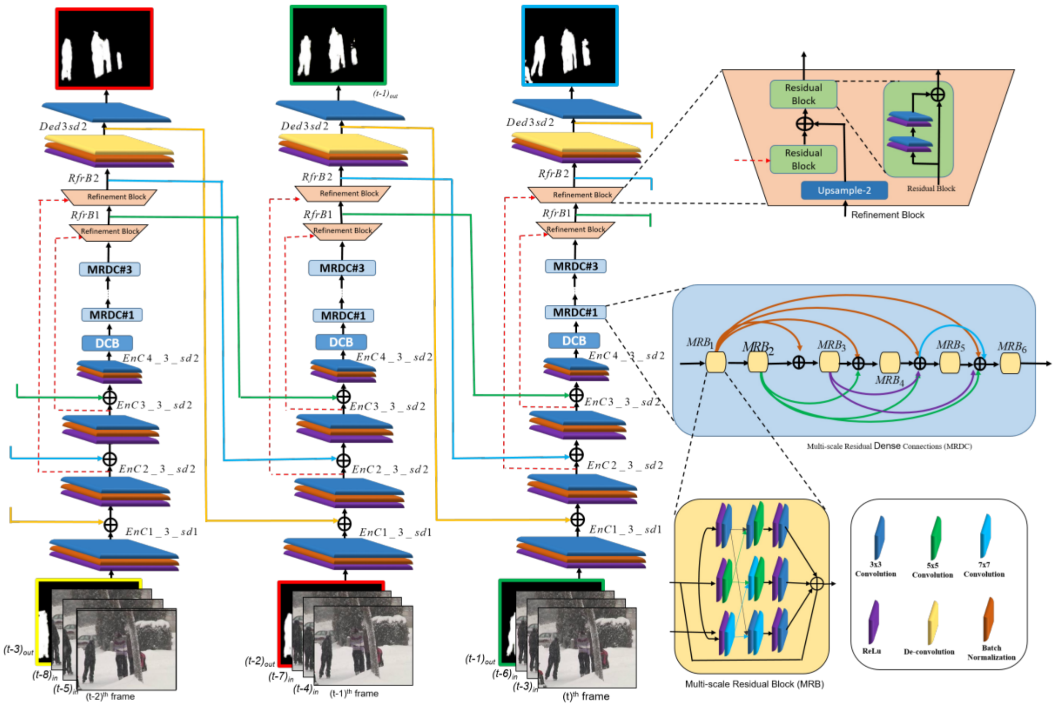


Figure 2.35: The generator architecture presented by Patil *et al.* [63].

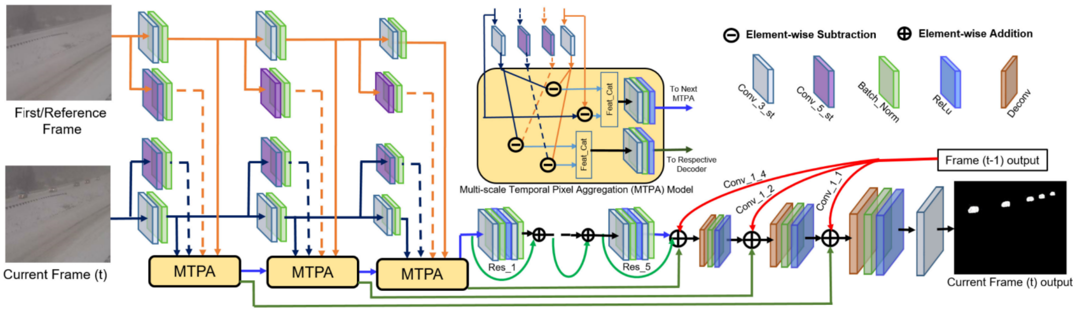


Figure 2.36: The generator architecture presented by Patil *et al.* [64].

Recently, Sultana *et al.* [65] proposed a novel approach named MOS-GAN. It is a background initialization method that uses three loss functions in the training phase and also includes a back-propagation in the testing phase that is also based on three loss terms. The Generator consists of a reshape step then five successive deconvolutional layers. On the other hand, the Discriminator consists of five convolutional layers followed by a fully connected layer. Figure 2.37 shows the training and the testing pipeline, and the architecture of the generator and the discriminator.

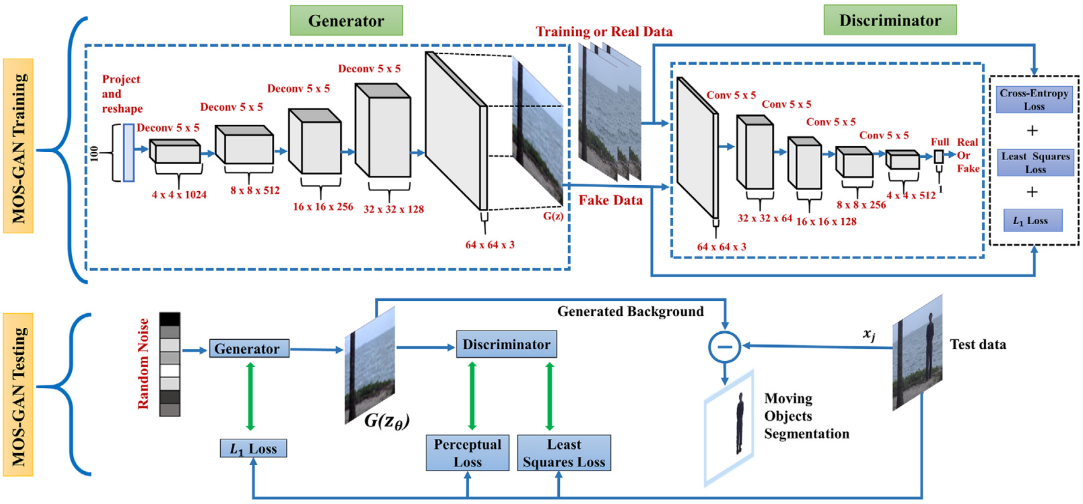


Figure 2.37: The MOS-GAN as presented by Sultana *et al.* [65].

Sultana *et al.* [65], proposed another GANs-based approach named M-cGAN. The generator of this approach is built based on a modified U-net, where the network input and output match the video frame size; and the network layers are contracted using convolutional (encoder) and deconvolutional (decoder) layers from down-sampling to up-sampling the path of the feature; in addition, skip connections are added between the encoder and the decoder; where the generator loss is a combination of Least Squares Loss (LSL), Reconstruction Loss (RL), and Content Loss (CL). On the other side, the discriminator is built of four convolutional layers to down-sample the input into a single feature map where the discriminator operates the final decision. The discriminator has only one loss function which, is the LSL. On the other side, the authors proposed what they called the RLNA, which provides the CL error to the generator network. The RLNA architecture is a down-sampling network constructed of four blocks; each block consists of two convolutional layers and one max-pooling layer (see Figure 2.38). The

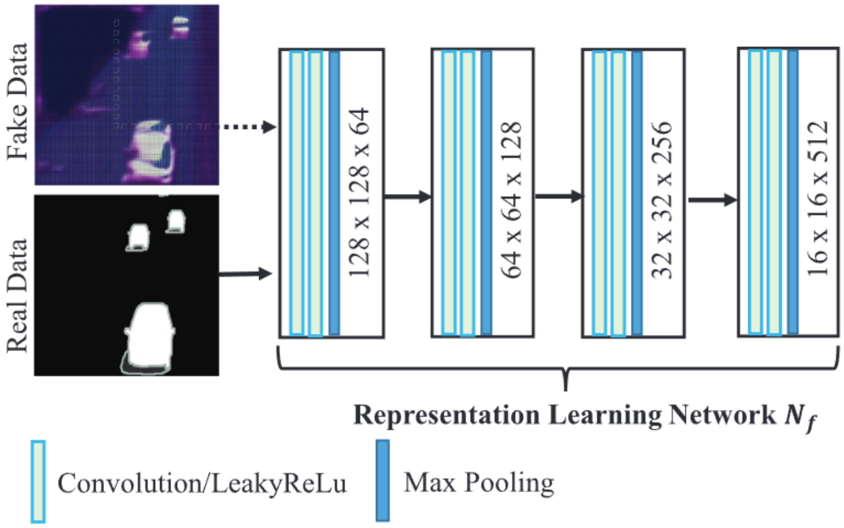


Figure 2.38: The RLNA architecture [65].

authors summarised their proposed framework in one figure, as illustrated in Figure 2.39.

However, despite the attention given to the CNN background subtraction approaches, few CNN-based approaches use depth information or color-depth (Red Green Blue Depth (RGBD)) fusion for background subtraction. To the best of our knowledge, we have found only three noticeable DL studies that give their attention to this approach.

Wang *et al.* [66] proposed BGSNet-D. This CNN-based method uses only depth information to extract the background subtraction. The proposed architecture is a down-sampling network that successively consists of three convolutional layers and three fully-connected layers. This network receives two images as inputs, the first is the current frame, and the second is a reference frame (in this case, it is the first frame of the video). This study also introduces a pre-processing technique to reduce depth data noise caused by the limitation of the depth sensors. An overview of the the BGSNet-D approach is presented in Figure 2.40.

Sultana *et al.* [67] presented an arXiv pre-print introducing a novel GANs-based approach that considers two information data (color and depth) for solving background subtraction. Following the steps illustrated in Figure 2.41, the proposed framework consists of two phases: training and testing. The first phase is divided into two main parts, one is devoted to the RGB generator training, and the second is for the Depth generator training separately. The second phase is divided into three parts; the first is to generate the background (4) and extract the BS (5) from a triplet that contains: the RGB

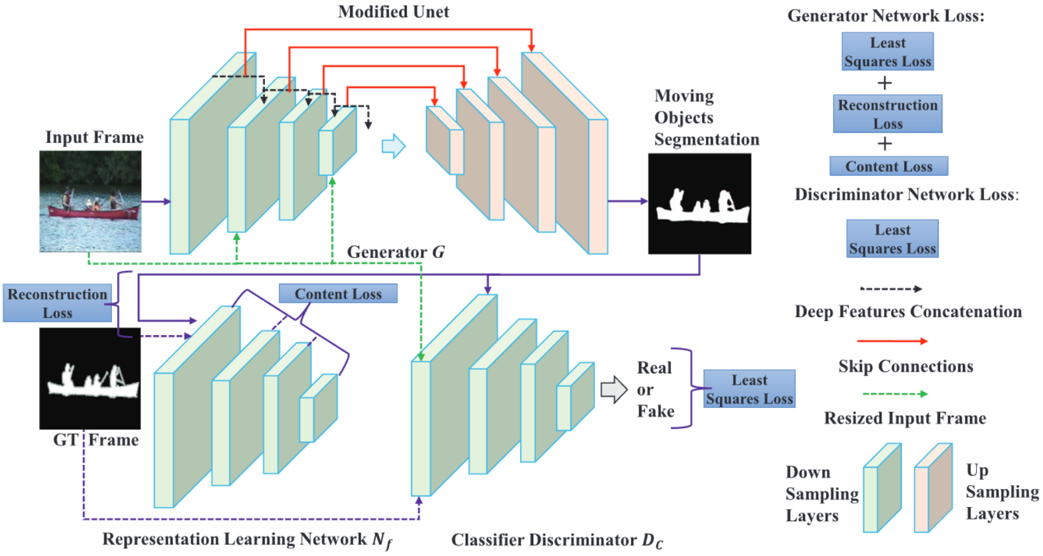


Figure 2.39: The M-cGAN framework as presented by Sultana *et al.* [65].

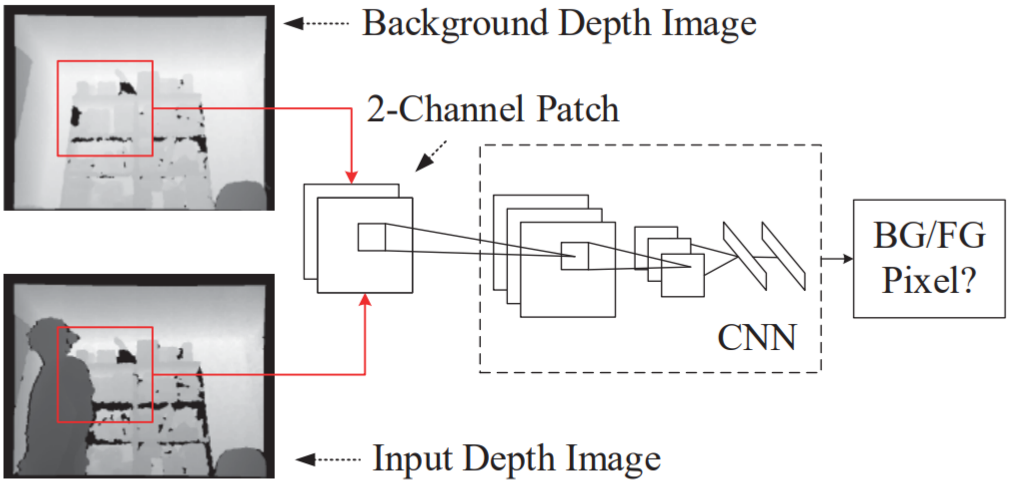


Figure 2.40: The BGSNet-D framework as presented by Wang *et al.* [66].

current frame (1), its Motion mask extracted via OF (2), and their multiplication (3); the second part is for generating the background (7) and extracting the BS (10) from the Depth frame; the third part of this phase is the fusion of the two extracted BS, (5) and (10), using pixel by pixel addition.

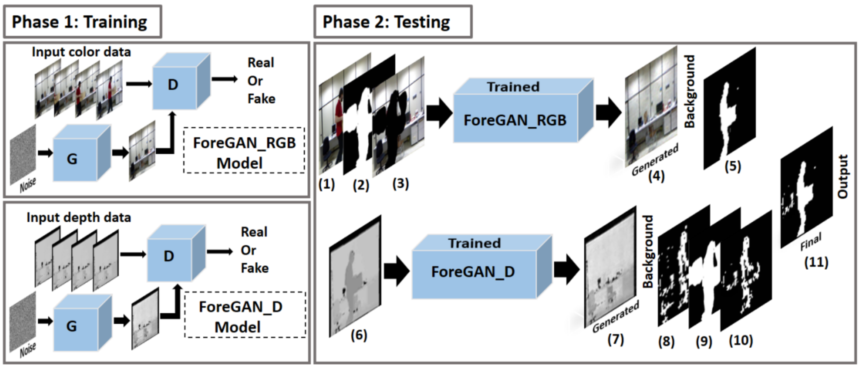


Figure 2.41: The ForeGAN-RGBD framework as presented by Sultana *et al.* [67].

The third method from the state-of-the-art that fuses color and depth information is named CGAN-RGBD. Sultana *et al.*[68] proposed a CGAN-based approach consisting of a Unet-kind architecture for the generator that contains a series of convolutional layers (encoder) followed by a series of deconvolutional layers (decoder); skip connections are added on different levels between the encoder and the decoder (see Figure 2.42); on the other hand, an FCN is built as a discriminator, which consists of four convolutional layers. In this study, the authors used a different technique compared to the previous one (ForeGAN-RGBD[67]); they used only one generator and one discriminator fed by the color and depth information of the current frame to extract the background subtraction. Figure 2.43 gives an overview of the proposed framework as presented in the original paper.

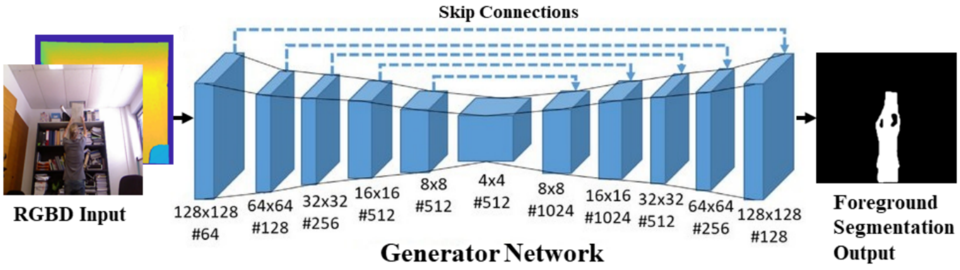


Figure 2.42: The CGAN-RGBD generator as presented by Sultana *et al.* [68].

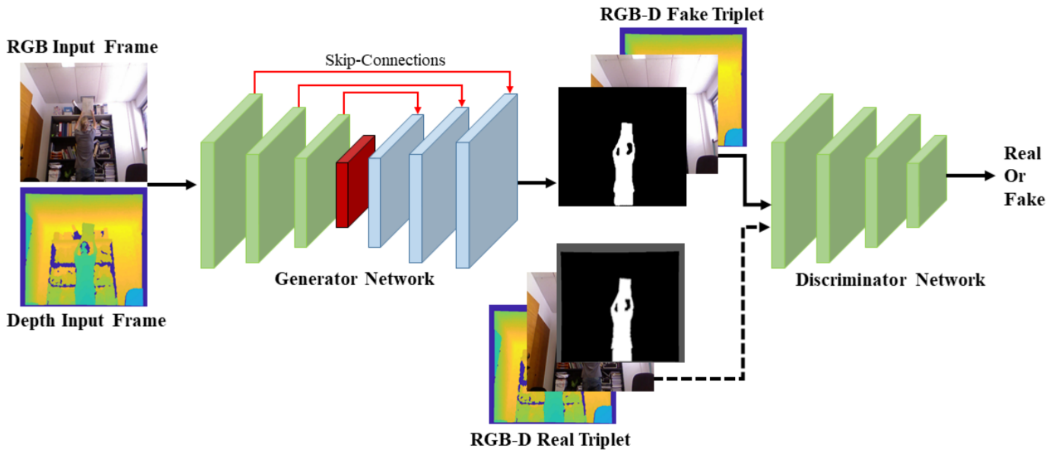


Figure 2.43: The CGAN-RGBD framework as presented by Sultana *et al.* [68].

