Anomaly detection and object tracking by future prediction using generative methods for transportation

Tuan-Hung Vu

IMT Lille Douai
Université Gustave Eiffel

École doctorale **Science Pour l'Ingénieur EDSPI**
Unité de recherche **CERI SN - IMT Lille Douai**

Thèse présentée par **Tuan Hung VU**

Soutenue le **28 juin 2021**

En vue de l'obtention du grade de docteur de l'IMT Lille Douai et de l'Université Gustave Eiffel

Discipline **Computer Science and Informatics**
Spécialité **Signal and Image Processing**

ANOMALY DETECTION AND OBJECT TRACKING BY FUTURE PREDICTION USING GENERATIVE METHODS FOR TRANSPORTATION

Thèse dirigée par

Abdelmalik TALEB-AHMED	directeur
Jacques BOONAERT	co-encadrant
Sebastien AMBELLOUIS	co-encadrant

Composition du jury

Rapporteurs	Cathérine ACHARD	professeur à l'UPMC	
	Amir NAKIB	MCF à l'U-PEC(HDR)	
Examinateurs	Atika RIVENQ-MENHAJ	professeur à l'UPHF	président du jury
	Abdenour HADID	professeur à l'UPHF	
	Yassine RUICHEK	professeur à l'UTBM	
Invité	Charles TATKEU	directeur de recherche à l'UGE	
Directeurs de thèse	Abdelmalik TALEB-AHMED	professeur à l'UPHF	
	Jacques BOONAERT	MCF à l'IMT Lille Douai	
	Sebastien AMBELLOUIS	chargé de recherche à l'UGE	

L'IMT Lille Douai et l'Université Gustave Eiffel n'entendent donner aucune approbation ni improbation aux opinions émises dans les thèses : ces opinions devront être considérées comme propres à leurs auteurs.

Mots clés : détection d'anomalies, apprentissage profond, modèle génératif, application au transport

Keywords: anomaly detection, deep learning, generative model, transportation application

Cette thèse a été préparée au

CERI SN - IMT Lille Douai

⟨*dédicace*⟩

⟨*dédicace*⟩

⟨*épigraphe*⟩

⟨*épigraphe*⟩

ANOMALY DETECTION AND OBJECT TRACKING BY FUTURE PREDICTION USING GENERATIVE METHODS FOR TRANSPORTATION

Résumé

Actuellement, le traitement automatiquement des problèmes de transport devient un sujet actif. Dans le cadre de ce travail, nous visons à relever un défi spécifique dans ce domaine : la détection et le suivi des anomalies. Notre objectif est de construire un système flexible et efficace produisant des performances élevées sur diverses bases de données publiques. Le contexte de notre recherche est l'amélioration des approches précédentes pour obtenir de meilleurs résultats. Nous traitons deux scénarios conduisant à deux méthodes mentionnées dans les parties suivantes : (1) la segmentation et le suivi des véhicules et des piétons par des prédictions utilisant des méthodes génératives classiques basées sur des descripteurs a priori (hand crafted) et sur l'estimation des flux optiques ; (2) la détection des anomalies par des prédictions utilisant des systèmes génératifs multicanaux profonds et l'apprentissage supervisé.

Notre première recherche vise à l'évaluation des performances de l'approche générative classique pour les prévisions et la détermination des ses capacités à améliorer la segmentation et le suivi d'objets. Récemment, divers détecteurs d'apprentissage profond ont été proposés *e.g*. Mask R-CNN qui permettent une approche efficace du problème de suivi : le suivi par détection. A l'exception de tout autre information visuelle, ce type de tracker rapide ne prend en compte que l'intersection-sur-union (IOU) entre les boîtes de délimitation pour apparier les objets. Ainsi, l'absence d'informations visuelles du tracker IOU combinée avec les possibles défaillances des détecteurs créent des trajectoires fragmentées. Nous proposons alors un tracker amélioré basé sur la détection par suivi et sur l'estimation du flux optique. Notre solution génère de nouvelles détections ou segmentations basées sur une translation temporelle en avant et en arrière des résultats des détecteurs CNNs en utilisant les vecteurs de flot optique. Cette étape permet de combler une première partie des lacunes des trajectoires. Les résultats qualitatifs montrent alors que notre solution a obtenu des performances stables avec différentes méthodes d'estimation du flot optique. Les lacunes résiduelles au sein des trajectoires sont traitées en utilisant des caractéristiques SURF. La base de données DAVIS est utilisée pour évaluer la meilleure façon de générer de nouvelles détections. Enfin, le tracker résultant est testé sur la base de données DETRAC. Les résultats qualitatifs montrent que notre approche diminue très significativement la fragmentation des trajectoires. Pour les travaux futurs associés à ce tracker, nous prévoyons d'appliquer les réseaux CGAN développés dans le cadre de la seconde partie de notre travail afin de proposer un système compétitif de suivi d'objet basé prévision.

Malgré les résultats tangibles de cette première approche, les méthodes classiques présentent des limitations importantes concernant la détection d'anomalies qui est l'un de nos objectifs principaux. La fréquence plus faible des événements anormaux donne un scénario déséquilibré et leurs caractéristiques ne suivent généralement aucune relation spatiale ou temporelle. Face à ces défis, la plupart des méthodes de l'état-de-l'art se basent sur des réseaux prédictifs et utilisent les erreurs entre informations générées et réelles comme caractéristiques de détection. Inspirés par cette approche, d'une part, nous proposons un cadre multicanal flexible pour générer des caractéristiques multitypes au niveau image. D'autre part, nous étudions la possibilité d'améliorer les performances de détection par un apprentissage supervisé. Notre système est ainsi basé sur quatre GAN conditionnels (CGAN) prenant en entrée différents types d'informations d'apparence et de mouvement et produisant des informations de prédiction. Ces CGAN représentent la distinction entre événements normaux et anormaux. Ensuite, la différence entre les informations générées et les vérité-terrains est encodée par le pic du rapport signal / bruit (PSNR). Nous classons alors ces caractéristiques dans un contexte supervisé en construisant un petit ensemble d'entrainement à partir de quelques échantillons anormaux de l'ensemble de test original. C'est un Séparateur à Vaste Marge (SVM) qui est appliquée pour la détection des anomalies au niveau trame. Enfin, nous utilisons Mask R-CNN comme détecteur pour effectuer la localisation d'anomalies centrées objet. Notre solution est largement évaluée sur les bases de données Avenue, Ped1, Ped2 et ShanghaiTech. Nos résultats démontrent que les caractéristiques de PSNR combinées avec le SVM supervisé sont meilleures que les cartes d'erreurs calculées par les méthodes précédentes. En particulier, pour la base de données la plus difficile qu'est Shanghaitech, notre modèle surpasse jusqu'à 9% l'état-de-l'art des methodes non-supervisées. En perspective, nous prévoyons de construire une base de données pour la détection d'anomalies dans un cadre semi-supervisé, et d'intégrer un classifieur one-class SVM pour proposer un système "de bout en bout".

Mots clés : détection d'anomalies, apprentissage profond, modèle génératif, application au transport

CERI SN - IMT Lille Douai

Abstract

Today, automatic solving transportation problem becomes active subject. In our PhD project, we aim to address a specific challenge in this domain: anomaly detection and tracking. Our ultimate goal is constructing a flexible and effective framework producing high performance on various public datasets. The context of our research is applying and improving previous successful approaches to achieve better results. We deal with two scenarios leading to two methods mentioned in following parts: (1) vehicles and road users segmentation and tracking by future predictions using classical hand-crafted generative methods based on optical flow estimation; (2) anomaly detection by future predictions using multi-channels deep generative frameworks and supervised learning.

Our first research is evaluating the performance of the classical hand-crafted generative approach in future prediction and its capability for improving segmentation and tracking. Recently, there existed various strong deep learning detectors *e.g.* Mask R-CNN lead to an effective approach for tracking problem: tracking-by-detection. This very fast type of tracker considers only the Intersection-Over-Union (IOU) between bounding boxes to match objects without any other visual information. In contrast, the lack of visual information of IOU tracker combined with the failure detections of CNNs detectors create fragmented trajectories. We propose an enhanced tracker based on tracking by-detection and optical flow estimation in vehicle tracking scenarios. Our solution generates new detections or segmentations based on translating backward and forward results of CNNs detectors by optical flow vectors. This task can fill in the gaps of trajectories. The qualitative results show that our solution achieved stable performance with different types of flow estimation methods. Then we match generated results with fragmented trajectories by SURF features. DAVIS dataset is used for evaluating the best way to generate new detections. Finally, the entire process is tested on DETRAC dataset. The qualitative results show that our methods significantly improve the fragmented trajectories. For future work, we plan to apply CGANs streams of second work for the first task to propose a new competitive process of future prediction for segmentation and tracking.

Despite the moderate success of the first work, there is significant limitations of classical approaches to deal with our main task: anomaly detection. The lower frequency of abnormal events leads to an unbalanced scenario and the features of abnormal events usually do not follow any spatial or temporal relationship. It is also difficult to pre-define the structure or class of abnormal events. Facing to those challenge, most of state-of-the-art (SOTA) anomaly detection methods are based on apparent motion and appearance reconstruction networks and use error estimation between generated and real information as detection features. These approaches achieve promising results by only using normal samples for training steps. In this thesis, our contributions are two-fold. On the one hand, we propose a flexible multichannel framework to generate multi-type frame-level features. On the other hand, we study how it is possible to improve the detection performance by supervised learning. The multi-channel framework is based on four Conditional GANs (CGANs) taking various types of appearance and motion information as input and producing prediction information as output. These CGANs provide a better feature space to represent the distinction between normal and abnormal events. Then, the difference between those generative and ground-truth pieces of information is encoded by Peak Signal-to Noise Ratio (PSNR). We propose to classify those features in a classical supervised scenario by building a small training set with some abnormal samples of the original test set of the dataset. The binary Support Vector Machine (SVM) is applied for frame-level anomaly detection. Finally, we use Mask R-CNN as a detector to perform object-centric anomaly localization. Our solution is largely evaluated on Avenue, Ped1, Ped2 and ShanghaiTech datasets. Our experiment results demonstrate that PSNR features combined with supervised SVM are better than error maps computed by previous methods. We achieve SOTA performance for frame-level AUC on Avenue, Ped1 and ShanghaiTech. Especially, for the most challenging Shanghaitech dataset, a supervised training model outperforms up to 9% the SOTA on unsupervised strategy. Furthermore, we keep in progress several promising ways: building a new dataset for semi-supervised anomaly detection containing both normal and abnormal samples in its training set and applying one-class SVM to propose an end-to-end framework.

Keywords: anomaly detection, deep learning, generative model, transportation application

Acknowledgment

Firstly, I would like to express my sincere gratitude to my director Prof. Abdelmalik Taleb-Ahmed and my supervisors Dr. Jacques Boonaert and Dr. Sebastien Ambellouis for their enthusiastic and responsible guidance during my PhD research as well as my integration into life in France. I can not successfully finish my thesis without their continuous support.

Besides my advisors, I would like to thank the two referees of my thesis : Prof. Catherine Achard, Assoc.Prof. Amir Nakib for accepting to examine my thesis. Their constructive comments and interesting suggestions help me significantly improve the quality of this manuscripts. I would like to express my thankfulness to all the examiners : Prof. Yassine Ruicheck, Prof. Atika Rivenq-Menhaj, Prof. Abdenour Hadid, for their significant help to enhance my research from various perspectives. My thanks also go to Prof. Charles Tatkeu for participating in our defense as a special guest.

We would like to acknowledge the funding contribution of IMT Lille Douai and University Gustave Eiffel, especially their extra-funding to help us continue working on COVID-19 pandemic. Our work is a part of the ORIO ELSAT2020 project, which is co-financed by the European Union with the European Regional Development Fund, the French state, and the Hauts-de-France Region Council.

My sincere thanks also go to the staff of IMT Lille Douai and SPI Doctoral school for supporting me in various administrative processes during my work at IMT Lille Douai.

Finally, I would like to give my deep gratitude to my family for their love and support during the good but also the hard moments. My thankfulness also goes to my dear friends for sharing many unforgettable moments.

Publication

Journal :

Vu, T-H. ; Boonaert, J. ; Ambellouis, S., Taleb-Ahmed, A. **Multi-channel generative framework and supervised learning for anomaly detection in surveillance videos**. Submitted to *MDPI Sensors* (Accepted in May 2021).

International conferences :

Vu TH., Boonaert J., Ambellouis S., Ahmed A.T. **(2020) Vehicles Tracking by Combining Convolutional Neural Network Based Segmentation and Optical Flow Estimation**. In : Blanc-Talon J., Delmas P., Philips W., Popescu D., Scheunders P. (eds) *Advanced Concepts for Intelligent Vision Systems. ACIVS 2020.* Lecture Notes in Computer Science, vol 12002. Springer, Cham.

Vu, T. ; Ambellouis, S. ; Boonaert, J. and Taleb-Ahmed, A. **(2020). Anomaly Detection in Surveillance Videos by Future Appearance-motion Prediction**. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* - Volume 5 : VISAPP, ISBN 978-989-758-402-2 ISSN 2184-4321, pages 484-490.

List of abbreviations

CNN : Convolutional Neural Network
FCN : Fully Convolutional Network
AUC : Area Under Curve
EER : Equal Error Rate
SVM : Support Vector Machines
GAN : Generative Adversarial Network
CGAN : Conditional Generative Adversarial Network
IOU : Intersection Over Union
mIOU : mean Intersection Over Union
LDOF : Large Displacement Optical Flow
ROC : Receiver Operating Characteristic
SURF : Speed Up Robust Feature
PSNR : Peak Signal to Noise Ratio
HOG : Histogram Oriented Gradients
HOF : Histogram Optical Flow
SHT : ShanghaiTech
CVAE : ConVolutional Auto Encode
BB : Bounding boxes
TP : True Positive
TPR : True Positive Rate
TN : True Negative
TNR : True Negative Rate
FP : False Positive Rate
FPR : False Positive Rate
FN : False Negative
FNR : False Negative Rate
ACC : Accuracy
AP : Average Precision
mAP : mean Average Precision
MSE : Mean Square Error

Sommaire

Liste des tableaux

Table des figures

Chapitre 1

Introduction

This chapter introduces a general overview about our research project. First, we present the scientific context and scenario of our works. This part shows the definition of abnormality and how we consider the tasks of anomaly detection. Then we present the objective of this project in term of scientific researches and its application in real-world context. In order to achieve those goals, various approaches will be largely taken into account for both classical hand-crafted features methods and modern deep learning approaches. On the one hand, we analyse and compare them to show the advantage of recent deep learning models that inspired us to apply. On the other hand, we find the limitations of those state of the art models. Our solutions to improve those drawbacks will be introduced in the next part. Then we highlight our contributions along this project. Finally, we end this chapter by presenting the structure of our PhD thesis.

1.1 Context and Scenario

In this work, we study the problem of discovering and modelling the interactions between users and transport infrastructures by the analysis of video surveillance streams. This challenging work leads to the ability to recognise particular and abnormal activities in traffic. That can facilitate various applications targeted to improve the safety of vulnerable road users, the security of transport networks or supporting the autonomous car driving. We cast our works as a visual action recognition and understanding problem and we apply different solutions that have been proved being effective in many computer vision tasks. We plan to benefit from the promising results of deep learning in detection and segmentation tasks to detect, segment and track all the elements participating in traffic networks. Thus, we entirely extract all their information for the purpose of constructing a new model for unlearned behaviours or interactions.

Our works address the problem of analysing the information obtained from video surveillance streams. This is an interesting problem because behaviours modelling of individuals and their interactions thanks to the analysis of audio and video streams acquired from surveillance systems is a very active research topic in the scientific community. It is directly related to "Big data" problem which is a common aspect to many application areas such as monitoring the health of elderly people while enforcing their safety and security. Within the area of this research project, we focus on the transportation field. The objects we are searching for are all the ones and all the things that participate to the traffic network : pedestrians, bicyclists, motorbikes, trucks, traffic lights and signs, etc. We particularly concentrate on vulnerable road users (VRU) and their interactions with other actors inside the transportation network.

Actually, there are a lot of successful researches in transportation domain [113, 10, 29, 3, 6, 11, 112], especially for autonomous driving cars. But they are almost only addressing the first step of video information analysis for answering the question about detection and localisation of all objects inside a video signal. In this thesis, we want to continue the next step that is using the information extracted from the first step to model the interactions between them. In a naive way, it can be considered as a visual action recognition and understanding task. But in this case, the set of actions needed to be classified is very challenging due to the very occlusive scenarios attached to traffic networks. Additionally, the occurrence of abnormal and randomly performed actions also raises the difficulty of our context because those actions are totally unlearned by traditional methods.

In details, we particularly work with two scenarios : (1) vehicles and road users segmentation and tracking by future predictions using classical hand-crated generative methods based on optical flow estimation ; (2) anomaly detection by future predictions using multi-channels deep generative frameworks and supervised learning. Both research scenarios are evaluated on various public benchmarks in the off-line working mode. To evaluate the performances of our approaches,

we consider both qualitative and quantitative measurements. The experimental setup for these two scenarios as well as our proposed methods for each one is precisely presented in the next chapters.

1.2 Objective

In term of scientific objectives, we aim to address a specific challenge in transportation domain : anomaly detection and tracking. Our ultimate goal is constructing a flexible and effective framework producing high performances on various public datasets. As a context of this research project, we focus on applying and improving previous successful approaches to achieve better results rather than proposing brand-new methods or innovative network architectures.

For the first scenario, our purpose is to evaluate the capability of classical hand-crafted generative methods for improving segmentation and tracking tasks as well as its potential to deal with further challenging *e.g*. future prediction and anomaly detection. Naturally, hand-crafted methods without training a complex deep learning models are always the first priority for each system due to its simple pipeline. Supposing that classical hand-crafted methods can treat well almost all tasks, we benefit from the reduction of time consumption and, particularly, we also avoid the preparation of a large dataset for deep learning. In contrast, it is necessary to take the deep learning approaches into consideration when classical ones can not provide sufficient performances to us.

For the second scenario, we cast this challenging problem as our main task. Our purpose is to build a flexible and strong deep learning model for anomaly detection at frame-level. Then we develop a suitable pipeline for abnormality localisation at object-centric level. On the one hand, the most important request for our work is obtaining high performances. We expect our models to surpass state-of-the-art performance in various public datasets. On the other hand, the flexibility of our approach is based on the fact we would like to build a model adapted for various types of input information (*i.e*. RGB, grayscale, optical flow). As a consequence, it could be easily extended or lightened for further developments. We also demand our model to create distinct features for representing the characteristics of abnormal activities that can be comfortably fed into various classification models.

1.3 Definition of Anomaly Detection

As previously defined in [89] and retrieved in [80], we define an abnormal event as **"the occurrence of unusual appearance or motion attributes or the occurrence of usual appearance or motion attributes in unusual locations or times"**. It clearly appears that the description of an abnormal event can be different from one scene to another because of the environment context

or because of the intrinsic and extrinsic camera parameters (location, pose, focal length *etc*.). Moreover, different types of anomalous events may occur. An appearance based event can be illustrated by a car or a pedestrian moving in a forbidden area. Short or long-term motion event can be respectively illustrated by a pedestrian jumping over a barrier or by a bicycle zigzagging around pedestrians. For these reasons, we easily imagine that it will be difficult to compute a unique model with high performance for all the situations and for all the environments.

Abnormal events occur (fortunately) rarely in the real-life, and it is unrealistic to video caption all anomalous events we could encounter. Even if it is always possible to come across anomalous events in video surveillance, huge resources are required (in terms of time and human work load) to manually detect and annotate each one. The common approach used in most of the state-of-the-art methods is to model the normal activities because they are easy to collect from the video camera. Moreover, the annotation task is reduced to making sure that the "normal events" dataset is not corrupted by anomalous events. We follow this assumption, and we propose to analyse the gain of adding a supervised learning step regarding abnormal events, the final objective being to reach the trade-off between performance and required resources linked to dataset annotation.

1.4 Classical approaches

Generally, there are relatively large differences between classical approaches and modern ones in terms of considering the anomaly detection problem. Before the existing of specific datasets for anomaly detection mentioned in some works [58, 53, 66], most of the researches considered abnormal actions as a particular set of normal activities. All the public datasets [44, 92] during this period was constructed for the classical action recognition problems by combining all types of actions without highlighting some abnormalities. By this way, the abnormality was similarly considered as normality. To solve this classical problem in computer vision, we can apply various techniques for the discriminative task such as : image classification, video detection, segmentation and tracking, *etc*. Following classical hand-crafted methods [104, 105, 16, 1], first, activities are densely tracked to create smooth and continuous trajectories. Then, points of interest are detected around objects and activities. Various types of feature (SIFT [57], HOG [17], HOF [9], MBH [104], *etc*.) are extracted at the points of interest's locations along trajectories to help representing the distinction between activities. Finally, they construct suitable inference models to classify the corresponding features. Applying this technique to the transportation domain leads our problem to be simply treated as a detection and tracking task. Intuitively, if we can detect all the users and vehicles involved in a scenario, then we can smoothly and continuously track them frame by frame. In turn, we can extract features around those objects such as velocity, motion vector, etc to model the interactions between them. The first part of our research is for the purpose of doing this work.

Thanks to the recent successful CNNs model for object detection and segmentation [85, 107, 29, 83] we can almost retrieve the full spatial information about all users and objects involved in a traffic networks related scene captured by a camera. Those strong CNN models outperform all the previous traditional methods that use hand-crafted features in combination with learning representation. Among them, Mask R-CNN [29] is our best candidate because we can benefit from various types of information (object class, localisation and segmentation at instance level) using only one network. We observe that object tracking is an important proxy task towards action recognition. Most of the effective methods for normal daily action recognition [105, 90, 108] starts with a tracking step for obtaining objects' trajectories. Intuitively, the more information we get from objects, the more effective solution we have for the tracking step. Thanks to the strong performances of Mask R-CNN, we have a simple and effective approach for object tracking : tracking-by-detection. This type of tracker is an active topic in object tracking and achieves many successful works from early researches [8, 2, 32]. Tracking-by-detection method first applies an object detector to each video frame then associates these detections to tracks. One of the most representative and popular tracker of that kind is IOU Tracker [7]. It is the core tracker of winning solutions for multi-objects tracking challenge in AVSS 2018 [65].

1.4.1 Limitations

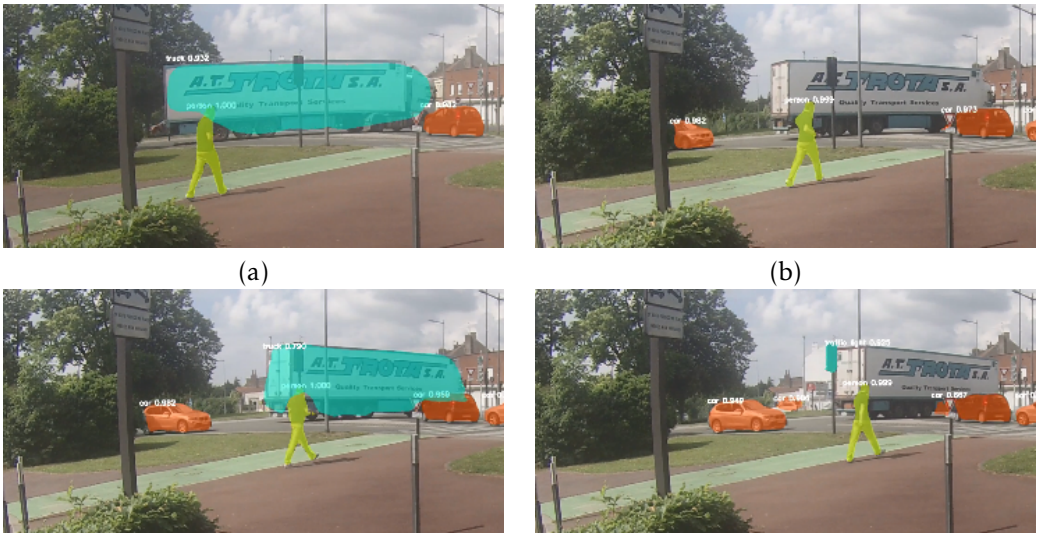
Instead of using visual information to match the objects locations, IOU Tracker takes into account the Intersection Over Union (IOU) between bounding boxes to associate them together. This feature first makes the performance of the tracker completely dependent from the performance of the detector itself. Second, the lack of visual information can lead to confusion between objects in overlapping cases.

$$IOU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{1.1}$$

Class	Person	Rider	Car	Truck	Bus	Train	Bicylce
AP	34.8	27.0	49.1	30.1	40.9	30.9	18.7

Tableau 1.1 – Performance of instance segmentation by Mask R-CNN on CityScapes dataset. Evaluation metric is the COCO-style mask AP (average precision on region level).All performances are state-of-the-art.

On the other hand, Mask R-CNN shows its drawback in the case of false negative error on usual objects (see Figure 1.1) although it yields strong performances on COCO dataset [50] and particularly on transportation related domain datasets such as the Cityscapes dataset [15] (see Table 1.1). These false negative results or missing detections from Mask R-CNN can lead to



(a) (b)

(c) (d)

FIGURE 1.1 – Common missing detection of Mask R-CNN in consecutive frames from (a) to (d) : (a) detection ; (b) non-detection ; (c) detection ; (d) non-detection

cases of failure in the next step of object's tracking because the object trajectories get "broken". Obviously, confusing detections between object's classes from Mask R-CNN can also produce wrong tracks, for instance when creating new tracking processes instead of maintaining existing ones.

1.4.2 Solution and contributions

Our work aims to build an improved tracker that can limit the drawback of fragmented trajectories by Mask R-CNN detector on IOU Tracker while keeping the advantage of being fast as a tracker. The first step is filling in the gaps of fragmented detections using a generative approach. The next step consists in combining the generated results with the current results from the detectors. Then, by applying the idea of IOU trackers, IOUs between the bounding boxes are used to eliminate the overlapping results. The final step is performed by associating the trajectories with SURF feature [5], which is a high-performance feature for image matching. Our contributions are the following :

— Introduction of a fast and efficient generative approach using optical flow for filling in the discontinuities of object tracking. Our solution is stable with different types of flow estimation.