2.5 Limitations and general thoughts

Despite the superiority of the latest methods created in the field of background subtraction, we ought to mention three common cons that have been widely arising in the state-of-the-art to create a common ground for the next generation of the background subtraction approaches:

- **Temporal history and background reference:** Here, the approaches need to be fed at least with one frame as a historical reference to build the background model, either with a background reference frame (Empty frame) that does not contain any foreground objects, which is often selected manually from among the video frames, or created using one of the background estimation techniques.

- **Seen and Unseen scenarios:** When the background methods based on models were trained on a set of frames, these frames are randomly or manually selected from the same video used for the testing phase. Like building a background model based on several images taken from the first frames of the testing video.

- **Robustness and efficiency:** The methods are built to solve one specific scenario (video-based or category-based) instead of building one model that solves many scenarios simultaneously.

The limitations of state-of-the-art techniques based on RGB information motivated us to consider also depth information, as some existing works, and build an RGBD model capable of improving background subtraction performance (overall and for specific challenges). On the other hand, the original protocol adopted by the state-of-the-art methods is saturated, reaching ~ 99% (except for two specific videos) over the *F-score* metric due to the overlapping between the training and testing splits and/or the use of reference frames. This fact led us to adopt the SIE protocol for the SBM-RGBD dataset, which will be explained later in the next chapters of this thesis. This new protocol offers a fair comparative evaluation that can be adopted by the next upcoming approaches.

2.6 Conclusion

In this chapter, we represented the most valuable datasets used in background subtraction. These datasets hand over the necessary scenarios to challenge the motion detection methods. Some of these datasets provide color information and some others provide depth information in addition, which can be beneficial to introducing RGBD-based techniques. Also, we explained the *F-score* evaluation metric. The *F-score* metric is the most common metric used to evaluate the background subtraction performance. The third section of this chapter was dedicated to the existing work related to our study. We divided this section into three main subsections: Traditional-based techniques, Deep-learning-based techniques, and then we explained the limitations, talked about some new ideas, and proposed some solutions thoughts.

CHAPTER 3

TRADITIONAL APPROACHES: MULTI-SCALE RESOLUTIONS & RGBD-BASED TECHNIQUES

Contents

Page

3.1 Introduction 51

3.2 Multi-Scale Resolution for background subtraction 51

3.2.1 The SuBSENSE Approach 52

3.2.2 Multi-Scale Resolution (Image Pyramid) 52

3.2.3 Performance Evaluation 54

3.2.4 Conclusions 56

3.3 RGBD ViBe-based background subtraction 56

3.3.1 The proposed framework 56

3.3.2 Experiments and discussions 59

3.3.3 Conclusion 66

3.4 Conclusion 66

3.1 Introduction

This chapter contains a detailed description of the proposed approaches (related to traditional existing works) of our study. This chapter is divided into two main parts, the first section will be dedicated to the usage of MSR (or image pyramid) in background subtraction as a pre-processing and post-processing technique in order to improve the performance of future works. Then, we will introduce the proposed RGBD-based background subtraction framework, which is created as a robust method to improve object detection for several techniques. Both sections contain the approach explanation details, the tests and experiments, and also the results discussions.

3.2 Multi-Scale Resolution for background subtraction

In this section, we proposed to use the MSR in order to reduce the noise from the input frames and increase the frame rate of the background subtraction process. This approach is built to be used in future works along with color and depth information fusion. We chose the SuBSENSE technique as known as one of the most effective traditional approaches in background subtraction. Fig. 3.1 illustrates the proposed framework using MSR.

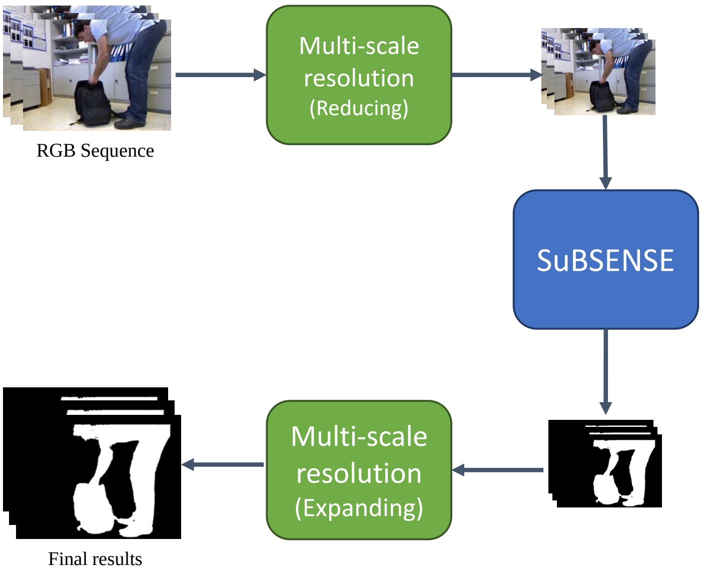

RGB Sequence
Multi-scale resolution (Reducing)
SuBSENSE
Multi-scale resolution (Expanding)
Final results

Figure 3.1: The proposed framework based on the MSR

3.2.1 The SuBSENSE Approach

Pierre-luc et al. [34] proposed a new method called SuBSENSE, which refers to (Self-Balanced SENsitivity SEgmenter). It is a pixel-level modeled using a spatiotemporal feature descriptor and is considered as a non-parametric approach while it uses a feedback strategy to adjust the input parameters. As referred in the original papers [34, 35], the SuBSENSE approach was inspired by ViBe [69], but instead of using only color information, the SuBSENSE uses color and Local Binary Similarity Pattern (LBSP) pixel representations to classify whether the pixel belongs to the background or foreground.

The equation (3.1) describes the way of storing the color-LBSP pixel representations in the background model. The background model B contains pixel models which each contain a set of N recent background samples [35].

$$B(x) = \{B_1(x), B_2(x), ..., B_N(x)\} \tag{3.1}$$

These samples are matched with their respective observation $I_t(x)$ on the input frame at time t, to classify the pixel as foreground (1) or background (0) at the x coordinate, as following:

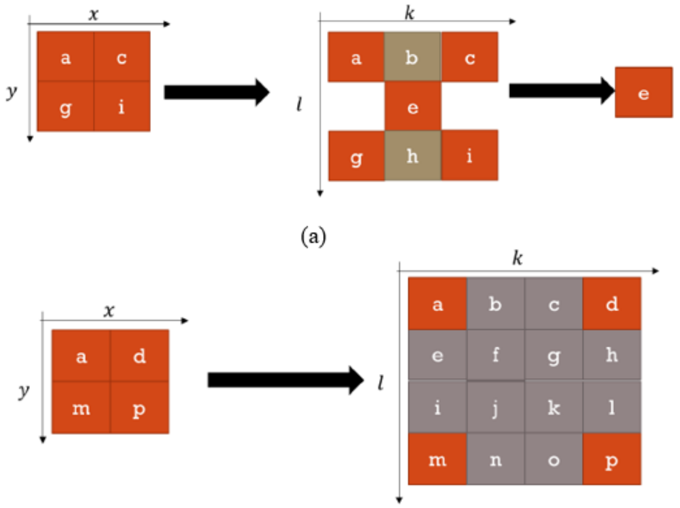
$$S_t(x) = \begin{cases} 1 & M\{dist(I_t(x), B_n(x)) < R, \forall n\} < M_{min} \\ 0 & otherwise \end{cases} \tag{3.2}$$

Where the output segmentation map is noted as $S_t(x)$, $dist(I_t(x), B_n(x))$ presents the distance between $I_t(x)$, the current observation, and $B_n(x)$, the given background model. While this distance will be compared to the maximum distance threshold R. The model will be very precise in classifying pixels as background successfully in the case of a small maximum distance threshold. On the other hand, a large maximum distance threshold will cause a complex detection when it comes to foreground objects that look very similar to the background, but also it will give us better resistance against irrelevant changes. M and M_{min} are respectively the number of matching counted and the minimum number of matches required for a background classification. For further information, we recommend readers to check [34, 35, 70].

3.2.2 Multi-Scale Resolution (Image Pyramid)

During the background subtraction process, we notice that the input images, in some cases, could have more details than we want; these details should be eliminated to prevent the process's delay time and provide free memory space. In these cases, it is better to use a method that reduces such unnecessary details. One of these techniques, called Image Pyramid, is the main of this technique in our study, which is to reduce the image size so we will have the essential information and less noisy details to do our processing smoothly in real-time and with better performance. There are many kinds of image pyramids, such as the Gaussian pyramid, Laplacian pyramid, and image pyramid, using interpolation or filtering. For more details about the image pyramids and their types, please check [71–73].

In this study, based on state of the art, we chose to use the bilinear image pyramid. As we mentioned in the beginning, we use this method before and after the selected approach means that we reduce the size of the input frames, then after applying the process, we expand the size of the output frames to reach the original frame size, which depends on which level we chose to use. Thus, To reduce the image size is to create a new image pixel by pixel; each pixel of the new image represents an interpolation between a set of neighbor pixels from the original image. On the other way, to expand the image size is to create a new image pixel by pixel; each pixel of the new image represents an interpolation between four pixels from the original image, which also depends on how much we want to reduce or expand the image. We can resume all this as simply as described in Fig. 3.2, where we can see two examples of the two parts of the process, the first is how reducing four pixels into one pixel, and the second example is expanding four pixels to sixteen pixels.



(a)

(b)

Figure 3.2: MSR explanation (in general): (a) Reducing the size. (b) Expanding the size.

3.2.3 Performance Evaluation

In Fig. 3.1, we simplified the framework and the way of using the image pyramid. Each input frame will be down-scaled to a specific level and fed to the SuBSENSE to extract the background subtraction results; these preliminary results will be up-scaled to the original size to obtain the final results. We mentioned in previous sections to reduce the frame size for two reasons. The first is to reduce the noise caused by the small movements in the background of the scene. The second is to increase the frame rate of the original approach, which is one of the widely famous reasons for using image pyramids in computer vision applications. In our experiments, we used four scale levels (including the original level) to define the better level we could use for our work and reach the best performance with the highest frame rating possible. As we can notice in Fig. 3.3, the more we go down, the more blurry the image quality will get, which causes low performance.

Table 3.1: *F-score* calculation

Categories	**Original SuBSENSE**	**First level**	**Second level**	**Third level**
badWeather	0.86	0.86	0.82	0.70
baseline	**0.96**	0.94	0.78	0.27
cameraJitter	**0.83**	0.76	0.70	0.51
dynamicBackground	0.82	**0.89**	0.67	0.30
intermittentObjectMotion	**0.61**	0.54	0.25	0.03
lowFramerate	0.67	**0.71**	0.46	0.16
nightVideos	0.49	**0.51**	0.43	0.19
PTZ	**0.39**	0.11	0.09	0.21
shadow	0.95	0.95	0.89	0.53
thermal	**0.69**	0.63	0.61	0.41
turbulence	0.87	**0.90**	0.44	0.13
Average F-Score	**0.74**	**0.71**	**0.56**	**0.31**
Average processing speed (nf/s)	**3**	**12**	**47**	**161**

The dataset CDNET 2014 contains 11 categories; each category has 4 to 6 video sequences, and each sequence has its own ground truth to evaluate the results. Table 3.1, includes the *F-score* average calculation, category-based and overall. Also, we mentioned the total average frame number per second in order to see the difference in processing time compared to the original method. These results show that the more we reduce the size, the better our frame rate will achieve. But it also depends on the

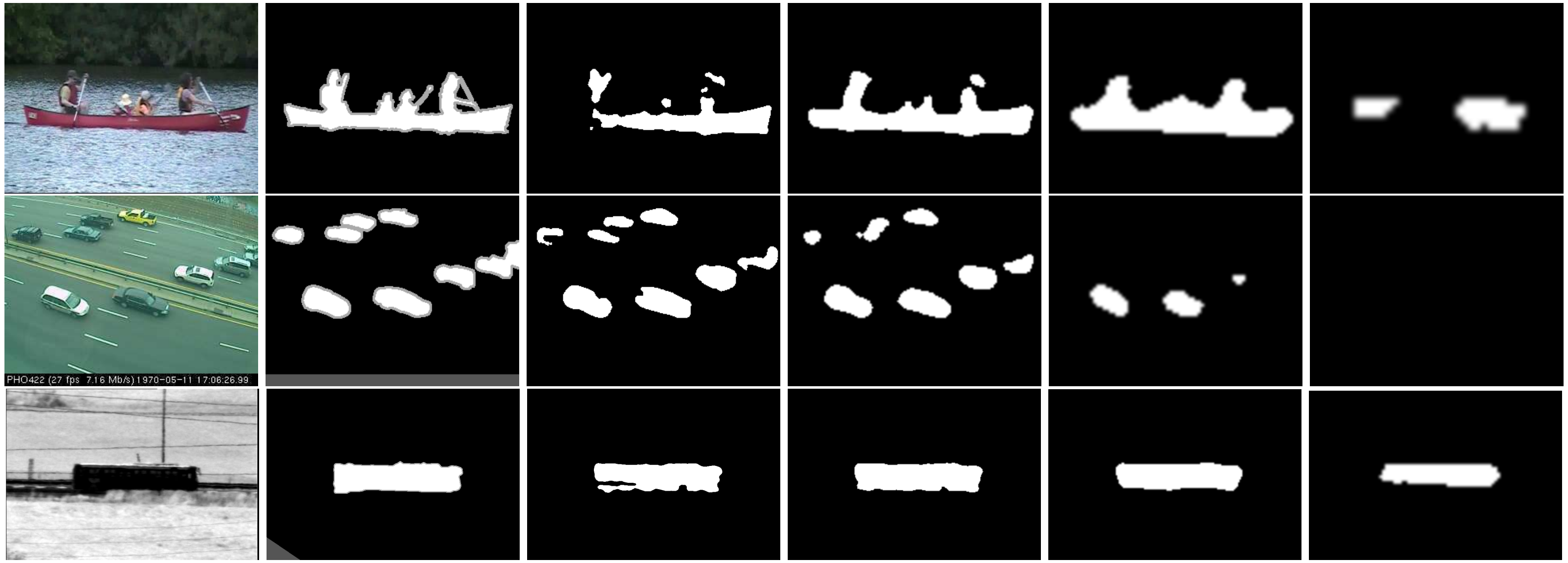


Figure 3.3: The visual comparison between the proposed method and the original SuBSENSE. Rows from top to bottom ranged to contain three frame examples from three categories, *dynamicBackground*, *lowFramerate*, and *turbulence* respectively. Columns from left to right ranged to contain input frame, groundtruth, SuBSENSE result (without MSR), first resolution level, second resolution level, and third resolution level result, respectively.

F-score calculation. In some categories such as *baseline*, *cameraJitter*, *intermittentObjectMotion*, *PTZ*, and *thermal*, the original method of SuBSENSE was performing better than all other levels. In the *badWeather* and *shadow*, the first resolution level was competitive with the original approach. In the remaining categories, *dynamicBackground*, *lowFramerate*, *nightVideos*, and *turbulence*. we notice that level one from our method outperforms the original SuBSENSE and all the other resolution levels.

3.2.4 Conclusions

This study proved that the use of the image pyramid, in some cases, can improve the performance and give us better results with good executing time compared with the original work of the SuBSENSE method. Our method performance was not as expected in some video categories from the CDNET dataset. On the other hand, the performance is acceptable compared to the original approach. The better results achieved using our technique are in the following categories: *turbulence*, *lowFramerate*, *nightVideos*, and *dynamicBackground*. This explains that the proposed framework using the image pyramid performs well in such scenes with some noise from small movements in the background, making it more static and smooth, leading to better detection.

3.3 RGBD ViBe-based background subtraction

In this section, we propose a framework for improving Background Subtraction techniques. This framework is based on two types of data, RGB and Depth. This study stands for obtaining preliminary results of the background segmentation using Depth and RGB channels independently, then using an algorithm to fuse them to create the final results.

3.3.1 The proposed framework

This framework is inspired by many state-of-the-art methods that use RGBD information to solve Background Subtraction, such as [44, 74, 75]. These RGBD-based studies prove the efficiency of using depth information in different cases using different approaches. The proposed framework provides another novel approach that is less complex and more efficient compared to the previous ones. Using the Depth maps along with the RGB channels affords more information about the scene, specifically about some challenges related to illumination, color overlapping, or reflections such as *ColorCamouflage*, *IlluminationChanges* categories where most of the RGB-based approaches have failed in solving such situations. The proposed framework is presented in Fig. 3.4.

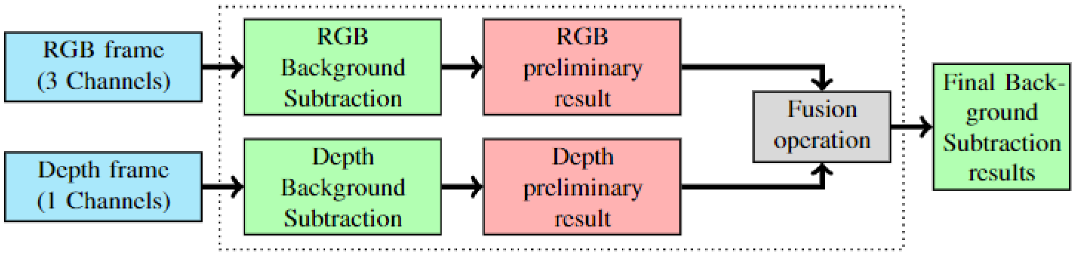


Figure 3.4: The proposed framework for background subtraction based on RGBD information.

The idea behind this approach is to extract the background-foreground segmentation using two different paths, one comes from using the RGB information and the other comes from using Depth information. These operations are applied separately by using one of the recent state-of-the-art techniques to obtain what we call, preliminary results. These preliminary results will be then fused pixel-to-pixel to generate the final results of the background subtraction.

This approach is made to benefit from the advantages of both information to improve the performance. As known, the RGB-based methods are performing better in such challenges as *DepthCamouflage*, where the targeted moving object and the background are at the same distance point from the camera sensor. This fools most algorithms that use only Depth information, leading them to detect these separated objects as one single object; also, Depth-based approaches are performing better in other challenges such as *ColorCamouflage*, *IlluminationChanges*, and *Shadows*. Where all these challenges are difficult to handle with RGB-based strategies. This difficulty comes from the following properties for each category: *ColorCamouflage*, which includes scenarios where the moving object has the same color as the background, which can be detected as one moving object or as background in some other cases; *IlluminationChanges*, which includes scenarios with unstable lighting situations. This will cause a lot of false positives as a result of detecting light as a new moving object in the scene; *Shadows*, including scenarios where the moving object has a raw shadow that can fool the model into seeing it as a moving object.

Figure 3.5 shows some examples of these challenges that we aim to handle using our proposed method. In the *ColorCamouflage* example, one can clearly notice that the Depth information is differentiating the grey box from the background compared to the RGB image. In the *DepthCamouflage* example, one can see that the RGB image shows the difference between the person and the chair, but from the Depth map, both of them have the same distance from the camera. The last example is from the *Shadows* category; the shadow of the box is detected by the RGB camera but is ignored by the Depth camera.

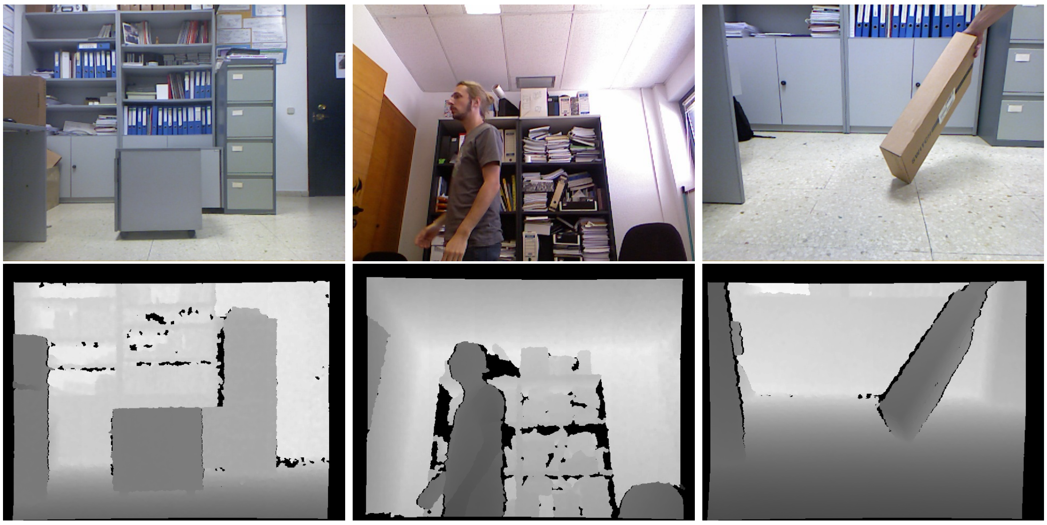


Figure 3.5: Some frame samples from SBM-RGBD Dataset, to illustrate the challenging categories. From left to right, *ColorCamouflage*, *DepthCamouflage*, and *Shadows*

Therefore, to handle such challenges, we study the possible cases that lead to better results. We created Algorithm. 3, aiming to solve this situation with minimum effort and less complexity. The algorithm stands for comparing pixel-to-pixel the preliminary results obtained from both RGB-based and Depth-based approaches (probably from the same method). Hence, four statuses will be concluded from this comparison, taking into account the binary kind of background subtraction results (Background = 0, Foreground = 1).

The first two statuses are occurring when the RGB and Depth background subtraction preliminary results have the same value (0 or 1). It means when the RGB and Depth lead to the same result in one specific pixel, either classify it as a background (0) or a foreground (1) pixel; The third and the fourth statuses are occurring when the RGB and Depth background subtraction preliminary results have opposite values.

Algorithm 3: The Fusion Operation

```
Data:
$rgb\_bg \leftarrow RGB\ Result\ Frame$;
$d\_bg \leftarrow Depth\ Result\ Frame$;
$[J,K] \leftarrow Size(d\_bg)$;
$bg\_mask \leftarrow Zeros([J,K])$;
for $j = 1:J$ do
    for $k = 1:K$ do
        if $rgb\_bg(j,k) == 1$ && $d\_bg(j,k) == 1$ then
            $bg\_mask(j,k) = a$;
        if $rgb\_bg(j,k) == 0$ && $d\_bg(j,k) == 0$ then
            $bg\_mask(j,k) = b$;
        if $rgb\_bg(j,k) == 1$ && $d\_bg(j,k) == 0$ then
            $bg\_mask(j,k) = c$;
        if $rgb\_bg(j,k) == 0$ && $d\_bg(j,k) == 1$ then
            $bg\_mask(j,k) = d$;
```

In order to give further clarification about **Algorithm** 3, we should mention that J and K are the size of the frame; $RGB(orD)\_BG(j,k)$ is the pixel value located in position (j,k) from the RGB (or Depth) background subtraction result; $bg\_mask(j,k)$ is the final pixel result located in the same position (j,k). In order to define the a, b, c, and d values, we studied the possibilities of the decisions that can be effective for our approach. First, we examined the a and b decision cases when the same values are given from both sides (RGB and Depth) to consider them as a confirmation of each other. Second, we run a test to define the values of c and d to conclude the final operation that can be applied between the background subtraction preliminary results of RGB and Depth. More details will be explained in the experiments where we will study all cases and conclude the chosen operation.

3.3.2 Experiments and discussions

This subsection describes three main parts, experiments, the illustration and discussion of the results, and the conclusion. In this subsection, we will explain the footsteps we followed, starting from the proposed framework presented at the beginning of this section, to choosing the correct values for Algorithm. 3 until the results extraction and evaluation. The experiments presented here are essential to understanding how this framework works and how we evaluated the methodology process.

Following the proposed framework presented in Fig. 3.4 and the Algorithm. 3 from Chapter 3, the statuses of a and b are considered as a co-validation between the RGB and Depth background results. Hence, if the pixel is detected as a background from both RGB and Depth, the final result will be a background. Vice versa, if the pixel is detected as a foreground from both RGB and Depth, the final result will be foreground. This hypothesis has been proved using the ViBe[69] method by calculating the error ratio between the cases where the RGB and Depth results are identical, and the groundtruth. Thus, we use the two following equations, Foreground Error Ratio (FErR) and Background Error Ratio (BErR):

$$FErR = \frac{FF}{TF} \cdot 100 \tag{3.3}$$

$$BErR = \frac{FB}{TB} \cdot 100 \tag{3.4}$$

where False Foreground (FF) is the number of the background miss-classification pixels reported in the final results; the False Background (FB) is the number of the foreground miss-classification pixels reported in the final results; the Total Foreground (TF) and Total Background (TB) are the total correct number of foreground and background pixels respectively.

Table 3.2 represents the calculation of FErR and BErR per category and overall the SBM-RGBD Dataset. The calculations prove that the risk of having wrong decisions linked to the co-validation based on the a and b values is small compared to having a correct decision. On the other hand, where the preliminary results of RGB and Depth have opposite decisions, we run a test that considers the *F-score* as feedback information to our system while trying the possibilities on c and d from Algorithm. 3 to conclude the final form of the fusion operation in the proposed framework. The test results are illustrated in Table 3.3 proving that using a fixed value for c and d equal to 1 is the outstanding option for the proposed framework. These results surprisingly conclude that the fusion operation in our system is the *OR* logical operator.

The visual comparison presented in Fig. 3.6 and Fig. 3.7 show the effectiveness of the proposed framework on the ViBe and PAWCS, respectively. The quantitative evaluation using *F-score* illustrated in Table 3.3 proves the improvement provided by the proposed approach in overall and category-based evaluation using the SBM-RGBD dataset. The

Table 3.2: Possibilities of foreground-foreground and background-background effects.

Dataset categories	FF	TF	FErR	FB	TB	BErR
Bootstrapping	640888	1595516	40.17	132098	10680989	1.24
ColorCamouflage	23222	2146839	1.08	48618	21143837	0.23
DepthCamouflage	12070	2882381	0.42	8403	49345914	0.017
IlluminationChanges	91983	1973768	4.66	158969	23563132	0.67
IntermittentMotion	1782325	5685471	31.35	102816	81317838	0.13
OutOfRange	296220	2974936	9.96	13336	49510346	0.03
Shadows	101033	3014480	3.35	5388	21962588	0.03
Overall	**2947741**	**20273391**	**14.54**	**469628**	**257524644**	**0.18**

Table 3.3: The *F-Score* Calculation for the four *c/d* possible combinations (tested on ViBe)

Dataset Categories	RGB	Depth	RGB-Depth Fusion			
			c=1 and d=1	c=1 and d=0	c=0 and d=1	c=0 and d=0
Bootstrapping	0.4045	0.2958	0.3813	0.5270	0.5345	0.5035
ColorCamouflage	0.4063	0.8271	0.8808	0.8511	0.6825	0.7908
DepthCamouflage	0.6452	0.4520	0.8047	0.6194	0.6953	0.6847
IlluminationChanges	0.3915	0.4278	0.4492	0.4246	0.3941	0.3895
IntermittentMotion	0.5695	0.4478	0.6084	0.6096	0.3899	0.3768
OutOfRange	0.7928	0.3926	0.6828	0.4158	0.7165	0.3824
Shadows	0.6824	0.7663	0.8364	0.8551	0.6956	0.6490
Overall	**0.5633**	**0.5087**	**0.6573**	**0.6130**	**0.5806**	**0.5271**

Table 3.4: The *F-Score* Calculation for ViBe-based techniques on SBM-RGBD Dataset

Dataset categories	**ViBe**[69]		**LOBSTER**[70]		**SuBSENSE**[34]		**PAWCS**[35]	
	RGB	**RGBD**	**RGB**	**RGBD**	**RGB**	**RGBD**	**RGB**	**RGBD**
Bootstrapping	0.4045	0.3813	0.6026	0.4722	0.5076	0.3576	0.5895	0.5581
ColorCamouflage	0.4063	0.8808	0.5736	0.9219	0.6231	0.9053	0.6056	0.8237
DepthCamouflage	0.6452	0.8047	0.7955	0.8718	0.7484	0.8486	0.8796	0.8904
IlluminationChanges	0.3915	0.4492	0.4630	0.4596	0.4679	0.4603	0.4516	0.4529
IntermittentMotion	0.5694	0.6084	0.5412	0.6275	0.5809	0.6545	0.6382	0.6614
OutOfRange	0.7928	0.6828	0.8372	0.7506	0.8310	0.7694	0.8623	0.8194
Shadows	0.6824	0.8364	0.8850	0.8794	0.9246	0.9409	0.9447	0.9573
Overall (Average)	**0.5633**	**0.6573**	**0.6727**	**0.7057**	**0.6715**	**0.7007**	**0.7102**	**0.7376**

results evaluation explains that using the 1 value for *c* and *d* helps the system prevent the False Positive (*FP*) detection. These exciting results motivate us to test our proposed framework using some SOTA methods. Table 3.4 illustrates the *F-score* calculation of the results extracted using ViBe and three other methods: LOBSTER [70], SuBSENSE [33, 34], and PAWCS [35, 39]. These methods are inspired by the ViBe method, which we call the ViBe-based methods. The results evaluation confirms the efficiency of the proposed framework on all the tested techniques. The *F-score* measurement reached ~ 10% of improvement on ViBe and ~ 3% on the remaining methods. One can notice that the most significant enhancement is detected in the *ColorCamouflage* category, which is a fact that when the background has the same color as the moving object, it will cause a misunderstanding of the scene and cause false detection (precisely False Positive). However, one of the drawbacks of our framework is the misleading caused by the Depth information. This case is noted in the *OutOfRange* category where the moving object is too far (or too close) from the camera, which leads to massive false detection coming from the depth background subtraction results.

Table 3.5 shows the detailed *F-Score* measurement for each video on the SBM-RGBD dataset. The pattern seen through the mentioned table confirms the efficiency provided by the proposed approach. The targeted categories such *ColorCamouflage*, have proven the theoretical assumption that Depth can correct the behavior of the system for better results. *Shadows* category also shows the stability driven by the methods of our proposed framework if we compare the total average of all videos. However, the provided data from the SBM-RGBD dataset is not always stable. The Depth information is not accurate for all given frames. This factor has a negative impact on the proposed method's performance. As an example, we should drive attention to the *IlluminationChange* category, which is one of the targeted categories. The Depth maps of this category have many cases of the out-of-range depth sensor and also some depth camouflage scenes which switch the *IlluminationChange* from a target category to a very challenged category for the Depth-based approach. Further, one can notice that the proposed approach does not improve the performance in the *Bootstrapping* category. In fact, the four tested techniques (RGB and RGBD versions) are not performing well in such a category related to the background initialization. The ViBe-based approaches use the first frame to create the background model, while the *Bootstrapping* category has moving objects in all its frames. The *IntermittentMotion* is considered as an unbiased category to predict theoretically the performance using its data. However, the results proved the positive impact of the Depth information on this category except for a few

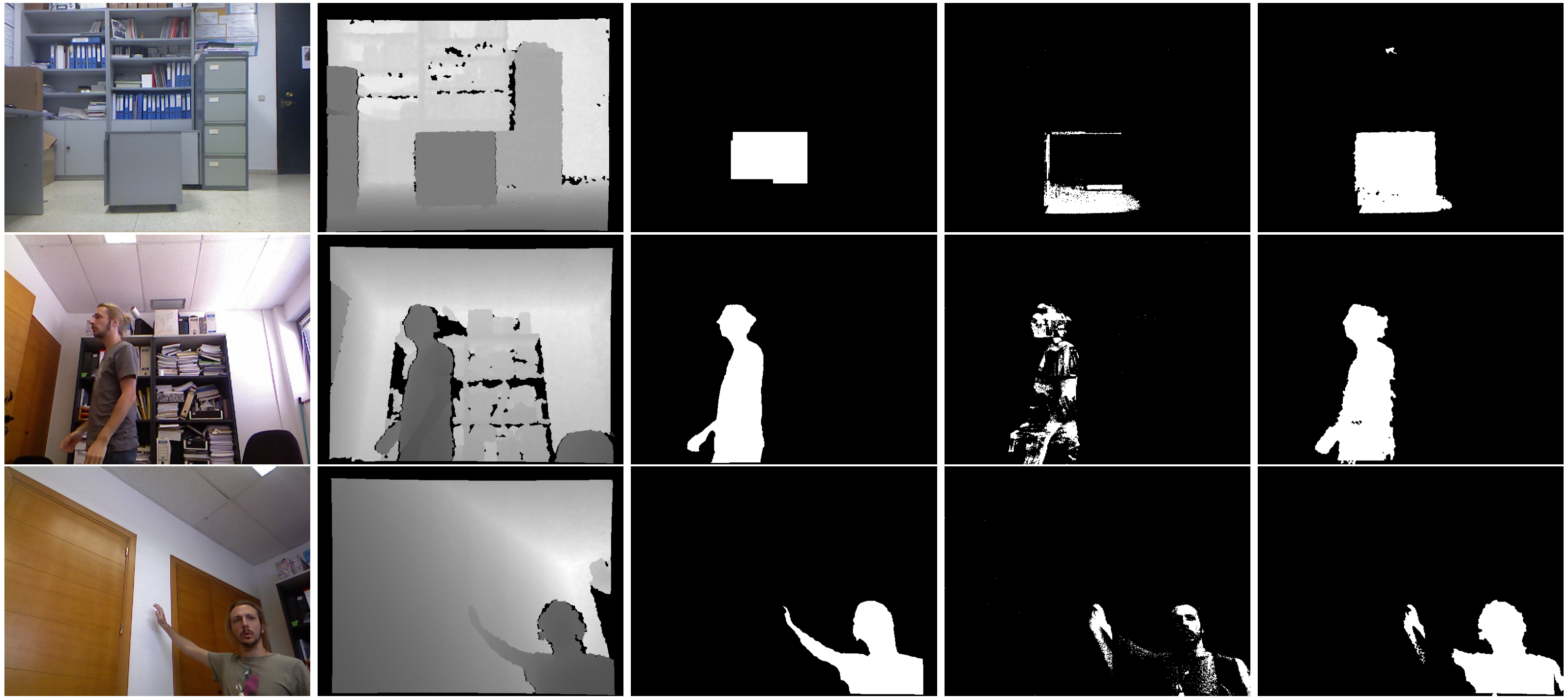


Figure 3.6: The visual comparison between the proposed method and the original ViBe approach. Rows from top to bottom represent three frame samples from the categories, *ColorCamouflage*, *DepthCamouflage*, and *Shadows*, respectively. Columns from left to right correspond to the input RGB frame, input Depth frame, groundtruth, RGB ViBe result, and the proposed RGBD framework result, respectively.