— An enhanced tracker integrating Mask R-CNN based IOU tracker with visual information.

1.5 Abnormality centric approaches

Naturally, in human behavior analysis, we might consider anomaly detection as an action recognition problem. But this classical point of view leads us to an unbalanced situation where the number of samples for each class is significantly different. Beside, it is difficult to predefine the structure of abnormal events because there is usually not any spatial and temporal relations between those events. Hence, we should tackle this challenge in a specific way. Generally, from the first successful works until now, they proposed three solutions : one-class classification based [109, 106], changing detection based [22, 27, 38] and future prediction based [72, 53].

One-class learning first constructs the representation for events then fit a model to data for which annotations are available only for a single class. In anomaly detection, those are labels for abnormal samples. This solution is only appropriated for binary classification, and it has limitations when we need further information as type and localisation. Changing detection is a classical way where each event is compared with its neighbours to find the most different ones. By this way, we could get trouble when abnormal event always or never happens in a sequence. The future prediction based techniques casts abnormal events as unpredicted events. A generative model to produce future information from previous frames is computed and a model is trained from normal frames and noisy ones ; usually, noisy frames are more blurred than the ground truth.

1.5.1 Limitations and Solutions

Most of recently successful researches [27, 62, 63, 38, 53, 72, 36] have tackled this challenge in specific unsupervised ways. They only use normal samples from training set to generate the standards for normal actions then try to enlarge the deviation between abnormal samples in test set and theirs standards. Using unsupervised strategy naturally follows the structure of popular benchmarks datasets containing only normal samples that do not require annotation task.

The state of the art proposes two approaches for defining the anomalous feature that are respectively based on changing detection [22, 27, 38] and reconstruction/prediction errors [53, 72, 36]. The first solution is a natural approach where each event is compared with its neighbours to find the most different ones. The weakness of this solution appears when abnormal event always or never happens in a sequence. Besides, this solution is mainly appropriated for binary classification, and it has the limitations when we need further information as type and localisation. The second approach deals with these limitations and achieves state-of-the-art performance. Technically, they generate the future prediction or reconstruct current information for each action by GANs models and Convolutional Auto Encoder-decoder (CAE) models. Intuitively, the models trained by only normal samples from the training set will reconstruct better images

for normal frames than abnormal ones in test set. Inspired by the promising results following this approach and the benefits from the rise of CGANs and CAE models, we continue to extend the architecture of reconstructing models by integrating four pix-to-pix Image Translation CGAN models [39] as four parallel channels. This 4-channels framework processes both appearance and motion information in grayscale or color format.

In term of features extractions and descriptors encoding, previous anomaly detection frameworks proposed various reasoning approaches related to the features types : at object-centric or local level [66, 58, 62, 36], at frame level [53, 72] or at both [27]. On the one hand, some methods encode their features to calculate abnormality scores ; the decision is done thanks to a threshold or a peak estimation [53, 72]. In fact, by integrating all features into one score value, spatial and temporal information are discarded and a simple comparing decision model is learned. On the other hand, some keep their large scale features [38, 36] to train a classifier. If these solutions maintain the wealth of information of the features, the learning process is more complex and takes a long time to converge. To solve this problem, we propose to integrate the complete wealth of the features produced by each of the channels of our framework into a 10-dimensions vector by PSNR technique. The implementation details will be introduced in subsection 4.1.5. Our descriptors bring specific information of each channel but are light enough for fast learning models.

1.5.2 Contributions

In this work we first have modified the state-of-the-art unsupervised GAN-based abnormality detection approach by increasing the feature space. Second, we have added a supervised final step to improve the detection rate. In summary, our contributions are as follows :

— We introduce a flexible and powerful framework containing a multi-channel CGANs (4 streams with 9 output channels in our case) to generate multi-type future appearance and motion information. Our architecture considers more consecutive frames forward translation from t to $(t+2)$ than previous methods with only encode-decode reconstruction [36] or translation from t to $(t+1)$ [72, 53]. The number of channels can be freely inserted or removed for multiple purposes.

— We experimentally prove the effectiveness of PSNR in image comparison task. Based on PSNR method, a useful descriptor of output from our generative framework and ground-truth is proposed. The size of the descriptor is also flexible, as well as the number of channels and quite small to adapt to fast classifiers.

— We demonstrate the improvement we get on anomaly detection by adding a supervised stage exploiting the feature extractor proposed by our unsupervised architecture. We add a SVM stage that we train on a subset of abnormal samples from the test dataset. We achieve at least competitive performances on all benchmarks : Avenue, Ped1, Ped2 in term of frame-level AUC and a huge improvement on the most challenging datasets such

as Shanghaitech.

1.6 Manuscript Organization

Our thesis manuscript is organized by following a classical structure. It contains five chapters : Introduction, Related Works, Proposed Methods, Experiments and Conclusion.

After the first part introducing general information of our thesis, we present our results of searching for related work. We begin with the background of recent state-of-the-art deep learning models for all classical tasks of computer vision applying in transportation domain. The basis of the convolutional neural network following with various effective models for action recognitions, object detection, segmentations, optical flow estimation is introduced in both supervised and unsupervised contexts. Besides the presentation of existing discriminative models, we focus on the development of recent deep learning generative models to search for their capability of solving our problems. To improve segmentation and tracking framework, we mention two state-of-the-art models : Mask R-CNN and IOU object tracker as baseline methods. Then, we propose some useful add-on modifications to minimise its drawbacks. Then we continue with state-of-the-art solutions for our main task : anomaly detection. As explained before, we follow recent successful generative models for future prediction. We highlight their strengths and limitations. We end this chapter with a brief conclusion on the state-of-the-art models and the solutions that we apply.

The third chapter presents the essentials of our first work in detail. In the proposed methods section, we sequentially introduce two frameworks corresponding to two stages : (i) Generating object segmentation by Mask R-CNN and Optical flow estimation and (ii) improving IOU Tracker with generated information and SURF features. We begin with theoretical presentation of Mask R-CNN [29], Optical flow estimation by LDOF [9], Full Flow [13] and SURF features [5]. Then we introduce our methods to generate future detection and segmentation. We apply this classical hand-crafted generative information to improve detection and tracking. In the experiments section, we present our implementation of generating object segmentation and the quantitative results on DAVIS dataset we get using it. Then we show qualitative improvements of enhanced trackers with generated information and optical flow on UA-DETRAC dataset. We end this chapter with some analysis about strengths and limitations of our first work.

The fourth chapter introduces the second research : anomaly detection. This is our main task and our most important achievements are included. We begin this chapter with the basis of GAN [25], Conditional-GAN [39] and U-Net models [87] . Then, we precisely describe our multi-channels network architecture to learn abnormality centric features as well as supervised inference models by SVM binary classifier. We end the first section with some conclusion and analysis on strengths and limitations of our proposed methods. In the second section, we begin by

briefly introducing 4 of the public datasets and some evaluation metrics for anomaly detection. Then, we provide some tutorials of our network implementations in a step by step manner. We precisely show our results in terms of both quantitative and qualitative performances to illustrate the effectiveness of our approach for all of 4 datasets. We end this chapter with some discussion related to experimental strengths and weaknesses of our frameworks.

The last chapter provides the global conclusion of our thesis and gives tracks for our future work. We mention the significant contributions of our two frameworks and give some final evaluation regarding its advantages and limitations. We end this chapter as well as this thesis with some promising idea to continue improving this research in the future.

Chapitre 2

Related work

This chapter largely presents the successful state-of-the-art models related to our problems : (1) Improving segmentation and tracking with classical generative methods and (2) Anomaly detection with deep learning generative methods and supervised inference models.

We begin with the background of successful works in both classical and deep learning methods. We introduce the principle of Convolutional Neural Network (CNN) then all classical hand-crafted methods and modern deep learning models for popular computer vision tasks in transportation domain are mentioned in the next sections.

After that, we focus on the baseline methods that provide us promising solution to deal with our challenging problem : future prediction. There are two approaches we are dealing with : (i) classical hand-crafted future generating based on optical flow and (ii) deep learning multi-channels Conditional Generative Adversarial Network (C-GAN) for future prediction. We analyse their successful performances as well as their drawbacks to find suitable candidates and to propose some improvements to go beyond its limitations.

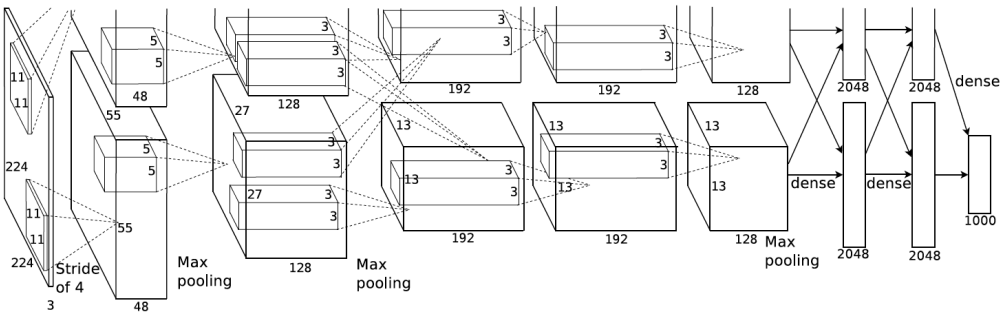


FIGURE 2.1 – AlexNet architecture [43] - The first successful CNN model with 7 layers. This architecture is a basic model for most of the recent innovative CNN architectures.

2.1 Background

2.1.1 Convolutional neural network

By now, video information analysis is an important challenge in computer vision and machine learning research. This challenging research domain has been largely driven by the advances in various topics : action recognition and localisation ; object detection, segmentation, and localisation, etc. The rise of Convolutional Neural Networks (CNNs) recently provides us lots of efficient approaches for this problem, not only for learning task but also for robustly extracting representation information for elements in videos. In comparison with traditional approaches without CNNs architectures, CNNs achieve outstanding performances in almost all the ways : they are faster, stronger, more flexible, while being adapted to massive data challenges.

The first successful CNN model is AlexNET [43] with 8 layers (Figure 2.1) proposed by Alex Krizhevsky *et al*. that won the ImageNet ILSVRC 2012 challenge. Then, CNNs architecture becomes deeper and stronger with VGG-Net [91] featuring 16 an 19 layers, GoogLeNet [95] providing 22 layers. The state-of-the-art ResNet [31] contains more than a hundred layers, and gives us a powerful model for image classification tasks.

To build AlexNet [43], Krizhevsky *et al*. trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, they achieved top-1 and top-5 error rates of 37.5% and 17.0% which was considerably better than the previous state-of-the-art results. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. To make training faster, they used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully

connected layers, they employed a regularisation method called "dropout", the was recently developed at that time, that proved to be very effective. They also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to the 26.2% achieved by the second-best entry.

In VGG-Net[91], Simonyan *et al*. investigated the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Their main contribution was a thorough evaluation of networks of increasing depth using an architecture with very small (3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of their ImageNet 2014 Challenge submission, where their team secured the first and the second places in the localisation and classification tasks respectively. They also shown that their representations generalises well to other datasets, where they achieve state-of-the-art results. They have made their two best-performing ConvNet models publicly available to facilitate further researches on the use of deep visual representations in computer vision.

Continuing to extend the depth of layers, Szegedy *et al*. proposed GoogLeNet [95]. They introduced a deep convolutional neural network architecture called Inception that achieved the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition 2014 Challenge (ILSVRC14). The main hallmark of this architecture is the improved use of the computing resources inside the network. By a carefully crafted design, they increased the depth and width of the network while keeping the computational budget constant. To optimise quality, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular incarnation used in their submission for ILSVRC14 is called GoogLeNet, a 22 layers deep network, the quality of which is assessed in the context of classification and detection.

Recently, He *et al*. introduced a very deep residual learning architecture called ResNet [31]. Because deeper neural networks are more difficult to train, they presented a residual learning framework to ease the training of networks that are substantially deeper than those used previously. They explicitly reformulated the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. They provided comprehensive empirical evidence showing that these residual networks are easier to optimise and can gain accuracy from considerably increased depth. On the ImageNet dataset they evaluated residual nets with a depth of up to 152 layers, about 8× deeper than VGG-Nets, while still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1^{st} place on the ILSVRC 2015 classification task. They also presented analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to their extremely deep representations, they obtained

a 28% relative improvement on the COCO object detection dataset. Deep residual nets are the foundations of their submissions to ILSVRC and COCO 2015 competitions, where they also won the 1^{st} places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

What is the reason of such very strong performances of CNNs models ? Zeiler and Fergus [114] in their work tried to visualise and understand convolutional networks activations (Figure 2.2). They studied the characteristics of the represented information in each layer. Investigating 7-layers CNNs models inspired by AlexNET, they use the information in each layer as features then use non-CNNs learning models such as Support Vector Machine (SVM) to classify as a normal visual recognition task. According to their results, the first layer relates to simple low-level characteristics such as simple edges or color. The second layer provides more complex low level information as corner or center-surround. The semantic information exists from the third layer. It roughly corresponds to various objects' parts. In the fourth and fifth layers, we can find selective units for entire objects or large parts of them. Generally speaking, the deeper the layer, the more semantic the information. Comparing with low-level hand-crafted features whose filters are fixed before the learning tasks, the CNNs features are more adaptive to the learning tasks because all filters parameters are updated at each iteration during the training task.

2.1.2 Action understanding

In terms of naive approach, we can consider our problem as a traditional visual action recognition task. A classical solution scheme begins with features extraction, continues with information aggregation to have spatio-temporal representation then finishes with model learning. From first success of Space Time Interest Point - STIP [46] to the state-of-the-art low-level dense trajectories descriptors method [104], they proposed various types of hand-crafted features : SIFT [57], HOG [17], HOF [9], MBH [104], MpegFlow [40], SURF [5], HOG3D [42], *etc*. Especially, by adding some techniques to improve dense trajectories by explicit camera motion estimation, detecting human to remove outliers matches for homography estimation, and stabilising optical flow to eliminate camera motion, Wang *et al*. [105] achieved state-of-the-art performances with a low-level descriptor on Hollywood2 [67], HMDB51 [44] and UCF101 [92].

Recently, there existed some successful models for action recognition using CNNs architecture (Figure 2.3). Simonyan *et al*. [90] proposed a two-streams CNNs models combining spatial stream as single image input and temporal stream as multi-frame optical flow input. Tran *et al*. [97] constructed a 3D convolutional network to learn spatio-temporal features. Wang *et al*. [108] continued to improve the dense trajectories method by trajectories pooled convolutional descriptors.

In the two-streams CNNs model [90], they investigated architectures of discriminatively

FIGURE 2.2 – Visualization of features in fully trained AlexNet model [114].

trained deep Convolutional Networks (ConvNets) for action recognition in video. The challenge is to capture the complementary information on appearance from still frames and motion between frames. They also aimed to generalise the best performing hand-crafted features within a data-driven learning framework. Their contribution was three-fold. First, they propose a two-stream ConvNet architecture that incorporates spatial and temporal networks. Second, they demonstrated that a ConvNet trained on multi-frame dense optical flow is able to achieve very good performance in spite of limited training data. Finally, they proved that multitasks learning, applied to two different action classification datasets, could be used to increase the amount of training data and improve the performance on both. Their architecture is trained and evaluated on the standard video actions benchmarks of UCF-101 and HMDB-51, where it is competitive with the state of the art. It also exceeds by a large margin previous attempts to use deep nets for video classification.

Similar to the idea of integrating CNNs features to action recognition, Tran *et al*. [97] learned spatiotemporal features with 3D Convolutional Networks. They proposed a simple, yet effective approach for the spatio-temporal features learning using deep 3-dimensional convolutional networks (3D ConvNets) trained on a large-scale supervised video dataset. Their findings were three-fold : first, 3D ConvNets were more suitable for the spatio-temporal features learning compared to 2D ConvNets ; second, a homogeneous architecture with small $3 \times 3 \times 3$ convolution kernels in all layers was among the best-performing architectures for 3D ConvNets ; and third, their learned features, namely C3D (Convolutional 3D), with a simple linear classifier outperformed state-of-the-art methods on 4 different benchmarks and are comparable with current best methods on the other 2 benchmarks. In addition, the features were compact : achieving 52.8% accuracy on UCF101 dataset with only 10 dimensions and also very efficient to compute due to the fast inference of ConvNets. Finally, they were conceptually very simple and easy to train and use.

To improve previous effective approaches of dense trajectories, Wang *et al*. [108] introduced Trajectory-Pooled Deep-Convolutional Descriptors. Because visual features are of vital importance for human action understanding in videos, they investigated a new video representation, called trajectory-pooled deep convolutional descriptor (TDD), which shares the merits of both hand-crafted features and deep-learned features. Specifically, they used deep architectures to learn discriminative convolutional feature maps and conducted trajectory-constrained pooling to aggregate these convolutional features into effective descriptors. To enhance the robustness of TDDs, they designed two normalisation methods to transform convolutional feature maps, namely spatiotemporal normalisation and channel normalisation. The advantages of their features came from : first, TDDs are automatically learned and contain high discriminative capacity compared with those hand-crafted features ; second, TDDs take into account the intrinsic characteristics of the temporal dimension and introduce the strategies of trajectories constrained

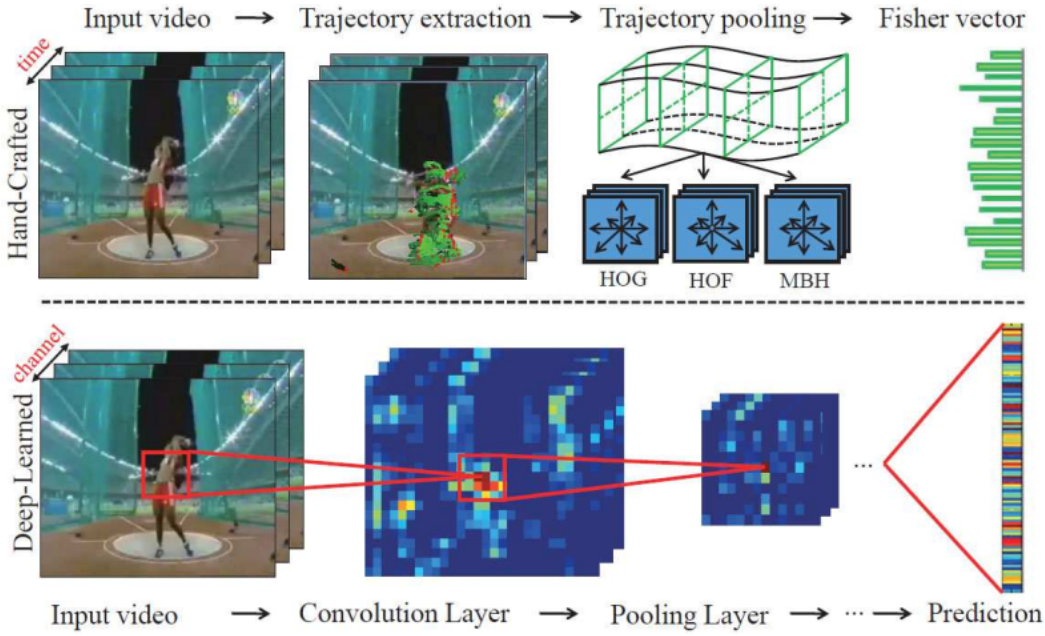


FIGURE 2.3 – Two types of features in action recognitions : Hand-crafted and Deep features [108].

sampling and pooling for aggregating deep learned features. They conduct experiments on two challenging datasets : HMDB51 and UCF101. Experimental results show that TDDs outperformed previous hand-crafted features and deep-learned features. Their method also achieved superior performance to the state of the art on these datasets.

Although all these methods achieved promising results, they focused only on common single activities with simple backgrounds and few participants. In contrast, our context is more challenging because many simultaneous actions appear in traffic networks with a disorderly scenario, and the activities involve multiple phases and multiple objects. A better solution is detection, segmentation and tracking of all the relevant elements in the video then using other specific techniques to model the interaction and to define the class of actions.

2.1.3 Object detection

Recent state-of-the-art object detectors are based on CNNs. R-CNN [24] casts the object detection task as a region-proposal classification problem based on features extracted from AlexNET pre-trained model and SVM classifiers. SPP-Net [30] provide a flexible input image size due to pooling features techniques. Fast R-CNN [23] extended previous works on R-CNN and SPP-Net to efficiently classify object proposals using an end-to-end architecture with several innovations to improve training and testing speed while also increasing detection accuracy. Fast

R-CNN trained the very deep VGG16 network 9x faster than R-CNN, is 213x faster at test-time, and achieves a higher mAP (that stands for mean Average Precision) on PASCAL VOC 2012. Faster R-CNN [85] extends this approach by generating bounding box proposals with a fully convolutional Region Proposal Network (RPN). RPN considers a set of densely sampled anchor boxes, that are scored and regressed. Moreover, it shares convolutional features with proposal classification and regression branches. These branches operate on fixed-size features obtained using a Region-of-Interest (RoI) pooling layer. In a similar spirit, YOLO [82] and SSD [51] also use a set of anchor boxes, which are directly classified and regressed without a RoI pooling layers. In YOLO, all scores and regressions are computed from the last convolutional feature maps, whereas SSD adapts the features to the size of the boxes. Features for predicting small-sized boxes come from early layers, and features for big boxes come from the latter layers, with larger receptive fields. All these object detectors rely on anchor boxes.

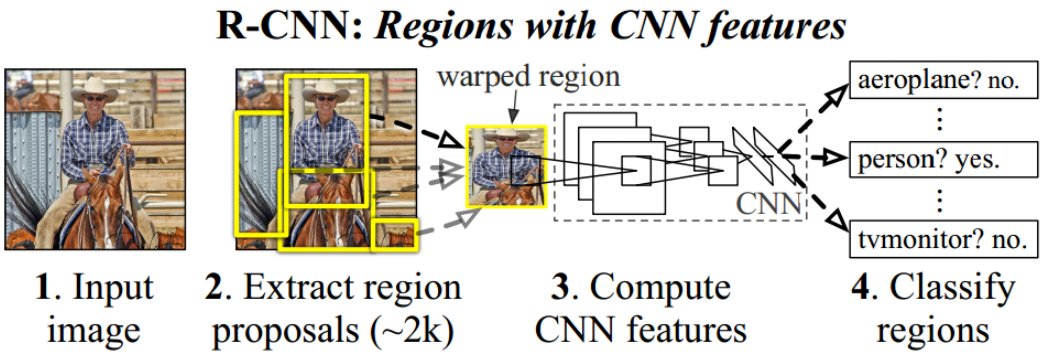


FIGURE 2.4 – The pipeline of R-CNN [24]. This is the first successful deep learning model for object detection.

In more details, the first successful model is R-CNN, proposed by Ross Girshick *et al*. The idea is quite natural by integrating CNNs features into interesting regions extracted by previous region proposal methods (Figure 2.4). Those CNNs features were easy to outperform all previous handcrafted features and show the promising possibilities of CNNs features when applied to the object detection task. Object detection performances, as measured on the canonical PASCAL VOC dataset, have plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In their works, they proposed a simple and scalable detection algorithm that improves the mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012 achieving a 53.3% mAP. Their approach combines two key insights : first, one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localise and segment objects and second, when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant

performance boost. Since they combined region proposals with CNNs, they call their method R-CNN : Regions with CNN features. They also compared R-CNN to OverFeat, a previous proposed sliding-window detector based on a similar CNN architecture. They found that R-CNN outperforms OverFeat by a large margin on the 200-class ILSVRC2013 detection dataset.

To reduce time consumption in R-CNN, He *et al*. proposed Spatial Pyramid Pooling architecture called SPP-Net [30]. Existing deep convolutional neural networks (CNNs) required a fixed-size (*e.g.*, 224×224) input image. This requirement was artificial and may reduce the recognition accuracy for the images or sub-images of an arbitrary size/scale. In this work, they equipped the networks with another pooling strategy, called "spatial pyramid pooling", to eliminate the above requirement. The new network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. The pyramid pooling is also robust to object deformations. With these advantages, SPP-net should in general improve all CNN-based image classification methods. On the ImageNet 2012 dataset, they demonstrated that SPP-net boosts the accuracy of a variety of CNN architectures, despite their different designs. On the Pascal VOC 2007 and Caltech101 datasets, SPP-net achieved state-of-the-art classification results using a single full-image representation and no fine-tuning. The power of SPP-net is also significant in object detection. Using SPP-net, they computed the feature maps from the entire image only once, and then pool features in arbitrary regions (sub-images) to generate fixed-length representations for training the detectors. Their method was up to a hundred times faster than the R-CNN method while achieving a better or comparable accuracy on Pascal VOC 2007. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, their method ranks 2 in object detection and 3 in image classification among all the 38 teams. This manuscript also introduces the improvement made for this competition.

Improving the depth and speed of R-CNN, Girshick proposed the Fast R-CNN architecture [23] with many significant contributions. This paper proposed a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection. Fast R-CNN was built on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employed several innovations to improve training and testing speed while also increasing detection accuracy. As we stated, Fast R-CNN trained the very deep VGG16 network 9× faster than R-CNN, is 213× faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3× faster, tests 10× faster, and is more accurate. The advantages of Faster R-CNN were : Higher detection quality (mAP) than R-CNN, SPPnet ; Training was single-stage, using a multi-task loss ; Training can update all network layers ; No disk storage was required for the feature caching.