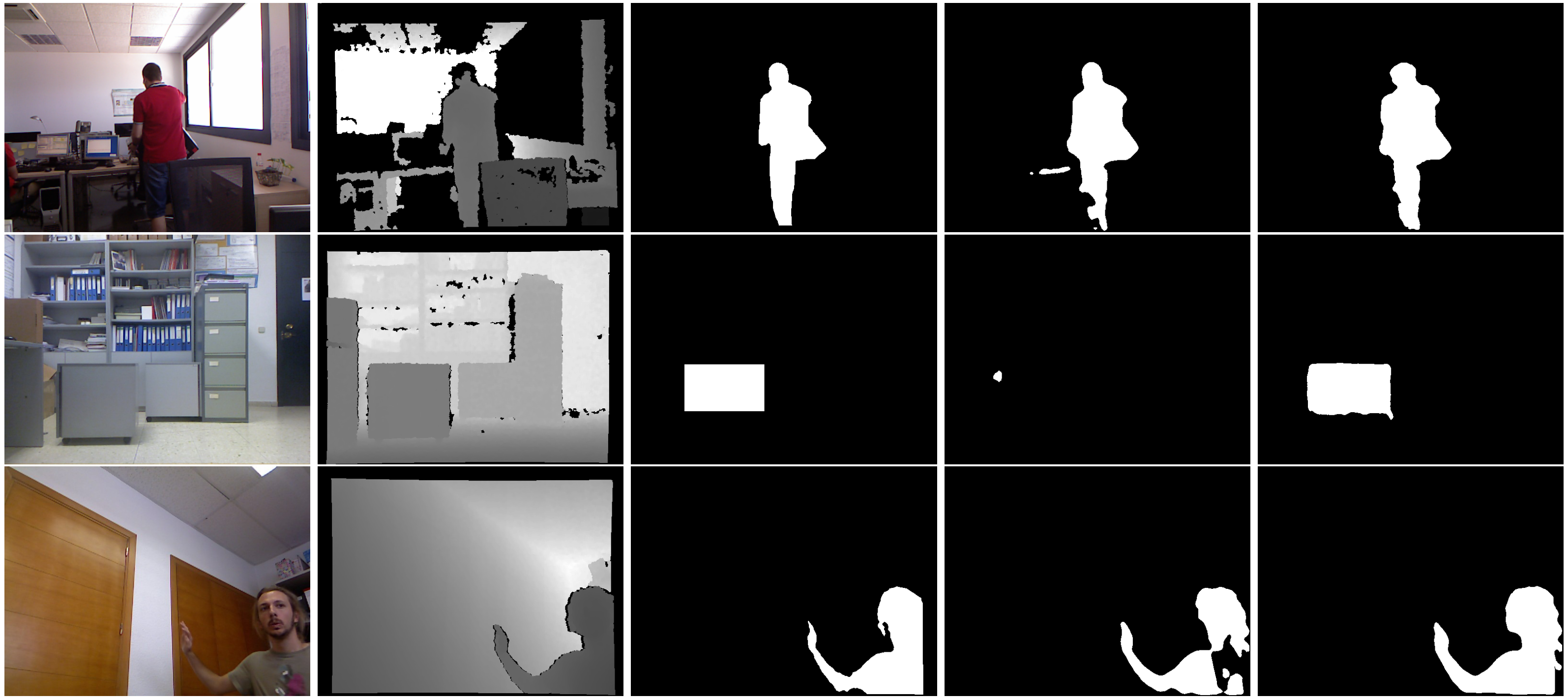


Figure 3.7: The visual comparison between the proposed method and the original PAWCS approach. Rows from top to bottom represent three frame samples from the categories, *IlluminationChanges*, *ColorCamouflage*, and *Shadows*, respectively. Columns from left to right correspond to the input RGB frame, input Depth frame, groundtruth, RGB PAWCS result, and The proposed RGBD framework result, respectively.

Table 3.5: The video-based *F-Score* calculation for ViBe-based techniques on SBM-RGBD Dataset

Dataset		**ViBe**[69]		**LOBSTER**[70]		**SuBSENSE**[34]		**PAWCS**[35]	
Category	**Video**	**RGB**	**RGBD**	**RGB**	**RGBD**	**RGB**	**RGBD**	**RGB**	**RGBD**
Bootstrapping	adl24cam0	**0.4232**	0.3251	**0.7081**	0.5309	**0.4096**	0.2803	**0.3147**	0.2130
	bear_front	0.4324	**0.4660**	0.6495	**0.6609**	0.3741	**0.4608**	0.4717	**0.4772**
	BootStrapping_ds	**0.3515**	0.2644	**0.4861**	0.2943	**0.6983**	0.3309	0.8831	**0.8900**
	fall01cam0	0.4726	**0.5110**	**0.5938**	0.4540	**0.6711**	0.4846	**0.7978**	0.7804
	fall20cam0	**0.3435**	0.3377	**0.4644**	0.4152	**0.3735**	0.2373	**0.4800**	0.4301
ColorCamouflage	Cespatx_ds	0.6203	**0.8423**	0.8748	**0.954**4	0.9488	**0.9518**	0.8524	**0.8759**
	colorCam1	0.1410	**0.9829**	0.1058	**0.9826**	0.0778	**0.9819**	0.0304	**0.6757**
	colorCam2	0.2463	**0.9654**	0.2442	**0.9493**	0.7758	**0.9413**	0.7751	**0.9309**
	Hallway	0.6125	**0.7328**	0.7300	**0.7999**	0.6888	**0.7475**	0.7643	**0.8123**
DepthCamouflage	DCamSeq1	0.8810	**0.9145**	0.8952	**0.9218**	0.8871	**0.9134**	0.8781	**0.8966**
	DCamSeq2	**0.7637**	0.7455	**0.8218**	0.7458	**0.7258**	0.6746	0.8774	**0.8856**
	Despatx_ds	0.5823	**0.9424**	0.9063	**0.9505**	0.9432	**0.9454**	0.8069	**0.8225**
	Wall	0.3508	**0.6193**	0.4077	**0.8723**	0.4310	**0.8637**	0.9560	**0.9567**
IlluminationChanges	ChairBox	0.6679	**0.8771**	0.8940	**0.9298**	0.9181	**0.9358**	0.8453	**0.8869**
	genSeq1	0.8978	**0.9190**	**0.9526**	0.9081	**0.9533**	0.9063	**0.9612**	0.9246
	Ls_ds	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	TimeOfDay_ds	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
IntermittentMotion	abandoned1	**0.3112**	0.1996	0.4383	**0.5292**	0.6565	**0.7181**	0.7333	**0.7363**
	abandoned2	0.9187	**0.9463**	**0.9365**	0.9001	0.3602	**0.8817**	0.4409	**0.7101**
	movedBackground1	**0.7103**	0.5590	**0.5834**	0.4424	**0.6670**	0.4485	**0.4338**	0.4114
	movedBackground2	0.4208	**0.4938**	0.4887	**0.5027**	**0.6959**	0.5577	**0.6861**	0.5804
	Shelves	0.4701	**0.6642**	0.4882	**0.5685**	0.4243	**0.4941**	**0.6498**	0.6422
	Sleeping_ds	0.5936	**0.7832**	0.8106	**0.8132**	0.7258	**0.8366**	0.8851	**0.8877**
OutOfRange	MultiPeople1	**0.9002**	0.4559	**0.9396**	0.4294	**0.8304**	0.4128	**0.9297**	0.6764
	MultiPeople2	**0.9577**	0.8435	**0.9675**	0.8670	**0.8771**	0.7922	**0.9420**	0.8665
	TopViewLab1	0.7156	**0.7662**	0.8231	**0.8633**	0.8211	**0.9154**	0.7711	**0.8610**
	TopViewLab2	0.7437	**0.7704**	0.8028	**0.8607**	0.8707	**0.9177**	0.9148	**0.9286**
	TopViewLab3	**0.6502**	0.5726	**0.7414**	0.7317	0.7482	**0.8081**	0.7538	**0.7642**
Shadows	fall01cam1	**0.7929**	0.4844	**0.9151**	0.6911	**0.9559**	0.9316	**0.9754**	0.9591
	genSeq2	0.8742	**0.9414**	0.9225	**0.9418**	0.9359	**0.9395**	**0.9618**	0.9577
	Shadows_ds	0.2981	**0.9252**	0.8422	**0.9485**	0.9074	**0.9592**	0.8159	**0.9595**
	shadows1	0.6859	**0.9287**	0.8282	**0.9242**	0.8619	**0.9450**	**0.9874**	0.9608
	shadows2	0.7583	**0.8984**	0.8757	**0.8922**	**0.9524**	0.9299	**0.9828**	0.9494

cases which we think it is related to the provided data. The *OutOfRange* category is expected to be one of the hardest challenges for our proposed method. The results are not clearly proving that expectation but the unstable behavior for all tested methods can prove that it is not the right choice to use Depth information in such a category at least to ensure a stable performance that can be improved lately. *DepthCamouflage* as well as one of the serious challenges for Depth-based methods, but in this study we notice that the performance either remains stable or performs better while using Depth information. However, these unexpected results can be explained by investigating the different scenarios of this category. It is hard to make this category without the interference of other challenges. It is noticeable that most frames provided in the *DepthCamouflage* category have color camouflage especially the *Wall* video which makes the more promising scene for our proposed method.

3.3.3 Conclusion

This section proposed a framework that uses two information (RGB and Depth) to solve the background subtraction. The experimental evaluation applied to the SBM-RGBD dataset proved the efficiency of this framework in overall-based and category-based evaluation. The remarkable improvement is noted in the *ColorCamouflage* category. This framework is also considered as a robust approach regarding the notable improvement of the results in different methods: ViBe, LOBSTER, SuBSENSE, and PAWCS. The evident limitation of this study is in the *OutOfRange* category, where the depth information is unreliable. In future work, the use of pre-processing and post-processing with the replacement of the depth maps with another kind of information (e.g., descriptors) can improve the performance. Using another mechanism that prevents the misleading of the Depth information can also outcome the limitations of the proposed framework.

3.4 Conclusion

In this chapter, a detailed description was given to explain two proposed approaches for background subtraction. These frameworks are considered traditional techniques, apart from the supervised CNN-based ones. The first system is based on the use of Multi-scale resolution. This technique is chosen in order to eliminate the noise from the scene and to speed up the process for real-time system support. The second approach is a framework that uses color and depth information to improve the accuracy of four ViBe-based approaches. This method is built to overcome some well-known drawbacks in the background subtraction field, such as color camouflage, illumination changes, and shadows categories.

CHAPTER 4

DEEP-LEARNING APPROACHES: DEEP MULTI-SCALE NETWORK (DMSN)

Contents

Page

4.1 Introduction 68
4.2 The proposed method: Deep Multi-Scale Network (DMSN) 68
4.2.1 Introduction 68
4.2.2 The proposed framework 68
4.2.3 The proposed architecture 69
4.3 The proposed protocols: Scene independent protocols 74
4.3.1 Introduction 74
4.3.2 Scene Independent Evaluation protocol for SBM-RGBD 74
4.3.3 Leave One Video Out protocol for GSM 76
4.4 Experiments and performance evaluation 77
4.4.1 Results and discussion 77
4.5 Conclusion 88

4.1 Introduction

In this chapter, the first section is a detailed description of a novel deep learning model for background subtraction. This deep learning network is built to use color and depth information to solve many challenging scenarios in the field. The following section of this chapter is devoted to explaining a proposed protocol for the SBM-RGBD and GSM datasets. Along with the description of the deep learning method and the new protocol, this chapter includes the experiments and the performance evaluation for the proposed approach tested on the original and the proposed protocol.

4.2 The proposed method: Deep Multi-Scale Network (DMSN)

4.2.1 Introduction

This section proposes a novel deep learning model called Deep Multi-Scale Network (DMSN) for Background Subtraction. This convolutional neural network is built to use RGB color channels and Depth maps as inputs with which it can fuse semantic and spatial information. In comparison with previous deep learning background subtraction techniques that lack information due to their use of only RGB channels, our RGBD version is aiming to overcome most of the drawbacks, especially in some particular kinds of challenges that are hard to handle by the RGB-based methods.

4.2.2 The proposed framework

Our proposed framework is illustrated in Fig. 4.1, it is divided into two phases: (i) The training phase and (ii) The testing phase. First, we must provide three inputs for the first phase: RGB frames extracted from a standard camera, their corresponding depth maps extracted from a Microsoft Kinect depth camera, and the ground truth. It is worth noting that we do not use the raw depth maps; they are pre-processed by normalizing the pixel values between 0 and 255. After preparing the data, we feed these inputs into our training system to build the background model of our proposed network. The testing phase follows, in which we use the trained model to apply the background subtraction operation on the input frames using color and depth information.

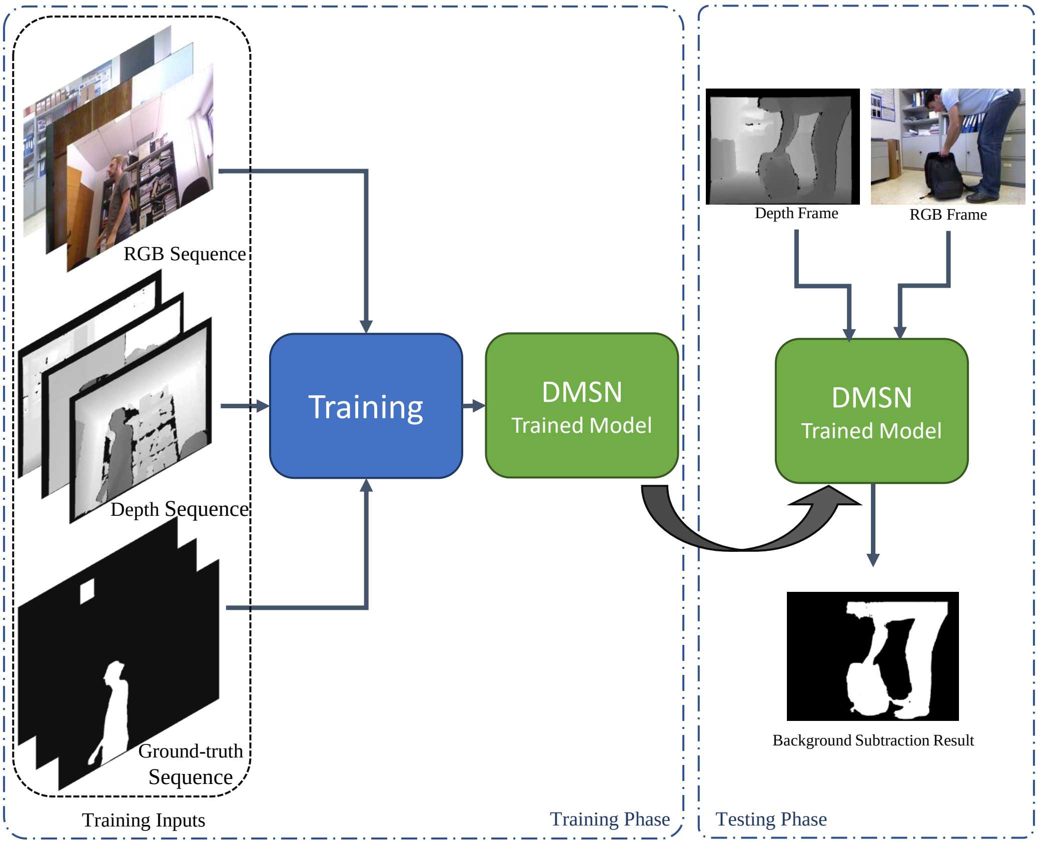


Figure 4.1: The proposed Framework

4.2.3 The proposed architecture

The proposed DMSN architecture is presented in Fig. 4.2. This architecture is inspired by several recent studies such as [54, 76, 77]. The main goal of this architecture is to extract different features from RGB channels and Depth maps, which will provide more discriminative information about the challenging scenes. The proposed network consists of two parallel encoders running together towards a decoder, with intermediate layers for multi-cross features between Encoder/Decoder. The two parallel encoders are built from VGG-16[78]. The first encoder is dedicated to the RGB channels, and the second is dedicated to the Depth map. These two VGG streams are similar; the only difference is the input size, i.e., the number of channels.

The VGG-based feature extraction proved to compute relevant feature maps thanks to its successive convolution with a maintained receptive field. It allows in-depth convolutional transformation without downsampling the input. The multi-modal background

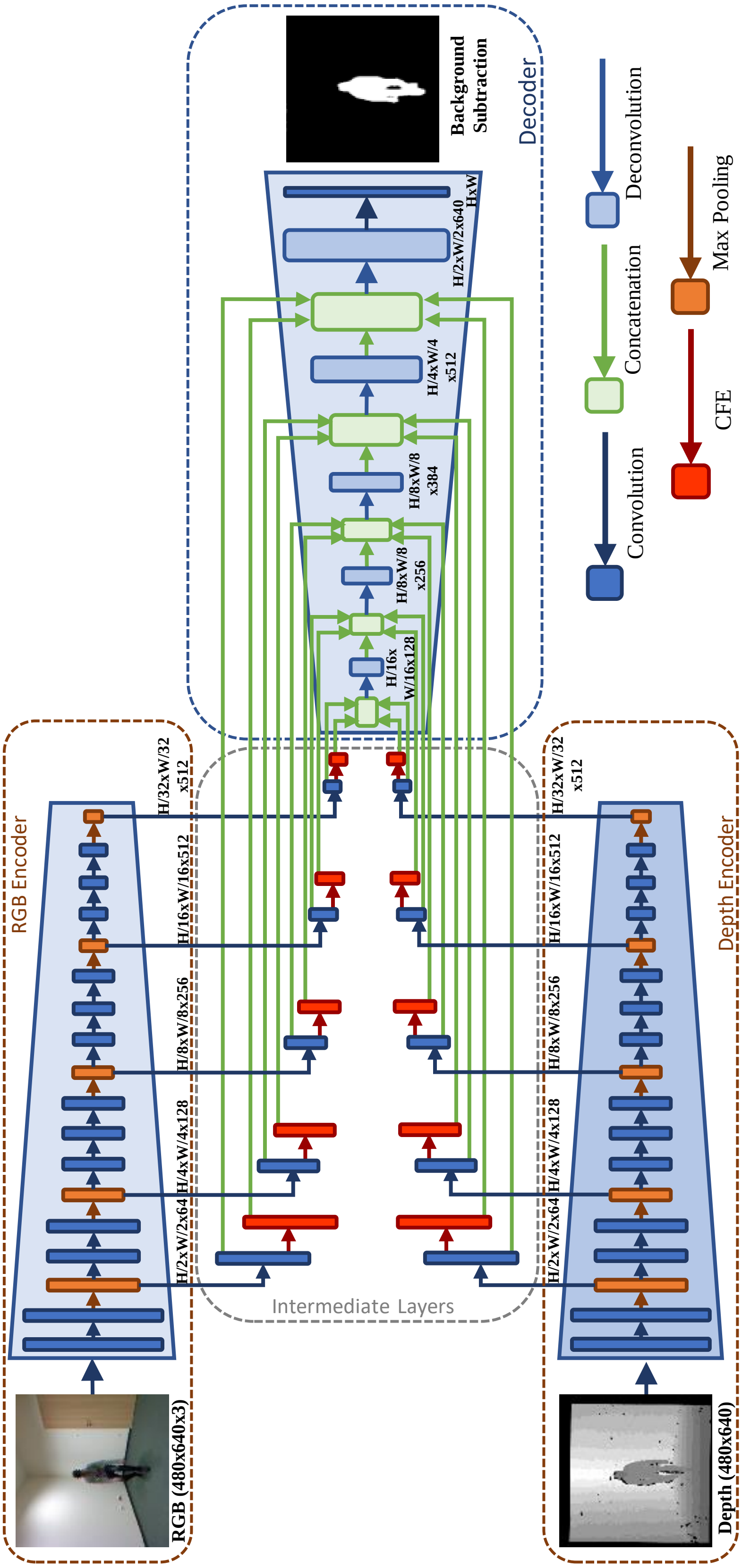


Figure 4.2: DMSN Architecture for Background Subtraction.

subtraction based on RGB and the corresponding depth can be implemented through stacking the inputs or feeding them to individual encoding streams. Stacking the inputs offers low computational complexity but at the cost of non-efficient feature extraction. In addition, both inputs will be encoded on the first convolution only, and the rest of the model processes it as one input leading to a waste of relevant features. In our model, we adopted dedicated streams to emphasize the feature extraction of each input independently from the other one since they do not share the same space. This fact helps each VGG encoder to compute semantic-wise representations of the RGB and Depth images on a set of low-resolution feature maps (512 filters of 1520). Afterward, our model merges the extracted RGB and depth maps combining the discriminant feature that can be found on the RGB space with depth ones for more accuracy. We believe that the depth information is helpful in solving many challenges related to foreground classification such as the camouflage, from which many RGB-approaches are suffering. The foreground extraction problem is addressed as a pixel-wise classification problem justifying the employment of the complete VGG-16 feature extraction stage.

Each encoder consists of five groups of Convolutional-MaxPooling layers. As known, the deeper we go into convolution, the more semantic information we learn about the pixel; this is called the "What is it!" information. On the other hand, going deeper into convolution will affect the learning of spatial details negatively. This spatial information is called the "Where is it!" information. In order to conserve both, we use the skip-connections through convolution and Contrast Feature Extractor (CFE) after each MaxPooling layer [54], where the CFE is, the input I minus the Average Pooling (AP) of the same input: $CFE(I) = I - AP(I)$. These features are placed straight into the decoder by using depth Concatenation layers. This operation leads to avoiding the lack of information, keeping the system aware of the pixel class, and saving more information about the pixel location in the frame. The use of adaptation-based convolution before the CFE comes from the non-homogeneity of the computed filter maps from RGB space compared to the depth one, which may confuse the model on the decoding stage. Moreover, both encoded filters will be adapted to fit the binary space targeted as output by our model. This fact helps the decoder to focus more on the up-scaling of the feature rather than making complex feature transformations. On the other hand, the CFE operation consists mainly in selecting the prominent information based on the intensity, which represents an effective transformation toward features binarization. Our decoder benefits from a light configuration thanks to the architecture of the encoding streams and their collaboration. Its task is focused on up-sampling the concatenated encoder-filters until reaching the targeted output size. Further analysis, through an ablation study, is provided in 4.4.1.3 to show the importance of the dual-encoders and Intermediate Layers.

The decoder section is a series of deconvolutional and depth concatenation layers. Each depth concatenation layer receives five inputs (four inputs for the first concatenation layer) as illustrated in Fig.4.2. In the end, we placed a Convolution layer followed by SoftMax activation function to extract the probability map, which is evaluated through a Cross-Entropy Loss (CEL) based on two classes (*background* and *foreground*) as a final step. The loss function is constructed as follows. Let us consider the split S of the inputs I_m and the predicted outputs O_m, as follows: $S = \{(I_m, O_m), m = 1, ..., M\}$, where $O_m = \{o_n, n = 1, ..., N\}$. M and N denote the total number of images in the Mini-batch size, and the total number of pixels in one image, respectively. The state of o_n can be one of the two categories: Background or Foreground. The loss function can be expressed as follows:

$$CEL = -\frac{1}{M}\sum_{m=1}^{M}(\alpha.\sum_{o\in O_f} log(P(O_m)) + (1-\alpha).\sum_{o\in O_b} log(1-P(O_m))) \tag{4.1}$$

Where α is the ratio between the number of background elements $|O_b|$ and the total number of elements of the label mask $|O|$, as well $(1-\alpha)$ will be dependent on the number of foreground elements $|O_f|$ and the total number of elements of the label mask $|O|$. Here, $\alpha = |O_b|/|O|$ and $(1-\alpha) = |O_f|/|O|$. P_m is the output of the *Softmax* operation of the last convolutional layer output. The training of the proposed model is done with the Adam optimization algorithm, using Mini-batches of 1 sample. The Initial-Learning-Rate is set to 10^{-4}, it drops every five epochs by a factor of 10^{-1}. The network weights are initialized using VGG-16 pre-trained model. The data is divided into two parts during training: 80% for training and 20% for validation, and shuffled for each epoch to avoid using the same data in the validation process. The training ends when the stability of the loss stays for 30 epochs of patience. For further details, Table 4.1 provides more information about each layer (Type, Stride, Padding, and Activations).

Table 4.1: DMSN Network Details

Encoder (RGB and Depth)

Type	Stride	Padding	Activations
Input RGB(Depth)	-	-	480x640x3(1)
Convolution	1	1	480x640x64
Convolution	1	1	480x640x64
Max Pooling	2	0	240x320x64
Convolution	1	1	240x320x128
Convolution	1	1	240x320x128
Max Pooling	2	0	120x160x128
Convolution	1	1	120x160x256
Convolution	1	1	120x160x256
Convolution	1	1	120x160x256
Max Pooling	2	0	60x80x256
Convolution	1	1	60x80x512
Convolution	1	1	60x80x512
Convolution	1	1	60x80x512
Max Pooling	2	0	30x40x512
Convolution	1	1	30x40x512
Convolution	1	1	30x40x512
Convolution	1	1	30x40x512
Max Pooling	2	0	15x20x512

Intermediate Layers (RGB and Depth)

Type	Stride	Padding	Activations
Convolution	1	1	240x320x128
Average Pooling	1	1	240x320x128
Convolution	1	1	120x160x128
Average Pooling	1	1	120x160x128
Convolution	1	1	60x80x128
Average Pooling	1	1	60x80x128
Convolution	1	1	30x40x128
Average Pooling	1	1	30x40x128
Convolution	1	1	15x20x128
Average Pooling	1	1	15x20x128

Decoder

Type	Stride	Padding	Activations
Depth Concatenation	-	-	15x20x512
Deconvolution	2	1	30x40x128
Depth Concatenation	-	-	30x40x640
Deconvolution	2	1	60x80x256
Depth Concatenation	-	-	60x80x768
Deconvolution	2	1	120x160x384
Depth Concatenation	-	-	120x160x896
Deconvolution	2	1	240x320x512
Depth Concatenation	-	-	240x320x1024
Deconvolution	2	1	480x640x640
Convolution	1	0	480x640x1

4.3 The proposed protocols: Scene independent protocols

4.3.1 Introduction

The protocols followed by most of the state-of-the-art methods and especially deep learning methods for background subtraction have been controversial recently. The selection of Training/Testing frames is making an exciting debate that concerns *Scene Dependent* and *Scene Independent* evaluation [57, 58, 79, 80]. This issue has been discussed in detail by Mandal and Vipparthi[79]. The authors named two kinds of scenarios: (i) Scene Dependent Evaluation (SDE), (ii) Scene Independent Evaluation (SIE). They proved that their proposed protocol (SIE) is more challenging and trustful due to its imperative to sustain non-overlapping between the training and testing splits. Moreover, our study is based on two datasets from which we listed in the literature review chapter. These two datasets are matching the requirements of our study. According to the literature, these datasets are the only ones that provide the two types of information needed in our work: RGB frames from a standard camera and Depth maps extracted from a Microsoft Kinect camera. In addition, each sequence in these datasets has its corresponding set of Groundtruth that is needed for supervised training and results evaluation.

4.3.2 Scene Independent Evaluation protocol for SBM-RGBD

This protocol stands for taking one video from each category for the test phase and using the remaining videos for the training phase. The SIE protocol scenario can be implemented regarding two strategies. The first one is *Category-wise training* which leads to one trained model for each category. The second one is *Complete dataset training*, which leads to only one trained model for the whole dataset. The second strategy is more challenging and computationally low as compared to the first one. Moreover, the *Category-wise training* requires a large amount of data per category that can not be granted all the time, as in the case of the SBM-RGBD dataset. On the other hand, the GSM dataset provides one video per category, allowing only the *Complete dataset training*. Therefore, this study considered the SIE protocol with *Complete dataset training* on the SBM-RGBD and GSM benchmarks.

To the best of our knowledge, this is the first study that applies the SIE protocol on the SBM-RGBD dataset. Since that, we proposed a strategy to set a united ground for

Table 4.2: SIE protocol for SBM-RGBD

Category	Video	S1	S2	S3	S4	S5	S6
Bootstrapping	adl24cam0	Tr	**Ts**	Tr	Tr	Tr	**Ts**
	bear_front	Tr	Tr	**Ts**	Tr	Tr	Tr
	BootStrapping_ds	**Ts**	Tr	Tr	Tr	Tr	Tr
	fall01cam0	Tr	Tr	Tr	**Ts**	Tr	Tr
	fall20cam0	Tr	Tr	Tr	Tr	**Ts**	Tr
ColorCamouflage	Cespatx_ds	**Ts**	Tr	Tr	Tr	**Ts**	Tr
	colorCam1	Tr	Tr	**Ts**	Tr	Tr	**Ts**
	colorCam2	Tr	Tr	Tr	**Ts**	Tr	Tr
	Hallway	Tr	**Ts**	Tr	Tr	Tr	Tr
DepthCamouflage	DCamSeq1	**Ts**	Tr	Tr	Tr	**Ts**	Tr
	DCamSeq2	Tr	**Ts**	Tr	Tr	Tr	**Ts**
	Despatx_ds	Tr	Tr	**Ts**	Tr	Tr	Tr
	Wall	Tr	Tr	Tr	**Ts**	Tr	Tr
IlluminationChanges	ChairBox	**Ts**	Tr	Tr	Tr	**Ts**	Tr
	genSeq1	Tr	Tr	Tr	**Ts**	Tr	**Ts**
	Ls_ds	Tr	**Ts**	Tr	Tr	Tr	Tr
	TimeOfDay_ds	Tr	Tr	**Ts**	Tr	Tr	Tr
IntermittentMotion	abandoned1	Tr	Tr	**Ts**	Tr	Tr	Tr
	abandoned2	Tr	Tr	Tr	**Ts**	Tr	Tr
	movedBackground1	Tr	Tr	Tr	Tr	**Ts**	Tr
	movedBackground2	Tr	Tr	Tr	Tr	Tr	**Ts**
	Shelves	**Ts**	Tr	Tr	Tr	Tr	Tr
	Sleeping_ds	Tr	**Ts**	Tr	Tr	Tr	Tr
OutOfRange	MultiPeople1	**Ts**	Tr	Tr	Tr	Tr	**Ts**
	MultiPeople2	Tr	**Ts**	Tr	Tr	Tr	Tr
	TopViewLab1	Tr	Tr	**Ts**	Tr	Tr	Tr
	TopViewLab2	Tr	Tr	Tr	**Ts**	Tr	Tr
	TopViewLab3	Tr	Tr	Tr	Tr	**Ts**	Tr
Shadows	fall01cam1	Tr	**Ts**	Tr	Tr	Tr	**Ts**
	genSeq2	Tr	Tr	**Ts**	Tr	Tr	Tr
	Shadows_ds	**Ts**	Tr	Tr	Tr	Tr	Tr
	shadows1	Tr	Tr	Tr	**Ts**	Tr	Tr
	shadows2	Tr	Tr	Tr	Tr	**Ts**	Tr

the following upcoming studies. This strategy is explained in Table 4.2. Our target is to test the maximum possible cases to avoid bias or manipulated evaluation. We set up six possible splits, where the number of splits is based on the category with the most number of videos (in this case, it is the *IntermittentMotion* category, which consists of six videos). The Testing(Ts)/Training(Tr) videos were selected randomly for each split, from S1 to S6. Therefore, each split consists of twenty-six videos for training and seven videos for testing.

4.3.3 Leave One Video Out protocol for GSM

The GSM dataset is considered as a small dataset compared with the SBM-RGBD. However, we choose to use it for our experiments in order to test our deep learning proposed method in the case of a small amount of data. This will explain if the model can be trained with this minor quantity of information or if it requires more data to be well-trained. The GSM dataset contains seven categories with one video per each. Hence, we followed the LOVO protocol of seven splits as listed in Table 4.3.

Table 4.3: LOVO protocol for GSM

Video	S1	S2	S3	S4	S5	S6	S7
BootStrapping_ds	**Ts**	Tr	Tr	Tr	Tr	Tr	Tr
Cespatx_ds	Tr	**Ts**	Tr	Tr	Tr	Tr	Tr
Despatx_ds	Tr	Tr	**Ts**	Tr	Tr	Tr	Tr
Ls_ds	Tr	Tr	Tr	**Ts**	Tr	Tr	Tr
Shadows_ds	Tr	Tr	Tr	Tr	**Ts**	Tr	Tr
Sleeping_ds	Tr	Tr	Tr	Tr	Tr	**Ts**	Tr
TimeOfDay_ds	Tr	Tr	Tr	Tr	Tr	Tr	**Ts**

4.4 Experiments and performance evaluation

This section shows in detail, the results obtained following the framework presented in the previous sections. The experiments and evaluations presented in this section are based on the original state-of-the-art protocol and then on the proposed protocol to explain the necessity of proposing a new protocol for the field.

4.4.1 Results and discussion

This section will aggregate two types of evaluations to check the efficiency of the proposed method: The *F-score* measurement for the quantitative evaluation, which focuses on pixel-wise information by comparing the results and the ground truth as explained in Section 2.3, using Equation(2.1); Then, the visual evaluation, which is not as effective as the quantitative evaluation to differentiate between the methods performing nearly at the same level, but in some cases, it can bring attention to many hidden spots that could explain much about the behave of the method.

In this study, together with the proposed method, we have also implemented four existing techniques in pursuance of fair comparative evaluation: *MFCN* [50], *FgSegNet_S*[54], *FgSegNet_M* [54] and *FgSegNet_v2*[55]. The reasons behind choosing these methods are: (i) They are the top four methods in the stat-of-the-art. (ii) These methods take into account one of the most critical limitations mentioned in Subsection 2.5; they are flexible in following the SIE protocol to avoid the training/testing overlapping problem. (iii) Less complexity for implementing them, thanks to the *MFCN* authors who gave a detailed explanation in their original paper, as well as the *FgSegNet* authors for sharing their code online.

4.4.1.1 Experiments on SBM-RGBD dataset

This subsection is divided into two parts: (i) Experiment using the original protocol (ii) Experiment using the SIE protocol. Table 4.4 shows the overall *F-score* based on the original protocol. The *F-score* values conclude two main points: 1) The proposed method performs perfectly and competitively compared to the top four SOTA methods; 2) The original protocol is saturated. The *F-score* reaches ~ 99% on all the videos except on *Ls_ds* and *TimeOfDay_ds* ones, which lead to *F-score* equal to zero due to the absence of foreground objects in these two videos. This saturation is caused by the overlapping between Training/Testing splits using the original protocol strategy. This problem has already been discussed in the Subsection 4.3.1. This situation requires a new protocol imposing non-overlapping between training/testing splits, and allowing the background model learns within a more realistic experimental setup. Following the SIE protocol (see Table 4.2), the background model will be tested on unseen videos that are different from the training ones. Considering the SIE protocol, we keep one video out from each category for the testing phase and use the remaining videos for training the models. The training phase generates one model (M_i) for each split (S_i) independently.