A huge improvement towards Real-Time Object Detection was introduced by Ren *et al*. in Faster R-CNN [85] using Region Proposal Networks. In their work, they proposed a Region Pro-

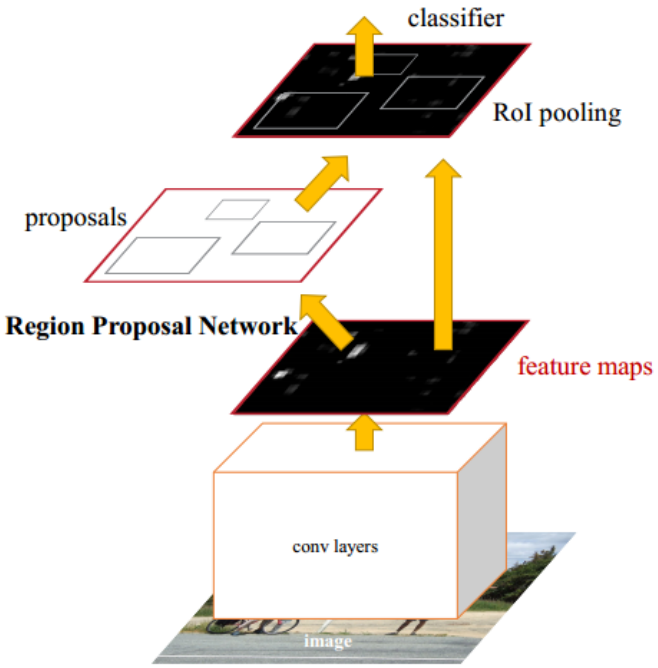


FIGURE 2.5 – Network architecture of Faster R-CNN [85]. RPN is a key module to construct an unified end-to-end model for object detection.

posal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals (Figure 2.5). An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. They further merged RPN and Fast R-CNN into a single network by sharing their convolutional features—using the recently popular terminology of neural networks with "attention" mechanisms, the RPN component tells the unified network where to look. For the very deep VGG-16 model, their detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image. In ILSVRC and COCO 2015 competitions, Faster R-CNN and RPN are the foundations of the 1^{st}-place winning entries in several tracks.

An alternative approach using anchor box and regression problem is YOLO [82]. Redmon *et al.* proposed YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, they framed object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation.

Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance. Their unified architecture was extremely fast. Their base YOLO model processes images in real-time at 45 frames per second. A smaller version of the network, Fast YOLO, processes an outstanding 155 frames per second rate while still achieving twice the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO made more localisation errors but is less likely to predict false positives in the background. Finally, YOLO learned very general representations of objects. It outperformed other detection methods, including DPM and R-CNN, when generalising natural images to other domains like artwork.

2.1.4 Object segmentation

Object detection is the tasks that lists the objects appearing in an image or a video sequence and that locates them. Intuitively, the first step of discovering the interactions between users and transports infrastructures is detecting and localising all users and vehicles existing in videos. Further steps, as modelling or classifying those interactions, need more information at pixel-level understanding. Assigning each pixel to an object or background category is the aim of object segmentation. Recent improved techniques of object segmentation proposed efficient approaches for semantic and instance segmentation of moving object, that is the priority input for learning model.

CNNs are driving advances in recognition, not only improving for whole-image classification [43, 91, 95, 31] but also making progress on local tasks with structured output. These include advances in bounding box object detection [24, 23, 85, 82, 51], part and keypoint prediction, and local correspondence. The natural next step in the progression from coarse to fine inference is to make a prediction at every pixel. FCNs [54] proposed by J.Long et al. in 2015 is the first end-to-end CNNs model achieving state-of-the-art performances without further machinery. Fully convolutional versions of existing networks predict dense outputs from arbitrary-sized inputs. Both learning and inference are performed whole-image-at-a-time by dense feedforward computation and backpropagation. In-network upsampling layers enable pixel-wise prediction and learning in nets with subsampled pooling.

Following the idea of making fully convolutional architectures, an improvement of the FCNs model was proposed by DeepLab [12]. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 was employed in a fully convolutional fashion, using atrous convolution to reduce the degree of the signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarged the feature maps to the original image resolution. A fully connected CRF was then applied to refine the segmentation result and better capture the object boundaries.

However, repeated subsampling operations like pooling or convolution striding in deep CNNs lead to a significant decrease in the initial image resolution. Lin *et al*. proposed RefineNet

[49], a generic multi-path refinement network that explicitly exploits all the information available along the down-sampling process to enable high-resolution prediction using long-range residual connections. In this way, the deeper layers that capture high-level semantic features could be directly refined using fine-grained features from earlier convolutions. The individual components of RefineNet employed residual connections following the identity mapping mindset, which allows for effective end-to-end training. Further, they introduced chained residual pooling, which captures rich background context in an efficient manner.

In terms of instance object segmentation, Mask R-CNN [29] achieved state-of-the-art performance in a general context. This method efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. It is extended from Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition (Figure 2.6). The technique for providing object masks can be considered as semantic segmentation that shared the same spirit with FCNs [54] inside a bounding box with binary classes : object or background.

Despite significant progressing steps in object segmentation in a static image, the problem of determining whether an object is in motion, irrespective of camera motion, is far from being solved. MP-Net [96] proposed by Tomakov *et al*. in 2017 addressed this challenging task by learning motion patterns in videos. The core of their approach is a fully convolutional network, which was learned entirely from synthetic video sequences, and their ground-truth optical flow and motion segmentation. This encoder-decoder style architecture first learned a coarse representation of the optical flow field features and then refined it iteratively to produce motion labels at the original high resolution. They further improve this labeling with an objectness map and a conditional random field, to account for errors in optical flow, and also to focus on moving "things" rather than "stuff". The output label of each pixel denotes whether it has undergone independent motion, *i.e*. irrespective of camera motion.

2.1.5 Optical flow

Optical flow is the apparent motion of brightness patterns in the image. Because the motion field is the projection of the 3D scene motion into the image, the optical flow would be ideally the same as the motion field. As a promising result of MP-Net, we consider using an optical flow map as an adding input channel for CNNs model in order to enhance information.

The first efficient way to solve this equation with two unknowns was proposed by B.Lucas and T.Kanade in 1981 [60]. Another algorithm was proposed by Horn and Schuck in 1981 [34] by adding smoothness constraint. In 2011, Brox and Malik proposed Large Displacement Optical Flow (LDOF) [9] estimation method by adding into the energy function a matching term that penalises the difference between flows and HOG matches. MDP-Flow 2 [64] proposed by Xu *et*

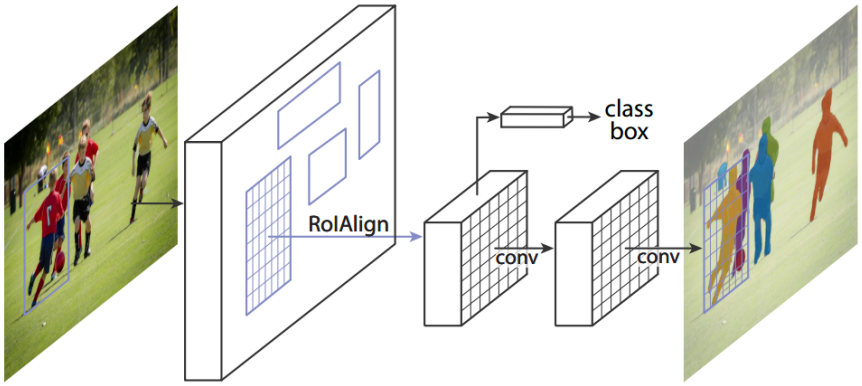


Figure 2.6 – The Mask R-CNN framework for instance segmentation [29].

al. in 2012 continued to extend this approach with the expensive fusion of matches (SIFT + PatchMatch) and estimation flow at each level. Inspired by the rise of CNNs models, many researchers proposed theirs approaches linked with CNNs architectures. In the case of DeepFlow [110] proposed by Weinzaepfel et al. in 2013, they combined deep matching and flow refinement with a variational approach. Feature information is aggregated from fine to coarse using sparse convolutions and max-pooling. However, it did not perform any learning and all parameters were set manually. The successive work of [86] termed EpicFlow had put even more emphasis on the quality of sparse matching as the matches from DeepFlow were merely interpolated to dense flow fields while respecting image boundaries.

Recently, FlowNet [19] only used a variational approach for the optional refinement of the flow field predicted by the convolutional net and did not require any handcrafted methods for aggregation matching and interpolation. P.Fischer *et al*. constructed appropriate CNNs which were capable of solving the optical flow estimation problem as a supervised learning task. They proposed and compare two architectures : a generic architecture and another one including a layer that correlates feature vectors at different image locations (Figure 2.7). However, FlowNet did not outperform traditional methods. The state of the art with regard to the quality of the flow had still been defined by traditional methods. Particularly on small displacements and real-world data, FlowNet cannot compete with variational methods. FlowNet 2.0 [35] proposed by Eddy Igg et al. in 2017 advanced the concept of end-to-end learning of optical flow and made it work really well. The large improvements in quality and speed are caused by three major contributions : first, they focused on the training data and show that the schedule of presenting data during training was very important. Second, they developed a stacked architecture that includes the warping of the second image with intermediate optical flow. Third, they elaborated on small displacements by introducing a subnetwork specializing in small motions. The lighter

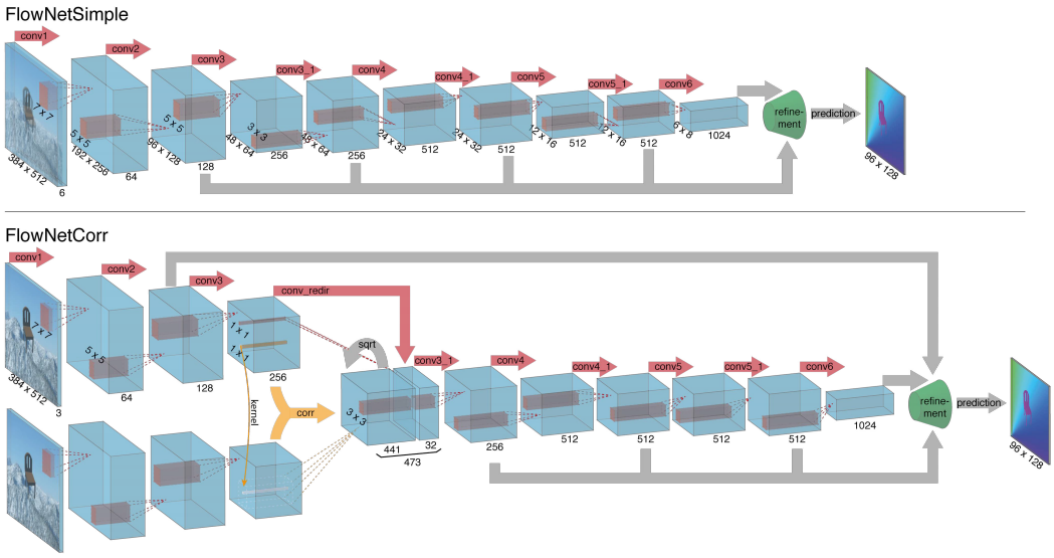


FIGURE 2.7 – Network architecture of Flow-Net [19]. This is the first successful deep learning model for optical flow estimation.

version of FlowNet 2.0 is PwCNet [94] with an encode-decode architecture.

Usually, optical flow is a basement for action recognition. On the other hand, Luc*et al.* [59] introduced a new utilisation of optical flow for predicting future segmentation. This first work suggests to us a promising way to exploit optical flow for generating the missing video information.

2.1.6 Supervised context Summary

Method	Description	Performance
AlexNET 2012 [43]	Image classification, single image, end-to-end architecture, 7 layer	17% top-5 err on ILSVRC12; 15.3% top-5 err on ILSVRC10
VGG 2014 [91]	Image Classification, single image, end-to-end architecture, 16 or 19 layers	8.43% top-5 err on ILSVRC14
GoogLeNet 2014 [95]	Image Classification, single image, end-to-end architecture, inception module, 22 layer	7.89% top-5 err on ILSVRC14

ResNet 2015 [31]	Image Classification, single image, end-to-end architecture, residual learning function, more than a hundred layers	4.49% top-5 err on ILSVRC14, 3.57% top-5 err on ILSVRC15
Dense trajectories 2012 (DT) [104]	Action Recognition, video level	83.5% ACC on UCF101, 58.2% on Hollywood2, 46.6% on HMDB51
Improved DT 2013 (IDT) [105]	Action Recognition, video level	85.9% ACC on UCF101, 66.8% on Hollywood2, 60.1% on HMDB51
2-streams CNNs 2014 [90]	Action Recognition, video level	88.0% ACC on UCF101, 59.4% on HMDB51
Convolutional 3D 2014 [97]	Action Recognition, video level	85.2% ACC on UCF101
TDD and IDT 2015 [108]	Action Recognition, video level	91.5% ACC on UCF101, 65.9% on HMDB51
R-CNN 2013 [24]	Object Detection, AlexNet architecture, multi stages pipeline	62.4% mAP PASCAL VOC12
Fast R-CNN 2015[23]	Object Detection, end-to-end, VGG-16 architecture	68.4% mAP PASCAL VOC12, test time 0.5 fps on GPU Titan X
Faster R-CNN 2015 [85]	Object Detection, end-to-end, ResNet architecture	73.8% mAP PASCAL VOC12, test time 5 fps on GPU Titan X
YOLOv2 2016 [83]	Object Detection, end-to-end	73.4% mAP PASCAL VOC12, test time 40 fps on GPU Titan X
SSD 2015 [51]	Object Detection, end-to-end	74.9% mAP VOC12, test time 19 fps on GPU Titan X
FCN 2014 [54]	Image Segmentation, sematic level	65.3% mIOU Cityscapes
SegNet 2015 [3]	Image Segmentation, semantic level	57.0% mIOU Cityscapes
DeepLab 2016 [12]	Image Segmentation, semantic level	70.4% mIOU Cityscapes

RefineNet 2016 [49]	Image Segmentation, semantic level	73.6% mIOU Cityscapes
MP-Net 2017 [96]	Moving Object Segmentation, only segment moving object, first successful	69.7% mIOU DAVIS
Mask R-CNN 2017 [29]	Image Segmentation, Instance level	58.1% AP_{50} Cityscapes
LDOF 2011 [9]	Optical Flow estimation	18.19 AEE KITTI15
DeepFlow 2013 [110]	Optical Flow estimation	10.63 AEE KITTI15
EpicFlow 2015 [86]	Optical Flow estimation	9.27 AEE KITTI15
FlowNet2 2016 [35]	Optical Flow estimation	8.94 AEE KITTI15

Tableau 2.1 – Supervised context summary

Top-5 error rate is the fraction of test images for which the correct label is not among the five labels considered most probable by the mode.

Accuracy (ACC) is the fraction between true prediction (both true positive and true negative) and total prediction.

For searching or detection problems, precision is the fraction of retrieved elements that are relevant to the searching query. Recall is the fraction of the elements that are relevant to the searching problem that is successfully retrieved. By computing precision and recall at every position, one can plot a precision-recall curve, plotting precision $p(r)$ as a function of recall r. Average precision (AP) computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$. Mean average precision (mAP) for a set of searching elements is the mean of the average precision scores for each searching element.

For image segmentation, Intersection over Union (IOU) score for each class is the fraction between true positive pixels and the sum of true positive, false negative and false positive pixels. The mean IoU (mIOU) is the mean of IOU for all classes. Especially, for instance-level segmentation, performance on this task is measured by the COCO-style mask AP (average precision on region level) and AP_{50} (average precision when overlap at region level is at least 50%).

For optical flow estimation, Endpoint Error (EE) is defined as the scalar length of different vector $\|V_{est} - V_{gt}\|$ between estimated optical flow vector V_{est} and ground- truth optical flow vector V_{gt}. Average Endpoint Error (AEE) is average of EE for all optical flow vector.

2.1.7 Unsupervised context

Despite the promising results in supervised context for action recognition and localization, the researches in unsupervised or weakly-supervised context are still challenging. Oneata *et al*. [74] extend 2D object proposal method to produce spatio-temporal proposals by mean of a randomized supervoxel merging process. Kwak *et al*. [45] tackle the discovery and localization problem using a part-based region matching approach [14] then extended to spatio-temporal tubes by tracking. Both methods used hand-crafted features with several complementary process so take long time compared with CNNs approaches. Besides, the context was simple with a single object or co-localization and the action was only localized by tracking.

Recently, we can benefit from a promising CNNs architecture adapted for unsupervised context : Generative Adversarial Networks (GANs) [25]. GAN is an unsupervised generative model, that learns to generate the true data distribution by implicit density estimation. GAN is composed of a generator and a discriminator. In training, the generator is trained in a way the discriminator cannot distinguish fake images produced by the generator from the real ones. Meanwhile, the discriminator learns to distinguish fake images from real images. Through this adversarial competition, the generated images from GAN become harder to distinguish from the reals. The successful results of GANs leads to various application in generative model : learning unsupervised representation [79], image translation [39] and synthesis [84], video gererating [100]. In spite of promising achievements in generative tasks, GANs have not been adapted to discriminative task such as the one of our context yet.

2.2 Future prediction

2.2.1 State of the art

Future video information prediction recently has became an active topic due to significant progresses in deep learning, especially in generative adversarial networks (GANs) and Convolutional Auto-Encode (Conv-AE) models. They predicted various types of future information for specific applications. [68] trained a classical 7-layers CNN to generate future frames given an input sequence. To deal with the inherently blurred predictions obtained from the standard Mean Squared Error (MSE) loss function, they proposed three different and complementary features learning strategies : a multi-scale architecture, an adversarial training method, and an image gradient difference loss function. [103] built a 7-layers CNN for predicting the future motion of each and every pixel in the image in terms of optical flow given a static image. [20] developed a Long-Short Term Memory (LTSM) based action-conditioned video prediction model that explicitly models pixel motion to learn about physical object motion without labels, by predicting a distribution over pixel motion from previous frames. Inspired by the same idea, [55] constructed the LTSM based PredNet network which learned to predict future frames in a video

sequence, with each layer in the network making local predictions and only forwarding deviations from those predictions to subsequent network layers. [98] built a deep neural network for the prediction of future frames in natural video sequences upon the Conv-AE and Convolutional LSTM for pixel-level prediction, which independently capture the spatial layout of an image and the corresponding temporal dynamics. In [73], the authors introduced an architecture based on recurrent Conv-AEs to deal with the network capacity and error propagation problems for future video prediction. It consisted of a series of bijective Gate Recurrent Unit (GRU) layers, which allowed for a bidirectional flow of information between input and output : they considered the input as a recurrent state and update it using an extra set of gates. [21] proposed an approach using Conv-AE that hallucinated the unobserved future motion implied by a single snapshot to help static-image action recognition. The key idea was to learn prior over short-term dynamics from thousands of unlabeled videos, infer the anticipated optical flow on novel static images, and then train discriminative models that exploit both streams of information. Obviously, most recent researchers build their model upon a Conv-AE model to reconstruct the future informations.

Generative modelling of future RGB video frames has recently been studied using a variety of techniques : prediction of future human pose [99], generative adversarial training model [68], forecasting convolutional features [59]. All their works focus on predicting future information which they never can not achieve. This context is a bit different than our work where we address the problem of generating only the missing information the detector failed to produce. We can freely generate backward and forward segmentation or bounding boxes based on some current results from the detectors. Despite the difference between those contexts, we apply the simple baseline approach from Luc *et al*. [59] that translated a segment on the basis of flow vectors.

2.2.2 Generative Adversarial Network

Generative Adversarial Network (GAN) proposed by Ian Goodfellow *et al*. [25] is a framework estimating generative models via an adversarial process. By now, GANs are built as deep neural network models and they are used to generate synthetic images. Generally, the architecture comprises two deep neural networks, a generator and a discriminator, which work against each other (thus, "adversarial"). The generator generates new data instances, while the discriminator evaluates the data for authenticity and decides whether each instance of data is "real" from the training dataset, or "fake" from the generator (Figure 2.8). GANs models focus to solve the difficulty of approximating many intractable probabilistic computations that arise in maximum likelihood estimation and related strategies, and difficulty of leveraging the benefits of piecewise linear units in the generative context.

Theoretically, let G denotes a generative model that captures the data distribution, and D denotes a discriminative model that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability

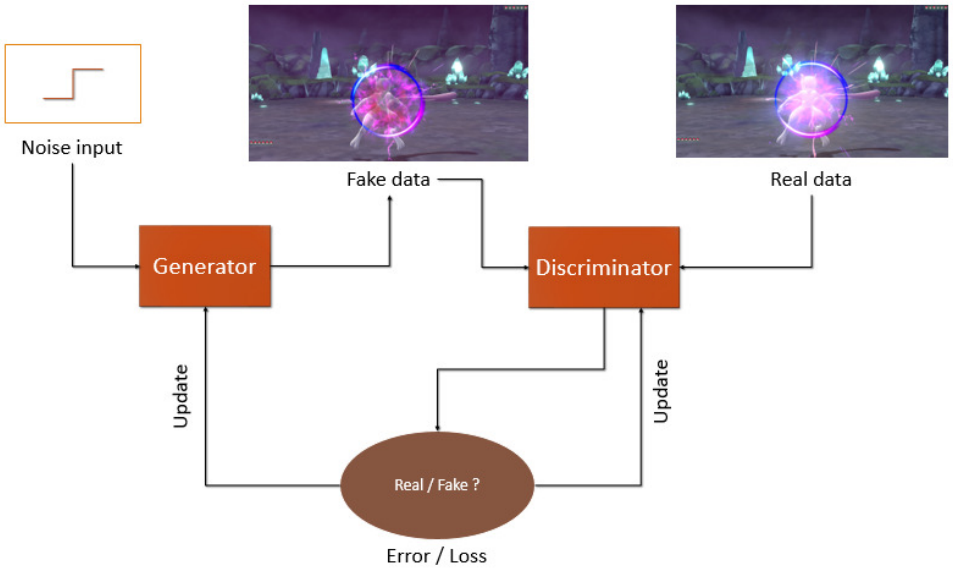


FIGURE 2.8 – GAN principle : The generator generates new data instances, while the discriminator evaluates the data for authenticity. The loss is applied to updated both generator and discriminator.

of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and the probability of D making a mistake being $\frac{1}{2}$ everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. In the proposed adversarial nets framework, the generative model is pitted against an adversary : a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution. The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are inseparable from the genuine articles.

2.3 Improving Detection and Tracking

2.3.1 Mask R-CNN

Mask R-CNN [29] is a conceptually simple, flexible, and general framework for objects instances segmentation. In principle, Mask R-CNN is an extension of Faster R-CNN [85] that constructs a third branch as FCN [54] for segmentation (Figure 2.9). Therefore, we find Mask R-CNN is an effective combination of elements from the classical computer vision tasks for

object detection and semantic segmentation. Mask R-CNN surpasses all previous state-of-the-art single-model results on the COCO dataset for both instance segmentation and bounding box detection tasks. Though, it has a limitation in implementation for other datasets. In particular, when applied to video sequences the detection can get unstable when the object is rapidly changing its appearance. Such events frequently occur in transportation based video sequences, for instance when a vehicle is turning right or left or when its apparent size increase or decrease owing to its relative move with respect to the camera.

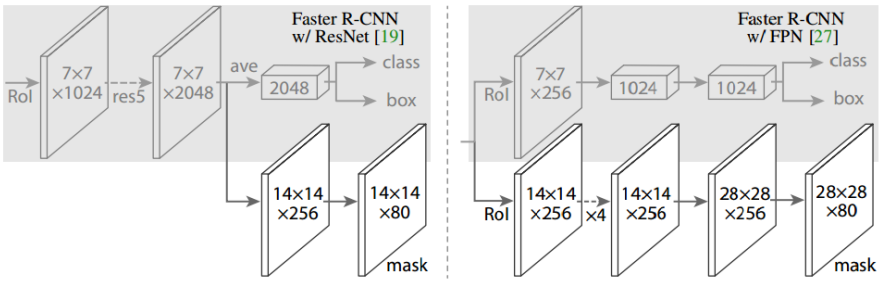


FIGURE 2.9 – Network architecture of Mask R-CNN based on very deep ResNet [29].

2.3.2 IOU object tracker

Before the rise of CNNs models, almost all successful trackers performed object location based on hand-crafted features : IHTLS [18], H2T [111], CMOT [4], etc. Because of the use of (so called) too simple concepts, these trackers show their limited performances when facing difficult scenarios from recent multi-object tracking challenges [65]. Recently, IOU Tracker [7] based on the strong performances of CNN based detector surpassed all previous methods in both easy and difficult scenarios. Bochinski *et al*. considered only the overlaps between bounding boxes obtained from detector to associate them (Figure 2.10). It benefits from the fast and strong performance of CNNs-based detector without any other visual information. As a consequence, this can be seen as a bit risky process due to the complete dependence of the tracking task to the accuracy of the detector : as discussed above, both false positive and negative error of the detector can cause fragmented trajectories.

2.4 Anomaly Detection

2.4.1 Evaluation metrics for Anomaly Detection

Generally, one usually uses Equal Error Rate (EER) and Area Under Curve (AUC) to evaluate quantitative performance of an anomaly detection model.

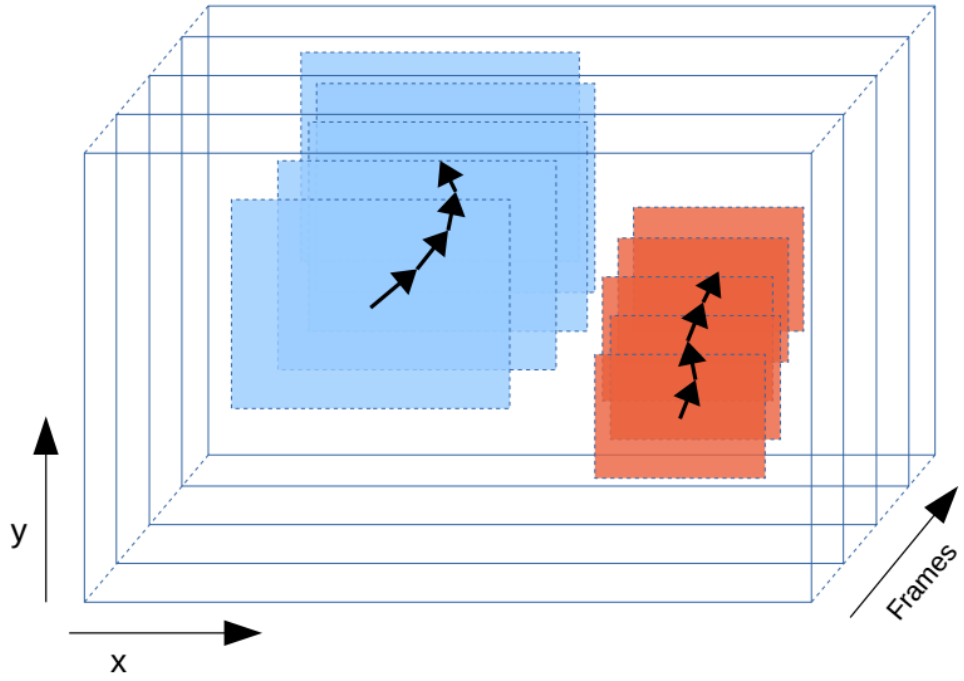


FIGURE 2.10 – IOU Tracker principle : Associating detections by their spatial overlap between time steps based on high accuracy detections at high frame rates [7].