Table 4.4: *F-score* on SBM-RGBD dataset (original protocol).

Method	***F-score*(Overall)**
BSABU[81]	0.85
RGBD-SOBS[43]	0.86
SCAD[82]	0.88
FgSegNet_S[54]	0.89
FgSegNet_M[54]	0.90
FgSegNet_v2[55]	0.91
MFCN[50]	0.91
DMSN (Proposed method)	0.91

We run the test on the forty-two *Ts* selected videos, using the corresponding model for each split to obtain the background subtraction for all sequences. Since the model yields are pixels with values that vary between 0 and 255, later we apply a threshold on each pixel value to obtain its final class, either Background or Foreground. However, it is notable in Fig. 4.3 that the threshold impact is not significant whereas the pixel output values converged to the edges (0 or 255). Nevertheless, the proposed network surpasses the tested methods over all the threshold values from 0.1 to 0.8. Table 4.5 introduces further details about quantitative evaluation from two perspectives, i.e., category-based and overall the SBM-RGBD dataset. As an accurate approach to measure the overall *F-score*, we insist to consider the average on all the threshold values (from 0.1 to 0.8), all the splits (from S1 to S6), and all the categories, respectively. We believe that this measurement approach is an outstanding way to distinguish the robust method from the others. The proposed method proves its efficiency compared to the state-of-the-art methods, where the improvement has reached ~ 3% in overall. At the Category-based evaluation level, we find the proposed method has outranked all the tested methods in three out of seven categories (*Bootstrapping*, *IlluminationChanges*, and *IntermittentMotion*), ranked the second in *ColorCamouflage*, and the third for remaining ones

(*DepthCamouflage*, *OutOfRange*, and *Shadows*). Indeed, these achievements validate the idea of including depth maps along with the RGB channels, since using only the color information will unequivocally occur many false detections in such circumstances, e.g., The sudden change in the illumination or an overlap of two (or more) objects having the same color, one is stationary, and the other is moving. Many of these drawbacks can be addressed by using depth information that is not affected by these scenarios. Besides, despite the non-use of temporal history or background reference, the proposed method still has a remarkable accomplishment in the *Bootstrapping* category compared to the other methods, where this category consists of videos that include foreground objects from the first frame till the last one, which considered as a challenging scene for all methods, including the techniques that utilize frame referencing from the same video of the testing. The only flop was in the *DepthCamouflage* category, which is logically due to the misleading information from the Depth maps; when two objects (or more) have almost the same distance from the camera, where one of them is moving the other is static, this will occur false detections.

Table 4.5: *F-score* on SBM-RGBD dataset.

Dataset categories	**FgSegNet_S** [54]	**FgSegNet_M** [54]	**FgSegNet_v2** [55]	**MFCN** [50]	**DMSN** (Proposed)
Bootstrapping	0.6639	0.6655	**0.7270**	0.7172	**0.7692**
ColorCamouflage	**0.4632**	0.3985	0.4307	0.4420	**0.4442**
DepthCamouflage	**0.8013**	**0.7855**	0.7245	0.7359	0.7383
IlluminationChanges	0.5341	0.5490	**0.5498**	0.5385	**0.5701**
IntermittentMotion	0.8061	0.8196	0.7210	**0.8336**	**0.8536**
OutOfRange	0.7760	**0.9233**	**0.9206**	0.8972	0.9118
Shadows	**0.7994**	0.6634	**0.7969**	0.7016	0.7783
Overall (Average)	0.6920	0.6864	0.6958	0.6952	**0.7236**

Furthermore, four samples from the tested videos are presented in Figure. 4.4. From a visual perspective, it is evident that the proposed method overcomes most of the false detections. We should mention that the bottom row illustrates a sample from the *TimeOfDay* sequence, which is included in the *IlluminationChanges* category; this sequence does not contain foreground objects which affects the *F-score* measurement making it equal to zero. However, referring to the stat-of-the-art and the dataset website, including this sequence in the quantitative evaluation is mandatory. This kind of video is essential to test the performance in illumination change, where it is clear that our proposed method performs well in such scenarios.

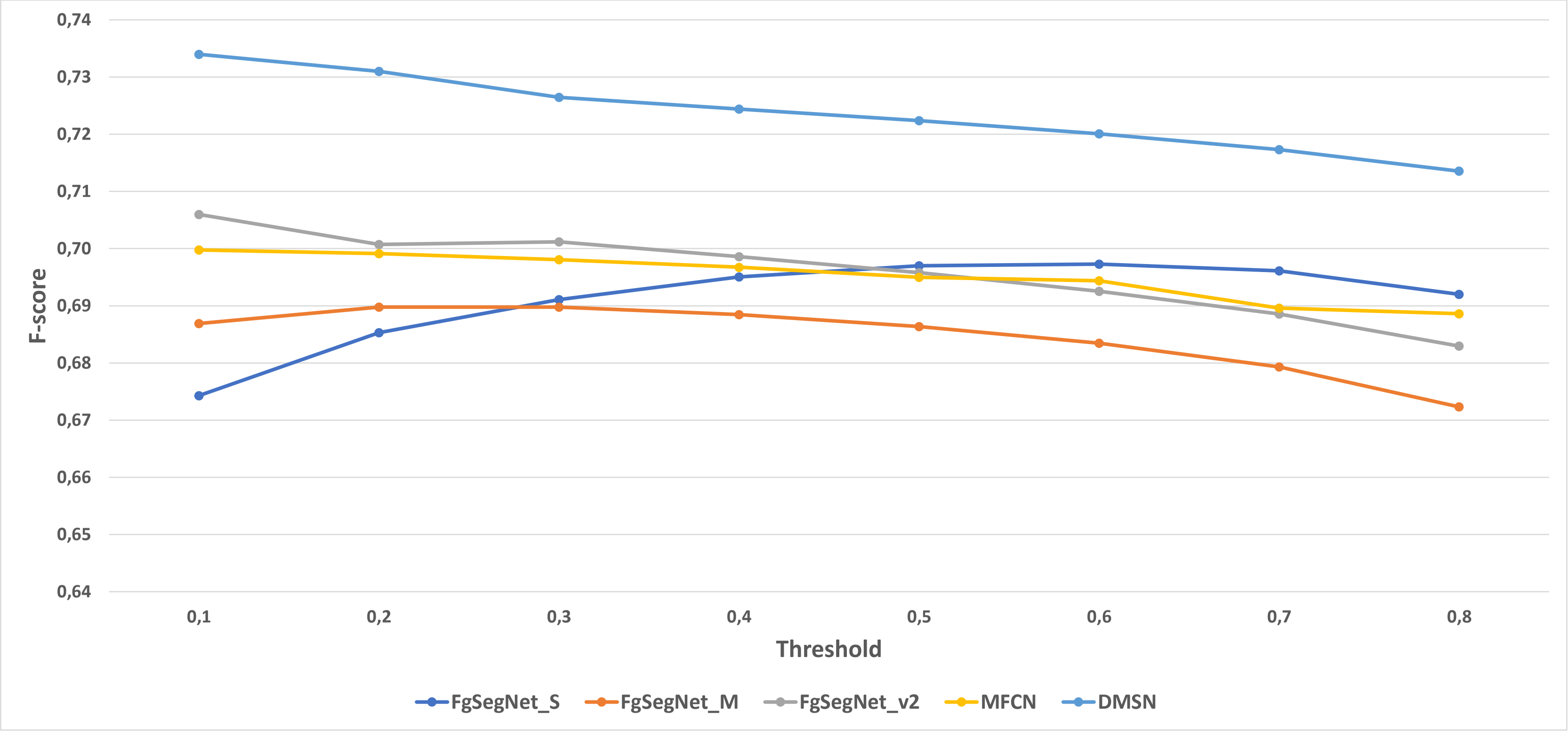


Figure 4.3: Average *F-score* at different thresholds on SBM-RGBD.

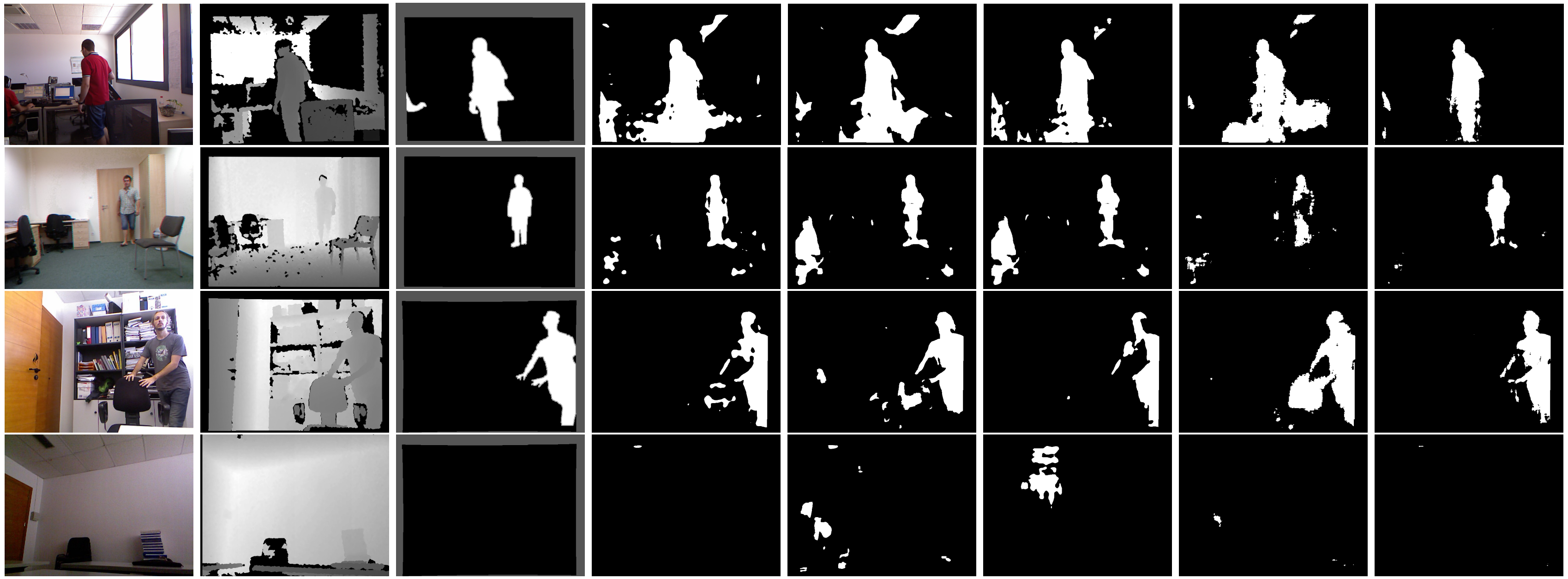
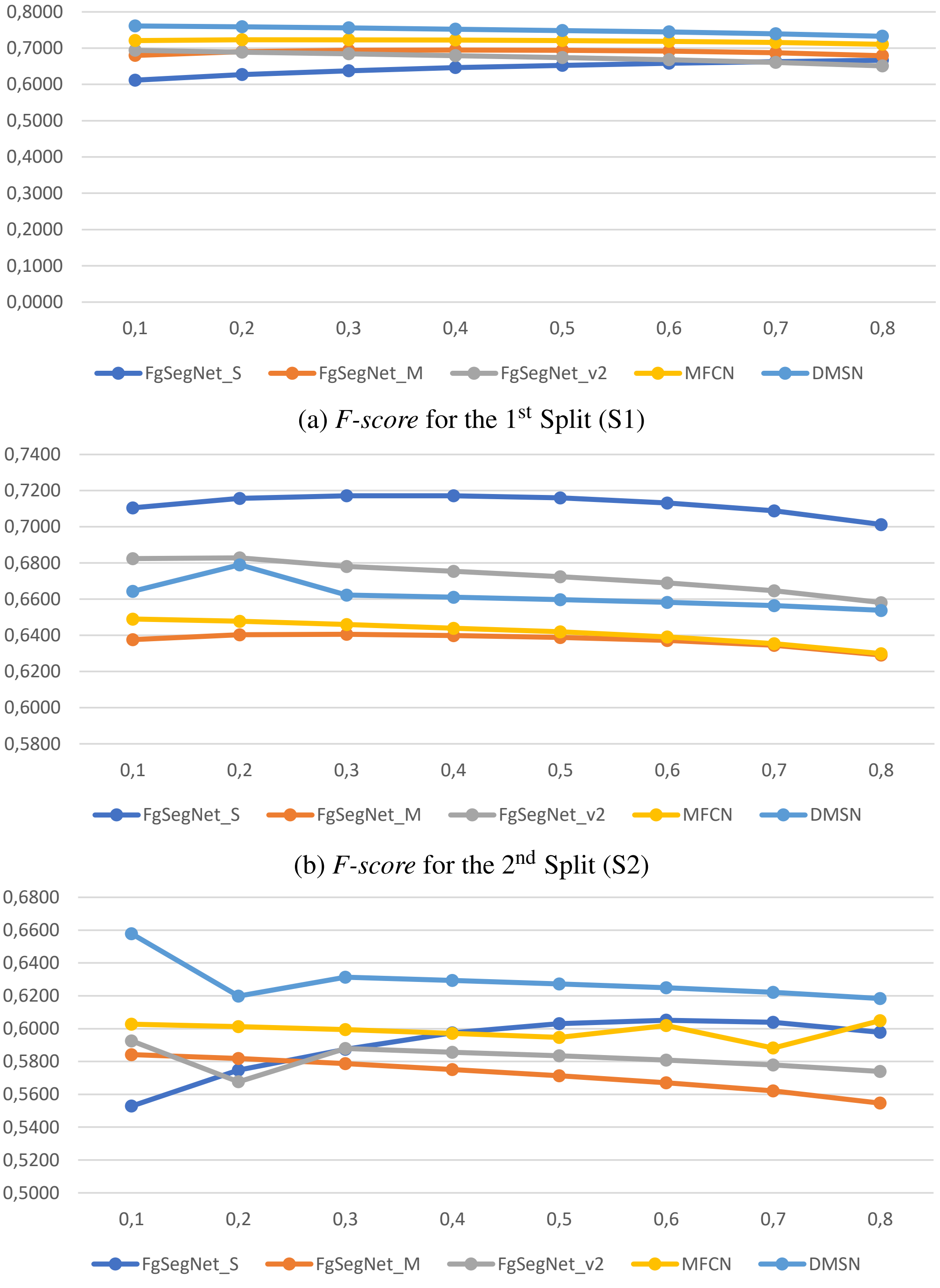


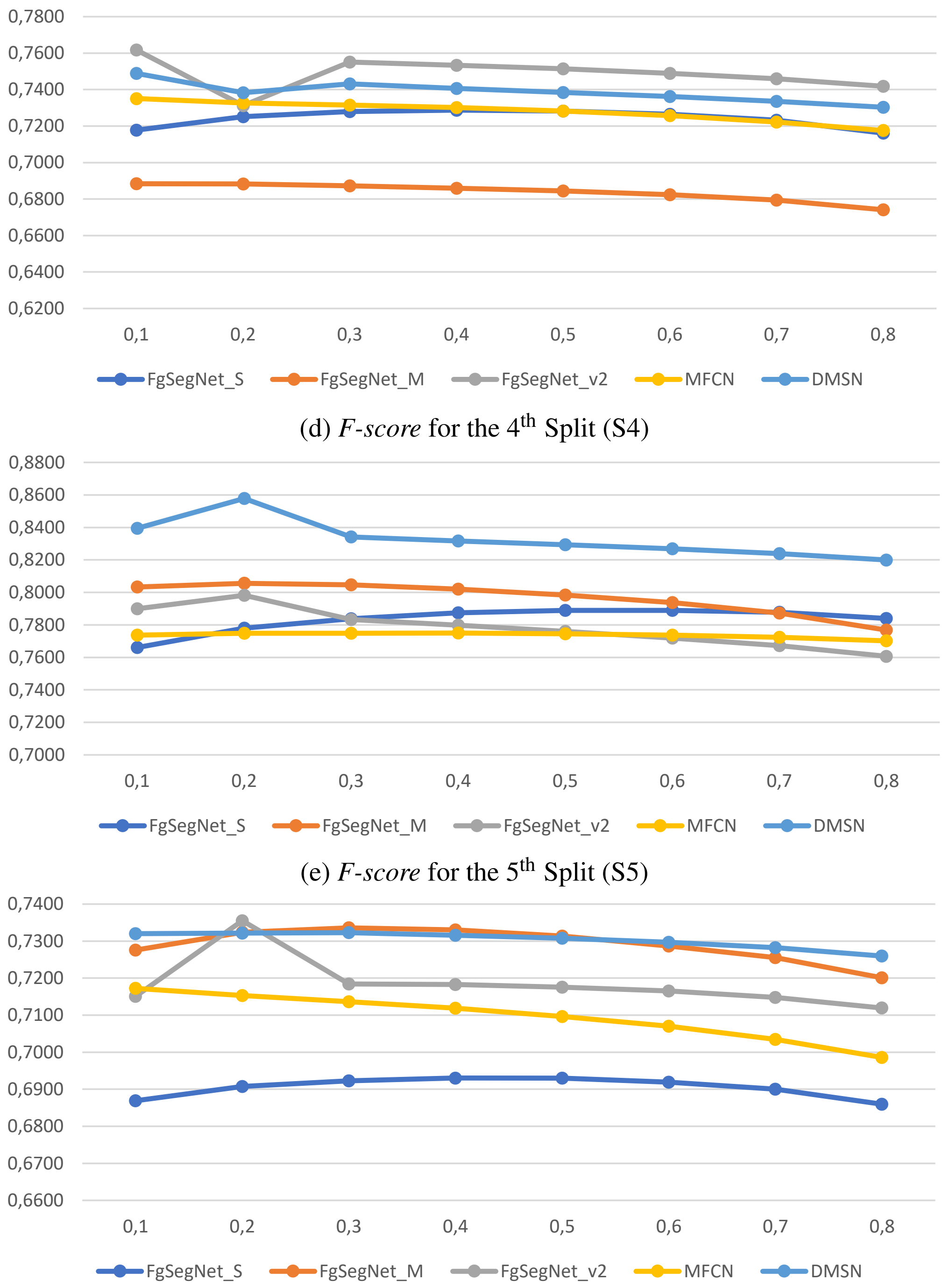
Figure 4.4: The visual comparison. Rows from top to bottom show four video examples from the SBM-RGBD dataset: *ChairBox*, *Fall01cam0*, *Sleeping_ds*, and *TimeOfDay*, respectively. Columns from left to right represent the RGB input, Depth, groundtruth, and example frames from the results of, *FgSegNet_S*[54], *FgSegNet_M*[54], *FgSegNet_v2*[55], *MFCN*[50], and the proposed method *DMSN*, respectively.