Equal error rate (EER) is a popular measurement metric in biometric security system that used to predetermine the threshold values for its false acceptance rate (FAR) and its false rejection rate (FRR). When the rates are equal, the common value is called the equal error rate. The value represents the point where the proportion of false acceptances is equal to the proportion of false rejections. The lower the equal error rate value, the higher the accuracy of the system.

Area Under Curve (AUC) of Receiving Operation Characteristic (ROC) curve is more popular than EER. ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on the y-axis and FPR is on the x-axis. AUC is the area under this curve. Higher the AUC, the better the model is. AUC is precisely presented in section 4.1.6.

From frame-level to object-centric level, we can also apply the AUC and EER as potential evaluation metric by adding some bounding box level conditions. Following the definition of [101, 58], a candidate frame will be considered as abnormal frame if the intersection between its detected boxes and the ground-truth boxes is greater than 40%. To have a fair comparison, we apply this evaluation metric. In contrast, we find that the bounding box AUC and EER is not enough good for evaluating object-centric level performance. It is still similar to frame-level rather than purely bounding box level. Hence, we also define ourself some alternative evaluation

metrics that more adaptive to object-centric level : $IOU_{50}Rate$, $IOU_{75}Rate$ and $mIOU$. Those metric are presented in detail in section 4.1.6.

2.4.2 Early works with hand-crafted features

Before Convolutional Neural Networks (CNNs) became popular, most of the early methods were based on the extraction of hand-crafted features to estimate the models of normal and abnormal events. Motion trajectories were used as the principal features [69, 115] because of their fast extraction and simple implementation. However, the single motion information was not sufficient to represent all the spectrum of abnormal events and the motion estimator was easily confused in crowed and complex scenes.

To improve these limitations, both appearance and motion were extracted along the trajectories. Kim *et al*. [41] used Histogram of optical flow to build space-time Markov Random Fields (MRF) graph for detecting abnormal activities in video. The nodes in the MRF graph corresponded to a grid of local regions in the video frames, and neighbouring nodes in both space and time were associated with links. They captured the distribution of typical optical flow with a mixture of probabilistic principal component analyzers to learn normal patterns of activity at each local node. For any new optical flow patterns detected in incoming video clips, they applied the learned model and MRF graph to compute a maximum a posteriori estimate of the degree of normality at each local node. Further, they proposed the incremental update of the current model's parameters as new video observations stream in, so that their model could efficiently adapt to visual context changes over a long period of time. Qualitative performances illustrated that their space-time MRF model robustly detected abnormal activities both in a local and global sense.

Mahadevan *et al*. [66] learned the Mixture of Dynamic Textures (MDT) during training then computed negative log-likelihood of the spatio-temporal patch at each region at test phase (Figure 2.11). They proposed three properties for designing their models : (1) joint modelling of appearance and dynamics of the scene, and the abilities to detect (2) temporal and (3) spatial abnormalities. The model for normal crowd behaviour was based on mixtures of dynamic textures, and outliers under this model were labelled as anomalies. Temporal anomalies were equated to events of low-probability, while spatial anomalies are handled using discriminant saliency. To evaluate their models, they presented new datasets : USCD Pedestrians Ped1 & Ped2. They achieved state-of-the-art performances on their dataset at the moment of publication.

Developing the capability of motion representations, Histograms of optical flow orientation (HOFO) was extracted by Wang *et al*. [109] to classify abnormal events by one-class SVM or kernel PCA. The details of the histogram of the optical flow orientation descriptor were illustrated for describing movement information of the global video frame or foreground frame.

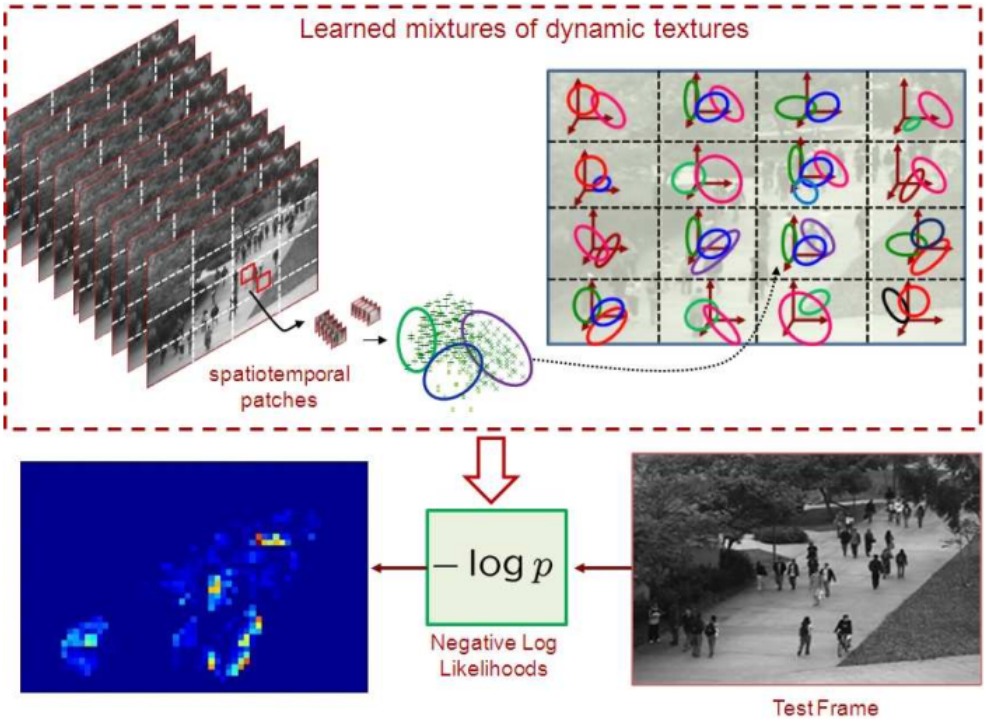


Figure 2.11 – Learning MDTs for temporal abnormality detection by Mahadevan *et al.* [66].

By combining one-class SVM and kernel PCA methods, the abnormal events in the current frame can be detected after a learning period characterizing normal behaviours. They achieve impressive performance $AUC = 0.99$on several parts of UMN dataset [76].

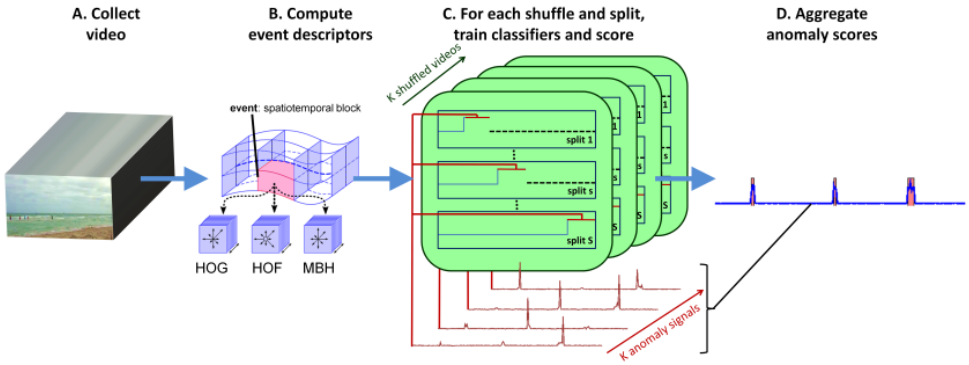


Figure 2.12 – Discriminative framework combining HOG, HOF, MBH for anomaly detection proposed by Giorno *et al.* [22].

More recently, Giorno *et al.* [22] built a combination of HOG, HOF, MBH to train their classi-

fiers then took the average classification scores to draw the output signal (Figure 2.12). They worked on a specific unsupervised scenario where training sequences were unavailable and anomalies were scored independently of temporal ordering. By defining anomalies as examples that can be distinguished from other examples in the same video, their definition inspired a shift in approaches from classical density estimation to simple discriminative learning. They also achieved state-of-the-art performance $AUC = 0.91$ on Avenue dataset [58] at the moment of their publication.

Generally, most of those methods achieved good performance on some simple datasets without changing the camera orientation, illumination or dealing with complex activities. For more challenging datasets, they just yielded moderate performance due to the limitation of hand-crafted features in case of large datasets and complex scenarios.

2.4.3 Recent successful models with Discriminative Deep learning model

Recently, the existing of powerful deep learning models leads to many successful approaches in anomaly detection. Hinami *et al.* [33] solved the problem of environment-dependent nature by integrating a generic Fast R-CNN model and environment-dependent anomaly detectors. They learned CNN with multiple visual tasks to exploit semantic information that is useful for detecting and recounting abnormal events and then appropriately plugged the model into anomaly detectors (Figure 2.13). They achieved $AUC = 0.898$ on Avenue and $AUC = 0.922$ on Ped2 dataset.

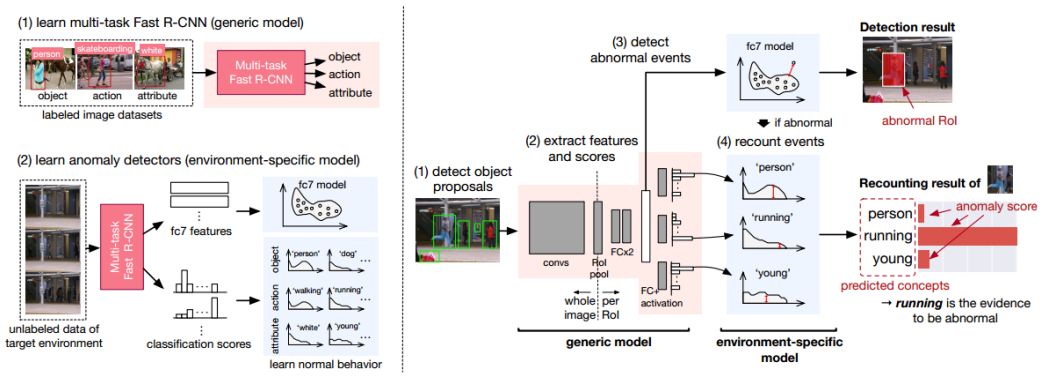


FIGURE 2.13 – Environment-dependent anomaly detectors proposed by Hinami *et al.* [33]. The left part illustrates learning procedures of two types of models : generic and environment-specific models, and right part shows testing procedure of joint detection and recounting abnormal events.

Ionescu *et al.* [38] introduced a framework without requirements of training data by applying unmasking techniques. They combined the motion features computed from 3D gradients at each

spatio-temporal cube with *conv5* layer of VGG-net with fine-tuning as appearance features. Then a binary classifier was trained to distinguish between two consecutive video sequences while removing at each step the most discriminant features. The higher training accuracy rates of the intermediately obtained classifiers represented abnormal events. They achieved $AUC = 0.806$ on Avenue dataset.

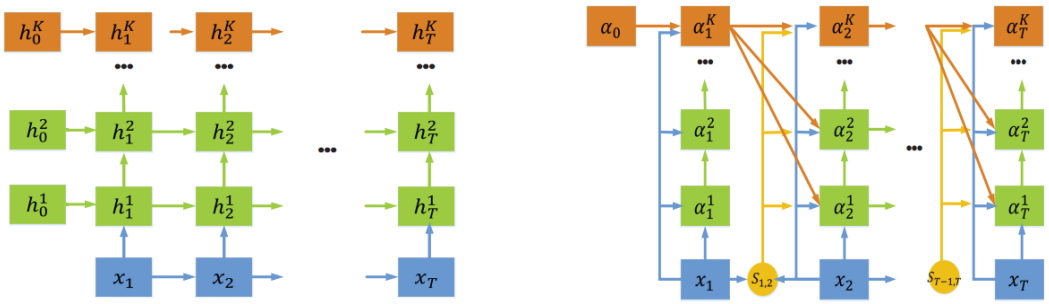


FIGURE 2.14 – Stack RCNN framework for anomaly detection proposed by Luo *et al.* [63]. The blue boxes represent the input of stacked RNNs. The green and orange boxes represent coding vectors. The yellow circles are similarities between neighbouring frames.

During this period, Luo *et al.* also proposed another model in [63] in which they mapped a Temporally-coherent Sparse Coding where they enforced similar neighbouring frames being encoded with similar reconstruction coefficients with a special type of stacked Recurrent Neural Network (sRNN). By taking advantage of sRNN in learning all parameters simultaneously, the nontrivial hyper-parameter selection to TSC could be avoided, meanwhile with a shallow sRNN, the reconstruction coefficients could be inferred within a forward pass, which reduced the computational cost for learning sparse coefficients (Figure 2.14). They achieved $AUC = 0.82$ on Avenue, 0.92 on Ped2 and 0.68 on ShanghaTech datasets.

Recently, Hamdi *et al.* [26] proposed an efficient method based on deep learning and handcrafted spatio-temporal feature extraction for anomaly detection using a pre-trained CNN and HOF (Histogram of Optical Flow) features. Abnormal motion was picked by relative thresholding. Then they trained One-class SVM with spatial features for robust classification of abnormal shapes. Moreover, they applied a decision function to correct the false alarms and the missed detections. Their method had a high performance in terms of speed and accuracy. It achieved anomaly detection with good efficiency in simple datasets ($EER = 0.145$ on USCD Ped2 and $EER = 0.035$ on UMN) and reduced computational complexity compared to state-of-the-art methods at the moment of publication.

2.4.4 State-of-the-art models with Generative Deep learning

Hasan *et al.* [27] learned all motion trajectories features (HOG, HOF, MBH) then built autoencoder to reconstruct the scene. They first leveraged the conventional handcrafted spatio-temporal local features and learned a fully connected autoencoder on them. Then they built a fully convolutional feed-forward autoencoder to learn both the local features and the classifiers as an end-to-end learning framework (Figure 2.15). Their model could capture the regularities from the multiple more challenging datasets (AUC = 0.80 on Avenue dataset and 0.61 on ShanghaiTech dataset). The idea of using reconstruction error to measure the regularity score was promising and has been extended by almost recent state-of-the-art methods.

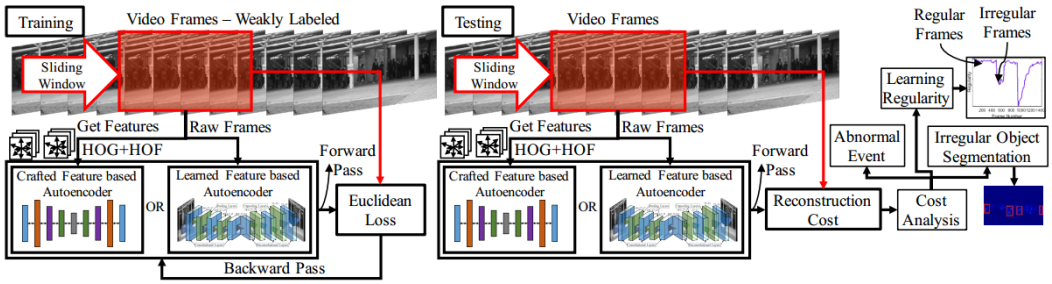


FIGURE 2.15 – Generative framework combining motion features or learned features with autoencoder to reconstruct the scene for anomaly detection proposed by Hasan *et al.* [27].

Luo *et al.* [62] integrated a ConvNet encoding appearance features for each frame and a ConvLSTM memorising motion features for all past frames with Auto-Encoder to learn the regularity of appearance and motion for the ordinary moments. Compared with 3D Convolutional Auto-Encoder based anomaly detection, their main contribution lied in that they propose a ConvLSTM-AE framework which better encodes the change of appearance and motion for normal events, respectively. They achieved AUC = 0.77 on Avenue, 0.88 on Ped2 and 0.75 on Ped1 datasets.

At the same time, Liu *et al.* [53] introduced a first work of future prediction based anomaly detection. They adopted CGAN techniques with U-Net model as generator to predict the next frame (Figure 2.16). To generate high-quality image, they made the constraints in terms of appearance (intensity loss and gradient loss) and motion (optical flow loss). Then the difference between a predicted future frame and its ground truth was used to detect an abnormal event. They achieved AUC = 0.85 on Avenue, 0.95 on Ped2, 0.83 on Ped1 and 0.73 on ShanghaiTech datasets.

Continuing of this approach, Ravanbakhsh *et al.* [81] proposed a GAN architecture for anomaly detection in particular crowd scenes. Their model is trained using normal frames and

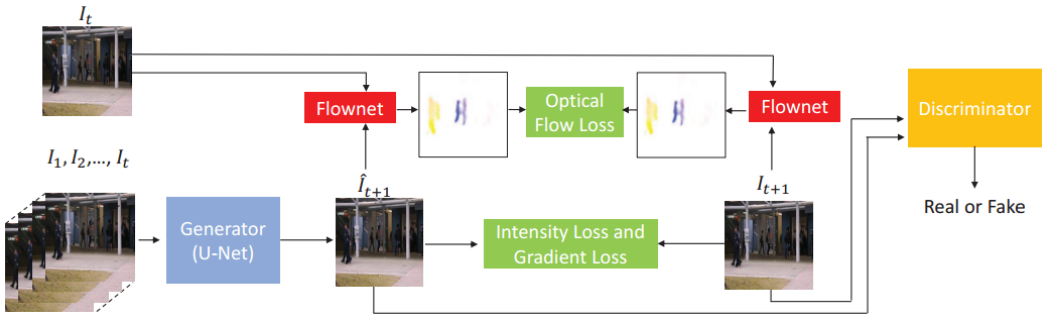


FIGURE 2.16 – Anomaly detection based on future frame prediction proposed by Liu *et al.* [53].

corresponding optical-flow images in order to learn an internal representation of the scene normality. Since their GANs are trained with only normal data, they are not able to generate abnormal events. At testing time, the real data were compared with both the appearance and the motion representations reconstructed by GANs and abnormal areas are detected by computing local differences. They achieved $AUC = 0.97$ on Ped1 and 0.93 on Ped2 dataset.

Developing this approach, Nguyen *et al.* [72] designed a model as a combination of a reconstruction network and an image translation model that share the same encoder. The former sub-network determined the most significant structures that appear in video frames and the latter one attempted to associate motion templates to such structures. The training stage was performed using only videos of normal events and the model was then capable to estimate frame-level scores for an unknown input. They achieved $AUC = 0.869$ on Avenue and 0.96 on Ped2 dataset.

Matching a single autoencoder network with classification sub-network to build a hybrid network, Nguyen *et al.* [71] proposed a model adapted from a typical auto-encoder working on video patches under the perspective of sparse combination learning. Their model focused on unsupervised learning common characteristics of normal events with the emphasis of their spatial locations (by supervised losses). This was the first work that directly adapts the patch position as the target of a classification subnetwork. The model is capable to provide a score of anomaly assessment for each video frame. They achieved $AUC = 0.83$ on Avenue and 0.84 on Ped2 dataset.

Lee *et al.* [47] used ConvLSTM to build spatio-temporal adversarial networks (STAN). They installed a spatio-temporal generator which synthesized an inter-frame by considering spatio-temporal characteristics with bidirectional ConvLSTM. A proposed spatio-temporal discriminator determined whether an input sequence was real-normal or not with 3D convolutional layers. Then they trained these two networks in an adversarial way to effectively encode spatio-temporal features of normal patterns. After the learning, they independently use the generator and the

discriminator as detectors, and deviations from the learned normal patterns were detected as abnormalities. They achieved $AUC = 0.87$ on Avenue, 0.965 on Ped2, 0.82 on Ped1 datasets.

Expecting to build an end-to-end generative model for anomaly detection, Sabokrou *et al.* [88] proposed a mix architecture of GAN and CNN for one-class classification. Their architecture contained two deep networks, each of them trained by competing with each other while collaborating to understand the underlying concept in the target class, and then classified the testing samples. One network worked as the novelty detector, while the other supported it by enhancing the inlier samples and distorting the outliers. The intuition was that the separability of the enhanced inliers and distorted outliers were much better than deciding on the original samples. They achieved $EER = 0.16$ on Ped2 dataset.

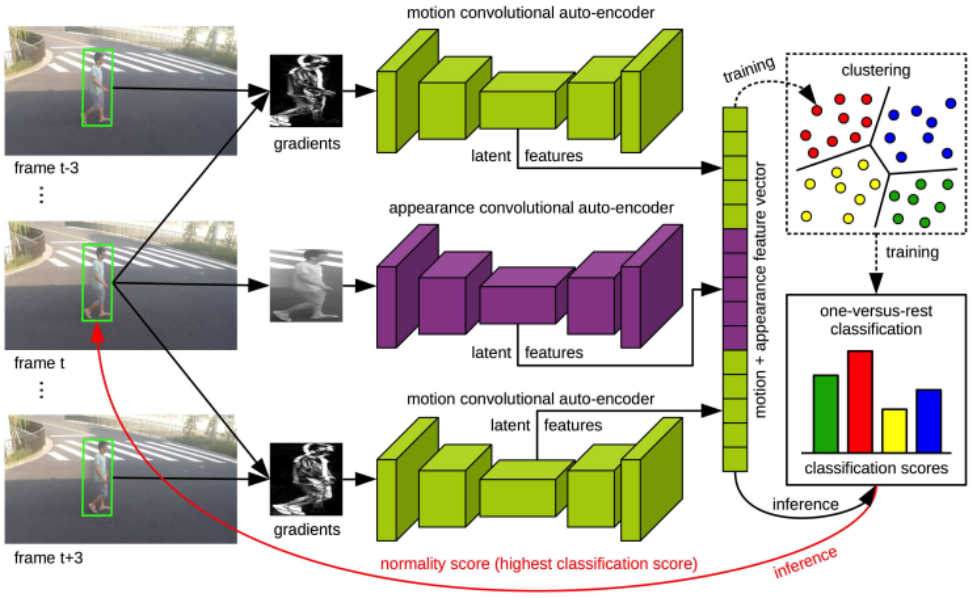


FIGURE 2.17 – Object-centric framework for anomaly detection proposed by Ionescu *et al.* [36].

Recently, Ionescu *et al.* [36] achieved state-of-the-art performance on various popular benchmarks [66, 58, 53] by building a reconstruction error model that learned both appearance and temporal gradient feature at object-centric levels then combined K-means clustering with SVMs techniques to produce abnormality scores (Figure 2.17). First, they installed an unsupervised feature learning framework based on object-centric convolutional auto-encoders to encode both motion and appearance information. Then, they proposed a supervised classification approach based on clustering the training samples into normality clusters. They applied one-versus-rest abnormal event classifier to separate each normality cluster from the rest. For the purpose of training the classifier, the other clusters is treated as dummy anomalies. During inference, they

decided an object as abnormal if the highest classification score assigned by the one-versus-rest classifiers was negative.

Continuing the approach of using an additional classifier layer to inference abnormality, instead of training SVMs with only normal samples from training dataset, Liu *et al*. proposed an interesting alternative supervised scenario [52]. Classical semi-supervised video anomaly detection assumes that only normal data are available in the training set because of the rare and unbounded nature of anomalies. It is obviously, however, these infrequently observed abnormal events can actually help with the detection of identical or similar abnormal events, a line of thinking that motivates us to study open-set supervised anomaly detection with only a few types of abnormal observed events and many normal events available. Under the assumption that normal events can be well predicted, they propose a Margin Learning Embedded Prediction (MLEP) framework. There are three features in MLEP- based open-set supervised video anomaly detection : i) they customize a video prediction framework that favors the prediction of normal events and distorts the prediction of abnormal events ; ii) The margin learning framework learns a more compact normal data distribution and enlarges the margin between normal and abnormal events. Since abnormal events are unbounded, their framework consequently helps with the detection of abnormal events, even for anomalies that have never been previously observed. Therefore, our framework is suitable for the open-set supervised anomaly detection setting ; iii) their framework can readily handle both frame-level and video-level anomaly annotations. Considering that video-level anomaly detection is more easily annotated in practice and that anomaly detection with a few anomalies is a more practical setting, their work thus pushes the application of anomaly detection towards real scenarios.

Inspired by the idea of adding supervised learning classifiers into the unsupervised features extraction framework of Liu *et al*. [52], we would like to go further by three aspects. Firstly, we propose a strong and flexible framework that provide more distinctive features for the following classifiers. Secondly, we investigate the effect of supervised classifiers corresponding to the number of abnormal samples transfered from original test set into training set. Thirdly, instead of going from frame-level to video-level, we have an local approach from frame-level to object-centric level. All of those aspects have their own significant improvements for our performances.

2.5 Conclusion

In this chapter, we largely present the state-of-the-art methods targeted to solve our two main tasks : (1)Improving segmentation and tracking with classical generative methods and (2) Anomaly detection with deep learning generative methods and supervised inference models. We

also provide the necessary knowledge of fundamental convolutional neural network models for various computer vision tasks in transportations domain. For the first problem, we focus on two baseline methods : Mask R-CNN and IOU Tracker. For the second problem, we are inspired by the modern approaches of future prediction for anomaly detection.

In the first part, although achieving impressive performance, both baseline methods have limitations in case of broken detection leading to fragmented trajectories. To deal with this drawback, we propose some interesting add-on modification based on simple hand-crafted generative methods. In the second part, we continue to develop the successful deep generative models of future prediction for anomaly prediction by proposing a flexible and strong framework to adapt with various challenging problems : the integration of both appearance and motion information in terms of RGB and grayscale at the input side ; the simplification of feature encoding dealing with supervised abnormality inference model, the capability of going from frame-level detection to object-centric level detection.

Chapitre 3

Improving detection and tracking using future prediction based on optical flow

In this chapter, we introduce the essential of our first contribution in details. The structure contains two parts : (1) Improving detection using future generated object segmentation based on optical flow, and (2) Enhanced Tracker with IOU-Tracker, Mask R-CNN and Optical flow. Then we present experimental results for both qualitative and quantitative evaluation.

Our first research is an initial work for evaluating the performances of the classical hand-crafted generative approach in future prediction and its usability for improving segmentation and tracking. Based on the two baseline methods of Luc *et al*. [59] to generate future segmentation by using optical flow vectors, we propose an extension beyond the original transforms. We investigate the generated information not only for both backward and forward directions but also for longer sequences of frames. This simple but useful generative framework leads to the improvement of the objects tracking task. Applying generated information with adding SURF features allow us to enhance the performances of IOU Tracker [7] by connecting fragmented pieces of trajectories.

The experiments are installed for evaluating two scenarios corresponding to two stages of the methods we proposed in previous parts. For each scenario, we begin with the introduction of the public datasets and evaluation metric where we setup the experiments on. Then we describe our experimental implementations in details : the basic framework for installing our models, the execution environment, the parameters setting, the evaluation metric, *etc*. Next, we introduce our experimental results in both qualitative and quantitative evaluation. Finally, we analyze our

strengths and our limitations regarding to our results.

3.1 Proposed methods

3.1.1 Instance segmentation by Mask RCNN

Mask R-CNN [29] proposed by He *et al*. is a state-of-the-art deep neural network for solving instance segmentation problem in computer vision. Given an input image, Mask R-CNN infers the object bounding boxes, classes and masks. Generally, there are two stages in Mask R-CNN (Figure 3.1). First, it generates region proposals where there might be an object based on the input image. Second, it predicts the object classes, refines the bounding box and generates a mask at the pixel level of the object based on the first stage proposal. Both stages are connected to the backbone structure.

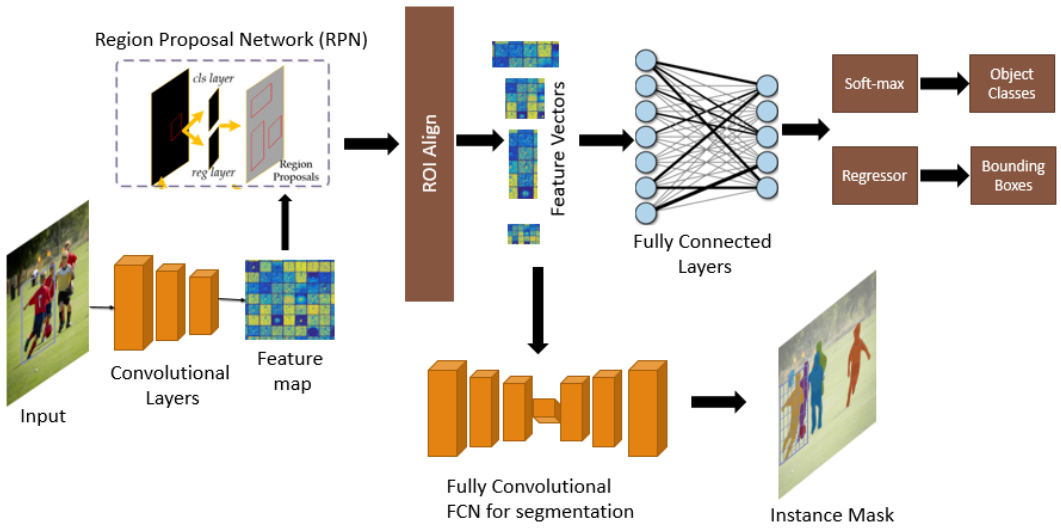


FIGURE 3.1 – General pipeline of Mask R-CNN. It contains two stages : First, it generates region proposals where there might be an object based on the input image. Second, it predicts the object classes, refines the bounding box and generates a mask at the pixel level of the object based on the first stage proposal.

First, input image is fed through the Convolutional layers to generate the feature maps. Region Proposal Network(RPN) uses a small CNN brand to generate the multiple Region of Interest (RoI) using a lightweight binary classifier. Generally, Mask R-CNN uses 9 anchors boxes over the image to detect multiple objects, objects of different scales, and overlapping objects in an image. This improves the speed and efficiency for object detection. Anchor boxes are a set of predefined bounding boxes of a certain height and width. These boxes are defined to capture the scale and aspect ratio of specific object classes needed to detect. To predict multiple objects or multiple instances of objects in an image, Mask R-CNN makes thousands of predictions. Final object detection is done by removing anchor boxes that belong to the background class and the

remaining ones are filtered by their confidence score. They choose the anchor boxes with IOU greater than 0.5. They applied Non-Max suppression strategy to select anchor boxes with the greatest confidence score. Non-Max Suppression removes all bounding boxes obtaining IOU less than or equal to 0.5 and take the bounding box with the highest value of IOU and suppress the other bounding boxes for identifying the same object.

Then the RoI Align network outputs multiple bounding boxes rather than a single definite one and warps them into a fixed dimension. Warped features vectors for each region are then fed into fully connected layers to make classification using softmax. At the same time, boundary box localization is further refined using the regression model. Warped features vectors are also fed into Mask Generator. It is another Fully Convolutional part acting like FCN [54] for segmentation, which consists of two CNN's to output a binary mask for each RoI. Mask generator allows the network to generate masks for every class without competition among classes.

3.1.2 Optical flow estimation by LDOF

Large Displacement Optical Flow (LDOF) [9] is an optical flow extraction method proposed by Brox *et al*. This method combines the advantages of two possible ways for establishing point correspondences between images with moving objects : one side, energy minimization methods that yield very accurate, dense flow fields, but fail as displacements get too large. Other side, there is descriptor matching that allows for large displacements, but correspondences are very sparse, have limited accuracy, and experience many outliers due to missing regularity constraints . LDOF establishes a region hierarchy for both images. Descriptor matching on these regions provides a sparse set of hypotheses for correspondences. These are integrated into a variational approach and guide the local optimization to large displacement solutions. The variational optimization selects among the hypotheses and provides dense and subpixel accurate estimates, making use of geometric constraints and all available image information.

Region segmentation : The first stage of LDOF is region computation. To create the image region, Brox *et al*. relied on a segmentation methods based on the boundary detector instead of simple edge detector. The advantage of using boundary detector over simple edge detection is that it takes texture into account. Boundaries due to repetitive structures are damped whereas strong changes in texture create additional boundaries. Consequently, boundaries are more likely to correspond to objects or parts of objects. Then, it increases the stability of the regions to be matched.

The segmentation method returns a boundary map $g(x)$ as shown in Figure 3.2. Strong edges correspond to more likely object boundaries. It further returns a hierarchy of regions created from this map. Regions with weak edges are merged first, while separations due to strong edges persist for many levels in the hierarchy. They generally take the regions from all the levels in the

hierarchy into account. From the regions of the first image, however, they only keep the most stable ones, *i.e.* those which exist in at least 5 levels of the hierarchy. Unstable regions are usually arbitrary subparts of large regions. They also ignore extremely small regions (with less than 50 pixels) from both images. These regions are usually too small to build a discriminative enough descriptor for reliable matching.

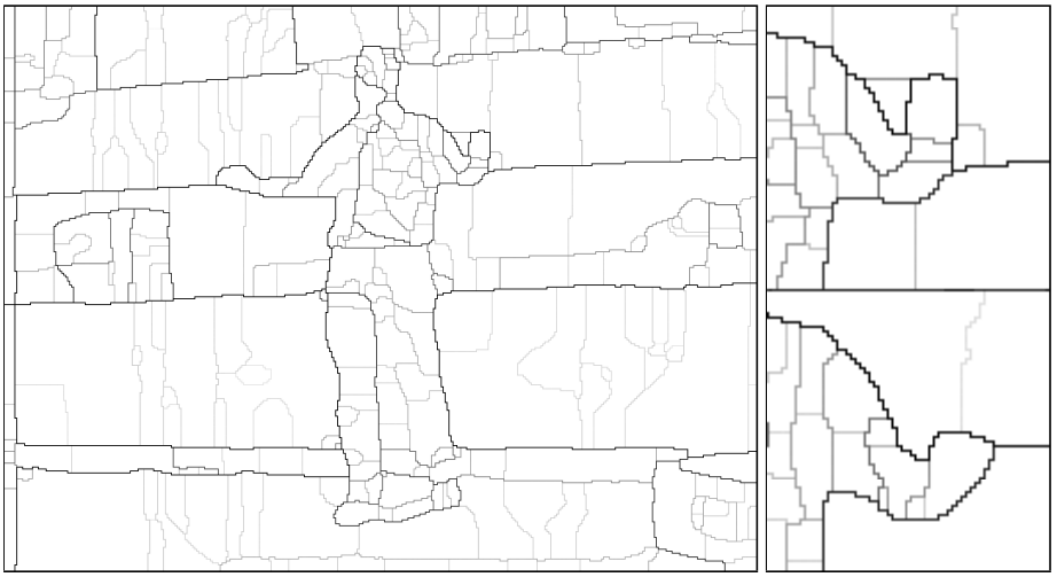


Figure 3.2 – Region segmentation in LDOF [9]. Left : Segmentation of an image. A region hierarchy is obtained by successively splitting regions at an edge of certain relevance. Dark edges are inserted first. Right : Zoom into the hand region of two successive images.

Region descriptor and matching : To each region they fit an ellipse and normalize the area around the centroid of each region to a 32×32 patch. The normalized patch then serves as the basis for a descriptor. They build two descriptors S and C in each region. S consists of 16 orientation histograms with 8 bins, like in SIFT [56]. C comprises the mean RGB color of the same 16 subparts as the SIFT descriptor. While the orientation histograms consider the whole patch to take also the shape of the region into account, the color descriptor is computed only from parts of the patch that belong to the region.

Correspondences between regions are computed by nearest neighbours matching. We compute the Euclidean distances of both descriptors separately and normalize them by the sum over

all distances :

$$d^2(S_i, S_j) = \frac{\|S_i - S_j\|_2^2}{\frac{1}{N}\sum_{k,l}\|S_k - S_l\|_2^2} \tag{3.1}$$

$$d^2(C_i, C_j) = \frac{\|C_i - C_j\|_2^2}{\frac{1}{N}\sum_{k,l}\|C_k - C_l\|_2^2} \tag{3.2}$$

where N is the total number of combinations (i, j). This normalization allows to combine the distances such that both parts in average have equal influence :

$$d^2(i, j) = \frac{1}{2}(d^2(C_i, C_j) + d^2(S_i, S_j)) \tag{3.3}$$

They can exclude potential pairs by adding high costs to their distance. They do this for correspondences with a displacement larger than 15% of the image size or with a change in scale that is larger than factor 3. Depending on the needs of the application, these numbers can be adapted. Smaller values obviously produce fewer false matches, but restrict the allowed image transformations.

Hypotheses refinement by deformed patches : Rather than deciding on a fixed correspondence at each keypoint, which could possibly be an outlier (Figure 3.3), they propose to integrate several potential correspondences into the variational approach. For this purpose, a good confidence measure is of great importance. They found that the distance between patches globally separates good and bad matches much better than the above descriptors. The main problem with direct patch comparison (classical block matching) is its sensitivity to small shifts or deformations. Once the deformation's corrected, the Euclidean distance between patches is very informative, particularly when considering only pixels from within the region.

The optimum shift and deformation needed to match two patches can be estimated by minimizing the following cost function :

$$E(u, v) = \int (P_2(x + u, y + v) - P_1(x, y))^2 dxdy + \alpha \int (|\nabla u|^2 + |\nabla v|^2) dxdy \tag{3.4}$$

where P_1 and P_2 are the two patches, $u(x, y), v(x, y)$ denotes the deformation field to be estimated, and $\alpha = 10000$ is a tuning parameter that steers the relative importance of the deformation smoothness. The energy is a non-linearized, large displacement version of the Horn and Schunck energy and is sufficient for this purpose. The regularizer gets a very high weight in this case, as without regularization every patch can be made sufficiently similar to any other.

As the patches are very small and a simple quadratic regularizer is applied, the estimation is quite efficient. Nevertheless, it would be a computational burden to estimate the deformation for

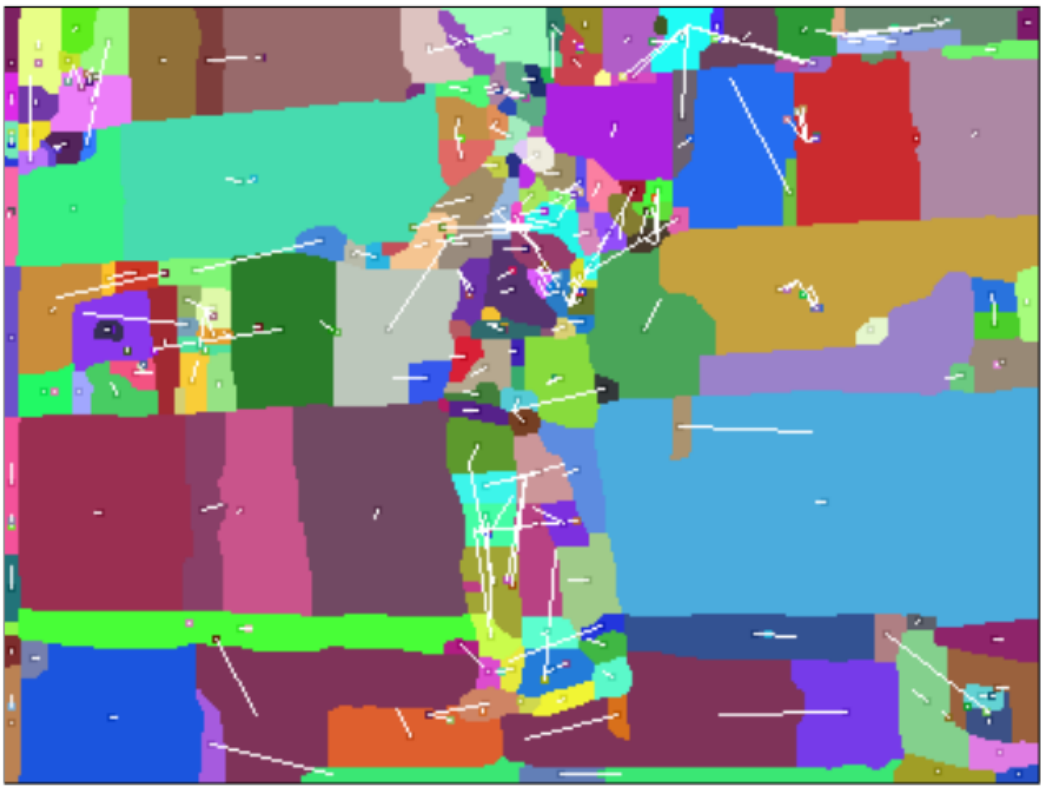


FIGURE 3.3 – Region matching with outliers existing in LDOF [9]. Displacement vectors of the matched regions drawn at their centroids. Many matches are good, but there are also outliers from regions that are not descriptive enough or their counterpart in the other image is missing.

each region pair. To this end, they preselect the 10 nearest neighbours for each patch using the distance from the previous section and compute the deformation only for these candidates. The five nearest neighbours according to the patch distance are then integrated into the variational approach described in the next part. Each potential match $j = 1, ..., 5$ of a region i comes with a confidence :

$$c_j(i) = \begin{cases} \dfrac{d^2(i) - d^2(i,j)}{d^2(i,j)} & \text{if } d^2(i) > 0 \\ 0 & \text{else} \end{cases} \tag{3.5}$$

where $d^2(i,j)$ is the Euclidean distance between the two patches after deformation correction and $d^2(i)$ is the average Euclidean distance among the 10 nearest neighbors. This measure takes the absolute fit as well as the descriptiveness into account. They restrict the distance to be computed only at patch positions within the region. Hence the changing background of a moving object part would not destroy similarity of a correct match.

Variational flow : Although most of the correspondences are correct, the flow field derived

from these by interpolation, is far from being accurate. This is because they have a hard decision to make when selecting the nearest neighbour. Moreover, a lot of image information is neglected and substituted by a smoothness prior. In order to obtain a more accurate, dense flow field, we integrate the matching hypotheses into a variational approach, which combines them with local information from the raw image data and a smoothness prior.

The energy following function is optimised with an additional data constraint that integrates the correspondence information :

$$\begin{aligned} E(w(X)) = & \int \Psi(|I_2(X+w(X)) - I_1(X)|^2) dX \\ & + \gamma \int \Psi(|\nabla I_2(X+w(X)) - \nabla I_1(X)|^2) dX \\ & + \beta \sum_{j=1}^{5} \int \rho_j(X) \Psi((u(X) - u_j(X))^2 + (v(X) - v_j(X))^2) dX \\ & + \alpha \int \Psi(|\nabla u(X)|^2 + |\nabla v(X)|^2 + g(X)^2) dX \end{aligned} \tag{3.6}$$

Here, I_1 and I_2 are the two input images, $w = (u, v)$ is the sought optical flow field, and $X = (x, y)$ denotes a point in the image. $(u_j, v_j)(X)$ is one of the motion vectors derived at position X by region matching (j indexing the 5 nearest neighbours). If there is no correspondence at this position, $\rho_j(X) = 0$. Otherwise, $\rho_j(X) = c_j$, where c_j is the distance based confidence in previous formula. $\alpha = 100$, $\beta = 25$, and $\gamma = 5$ are tuning parameters, which steer the importance of smoothness, region correspondences, and gradient constancy, respectively. They use the robust function $\Psi(s^2) = \sqrt{s^2 + 10^{-6}}$ in order to deal with outliers in the data as well as in the smoothness assumption. They also integrate the boundary map $g(X)$ in order to avoid smoothing across strong region boundaries. Rather than a straightforward three step procedure with (i) interpolation of the region correspondences, (ii) removal of outliers not fitting the interpolated flow field (iii) optical flow estimation initialized by the interpolated inlier correspondences, the above energy combines all three steps in a single optimization problem. The energy is non-convex and can only be optimized locally. They can compute the Euler-Lagrange equations, which state a necessary condition for a local optimum.

3.1.3 Optical flow estimation by Full Flow

Full Flow [13] proposed by Chen *et al*. is a global optimization approach to optical flow estimation. The approach optimizes a classical optical flow objective over the full space of mappings between discrete grids. No descriptor matching is used. The highly regular structure of the space of mappings enables optimizations that reduce the computational complexity of the algorithm's inner loop from quadratic to linear and support efficient matching of tens of thousands of nodes to tens of thousands of displacements.

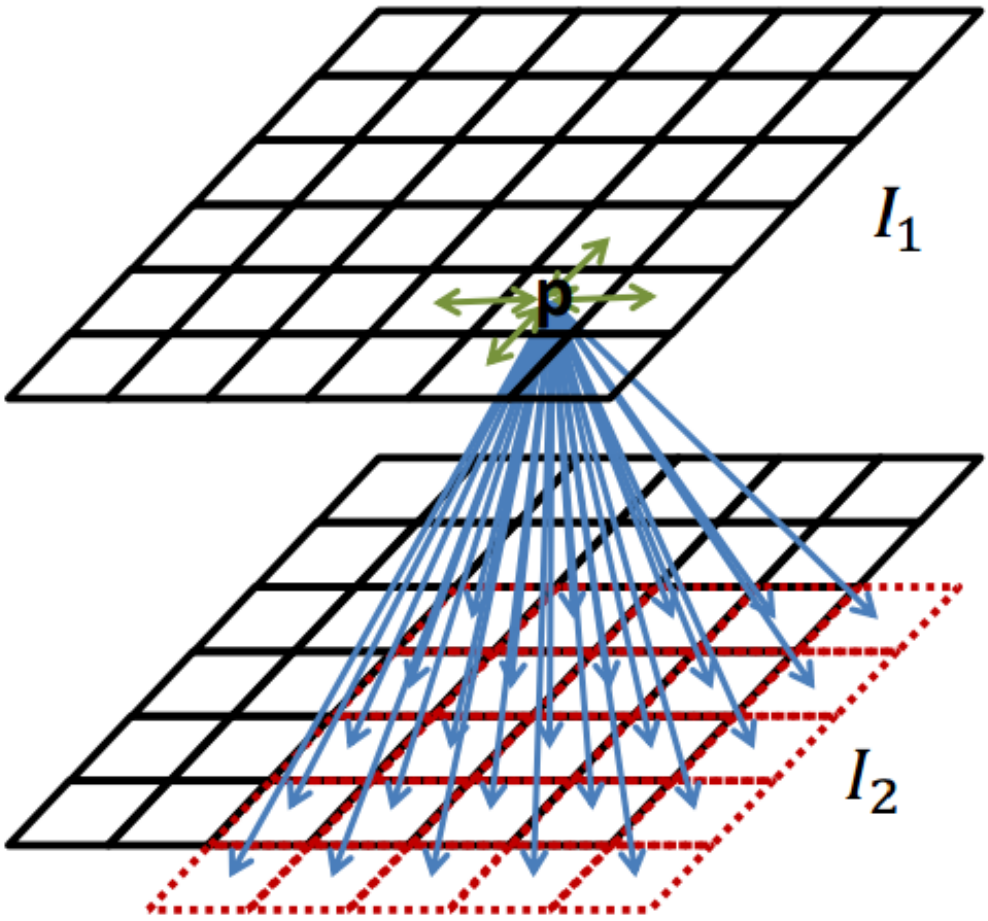


FIGURE 3.4 – Optical flow over regular grid in Full Flow method [13]. Each pixel p in I_1 is spatially connected to its four neighbors in I_1 and temporally connected to $(2\zeta+1)^2$ pixels in I_2.

Model : Let $I_1, I_2 : \Omega \to R^3$ be two color images, where $\Omega \subset Z^2$ is the image domain. Let $f = (f^1; f^2) : \Omega \to [-\zeta; \zeta]^2$ be a flow field that maps each pixel p in I_1 to $(p + f^p)$ in an augmented domain $\overline{\Omega} \supset \Omega$, which contains Ω and a large surrounding buffer zone. The buffer zone absorbs pixels that flow out of the visual field. The augmented domain $\overline{\Omega} \supset Z^2$ is the Minkowski sum of Ω and $[-\zeta; \zeta]^2 \cap Z^2$, where ζ is the maximal empirical displacement magnitude. The maximal empirical displacement magnitude is measured by taking the maximal displacement observed in a training set. For example, the maximal displacement magnitude on the KITTI training set is 242 pixels. They perform the optimization on1/3-resolution images, so $\zeta = 81$ for the KITTI dataset.

The objective function is :

$$E(f) = \sum_{p \in I_1} (\rho_D(p, f_p, I_1, I_2)) + \lambda \sum_{p,q \in N} w_{p,q} \rho_S(f_p - f_q) \tag{3.7}$$

where $N \subset \Omega^2$ is the 4-connected pixel grid illustrated in Figure 3.4. The data term $(\rho_D(p, f_p, I_1, I_2))$ penalizes flow fields that connect dissimilar pixels p and $(p + f_p)$. They use truncated normalized cross-correlation :

$$\rho_D(p, f_p, I_1, I_2) = 1 - \max(NCC, 0) \tag{3.8}$$

where NCC is the normalized cross-correlation between two patches, one centered at p in I_1 and one centered at $(p + f_p)$ in I_2, computed in each color channel and averaged. The truncation at zero prevents penalization of negatively correlated patches. If $(p + f_p)$ is in the buffer zone $\overline{\Omega}\ \Omega$, the data term is set to a constant penalty ζ.

Their optimization approach assumes that the regularization term has the following form :

$$\rho_S(f) = \min(\rho(f^1) + \rho(f^2), \tau) \tag{3.9}$$

where f^1, f^2 are the two components of vector f and $\rho(\cdot)$ is a penalty function, such as the $L1$ norm or the Charbonnier penalty. Their formulation and the general solution strategy can accommodate non-convex functions ρ, such as the Lorentzian and the generalized Charbonnier penalties. They apply the reduction of message passing complexity from quadratic to linear. The highly efficient min-convolution algorithm will assume that the function ρ is convex. Note that the regularization term couples the horizontal and vertical components of the flow. They apply a Laplace weight to attenuate the regularization along color discontinuities :

$$w_{p,q} = \exp(-\frac{\|I_1(p) - I_2(q)\|}{\beta}) \tag{3.10}$$

Optimization : Objective function 3.7 is a discrete Markov random field with a two-dimensional label space. The label space of the model is $[-\zeta; \zeta]^2 \cap Z^2$. To optimize the model, they use TRW-S, which optimizes the dual of a natural linear programming relaxation of the problem. They choose TRW-S due to its effectiveness in optimizing models with large label spaces. Note that TRW-S optimizes the dual objective and will generally not yield the optimal solution to the primal problem.

3.1.4 Generating object segmentation by Mask R-CNN and Optical flow

Luc *et al*. [59] proposed two baseline methods to generate future segmentation based on flow vector $F_{t-1 \to t}$ from frame $t-1$ to t and the current instance segmentation I_t at frame t. They were

called $Shift$ and $Warp$ (Figure 3.5a).

— **Warp** approach translates each pixel of instance mask independently using the flow vector at the corresponding position inside this mask. To yield the new mask, the object class and the confident score is copied from the previous one. The predicted mask and flow field are used to make the next prediction, and so on. This approach is suitable for longtime prediction because of the ability of rescaling objects based on flow vector. In contrast, predicted mask suffers from an accumulated error phenomenon and has many holes inside its boundaries. To fill in those gaps, a post-processing step with morphological operator is necessary.

— **Shift** approach, in brief, entirely shifts instance mask using the average flow vector calculated inside the mask. Then the object class and the confident score are also copied from the previous mask. While this approach can avoid accumulating errors and generating holes inside the mask, it is not really suitable for longtime prediction due to the non-rescaling of the mask.

We propose an extension beyond the two original baselines. Instead of only predicting future segmentation (forward), we also generate past segmentation (backward) by projecting the flow vector in opposite direction (Figure 3.5b). Intuitively, the missing detection of Mask R-CNN does not last along many frames so we can generate new segmentation in both forward and backward directions.