(a) *F-score* for the 1st Split (S1)

(b) *F-score* for the 2nd Split (S2)

(c) *F-score* for the 3ed Split (S3)



(d) *F-score* for the 4th Split (S4)

(e) *F-score* for the 5th Split (S5)

(f) *F-score* for the 6th Split (S6)

Figure 4.5: *F-score* calculation at different thresholds on SBM-RGBD for each split (S1 to S6).

Deeper into results illustration, Figure 4.5 shows the *F-score* measurement through the tested threshold values (from 0.1 to 0.8) for each split as divided and explained in Table 4.2. The results show that the proposed model DMSN is leading in most of the splits (S1, S3, S5, and S6). However, it is placed second and third in S4 and S2, respectively. These promising achievements explain the stability of the performance provided by the proposed model compared to the other tested models. The FgSegNet_S for example, shows a great achievement in S2, but its performance dramatically decreases in S6, which makes it less trustfully compared to our proposed method.

4.4.1.2 Experiments on GSM dataset

The same findings were observed through the evaluation analysis on the *GSM* dataset, which was challenging as we had to build our model from a small number of videos. Table 4.3 explains the Leave One Video Out (LOVO) protocol that has been followed in this study. We will generate seven different models referring to seven splits (from S1 to S7) in this case. Fig. 4.6 illustrates the *F-score* measurement at different threshold values, where the proposed method outperforms the state-of-the-art methods generally at all the threshold values. Table 4.6 reports the average of the *F-score* measurement considering all the threshold values. The proposed model surpasses the state-of-the-art methods in overall with a ratio difference of ~ 2% compared to the next best-performing method. One can notice the efficiency provided by the *DMSN* model on the *Bootstrapping_ds* video where the ratio difference reached ~ 9% compared to the next best performing method. Nevertheless, the potency is still competitive in the remaining videos compared to the tested techniques, this proves the robustness and performance stability provided by the *DMSN* model, in all kinds of scenarios. The visual evaluation presented in Fig. 4.7 gives more details about the model performance. Particularly, referring to the two videos (*Ls_ds* and *TimeOfDayds*) that can not be quantitatively evaluated as a result of having only one class along their frame sequences (*Background*). These two sequences do not contain any foreground objects, which affects the *F-score* measurement, as we mentioned earlier about the SBM-RGBD, referring to the *IlluminationChanges* category. The visual results show the precision of reducing the false detections, primarily false positives and providing better performance. The last two rows explain the behavior of the proposed method in the illumination change scenarios, in which the sample of the third row includes a scene of an office exposed to unpredictable lighting from the window, and the fourth row is an office that has a blinked LED lamp.

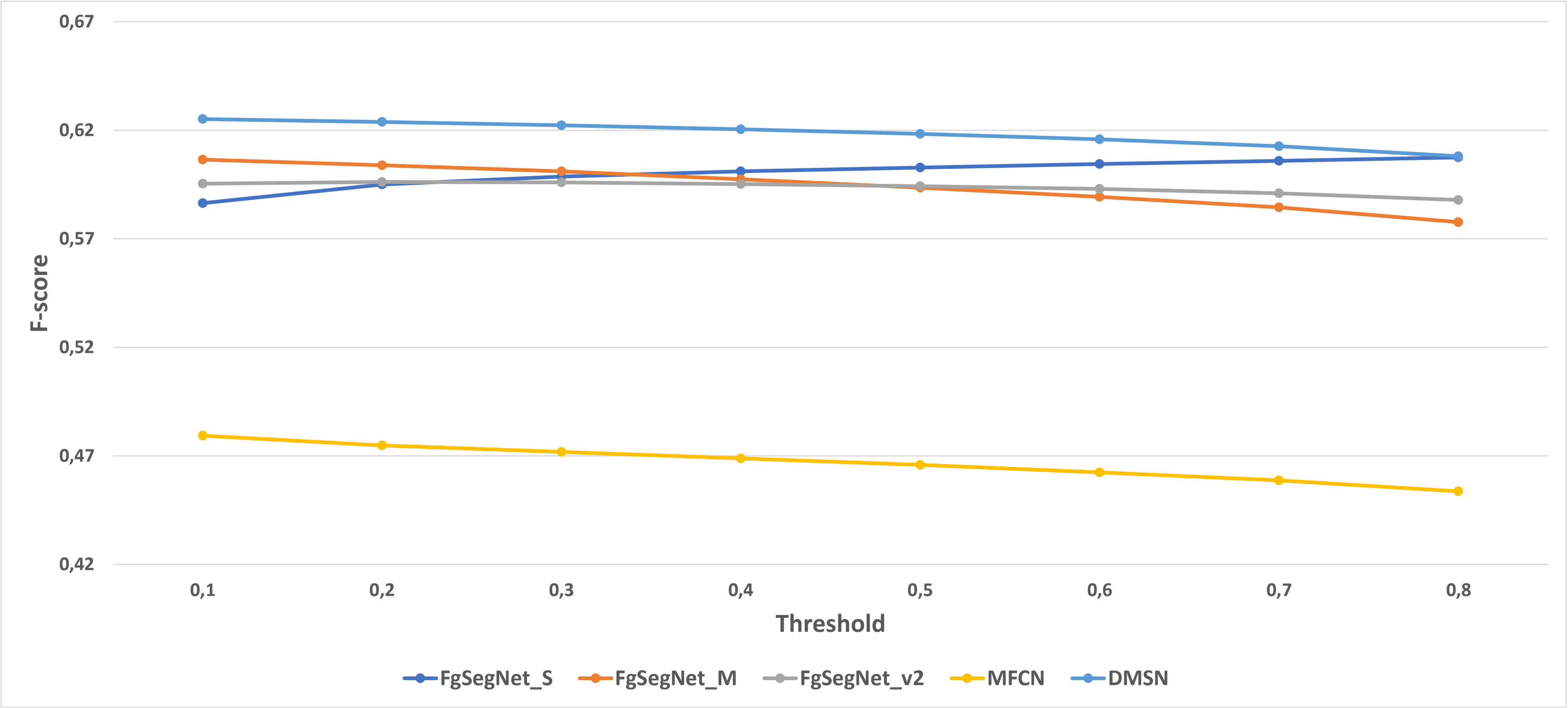


Figure 4.6: Average *F-score* at different thresholds on GSM dataset.

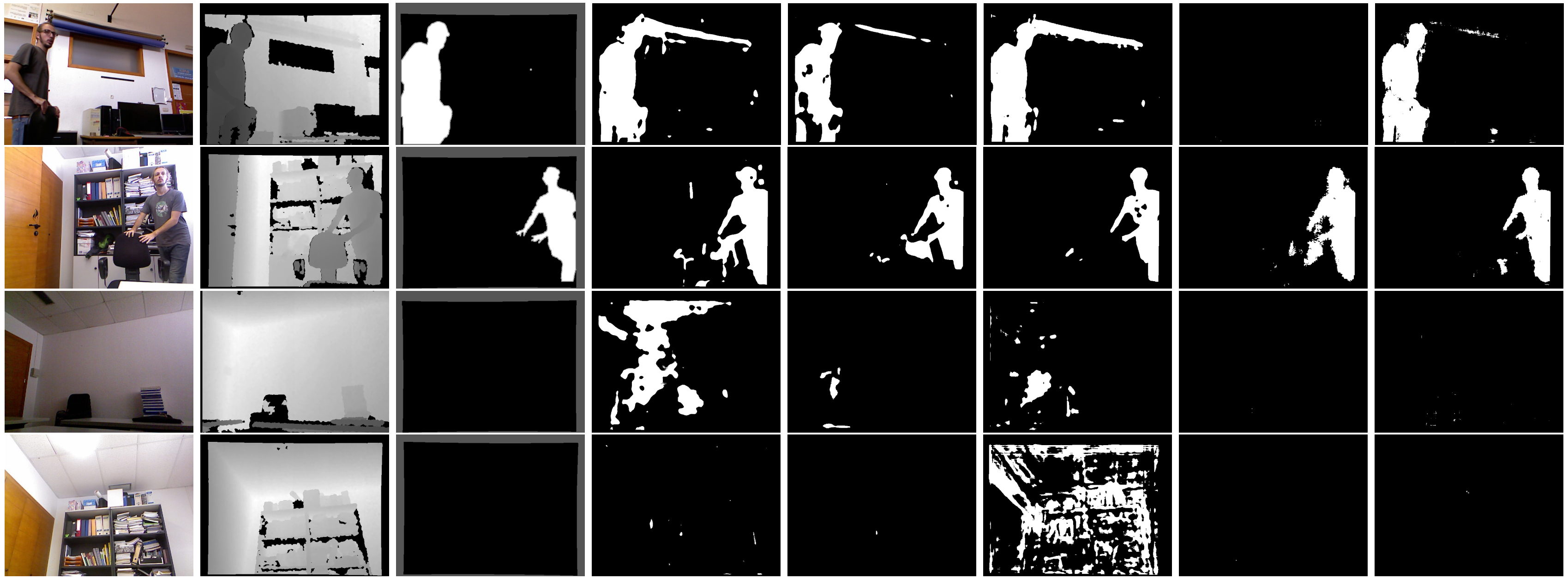


Figure 4.7: The visual comparison. Rows from top to bottom show four examples from the GSM dataset: *Bootstrapping_ds*, *Sleeping_ds*, *TimeOfDay*, and *LS_ds*, respectively. Columns from left to right represent the RGB input, Depth, groundtruth, and example frames from the results of, *FgSegNet_S*[54], *FgSegNet_M*[54], *FgSegNet_v2*[55], *MFCN*[50], and the proposed method *DMSN*, respectively.

Table 4.6: *F-score* on GSM dataset.

Dataset categories	FgSegNet_S [54]	FgSegNet_M [54]	FgSegNet_v2 [55]	MFCN [50]	DMSN (Proposed)
Bootstrapping_ds	0.6673	**0.7504**	0.6276	0.0001	**0.8461**
Cespatx_ds	0.9058	**0.9250**	**0.9211**	0.8847	0.9049
Despatx_ds	0.8784	0.9073	**0.9194**	0.8760	**0.9080**
Ls_ds	0.0000	0.0000	0.0000	0.0000	0.0000
Shadows_ds	**0.8237**	0.7091	**0.8063**	0.6526	0.7968
Sleeping_ds	**0.9264**	0.8679	**0.8880**	0.8551	0.8726
TimeOfDay_ds	0.0000	0.0000	0.0000	0.0000	0.0000
Overall (Average)	0.6002	0.5943	0.5936	0.4669	**0.6183**

4.4.1.3 The Ablation Study

This subsection presents two experiments to highlight the impact of different parts composing the proposed architecture. The first experiment is evaluating the Deep Multi-Scale Network with a Single Stream Encoder (DMSN_SSE), where the RGB and Depth inputs are concatenated on the channel dimension, then fed to one single VGG-16 encoder instead of two encoders as in the proposed DMSN. The second experiment concerns the Intermediate Layers illustrated in Fig. 4.1 and Table 4.1, considering three configurations. The first one is Deep Multi-Scale Network Without Intermediate Layers (DMSN_WIL), where the skipping connections are directly connected from the two encoders to the decoder without intermediate layers. The second configuration is Deep Multi-Scale Network, RGB Without Intermediate Layers (DMSN_RGB_WIL), where we removed the intermediate layers between the RGB encoder and the decoder while keeping them on the depth encoder side. The third configuration is Deep Multi-Scale Network, Depth Without Intermediate Layers (DMSN_D_WIL); it is the opposite of the DMSN_RGB_WIL.

Table 4.7 lists the recorded *F-scores* from experiment 1 and experiment 2 (with the three configurations) on both datasets.

The results from Experiment 1 show the importance of using two encoder streams for each input instead of merging them into one single encoder. We believe that using one single encoder may mislead the encoder, preventing the model from being correctly trained; this is due to the dissimilar types of the inputs, which should be encoded separately and differently, allowing more relevant feature extraction. On the other hand, the results related to Experiment 2 highlight the proposed method performances with and without the Intermediate Layers following the three configurations

Table 4.7: *F-score* from experiment 1 and experiment 2.

Experiments	**Configurations**	***F-score*(Overall)** SBM-RGBD	 GSM
Expt.1	**DMSN_SSE**	0.61	0.47
Expt.2	**DMSN_WIL**	0.27	0.35
	DMSN_RGB_WIL	0.70	0.57
	DMSN_D_WIL	0.72	0.50
Proposed	**DMSN**	0.72	0.62

as depicted in Table 4.7. The DMSN_WIL configuration suffered a major drop over the *F-score* on both datasets, proving the need for the Intermediate Layers within the skip-connections. Moreover, enabling them in only one of the encoders improves the performance significantly, as stated for DMSN_RGB_WIL and DMSN_D_WIL configurations. Furthermore, as proposed in our DMSN model, enabling them in both encoders guarantees more performance and stability. Therefore, we disclose that the Intermediate Layers placed between the two encoders and the decoder are conservative to the spatial features coming from each Max-Pooling layer.

4.4.1.4 Implementation and Processing Speed

Our experiments are implemented using Python 3.6, Keras 2.3, Tensorflow-gpu 2.1, and Cuda 10.1 installed on a machine built with Intel i9 CPU 9^{th} generation and GeForce RTX 2080 GPU under Windows 10 operating system. The processing speed of our tests is four frames per second for videos with a resolution of 640 x 480. We would also like to mention that the code source of the proposed method is available online at: `https://github.com/ihssanehouhou/DMSN`.

4.5 Conclusion

In this chapter, we proposed a novel deep learning model named DMSN (Deep Multi-Scale Network) that is VGG16-based. This model uses RGBD information and feature multi-scale extraction to solve diverse scenarios in background subtraction without using a background model initialization. We proposed also a new protocol for the SBM-RGBD dataset based on SIE in pursuance of a fair competitive analysis. The reported experiments show that our proposed model surpasses the state-of-the-art upon two datasets, SBM-RGBD and *GSM*. The proposed method is performing better in particular scenarios such as *Bootstrapping*, *IntermittentMotion* and, *IlluminationChanges*. However, *DepthCamouflage* scenes are more challenging for our model due to the misleading information provided by the depth maps. In future work, we will focus on improving the model to reduce the potential effect of the depth information misleading. Further, considering other pre-trained models instead of the VGG-16 could be beneficial in improving the overall performance.

CHAPTER 5

CONCLUSIONS AND FUTURE WORKS

Contents

Page

5.1 Conclusions 91

5.2 Future perspectives 93

This chapter concludes together the study by reviewing the main findings in regard to the research goals and queries, as well as the significance and contribution thereof. It also discusses the limitations and includes recommendations for further studies.

5.1 Conclusions

At the beginning of this thesis, we explained how important to use multiple kinds of input information to improve the background subtraction techniques. Our focus area was targeting the use of color and depth information along. Based on the literature, some of the traditional approaches have already adopted this idea and proved its efficiency. However, to the best of our knowledge, the use of Depth maps along with RGB channels through double-stream encoders has not been included in deep-learning-based techniques.

In this study, we aimed to propose novel RGBD-based frameworks in both branches of this ax of research, traditional and deep learning. The results of this study indicate that depth information (Depth maps) has a significant effect when paired with color information (RGB channels). Moreover, this research also aimed for a better evaluation of the background methods, which led us to propose a new protocol based on SIE. The findings show that based on this protocol, we can build models that can be more practical in real scenarios.

Our motivation to work on traditional approaches was the fact that they are unsupervised methods. At this point, we proposed a paradigm that employs two types of information (RGB and Depth). This framework is dedicated to the ViBe-based traditional background subtraction methods. The experimental assessment of the SBM-RGBD dataset demonstrated the framework's effectiveness in both overall and category-based. The ColorCamouflage category is one of the targeted categories; it shows a significant improvement due to the additional information provided by the depth maps. This framework is also regarded as a reliable strategy for significantly improving outcomes in a variety of techniques, including ViBe, LOBSTER, SuBSENSE, and PAWCS.