Furthermore, we also consider the development of flow vector beyond one time step. Given instance results I_t, I_{t+N} of frame t and $t+N$ achieved from Mask R-CNN detector, we have (Figure 3.5c) :

$$I_{t+j;0<j<N} = \sum p_{t+j} \text{ w.r.t } \mathbf{F}_{t\to t+j}(p_t) = p_{t+j} \text{ and } \mathbf{F}_{t+j\to t+N}(p_{t+j}) = p_{t+N} \tag{3.11}$$

where p_i is the pixel of instance I_i, $\mathbf{F}_{i\to l}$ is a translation by flow vectors from frame i to frame l. To avoid the accumulated errors of optical flow estimation methods, we minimize $N = 2$, $j = 1$.

This condition maintain the continuity of trajectories along the flow vectors when we generate new masks. We apply this extension in both forward and backward directions with both $Shift$ and $Warp$ approaches. Therefore, we have 8 ways to generate object segmentation to fill in the gaps of trajectories.

3.1.5 Extracting SURF feature

Speed Up Robust Feature (SURF) [5] is a scale and rotation invariant interest point detector and descriptor proposed by Bay *et al*. It respects to repeatability, distinctiveness, and robustness. Moreover, SURF can be computed and compared much faster.

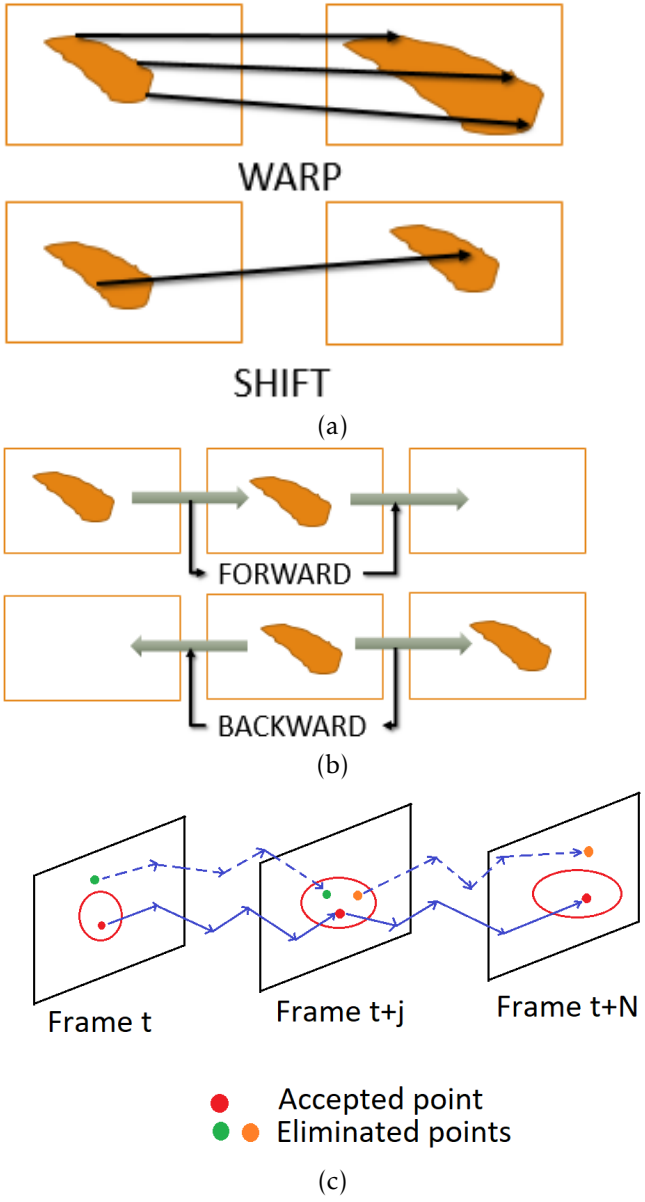


FIGURE 3.5 – Generating object segmentation (a) Shift and Warp translation ; (b) Forward and backward translation ; (c) Combined results beyond one time step

SURF's interesting performance is achieved by relying on integral images for image convolutions ; by building on the strengths of the leading existing detectors and descriptors they use a Hessian matrix-based measure for the detector and a distribution-based descriptor. Simplifying these methods to the essential, this leads to a combination of novel detection, description, and matching steps.

Fast Hessian detector : Given a point $p = (x, y)$ in an image I, the Hessian matrix $H(p, \sigma)$ in p at scale σ is defined as follows :

$$H(p, \sigma) = \begin{bmatrix} L_{xx}(p, \sigma) & L_{xy}(p, \sigma) \\ L_{xy}(p, \sigma) & L_{yy}(p, \sigma) \end{bmatrix} \tag{3.12}$$

where $L_{xx}(p, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point p, and similarly for $L_{xy}(p, \sigma)$ and $L_{yy}(p, \sigma)$.

Generally, SURF is a speed up version of SIFT [56] proposed by David Lowe. Gaussians are optimal for scale-space analysis. In practice, however, the Gaussian needs to be discretized and cropped (Figure 3.6 left half) and even with Gaussian filters aliasing still occurs as soon as the resulting images are sub-sampled. As Gaussian filters are non-ideal in any case, and given Lowe's success with LoG approximations [56] in extracting SIFT, Bay *et al*. pushed the approximation even further with box filters (Figure 3.6 right half). These approximate second order Gaussian derivatives, and can be evaluated very fast using integral images, independently of size. One big advantage of this approximation is that, convolution with box filter can be easily calculated with the help of integral images. And it can be done in parallel for different scales. Also the SURF rely on determinant of Hessian matrix for both scale and location.

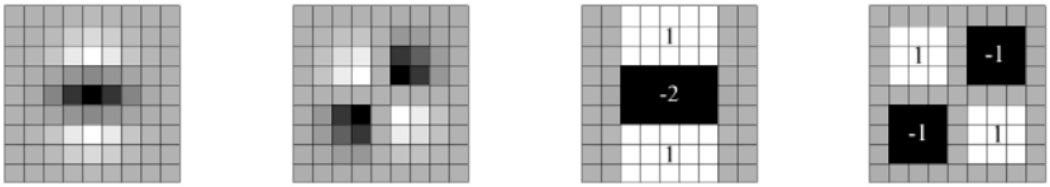


Figure 3.6 – Basic of Laplacian of Gaussian with Box Filter in SURF detector with Hessian matrix [5]. Left to right : The (discretized and cropped) Gaussian second order partial derivatives in y-direction and xy-direction, and approximations thereof using box filters. The grey regions are equal to zero.

In practice, the 9×9 box filters in Figure 3.6 are approximations for Gaussian second order derivatives with $\sigma = 1.2$ and represent highest spatial resolution). Denote the approximations by D_{xx}, D_{yy}, D_{xy}, the weights applied to the rectangular regions are kept simple for computational efficiency, but they need to further balance the relative weights in the expression for the Hessian's determinant with coefficient :

$$\frac{|L_{xy}(1.2)|_F |D_{xx}(9)|_F}{|L_{xx}(1.2)|_F |D_{xy}(9)|_F} \approx 0.9 \tag{3.13}$$

where $|x|_F$ is the Frobenius norm. Finally, they achieved :

$$\det(H_{approx}) = D_{xx} D_{yy} - (0.9 D_{xy})^2 \tag{3.14}$$

Furthermore, the filter responses are normalized with respect to the mask size. This guarantees a constant Frobenius norm for any filter size. Scale spaces are usually implemented as image pyramids. The images are repeatedly smoothed with a Gaussian and subsequently sub-sampled in order to achieve a higher level of the pyramid.

In order to localize interest points in the image and over scales, a non-maximum suppression in a $3\times3\times3$ neighborhood is applied. The maximum of the determinant of the Hessian matrix are then interpolated in scale and in image space. Scale space interpolation is especially important in this case, as the difference in scale between the first layers of every octave is relatively large.

Orientation assignment : The proposed SURF descriptor is based on similar properties as SIFT, with a complexity stripped down even further. The first step consists in fixing a reproducible orientation based on information from a circular region around the interest point. SURF uses wavelet responses in horizontal and vertical direction (Figure 3.7)for a neighbourhood of size $6s$, with s the scale at which the interest point was detected. Also the sampling step is scale dependent and chosen to be s. In keeping with the rest, also the wavelet responses are computed at that current scale s. Accordingly, at high scales the size of the wavelets is big. Therefore, they use again integral images for fast filtering. Only six operations are needed to compute the response in x or y direction at any scale. The side length of the wavelets is $4s$.

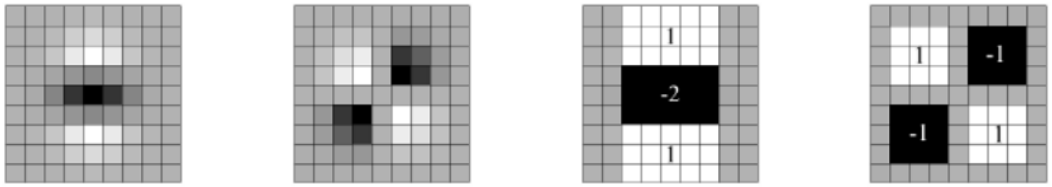


Figure 3.7 – Localizing interest points and using Haar wavelet for orientations assignments and descriptor extraction in [5]. Left : Detected interest points for a Sunflower field. This kind of scenes shows clearly the nature of the features from Hessian-based detectors. Middle : Haar wavelet types used for SURF. Right : Detail of the Graffiti scene showing the size of the descriptor window at different scales.

Once the wavelet responses are calculated and weighted with a Gaussian ($\sigma = 2.5s$) centered at the interest point, the responses are represented as vectors in a space with the horizontal response strength along the abscissa and the vertical response strength along the ordinate. The dominant orientation is estimated by calculating the sum of all responses within a sliding orientation window of angle 60 degrees. Interesting thing is that, wavelet response can be found out using integral images very easily at any scale. For many applications, rotation invariance is not required, so no need of finding this orientation, which speeds up the process. SURF provides such a functionality called Upright-SURF or U-SURF. It improves speed and is robust

up to ±15°. If it is 0, orientation is calculated. If it is 1, orientation is not calculated and it is faster.

Feature descriptor components : the first step consists in constructing a square region centered around the interest point, and oriented along the orientation selected in the previous section. For the upright version, this transformation is not necessary. The size of this window is $20s$. Examples of such square regions are illustrated in Figure 3.7.

In detail, a neighbourhood of size $20s \times 20s$ is taken around the key point where s is the size. It is divided into 4×4 subregions. For each subregion, horizontal and vertical wavelet responses are taken and a vector is formed like this, $v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$. This when represented as a vector gives SURF feature descriptor with a total of 64 dimensions. Lower the dimension, higher the speed of computation and matching, but provide lower distinctiveness of features.

Then, for more distinctiveness, SURF feature descriptor has an extended 128 dimensions version. The sums of d_x and $|d_x|$ are computed separately for $d_y < 0$ and $d_y \geq 0$. Similarly, the sums of d_y and $|d_y|$ are split up according to the sign of d_x , thereby doubling the number of features. It does not add much computation complexity.

Another important improvement is the use of the sign of the Laplacian (trace of Hessian Matrix) for underlying interest point. It adds no computation cost since it is already computed during detection. The sign of the Laplacian distinguishes bright blobs on dark backgrounds from the reverse situation. In the matching stage, they only compare features if they have the same type of contrast. This minimal information allows for faster matching, without reducing the descriptor's performance.

3.1.6 Improving IOU Tracker with generated informations and SURF features

Intuitively, the more boxes we have, the more accurate the IOU Tracker is. Despite the strong performances of Mask R-CNN, the missing detections in some frames are inevitable. The technique of generating new segment can be similarly applied for generating new bounding boxes. We proposed 2 methods of translation (Shift and Warp) with 2 directions (backward, forward) and 2 optical flow methods (LDOF, FullFlow), so we have 8 methods to generate bounding box. Each box is generated by one of the 8 methods. We consider each box as a special mask which contains only four pixel corresponding to four vertices of box and then the process is repeated. After applying generative approach, we get a situation where we have more overlapping boxes. To eliminate those redundant boxes, we apply the idea of IOU tracker that takes into account the IOU between boxes. Let B_{tG} denote the set of generated boxes b^i_{tG} and B_{tM} denote the set of instance boxes b^i_{tM} directly detected by Mask R-CNN at frame I_t (Figure 3.8). To combine the two sets, we compare the overlapping region with a threshold σ_{IOU} by running the algorithm 1 :

Algorithm 1 Eliminate overlapping boxes

for $i = 1 \rightarrow \|B_{tG}\|$ **do**
  $count \leftarrow 0$
  for $j = 1 \rightarrow \|B_{tM}\|$ **do**
    if $IOU(b^i_{tG}, b^i_{tM}) \geq \sigma_{IOU}$ and $class(b^i_{tG}) = class(b^i_{tG})$ **then**
      $count \leftarrow count + 1$
    end if
  end for
  if $count = 0$ **then**
    add b^i_{tG} to B_{tM}
  else
    discard b^i_{tG}
  end if
end for

It means that we trust more in the results provided by the Mask R-CNN detector (this assertion is based on the analysis of the experimental results described in section 5.1). After this step, we have only one set B_t containing the box b^i_t of frame I_t.

The next stage consists in applying the IOU Tracker. Each box b^i_t is compared with all the boxes from the L previous frames, where L is the length of the tracker. We associate a box to the previous box which obtains the same class and the maximum overlapping region. If we can not find any IOU value greater than the threshold σ_{IOU} along all L frames, the current tracking process is terminated and we start a new track.

To enhance our tracker, we propose to use SURF [5] to verify the matching boxes. This feature have shown its efficiency on image matching problem. By applying SURF to match the similar boxes, we can avoid the fragmented trajectories. We extract SURF points for each bounding boxes. Each box b^i_t is compared with all boxes of L previous frames. If we find two boxes associated to two different classes but with their IOU value and the number of matching SURF points greater than threshold σ_{IOU} and σ_{SURF}, respectively, we associate them together. To avoid the negative effect caused by extracting SURF features from a too small box, we set another threshold σ_S to select the suitable boxes which exhibit a value greater than this threshold. For all the boxes with a value smaller than σ_S, we do not extract SURF and we only take into account IOU value for the association step. The whole process is illustrated by Figure 3.8.

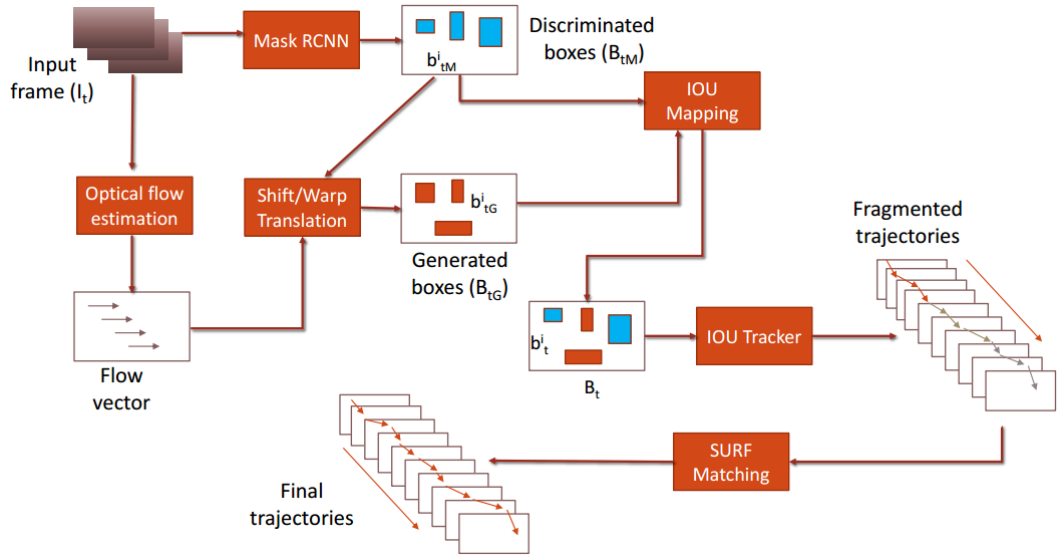


FIGURE 3.8 – Improving IOU-Tracker with generated information and SURF features. To combine the two sets : generated set B_{tG} and Mask R-CNN instances set B_{tM}, we compare the overlapping region with a threshold σ_{IOU}. After this step, we have only one set B_t containing the box b^i_t of frame I_t. Then Each box b^i_t is compared with all boxes of L previous frames. If we find two boxes associated to two different classes but with their IOU value and the number of matching SURF points greater than threshold σ_{IOU} and σ_{SURF}, respectively, we associate them together.

3.2 Experiments

We separately evaluate each stage using different datasets. First, we compare the techniques for generating objects segmentation stages using the DAVIS dataset [77]. Then, we chose the most suitable methods to apply for the tracking stage. The qualitative evaluation regarding the improvement of the IOU Tracker is performed using the UA-DETRAC dataset [65]. Some samples of those datasets is illustrated in Figure 3.9.

FIGURE 3.9 – Illustration of several samples in DAVIS dataset [77] and UA-DETRAC dataset [65]. First row : DAVIS dataset. Second row : UA-DETRAC dataset.

Method	Bus				Car			
	S	W	W-M	R	S	W	W-M	R
Backward	79.31	75.59	**81.76**	89.46	65.13	59.21	**70.92**	92.10
Forward	79.05	75.84	**82.07**	89.46	65.04	60.33	**69.15**	92.10
Combine-B	**79.35**	71.47	77.35	89.46	**66.87**	46.91	57.86	92.10
Combine-F	**79.03**	71.76	77.35	89.46	**65.36**	48.44	56.37	92.10

Tableau 3.1 – Results of generating new segmentation on DAVIS 2016 by LDOF with "car" and "bus" classes. S : Shift ; W : Warp without morphological post-processing ; W-M : Warp with morphological post-processing ; R : True masks of Mask R-CNN which are eliminated to make the missing detection context ; Combine-B : Extension case in backward direction ; Combine-F : Extension case in forward direction

3.2.1 Improving detection using future generated object segmentation based on optical flow

DAVIS 2016 consists of fifty high quality, Full HD video sequences, spanning multiple occurrences of common video object segmentation challenges such as occlusions, motion blur and appearance changes. Each video is accompanied by densely annotated, pixel-accurate and per-frame ground truth segmentation. Because the UA-DETRAC only takes into account three classes (bus, van and car) we restrict the use of the DAVIS sequences to the ones that deal with the same types of vehicles. Unfortunately, only a few sequences within the DAVIS dataset satisfy this constraint. Worst, those sequences are not challenging enough to generate the missing detections from Mask R-CNN we want to cope with. Thus, in order to prepare the situation of false negative errors, we run Mask R-CNN for each frames, then randomly discard some of the generated detection. The original annotation of discarded frames are utilized as ground-truth to compare with the generated results. The performance is measured with the mean Intersection-over-Union (mIOU) metric, which first computes the IOU for each semantic class and then computes the average over classes. IOU is defined as follows :

$$IOU = \frac{TruePositive(TP)}{TruePositve(TP) + FalseNegative(FN) + FalsePositive(FP)} \tag{3.15}$$

First, Mask R-CNN is applied to each frames. Next, we use LDOF [9] and Full Flow [13] to extract the optical flow vectors from pairs of frames. Then, we do the **Shift** and **Warp** translation in both backward and forward directions. Morphological operators are only applied as a post-processing step with the **Warp** method. The extension beyond more than one step time that we call "combined results" in Figure 3.5c is also evaluated in both directions. Here, we choose the simplest case with $N = 2$ and $j = 1$. Our results with LDOF optical flow is illustrated in Table 3.1 and Table 3.2 ; with Full Flow optical flow in Table 3.3 and Table 3.3 .

Quantitative evaluation : We observe that the morphological operators are important with the **Warp** method. After adding this post-processing, the performances are significantly increa-

Method	Bus				Car			
	S	W	W-M	R	S	W	W-M	R
Backward	79.24	76.72	**82.36**	89.46	65.66	60.16	**70.55**	92.10
Forward	79.05	76.63	81.96	89.46	65.34	61.00	**68.94**	92.10
Combine-B	**79.31**	72.42	77.57	89.46	**66.41**	48.03	57.33	92.10
Combine-F	**79.06**	72.64	77.48	89.46	**65.63**	49.19	56.66	92.10

Tableau 3.2 – Results of generating new segmentation on DAVIS 2016 by Full Flow with "car" and "bus" classes on missing frames *i.e.* discarded frames.

Method	LDOF			Full Flow			Mask R-CNN
	Shift	Warp	Warp-M	Shift	Warp	Warp-M	
Backward	72.22	67.40	**76.34**	72.45	68.44	**76.46**	90.78
Forward	72.05	68.08	**75.61**	72.19	68.81	**75.50**	90.78
Combine-B	**73.11**	59.19	67.60	**72.86**	60.23	67.45	90.78
Combine-F	**72.20**	60.10	66.86	**72.35**	60.192	67.07	90.78

Tableau 3.3 – Average performance of generating new segmentation on DAVIS 2016 with "car" and "bus" classes on missing frames *i.e.* discarded frames.

sed. While **Shift** works stably for all of the methods, this is not the case with the **Warp** based approach. In a simple case, where we only consider optical flow within two consecutive frames in both direction, **Warp** significantly outperforms **Shift**. Conversely, when we take into account the development of optical flow along many frames, **Shift** gives us better results. This difference can be explained by the accumulated error of optical flow. The more frames we process, the more optical flow error we integrate. **Shift** is chosen to reduce this issue. Furthermore, we did not see any significant differences for each method when we performed LDOF or Full Flow optical flow estimation. Most of the average results in Table 3.3 are similar (Full Flow is slightly better) despite Full Flow significantly outperforms LDOF in the original task of optical flow estimation for both MPI Sintel and KITTI Dataset. We draw that the performance of generating new segmentation depends on how we use the flow vector, not on the type of flow. Therefore, our methods allow us to get a benefit regardless of the optical flow estimation methods, then fast methods are highly prioritized. On the other hand, we find that the original masks created by Mask R-CNN always give us better accuracy than the optical flow generated masks. This analysis suggests us to put more confidence in the discriminative results from Mask R-CNN than in the generative results from optical flow when we combine those results for the next tracking stages.

Qualitative evaluation : Figure 3.10 and Figure 3.11 show the comparison of qualitative performance on optical flow estimation between LDOF and Full-Flow methods. We can see the significant differences in both bus and car classes. Despite the obvious out-performance of Full Flow over LDOF, the performance of generating new segmentation is almost stable.

FIGURE 3.10 – Comparison of optical flow estimation performance between LDOF and Full-Flow. Results are reported on bus class of DAVIS dataset. Obviously, the difference of quality between two methods is significant, not only for flow localization but also for flow intensity and direction. Best viewed in color

This interesting qualitative results is illustrated in Figure 3.12 and Figure 3.13. Once again, we draw that the performance of generating new segmentation depends on how we use the flow vector, not on the type of flow calculation. On the other hands, the importance of morphological operators integrated in Warp translation is confirmed by qualitative illustrations in both bus and car classes.

FIGURE 3.11 – Comparison of optical flow estimation performance between LDOF and Full-Flow. Results are reported on car class of DAVIS dataset. Obviously, the difference of quality between two methods is significant, not only for flow localization but also flow intensity and direction. Best viewed in color

3.2.2 Enhanced Tracker with IOU-Tracker based Mask R-CNN and Optical flow

Based on the results from previous stages, we choose the **Shift** generator to create the new bounding boxes from optical flow. Although *Combine backward* and *Combine forward* are better for Shift translation, the difference between those performances and *Forward* method is not significant. Additionally, tracking only in forward direction is more natural and simpler. Thus, we perform the **Shift** generator in forward direction. Full Flow [13] is used for estimating optical

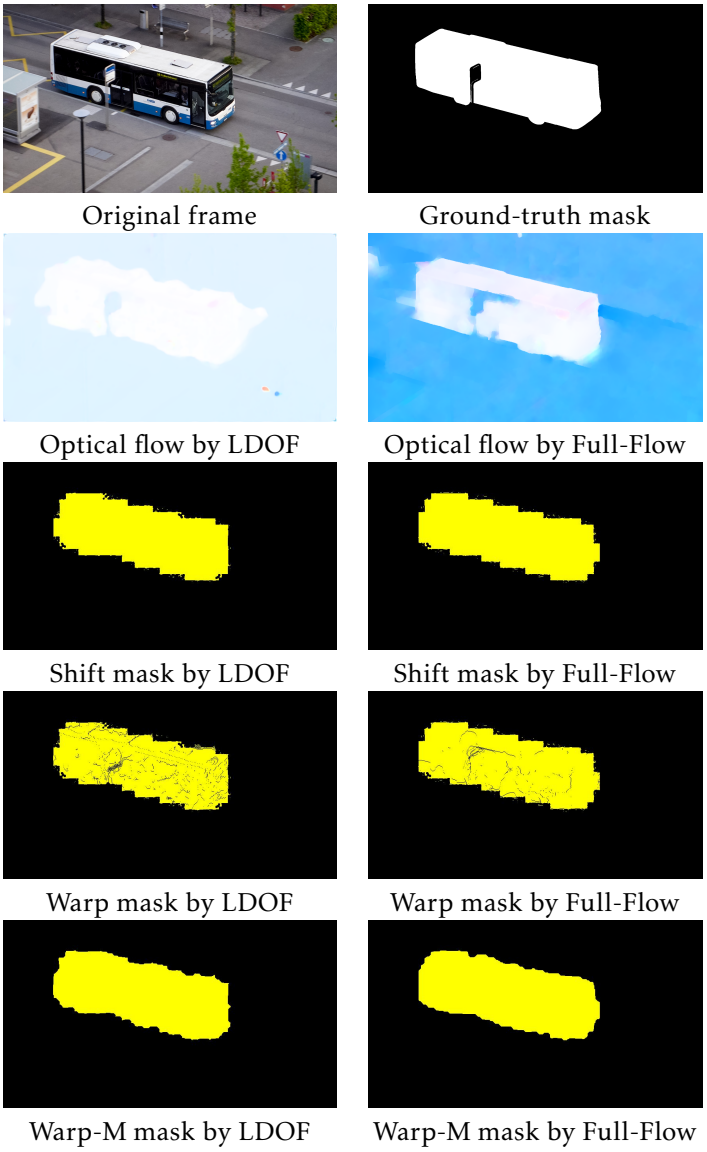


FIGURE 3.12 – Qualitative performance of generating object segmentation for bus class along forward direction. Optical flow is estimated by Full flow and LDOF. Warp-M denotes the Warp translation with Morphological operator. Despite the obvious out-performance of Full Flow over LDOF in optical flow estimation, the performance of generating new segmentation is almost stable. We draw that the performance of generating new segmentation depends on how we use the flow vector, not the type of flow. On the other hands, the importance of morphological operators integrating in Warp translation is confirmed by filling in all holes in Wrap mask.

flow vectors. The discriminative boxes of Mask R-CNN and the generative boxes of **Shift** are combined thanks to a IOU mapping (and its σ_{IOU}) to discard all the generative boxes matching

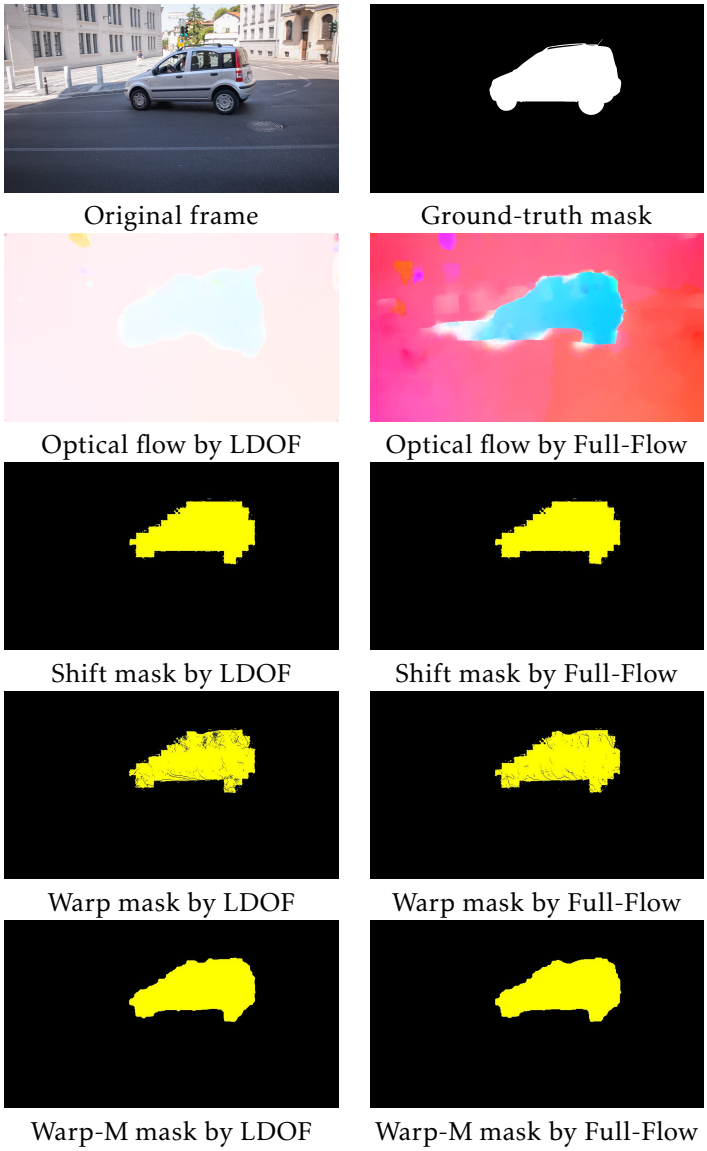


FIGURE 3.13 – Qualitative performance of generating object segmentation for car class along forward direction. Optical flow is estimated by Full flow and LDOF. Warp-M denotes the Warp translation with Morphological operator. Despite the obvious out-performance of Full Flow over LDOF in optical flow estimation, the performance of generating new segmentation is almost stable. We draw that the performance of generating new segmentation depends on how we use the flow vector, not the type of flow. On the other hands, the importance of morphological operators integrating in Warp translation is confirmed by filling in all holes in Wrap mask.

a discriminative box w.r.t its location and its object class. As discussed above, we trust more in discriminative boxes provided by Mask R-CNN. For the IOU tracking step, we choose $L = 5$

and use the same parameter σ_{IOU} as for the previous step. All the parameters σ_{IOU}, σ_{SURF}, are determined by experiment. The results show us that $\sigma_{IOU} = 0.5$, $\sigma_{SURF} = 1$ and $\sigma_S = 50 \times 50$ are the best choices. The qualitative results are shown in Figure 3.14. We observe that the trajectories are less fragmented after applying our techniques.

3.3 Conclusion

In this chapter, we presented our proposed methods for improving vehicle tracking in detail. The first section focuses on classical hand-crafted generative methods based on optical flow estimation while the second section presents anomaly detection based on future prediction framework. Generally, our first work is an extension of previous state-of-the-art segmentation and tracking frameworks for going beyond their limitations. In this part, our interesting add-on almost help them improving the qualitative performance aspect. We try to keep the source algorithms being stable then our extensions can be easily integrated into original algorithms.

The strengths of this first work is to provide several simple but interesting techniques to enhance the qualitative performances in some aspect of two classical tasks in computer vision. The computational complexity is quite low due to the hand-crafted solutions obtained without training or fine-tuning any deep network. In contrast, this work has limitations in term of proposing innovative framework. Besides, the possibility to reapply this methods as it is to specific task such as anomaly detection is not significant enough.

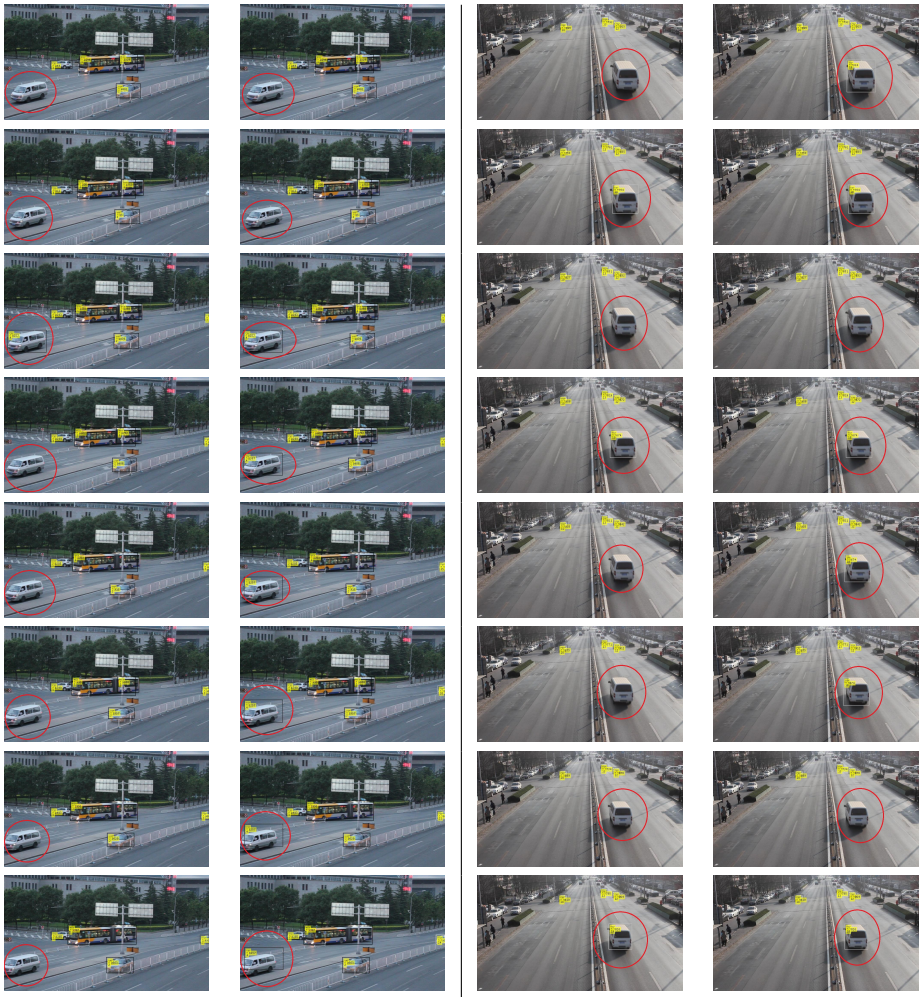


FIGURE 3.14 – Qualitative results of our solution for improving IOU Tracker based Mask R-CNN detector ; the trajectories of the car inside the red circle are improved : (1a)(2a) Original IOU Tracker based Mask R-CNN detector, for sequence 1a we can track only one frame over 8 frames and sequence 2a achieves 3 frames over 8 frames ; (1b)(2b) Improved IOU Tracker based Mask R-CNN detector with Shift generator and SURF matching, both sequences now achieve 6 frames over 8 frames. Those sequences are illustrated in large size in Annexe A.

Chapitre 4

Anomaly detection by future prediction using multi-channels generative framework and supervised learning

Our second work is the main contribution and is directly targeted to solve our ultimate goal : anomaly detection. We distinguish between anomaly recognition and classical action recognition by considering abnormal activities as unpredictable activities. Due to the difficulties of anomaly detection in case of unbalancing scenarios and the unavailable pre-defined spatial-temporal structure of abnormal activities, classical methods have various limitations in practical experiments. By now, most of the state-of-the-art [53, 36] anomaly detection methods are based on apparent motion and appearance reconstruction networks like Generative Adversarial Network (GAN). These methods use the error estimation between generated and real information as a detection feature.

In this thesis, our contributions are two-fold. On the one hand, we propose a flexible multichannel framework to generate multi-type frame-level features. Our method is based on the prediction of the motion and the appearance of image streams and exploits errors of prediction to detect the abnormalities. We define a multi-channel framework based on Conditional GAN (CGAN) to produce feature maps for better-representing appearance and motion. On the other hand, we study how it is possible to improve the detection performance by supervised learning *i.e.* by adding some labeled data and a SVM based classifier.

In the first part of this chapter, we present the generalities about GAN and CGAN and introduce the pix2pix CGAN we implement in our architecture. In the second part, we present

how we add supervised strategy in the last step to improve the detection rate and how the objects related to the abnormality are localized in each frame. The last part is dedicated to the results of our experiments. We begin with the introduction of the public datasets and the evaluation metric we use. Next, we introduce our experimental results in both qualitative and quantitative evaluations with regards to the training parameters. Finally, we analyze our strengths and our limitations regarding to our results.

4.1 Proposed generative backbone architecture

Our general pipeline is illustrated in Figure 4.1. It is based on 4 conditional GANs (CGAN) which inputs are defined in the caption section of the figure. Conditional GAN will be presented later in the chapter. A GAN is a machine learning framework that produces a generative model. It is an unsupervised method that does not need labeled training data. Its goal is to model how the training data look like to be able to generate new examples of what it has learned. In our anomaly detection application, by presenting input from *normal* image sequences, the GAN should be able to construct (or predict) what should be the next image content for *normal* context.

We assume that a situation can be qualified as a normal or abnormal situation by analyzing the objects types moving in the scene and their dynamics. Both informations are carried by optical flow and appearance : thus, our architecture is predicting optical flow and image appearance. To detect abnormality in a sequence, we make the assumption that the prediction (output of the GAN) will sufficiently differ from the *real* image data for the abnormal situations. As introduced in figure 4.1, our architecture is based on conditional GAN that is a type of GAN that involves the conditional generation of images by a generator model [70]. CGAN is better adapted to our application because it uses additional information (annotated input data w.r.t. random noise input of GAN)to generate new predicted data. By training a CGAN on images acquired from "normal" situations, the generative model is able to predict the images and the optical flow at time $t+1$ by inputting images and optical flow acquired at time t. In the first experimental section, we will show that our CGAN based backbone yields state-of-the-art performance.

4.1.1 From GAN to Conditional GAN

Generative Adversarial Nets [25] were recently presented as an effective way to train generative models. For the purpose of taking advantage of labels during the training process, they feed the conditional data to both the generator and the discriminator. The term label here is not the ultimate *abnormal label* but the meaning of pixel of training images. This model is called Conditional GAN (CGAN) [70], a conditional version of generative adversarial nets.

GAN consists of two adversarial models : a generative model G that represents the data distribution, and a discriminative model D that calculates the probability when a sample came from the training data (*i.e.* real sample) rather than G (*i.e.* fake sample). Both G and D could be a non-linear mapping function, such as a multi-layer convolutional neural networks. To learn the distribution p_g over data x, the generator builds a mapping function from a prior noise distribution $p_z(z)$ to data space as $G(z)$. And the discriminator, $D(x)$, produces a single scalar representing the probability when x is real samples rather than p_g. G and D are both trained simultaneously : adjusting parameters for G to minimize $\log(1 - D(G(z))$ and adjusting parameters for D to minimize $\log D(x)$. They are following the two-player min-max game with value function $V(G;D)$:

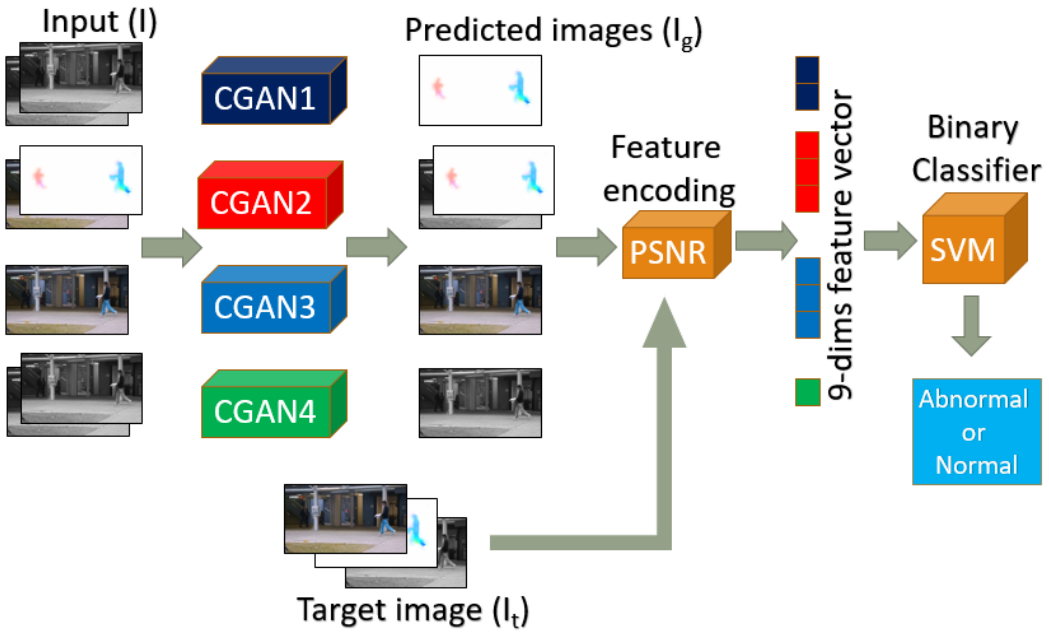


FIGURE 4.1 – Our multi-channels pix2pix-CGANs framework for anomaly detection. In each CGAN stream, the number of channels and the type of each channel for input image (*i.e.* source image I) and output image (*i.e.* generated or predicted image I_g) are different. CGAN1 takes 2 grayscale channels as input and generates 2-dimensional optical flow channels as output. CGAN2 takes 1 grayscale channel and a 2-dimensional optical flow channel as input and generates 1 grayscale channel and 2-dimensional optical flow channel as output. CGAN3 takes a 3-dimensional RGB channels as input and 3-dimensional RGB channel as output. CGAN4 takes 2-dimensional grayscale channels as input and generates 1-dimensional grayscale channel as output. The configuration is described in detail in Table 4.1. The channels of generated images are similar to the channels of ground-truth images (*i.e.* target images I_t). Best viewed in color.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_g(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{4.1}$$

Generative adversarial nets can be extended to a conditional model by letting both the generator and the discriminator be conditioned on some extra information y. The extra y could be any kind of auxiliary information, such as class labels. They can perform the conditioning by feeding y into both the discriminator and the generator as additional input layer. In the generator the prior input noise $p_z(z)$ and y are combined in joint hidden representation, and the adversarial training framework allows for considerable flexibility in how this hidden representation is composed. In the discriminator x and y are presented as inputs and to a discriminative function. In this conditional context, the objective function of a two-player minimax game would be as

following :

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_g(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y)))] \tag{4.2}$$

The Conditional GAN allows us not only to reconstruct from input noise to output image but also to extend to image-to-image translations. It means that we are capable of generating a target image which is not only expected to be similar to a training image but also expected to be similar to the transformation of a training image. This ability of CGAN is the key feature to help us constructing a framework of future predictions. We need the generated image to be similar to the future predicted image, not to the current one. For this reason, we choose CGAN as a suitable model.

Owing to its capacity to generate image, we a take U-Net architecture into account to construct the Generator G of the CGAN model. The U-Net architecture is shortly presented in the next section. The description of our network structure is presented in detail in section 4.1.3.

4.1.2 U-Net architecture

U-Net [87] is the convolutional neural network that was developed by Ronneberger *et al*. for biomedical image segmentation. U-Net is based on the fully convolutional network (FCN) [54] and its architecture was developed to work with less training samples and to obtain a higher segmentation quality.

By replacing pooling layers of usual CNN by upsampling operators, the main idea of U-Net is to supplement a usual contracting network by successive layers. Hence, these layers increase the resolution of the output. Furthermore, a successive convolutional layer can then learn to assemble a finer output based on this information.

There are a large number of feature channels in the upsampling part. This modification allows the network to propagate context information to higher-resolution layers. As a consequence, the expansive path is more or less symmetric to the contracting part and yields a u-shaped architecture (Figure 4.2). The network does not apply any fully connected layers but only uses the valid part of each convolution. The missing context is extrapolated by mirroring the input image to predict the pixels in the border region of the image. This tiling strategy is important to apply the network to large images, since otherwise the resolution would be limited by the GPU memory.

The U-Net architecture contains a contracting path and an expansive path which give it the U-shaped architecture. The contracting path is a typical convolutional neural network with successive typical blocks containing convolutions operations, each followed by a rectified linear unit (ReLU) and a max pooling operation. During the contraction, the spatial information is reduced while feature information is increased. In contrast, the expansive pathway joins the

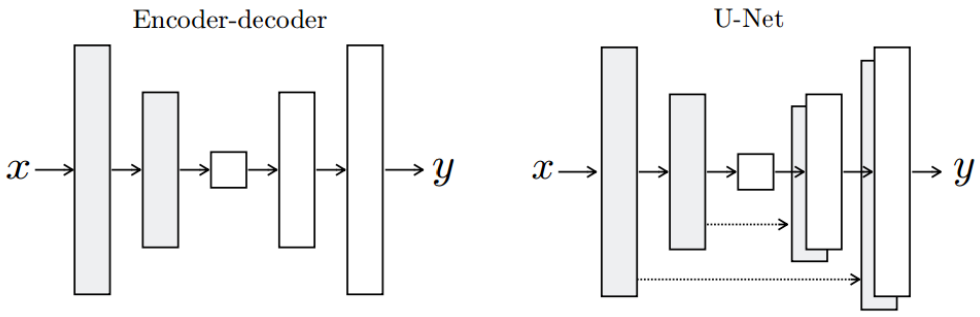


FIGURE 4.2 – Two possible network architecture of CGAN generator : Encode-Decode and U-Net with skip connection [39].

feature and spatial information through a sequence of upsampling operations and concatenations with high-resolution features from the contracting path.

4.1.3 Fundamental of Pix2pix CGAN framework

Image-to-Image Translation with Conditional Adversarial Networks (pix2pix CGANs) is a generative model proposed by Philip Isola *et al*. [39]. They investigated conditional adversarial networks as a general-purpose solution to image-to-image translation problems. These networks not only learn the mapping from input image to output image, but also learn a loss function to train this mapping. This makes it possible to apply the same generic approach to problems that traditionally would require very different loss formulations.

For the network architecture of the generator, many previous solutions to the problems in this area have used an encoder-decoder network (Figure4.2). In such a network, the input was passed through a series of layers that progressively downsample, until a bottleneck layer, at which point the process is reversed. Such a network required that all information flow passes through all the layers, including the bottleneck. For many image translation problems, there is a great deal of low-level information shared between the input and output, and it would be desirable to shuttle this information directly across the net. For example, in the case of image colorization, the input and output share the location of prominent edges. To give the generator a mean to circumvent the bottleneck for information like this, they add skip connections, following the general shape of a "U-Net". Specifically, they add skip connections between each layer i and layer $n-i$, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer $n-i$.

Theoretically, GANs are generative models that learn a mapping from random noise vector x

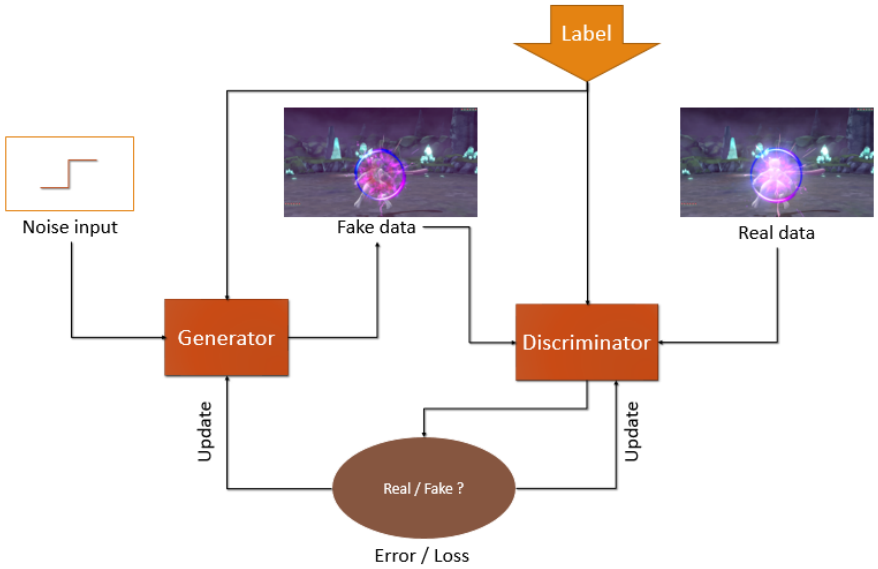


FIGURE 4.3 – Conditional-GAN principle : The generator - Given a label and random noise as input, this model generates data with the same structure as the training data observations corresponding to the same label. The discriminator - Given batches of labeled data containing observations from both the training data and the generated data from the generator, this network attempts to classify the observations as "real" or "fake".

to output image z, $G : x \rightarrow z$. In contrast, conditional GANs learn a mapping from observed image y and random noise vector x, to z, $G : x|y \rightarrow z$. The generator G is trained to produce outputs that cannot be distinguished from “real” images by an adversarially trained discriminator, D, which is trained to do as well as possible at detecting the generator’s “fakes” (Figure 4.3).

The objective of a conditional GAN can be expressed as :

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{y,x}[\log D(x|y)] + \mathbb{E}_{y,z}[\log(1 - D(x, G(z|y)))] \tag{4.3}$$

where G tries to minimize this objective against an adversarial D that tries to maximize it, *i.e.* $G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D)$. To test the importance of conditioning the discriminator, we also compare to an unconditional variant in which the discriminator does not observe y :

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_{z}[\log D(z)] + \mathbb{E}_{y,z}[\log(1 - D(G(z|y))] \tag{4.4}$$

Previous approaches have found it beneficial to mix the GAN objective with a more traditional loss, such as $L2$ distance . The discriminator’s job remains unchanged, but the generator is tasked to not only fool the discriminator but also to be near the ground truth output in an $L2$ sense. We

also explore this option, using L1 distance rather than $L2$ as $L1$ encourages less blurring :

$$\mathcal{L}(G) = \mathbb{E}_{x,y,z}[\|z - G(x|y)\|_1] \tag{4.5}$$

The final objective is

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \tag{4.6}$$

4.1.4 Multi-channel pix2pix-CGAN framework

Our framework combines 4 parallel streams that represent 9 image channels as output. Each CGAN stream is based on a pix2pix-CGANs [39] architecture described in Figure 4.4.

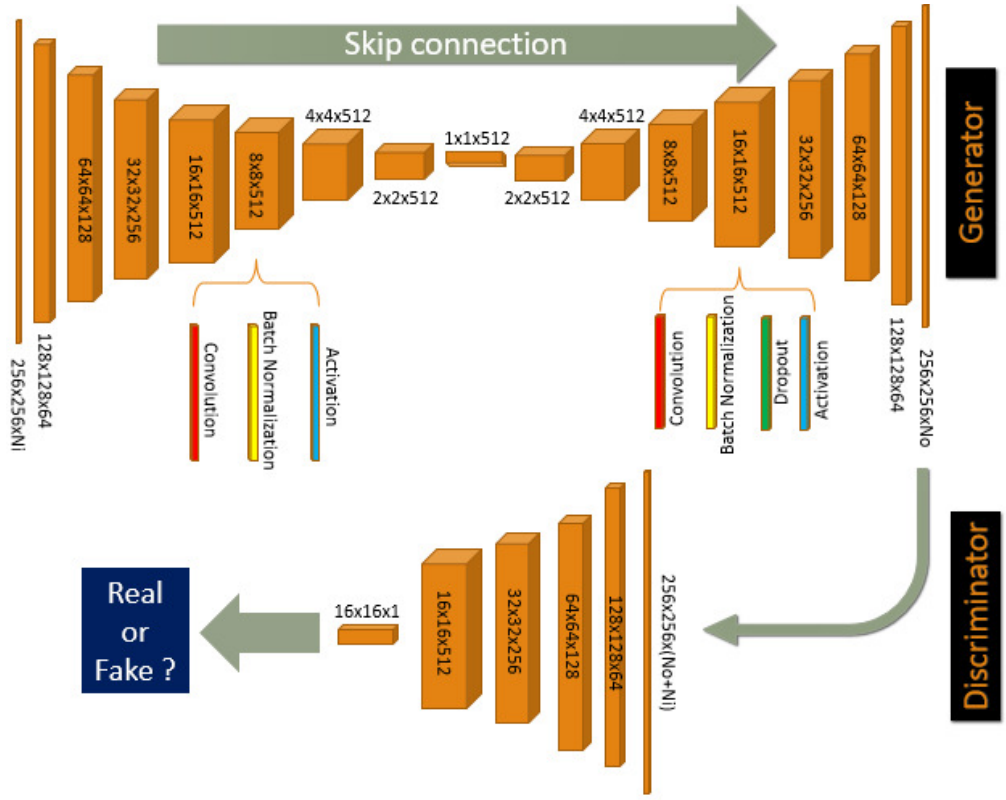


FIGURE 4.4 – Our pix2pix-CGAN architecture. Encoder blocks CEk are from first block $256 \times 256 \times N_i$ to bottleneck block $1 \times 1 \times 512$ of Generator. Decoder blocks CDk are from bottleneck block $1 \times 1 \times 512$ to last block $256 \times 256 \times N_o$ of Generator. N_i, N_o denote the number of input and output image channels of Generator. Discriminator blocks Ck are from block $256 \times 256 \times (N_i + N_o)$ to block $16 \times 16 \times 1$. Because of the skip connection, the total channels pass to Discriminator is $N_i + N_o$. Best viewed in color.