However, the achievements that we have accomplished using this framework still have some limitations. The traditional approaches have more complexity and a low frame rate as an inverse relation compared to the performance. Moreover, traditional techniques' efficiency is related to the use of reference frame(s), which drives the method to perform only at a specific scenario, i.e., the one linked to the reference frame. The evident limitation of the proposed traditional framework is the performance on the

OutOfRange category, where the Depth information here is unreliable.

Some proposed solutions can be helpful for developing this approach to overcome the limitations. Using low frame resolution will speed up the process. At the same time, it can reduce the unnecessary information that can be destructive in this case. Other data, like descriptors, or another depth extractor device/technique, will overcome the misleading prevented by the depth information using the Microsoft-Kinect camera in some scenes, such as the previously mentioned category (OutOfRange).

Despite the mentioned limitations, the traditional methods can always be valuable and practical for non-complex applications which do not include frequent moving backgrounds or require fast decisions.

As the second major objective of this study, we shifted our attention to CNN-based approaches. We aimed for a model to solve the background subtraction problem using color and depth information. The results of the original protocol indicated that our model achieves auspicious performance compared to the literature. On the other hand, our aim to build an effective model is related to the evaluation technique that we follow, which directed us to propose a new protocol to cope with realistic-kind challenges. This new protocol allows a fair evaluation and builds a competitive platform for the next generation of CNN-based background subtraction methods.

This section suggested a new VGG16-based deep learning model called DMSN (Deep Multi-Scale Network). This model solves several cases in background subtraction without requiring frame references by using RGBD information. This approach is premised on the principle of extracting multi-scale features. In the interest of fair, competitive analysis, and due to the performance saturation caused by the original protocols, we also presented a novel protocol based on SIE. This protocol is dedicated to the deep-learning-based approaches to set up a competitive platform that includes more challenging situations and avoids the original protocol gaps.

The experiments reveal that our suggested model outperforms the state-of-the-art in several categories, such as Bootstrapping, IntermittentMotion, and IlluminationChanges. These two first categories contain the most challenging scenarios for the methods that are based on frame references to initialize the background model. This makes our proposed method perform better in such scenes due to the new approach we proposed to create our model using the SIE-based protocol. On the other hand, using depth information along with color information helps the model solve the IlluminationChanges category challenges better than the RGB-based techniques from state-of-the-art. Besides, based on our literature study, no deep learning model was built to use color and depth information in parallel to solve background subtraction. Hence this study was the first to initiate this kind of CNN-based model using RGBD with double-stream encoders in this field.

However, the proposed model shows less efficiency in DepthCamouflage and OutOfRange scenarios. These categories are more problematic for our model due to the deceptive data provided by the depth maps. For this matter, we can propose a pre-processing for the depth maps to overcome this limitation. This pre-processing should determine the state of the depth information by indicating if the provided data include some properties of these mentioned categories. Thereafter, we can make the model decision rely more on the color information. Moreover, using other sources to extract reliable depth maps is also considered as a practical option to solve this issue.

5.2 Future perspectives

This research can be further developed by considering some of the following suggestions.

We believe that the proposed approaches can be applied to outdoor scenes, but this depends on the availability of the required data for this task. Since our proposed methods target indoor scenarios, researchers interested in our study can also implement them for outdoor scenarios. Moreover, autonomous cars seem like a fascinating topic for future works, with additional challenges besides the outdoor environment, such as moving camera scenarios and object-kind variety.

From a theoretical perspective, we expect the two proposed approaches (traditional and deep-learning) to be as effective and compatible with other applications and/or with the use of different kinds of data as well. The success of this hypothesis only depends on providing two types of information. As we built our proposed methods on color and depth information, other research can consider using another kind of combination between descriptors (Local Binary Patterns (LBP), Local Phase Quantization (LPQ), or Binarized Statistical Image Features (BSIF), etc.), color, and/or depth. Also, different color spaces or some recent feature extractors may be recommended to have better performance.

The fact that the proposed protocol based on the SIE opened up room for more improvement in deep learning approaches and the two-stream concept presented in the CNN-based approach proves its effectiveness, our following study will invest in utilizing the double streams in deep learning networks. Since our CNN-based method uses only one scene out of two types of information, our future work will consider two consecutive frames instead of a color-depth combination. This will allow the model to learn from the variation between these two frames and make a decision based on that. This proposition may include the previous suggestions mentioned earlier in this section.

BIBLIOGRAPHY

[1] E. R. Davies, *Computer vision: principles, algorithms, applications, learning*. Academic Press, 2017.

[2] J. L. Potter, “Velocity as a cue to segmentation,” *IEEE Transactions on Systems, Man, and Cybernetics*, no. 3, pp. 390–394, 1975.

[3] G. L. Walls, “The vertebrate eye and its adaptive radiation,” 1944.

[4] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2, pp. 246–252, IEEE, 1999.

[5] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, “Changedetection. net: A new change detection benchmark dataset,” in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 1–8, IEEE, 2012.

[6] 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*, pp. 387–394, 2014.

[7] C. Cuevas, E. M. Yáñez, and N. García, “Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta,” *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, 2016.

[8] D. D. Bloisi, L. Iocchi, A. Pennisi, and L. Tombolini, “Argos-venice boat classification,” in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2015.

[9] L. Maddalena and A. Petrosino, “Towards benchmarking scene background initialization,” in *International conference on image analysis and processing*, pp. 469–476, Springer, 2015.

[10] A. Petrosino, L. Maddalena, and T. Bouwmans, “Scene background modeling and initialization,” *Pattern Recognition Letters*, vol. 96, pp. 1–2, 2017.

[11] P.-M. Jodoin, L. Maddalena, A. Petrosino, and Y. Wang, “Extensive benchmark and survey of modeling methods for scene background initialization,” *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5244–5256, 2017.

[12] M. Camplani, L. Maddalena, G. M. Alcover, A. Petrosino, and L. Salgado, “A benchmarking framework for background subtraction in rgbd videos,” in *International Conference on Image Analysis and Processing*, pp. 219–229, Springer, 2017.

[13] G. Moyà-Alcover, A. Elgammal, A. Jaume-i Capó, and J. Varona, “Modeling depth for nonparametric foreground segmentation using rgbd devices,” *Pattern Recognition Letters*, vol. 96, pp. 76–85, 2017.

[14] E. J. Fernandez-Sanchez, J. Diaz, and E. Ros, “Background subtraction based on color and depth using active sensors,” *Sensors*, vol. 13, no. 7, pp. 8895–8915, 2013.

[15] S. Song and J. Xiao, “Tracking revisited using rgbd camera: Unified benchmark and baselines,” in *Proceedings of the IEEE international conference on computer vision*, pp. 233–240, 2013.

[16] M. Camplani and L. Salgado, “Background foreground segmentation with rgb-d kinect data: An efficient combination of classifiers,” *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 122–136, 2014.

[17] B. Kwolek and M. Kepski, “Human fall detection on embedded platform using depth maps and wireless accelerometer,” *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[18] P. KaewTraKulPong and R. Bowden, “An improved adaptive background mixture model for real-time tracking with shadow detection,” in *Video-based surveillance systems*, pp. 135–144, Springer, 2002.

[19] Z. Zivkovic, “Improved adaptive gaussian mixture model for background subtraction,” in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2, pp. 28–31, IEEE, 2004.

[20] D.-S. Lee, “Effective gaussian mixture learning for video background subtraction,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 1, no. 5, pp. 827–832, 2005.

[21] Z. Zivkovic and F. Van Der Heijden, “Efficient adaptive density estimation per image pixel for the task of background subtraction,” *Pattern recognition letters*, vol. 27, no. 7, pp. 773–780, 2006.

[22] Z. Zhao, T. Bouwmans, X. Zhang, and Y. Fang, “A fuzzy background modeling approach for motion detection in dynamic backgrounds,” in *International Conference on Multimedia and Signal Processing*, pp. 177–185, Springer, 2012.

[23] A. Darwich, P.-A. Hébert, A. Bigand, and Y. Mohanna, “Background subtraction based on a new fuzzy mixture of gaussians for moving object detection,” *Journal of Imaging*, vol. 4, no. 7, p. 92, 2018.

[24] A. Elgammal, D. Harwood, and L. Davis, “Non-parametric model for background subtraction,” in *European conference on computer vision*, pp. 751–767, Springer, 2000.

[25] H. Wang and D. Suter, "A consensus-based method for tracking: Modelling background scenario and foreground appearance," *Pattern recognition*, vol. 40, no. 3, pp. 1091–1105, 2007.

[26] B. Wang and P. Dudek, "A fast self-tuning background subtraction algorithm," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 395–398, 2014.

[27] G. Ramirez-Alonso and M. I. Chacon-Murguia, "Auto-adaptive parallel som architecture with a modular analysis for dynamic object segmentation in videos," *Neurocomputing*, vol. 175, pp. 990–1000, 2016.

[28] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Background modeling and subtraction by codebook construction," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, vol. 5, pp. 3061–3064, IEEE, 2004.

[29] J. Murgia, *Segmentation d'objets mobiles par fusion RGB-D et invariance colorimétrique*. PhD thesis, Université de Technologie de Belfort-Montbeliard, 2016.

[30] O. Barnich and M. Van Droogenbroeck, "Vibe: a powerful random technique to estimate the background in video sequences," in *2009 IEEE international conference on acoustics, speech and signal processing*, pp. 945–948, IEEE, 2009.

[31] S. Noh and M. Jeon, "A new framework for background subtraction using multiple cues," in *Asian Conference on Computer Vision*, pp. 493–506, Springer, 2012.

[32] J. Murgia, C. Meurie, and Y. Ruichek, "An improved colorimetric invariants and rgb-depth-based codebook model for background subtraction using kinect," in *Mexican International Conference on Artificial Intelligence*, pp. 380–392, Springer, 2014.

[33] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2014.

[34] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Flexible background subtraction with self-balanced local sensitivity," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 408–413, 2014.

[35] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 990–997, IEEE, 2015.

[36] Y. Wang, H. Lu, R. Gao, and Y. Wang, "V-vibe: A robust roi extraction method based on background subtraction for vein images collected by infrared device," *Infrared Physics & Technology*, vol. 123, p. 104175, 2022.

[37] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[38] G.-A. Bilodeau, J.-P. Jodoin, and N. Saunier, “Change detection in feature space using local binary similarity patterns,” in *2013 International Conference on Computer and Robot Vision*, pp. 106–112, IEEE, 2013.

[39] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, “Universal background subtraction using word consensus models,” *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4768–4781, 2016.

[40] B. Han and L. S. Davis, “Density-based multifeature background subtraction with support vector machine,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 1017–1023, 2011.

[41] L. Maddalena and A. Petrosino, “A self-organizing approach to background subtraction for visual surveillance applications,” *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008.

[42] L. Maddalena and A. Petrosino, “A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection,” *Neural Computing and Applications*, vol. 19, no. 2, pp. 179–186, 2010.

[43] L. Maddalena and A. Petrosino, “Exploiting color and depth for background subtraction,” in *International Conference on Image Analysis and Processing*, pp. 254–265, Springer, 2017.

[44] J. Leens, S. Piérard, O. Barnich, M. Van Droogenbroeck, and J.-M. Wagner, “Combining color, depth, and motion for video segmentation,” in *International Conference on Computer Vision Systems*, pp. 104–113, Springer, 2009.

[45] S. Ottonelli, P. Spagnolo, P. L. Mazzeo, and M. Leo, “Improved video segmentation with color and depth using a stereo camera,” in *2013 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1134–1139, IEEE, 2013.

[46] J. Huang, H. Wu, Y. Gong, and D. Gao, “Random sampling-based background subtraction with adaptive multi-cue fusion in rgbd videos,” in *2016 9th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, pp. 30–35, IEEE, 2016.

[47] V. Mahadevan and N. Vasconcelos, “Spatiotemporal saliency in dynamic scenes,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 1, pp. 171–177, 2009.

[48] M. Braham and M. Van Droogenbroeck, “Deep background subtraction with scene-specific convolutional neural networks,” in *2016 international conference on systems, signals and image processing (IWSSIP)*, pp. 1–4, IEEE, 2016.

[49] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

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

[51] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[52] C. Zhao, T.-L. Cham, X. Ren, J. Cai, and H. Zhu, “Background subtraction based on deep pixel distribution learning,” in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2018.

[53] C. Zhao and A. Basu, “Dynamic deep pixel distribution learning for background subtraction,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[54] L. A. Lim and H. Y. Keles, “Foreground segmentation using convolutional neural networks for multiscale feature encoding,” *Pattern Recognition Letters*, vol. 112, pp. 256–262, 2018.

[55] L. A. Lim and H. Y. Keles, “Learning multi-scale features for foreground segmentation,” *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1369–1380, 2020.

[56] R. Liu, Y. Ruichek, and M. El Bagdouri, “Multispectral background subtraction with deep learning,” *Journal of Visual Communication and Image Representation*, vol. 80, p. 103267, 2021.

[57] O. Tezcan, P. Ishwar, and J. Konrad, “Bsuv-net: a fully-convolutional neural network for background subtraction of unseen videos,” in *The IEEE Winter Conference on Applications of Computer Vision*, pp. 2774–2783, 2020.

[58] M. O. Tezcan, P. Ishwar, and J. Konrad, “Bsuv-net 2.0: Spatio-temporal data augmentations for video-agnostic supervised background subtraction,” *IEEE Access*, vol. 9, pp. 53849–53860, 2021.

[59] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

[60] M. C. Bakkay, H. A. Rashwan, H. Salmane, L. Khoudour, D. Puigtt, and Y. Ruichek, “Bscgan: deep background subtraction with conditional generative adversarial networks,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4018–4022, IEEE, 2018.

[61] M. Sultana, A. Mahmood, S. Javed, and S. K. Jung, “Unsupervised deep context prediction for background estimation and foreground segmentation,” *Machine Vision and Applications*, vol. 30, no. 3, pp. 375–395, 2019.

[62] W. Yu, J. Bai, and L. Jiao, “Background subtraction based on gan and domain adaptation for vhr optical remote sensing videos,” *IEEE Access*, vol. 8, pp. 119144–119157, 2020.

[63] P. W. Patil, A. Dudhane, and S. Murala, “Multi-frame recurrent adversarial network for moving object segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2302–2311, 2021.

[64] P. W. Patil, A. Dudhane, S. Murala, and A. B. Gonde, "Deep adversarial network for scene independent moving object segmentation," *IEEE Signal Processing Letters*, vol. 28, pp. 489–493, 2021.

[65] M. Sultana, A. Mahmood, and S. K. Jung, "Unsupervised moving object segmentation using background subtraction and optimal adversarial noise sample search," *Pattern Recognition*, vol. 129, p. 108719, 2022.

[66] X. Wang, L. Liu, G. Li, X. Dong, P. Zhao, and X. Feng, "Background subtraction on depth videos with convolutional neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, IEEE, 2018.

[67] M. Sultana, A. Mahmood, S. Javed, and S. K. Jung, "Unsupervised rgbd video object segmentation using gans," *arXiv preprint arXiv:1811.01526*, 2018.

[68] M. Sultana, T. Bouwmans, J. H. Giraldo, and S. K. Jung, "Robust foreground segmentation in rgbd data from complex scenes using adversarial networks," in *International Workshop on Frontiers of Computer Vision*, pp. 3–16, Springer, 2021.

[69] O. Barnich and M. Van Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2010.

[70] P.-L. St-Charles and G.-A. Bilodeau, "Improving background subtraction using local binary similarity patterns," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 509–515, IEEE, 2014.

[71] P. J. Burt, "Fast filter transform for image processing," *Computer graphics and image processing*, vol. 16, no. 1, pp. 20–51, 1981.

[72] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[73] M. Strengert, M. Kraus, and T. Ertl, "Pyramid methods in gpu-based image processing," *Proceedings vision, modeling, and visualization 2006*, pp. 169–176, 2006.

[74] X. Zhou, X. Liu, A. Jiang, B. Yan, and C. Yang, "Improving video segmentation by fusing depth cues and the visual background extractor (vibe) algorithm," *Sensors*, vol. 17, no. 5, p. 1177, 2017.

[75] R. Trabelsi, I. Jabri, F. Smach, and A. Bouallegue, "Efficient and fast multi-modal foreground-background segmentation using rgbd data," *Pattern Recognition Letters*, vol. 97, pp. 13–20, 2017.

[76] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[77] M. Afifi, "11k hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, vol. 78, no. 15, pp. 20835–20854, 2019.

[78] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[79] M. Mandal and S. K. Vipparthi, “Scene independency matters: An empirical study of scene dependent and scene independent evaluation for cnn-based change detection,” *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[80] J. H. Giraldo and T. Bouwmans, “Semi-supervised background subtraction of unseen videos: Minimization of the total variation of graph signals,” in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3224–3228, IEEE, 2020.

[81] N. Dorudian, S. Lauria, and S. Swift, “Moving object detection using adaptive blind update and rgb-d camera,” *IEEE Sensors Journal*, vol. 19, no. 18, pp. 8191–8201, 2019.

[82] T. Minematsu, A. Shimada, H. Uchiyama, and R.-i. Taniguchi, “Simple combination of appearance and depth for foreground segmentation,” in *International Conference on Image Analysis and Processing*, pp. 266–277, Springer, 2017.