Pix2pix-CGAN : The Generator model is based on an encode-decode U-Net architecture with skipped connections. The Discriminator is based on 4 convolutional blocks to provide a real or

fake decision. Each convolutional block of the Generator encoder and the Discriminator includes a convolution, a batch normalisation and an activation layer. One extra dropout layer with a dropout rate of 50% is added to define one block of the decoder. Let Ck, CEk and CDk denote respectively a Discriminator block, an Encoder block and a Decoder block with k filters. The model architecture is defined as :

— Encoder : CE64-CE128-CE256-CE512-CE512-CE512-CE512-CE512
— Decoder : CD512-CD512-CD512-CD512-CD512-CD256-CD128-CD64
— Discriminator : C64-C128-C256-C512

The last CE512 block only has no batch normalization layer. All the convolutional layers of Encoder are based on 4×4 kernel filters with stride 2 to downsample the input source image to the bottleneck layer. Then, Decoder uses transpose convolutional layers for upsampling from bottleneck output size to the predicted output size. We also add skip connections between the layers of Encode-Decode corresponding to the same size of feature maps. The source image is considered as the input of the Generator and it is concatenated with the target image to produce the first input for the Discriminator. The output image of the Generator concatenated with the source image is fed to the Discriminator as second input. We apply $L1$ loss to measure the distance between target image I_t and generated image I_g from source image I :

$$\mathfrak{L}_G = \mathbb{E}_g \|I_t - I_g\|_1 \tag{4.7}$$

The adversarial loss of Discriminator $\mathfrak{D}$ is calculated using the Conditional GAN strategy ($\mathcal{L}_{cGAN}(G,D)$) :

$$\mathfrak{L}_D = \mathbb{E}_t \log \mathfrak{D}(I_t|I) + \mathbb{E}_g \log(1 - \mathfrak{D}(I_g|I)) \tag{4.8}$$

where $\mathfrak{D}(I_t|I)$ is the discriminator's estimate of the probability that target image I_t is real w.r.t. input image I (*i.e.* the image contains $N_t + N_i$ channels) ; $\mathfrak{D}(I_g|I)$ is the discriminator's estimate of the probability that predicted image I_g is real w.r.t. input I (*i.e.* the image contains $N_o + N_i$ channels).

The final loss is the sum of both loss with ponderation factors λ_G and λ_D :

$$\mathfrak{L} = \lambda_D \mathfrak{L}_D + \lambda_G \mathfrak{L}_G \tag{4.9}$$

In our systems, we apply the default weights for λ_G and λ_D as in the original architecture of [78].

Multi-channel pix2pix-CGAN : In order to achieve richer and more sensitive features of appearance and motion to the anomaly to detect, we propose 4 parallel pix2pix-CGANs with different input and output configuration as mentioned in Table 4.1. We separately investigate the temporal evolution of motion and appearance (in both grayscale and RGB format) by CGAN-1, CGAN-3 and CGAN-4 while CGAN-2 explores the relation between appearance and motion between two consecutive time stamps t and $t+1$. About temporal length modeling, CGAN-1

Channel	Input (I)	N_i	Output I_g & Target I_t	$N_o = N_t$
CGAN-1	G_t, G_{t+1}	2	$F^{x,y}_{t\to t+1}$	2
CGAN-2	G_t, $F^{x,y}_{t\to t+1}$	3	G_{t+1}, $F^{x,y}_{t+1\to t+2}$	3
CGAN-3	RGB_t	3	RGB_{t+1}	3
CGAN-4	G_t, G_{t+1}	2	G_{t+2}	1

Tableau 4.1 – Configuration of input I, output (*i.e.* predicted images I_g) and target images I_t for each pix2pix-CGAN stream. N_i, N_o, N_t denote the dimension of the input, output and target image channels. G_t, $F^{x,y}_{t\to t+1}$ and RGB_t equal to grayscale image at frame t, optical flow from frame t to $t+1$ along axis x, y and RGB color frame at frame t. All channels are taken into account for calculating loss functions.

and CGAN-3 learn from the current frame to the next frame while CGAN-2 and CGAN-4 study the evolution from t to $t+2$.

Before passing into CGAN streams, all source image channels are resized to 256×256 resolution. We also normalise optical flow maps along both x and y axis to range $[0,1]$ by the following equation :

$$F^{norm}_{m,n} = \begin{cases} 0.5 - F_{m,n} \times \dfrac{0.5}{F_{min}}, & \text{for } F_{m,n} \leqslant 0 \\ F_{m,n} \times \dfrac{0.5}{F_{max}} + 0.5, & \text{for } F_{m,n} > 0 \end{cases} \tag{4.10}$$

where $F_{m,n}$ is the flow value at pixel (m,n), F_{max} and F_{min} are the maximum and the minimum value of optical flow over all videos. Due to the optical flow normalization, negative flow values are mapped to $[0, 0.5]$ while positive values are mapped to $[0.5, 1]$. By this way, we can maintain the difference between motion directions. All pix2pix-CGAN streams are then trained by Adam optimizer [43] with learning rate of 0.0002 and momentum parameter $\beta_1 = 0.5$, $\beta_2 = 0.999$ for both the Generator and the Discriminator. We test with different activation functions (LeakyReLU, ClippedReLU, eLU) and other hyper-parameters such as mini-batch size to find the most suitable parameters values for each stream and dataset.

4.1.5 Feature extraction with PSNR

Each CGAN predicts the outputs for which it has been trained as described in the table 4.1. Then it is possible to measure the difference between the predicted output and the target data to decide if the situation is different of what it should be. Generally, in order to measure the distance between two images, we can use two metrics : Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR). In signal processing, PSNR is an expression defined by the ratio between the maximum possible power of a signal and the power of distorting noise that affects the quality of its representation. In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of a predictor or an estimator equals to the average of the squares of the errors *i.e.* the average squared difference between the predicted/estimated values and the real value.

Given the original image I_t containing N pixels without noise and its noisy version I_g, MSE and PSNR are calculated as :

$$MSE(I, I_s) = \frac{1}{N} \sum_{i=0}^{N} (I_t^i - I_g^i)^2 \tag{4.11}$$

$$PSNR(I_t, I_g) = \frac{[I_{max}^i]^2}{MSE(I_t, I_g)} \tag{4.12}$$

where we denote I_{max}^i as the maximum value of a pixel of original image I.

The work of Mathieu *et al*. [68] shows that PSNR is a promising way to compare the quality of a target image I_t with a generated I_g, a higher PSNR value indicating a generated image conformed to the target image we reach. The authors define the PSNR between I_t and I_g from equation 4.12 but in db as :

$$PSNR(I_t, I_g)(dB) = 10 \log_{10} \frac{[max(I_t)]^2}{\frac{1}{N} \sum_{i=0}^{N} (I_t^i - I_g^i)^2} \tag{4.13}$$

We follow the conclusions of Mathieu *et al*. and we apply PSNR metric to evaluate the difference between the generated output and the target one for each CGAN. The generated output are the predicted optical flows and the images predicted by each of the CGAN. The target data are the available measures i.e. the acquired frames and the optical flow calculated from two consecutive frames. Obviously, if we accumulate all PSNR values along all channels and all streams to learn a threshold, we might lose the benefits carried by each stream. So we apply a late fusion strategy by separately calculating PSNR for each one and we decide by analysing the obtained equivalent PSNR vector.

Our 4-streams architecture provides 9 output channels. Thus, we obtain 9 PSNR values encoded as 9-dimensional features vector. The vectors are normalised to range $[0, 1]$ for all 9 dimensions in all videos sequences. By considering all the training sequences, we obtain a feature map that defines a sub-space that corresponds to the 9-dimensional vectors related to normal situations. Consequently, each new features vector of this sub-space (*i.e*. a vector obtained from an unseen image sequence), could be linked to a normal event.

4.1.6 CGANs backbone loss evaluation

In this section, we analyse the quality of the training step through 4 popular datasets used for anomaly detection in the state-of-the-art : CUHK Avenue [58], USCD Pedestrian 1 and 2 [66] and ShanghaiTech [53]. Some samples of those datasets are illustrated in Figure 4.5. First of all, we briefly introduce those benchmarks. Then we present our step by step implementation of the multi-CGANs backbone.

Datasets and evaluation metric

CUHK Avenue : It contains 16 training sequences with some outliers and 21 testing videos containing 47 irregular events as throwing objects, loitering and running. The size of the people is changing because of the camera position and angle. The normal samples of the test set are more numerous than abnormal ones.

Several outlier frames exist in Avenue training set as presented in [58]. All the state-of-the-art works that use this dataset [53, 72, 36] have manually removed these outliers to avoid training the CGANs for normal cases with abnormal input data. We apply this task to be fair for comparison.

USCD Ped1 & Ped2 : Ped1 has 34 training and 36 testing videos with 40 abnormal events. Ped2 is smaller with 16 training and 12 testing videos. Almost abnormal events are related to moving vehicles. Ped1 seems to be more challenging than Ped2 due to the different camera angles used. Both have more abnormal events than normal ones in their test sets.

ShanghaiTech : This is a very large benchmark containing 13 scenes integrating complex lightning conditions and camera angles. There are 130 abnormal events and over 270000 training frames. Moreover, pixel level ground truth of abnormal events is also annotated. Normal samples are more numerous than abnormal ones in the test set.

FIGURE 4.5 – Illustration of several samples in CUHK-Avenue dataset [58], USCD-Pedestrian datasets [66] and ShanghaiTech dataset [53]. First row : Avenue dataset. Second row : Pedestrian dataset. Third row : ShanghaiTech dataset. ShanghaiTech is more challenging than other due to multiple captured views and flexible abnormal samples.

Implementation details

Implementation frameworks : To build each Pix2pix CGAN stream, we implement our architecture based on framework [78] on Matlab. Optical flow images are needed as input and as output (Ground-truth) during training. Optical flow is extracted by applying the Full Flow technique [13] for Avenue, Ped1 and Ped2 datasets. Because ShanghaiTech is much larger, we use the simple Lucas-Kanade [61] optical flow algorithm implemented in OpenCV to reduce time processing.

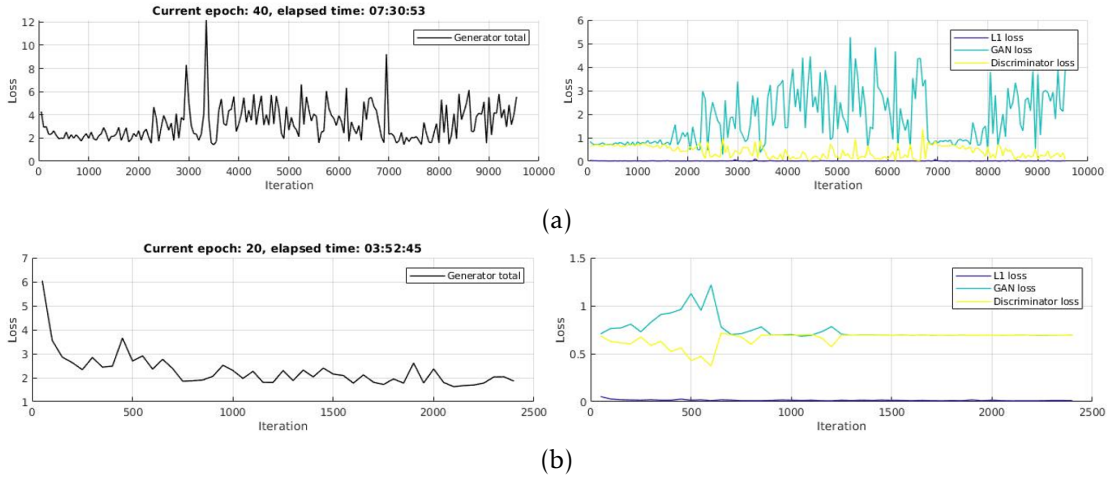


FIGURE 4.6 – Comparison of loss convergence between different parameters for CGAN-2 on Avenue dataset. Figure (a) shows the case where the loss are not convergent after 40 epochs, Minibatch size 64 and leakyReLU activation function while Figure (b) illustrates a good set of parameters $E = 20, M = 128, A = eLU$ corresponding to a good loss convergence. We find that the total loss is stable from the 1000^{th} iteration and each loss is almost convergent. Best viewed in color.

Loss evaluation during the CGAN training : For each CGAN stream, we investigate the effect of the Mini-batch size (M), the activation function (A) and the number of training epochs (E). We start from E = 20 for Avenue, Ped1 and Ped2 with M = $\{32, 64, 128, 256\}$ and A = $\{leakyRELU(\alpha = 0.2), clippedReLU(\alpha = 0.5), eLU(\alpha = 1)\}$. Then we increase E up to 30, 40 and observe the convergence of each loss function to choose the most suitable parameters according to the balance between time consumption and loss convergences. An example is illustrated on Figure 4.6. Particularly, because ShanghaiTech has a very large training set, we train at only 1 epoch and optimize parameters for CGAN-2 then adapt those parameters for the other CGANs to reduce processing time. We choose CGAN-2 as the reference stream for all optimization processes because CGAN-2 can learn the relation between both appearance and motion evolution. The final training results for all CGAN-streams are illustrated in Figures 4.7, 4.8, 4.9 and 4.10. Due

to the balance of layout for illustration of each part and for the purpose of clarity, we need to keep the figures as near as possible to their corresponding parts, so for the other datasets, the training convergences are shown in Annexe A.

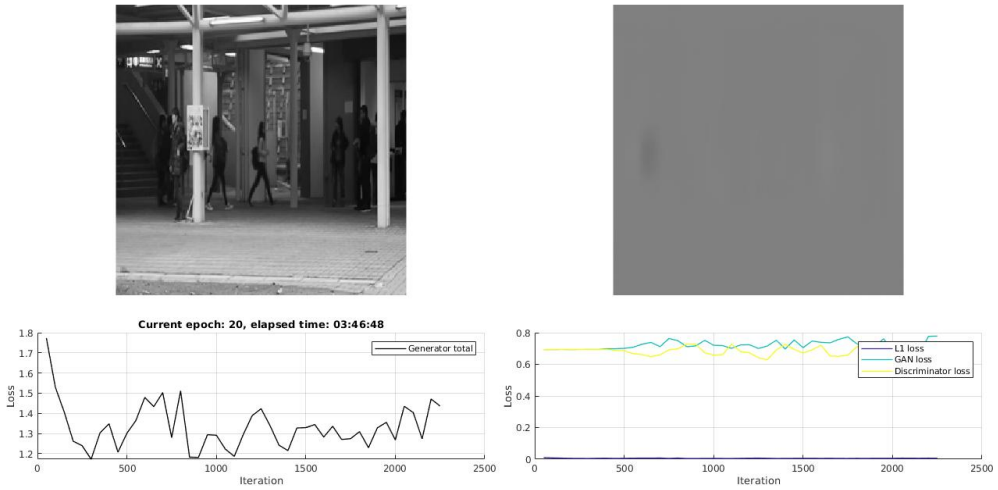


FIGURE 4.7 – Illustration of training CGAN-1 on Avenue dataset. We achieve good loss convergences for epoch $E = 20$, mini batch size $M = 128$ and activation function $A = eLU$. The total loss gets rapidly small and each loss is convergent from early iterations.

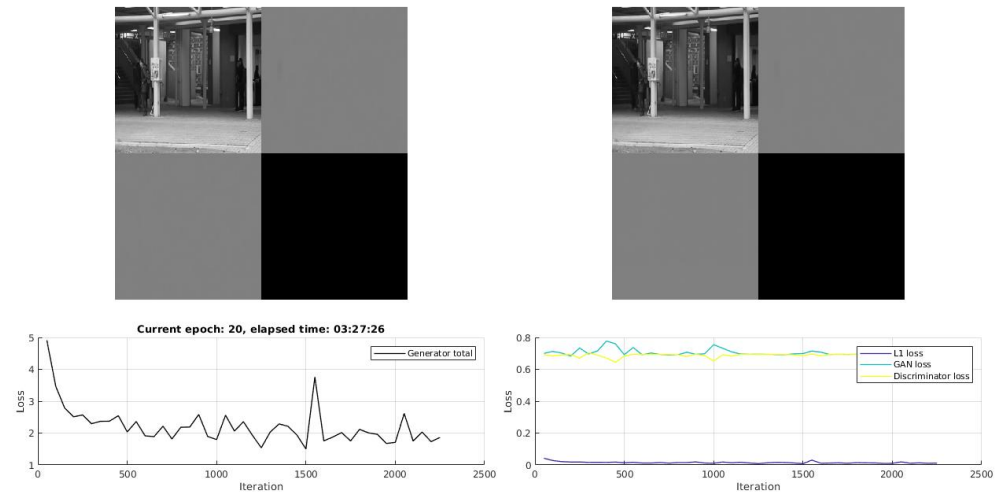


FIGURE 4.8 – Illustration of training CGAN-2 on Avenue dataset. We achieve good loss convergences at $E = 20$, $M = 128$ and $A = eLU$. We find that the total loss is stable from the 1700^{th} iteration while each loss is almost convergent from early iterations.

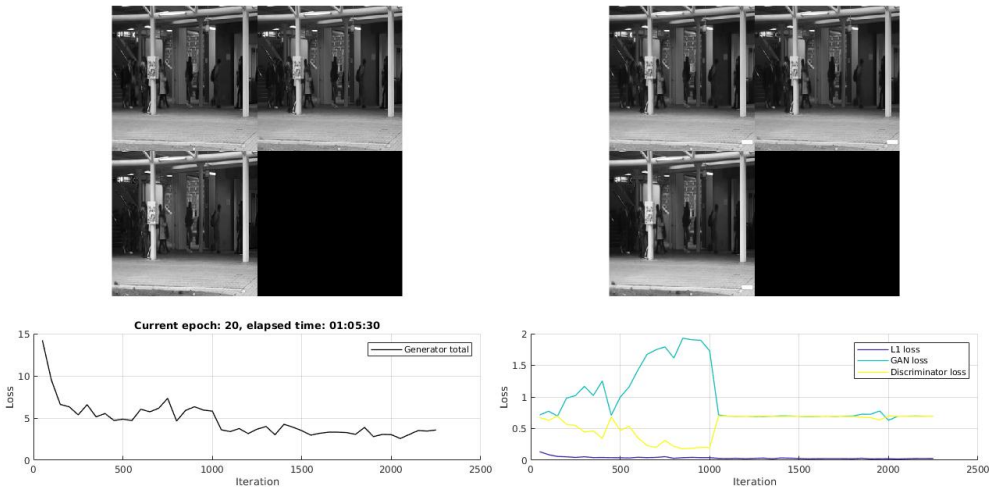


FIGURE 4.9 – Illustration of training CGAN-3 on Avenue dataset. We achieve good loss convergences at $E = 20$, $M = 128$ and $A = eLU$. We find that the total loss is stable and each loss is almost convergent from the 1000^{th} iteration.

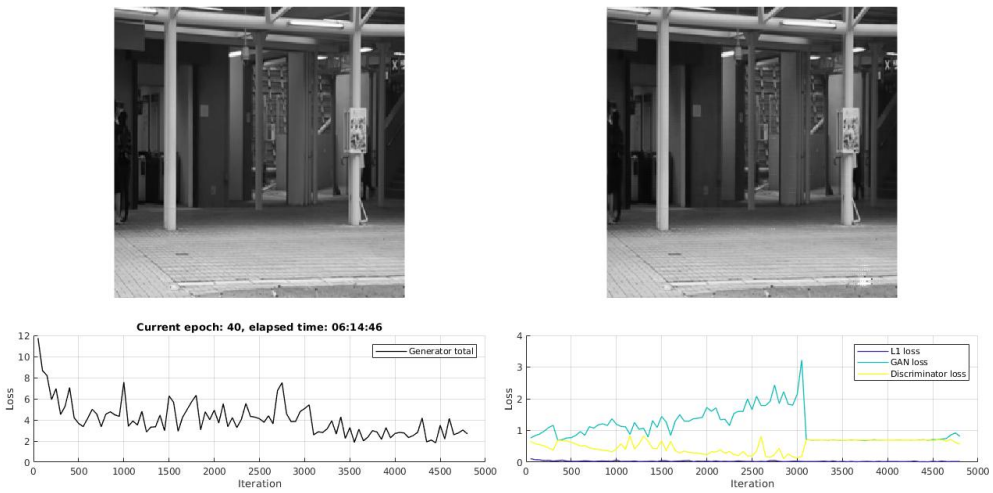


FIGURE 4.10 – Illustration of training CGAN-4 on Avenue dataset. We achieve good loss convergences at $E = 20$, $M = 128$ and $A = eLU$. We find that the total loss is stable and each loss is almost convergent from the 3000^{th} iteration.

Outliers removing on Avenue : As previously presented, some outlier frames have been manually removed from the training set to avoid degrading loss function convergence. As a first feasibility evaluation to detect abnormality, we use the trained CGAN-2 to automatically detect these outliers. For that, we accumulate MSE values calculated between the predicted output of

the CGAN-2 and the target data for the 3 channels (*i.e.* the 2-dimensional optical flow and the gray image) and we draw the values as presented on figure Figure 4.11. We can easily observe some peaks occurring exactly at the abnormal frames appearing on the training videos and that have been manually removed. This result shows clearly that CGAN-2 is able to highlight parts of a sequence that differ from the normal cases for which the CGAN-2 has been trained. But if the CGAN-2 has been sufficient for detecting the outliers it is because, it has been trained on the same sequence in which we aim at detecting the outliers. In an operational context, the goal is to detect the abnormalities in an unknown image sequence. To propose a better classification model, we have to analyse the distribution of features vectors for normal and abnormal frames. This analysis is presented in the next section.

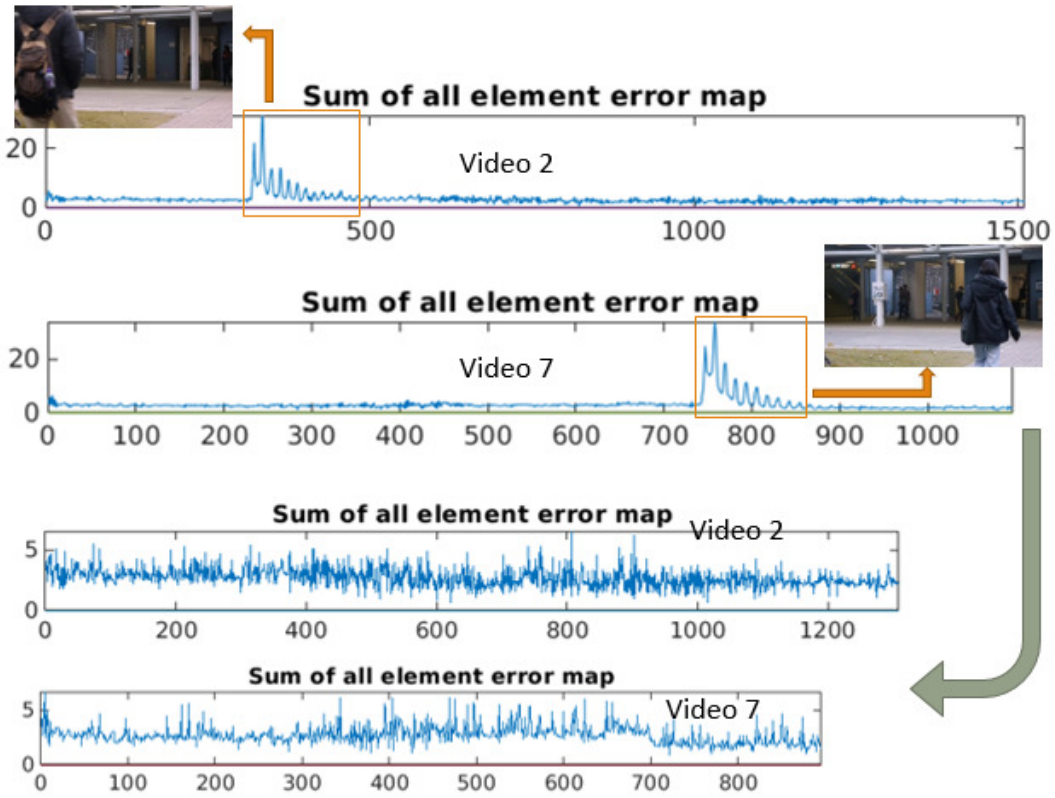


Figure 4.11 – Example of outliers removing on Avenue dataset. In order not to do it manually, we train CGAN-2 on training set to produce output images. Then we accumulate MSE values between output and ground-truth images for all of the 3 channels and we draw the corresponding values. We can easily observe some peaks occurring at the abnormal frames on training videos. We remove the outlier images corresponding to the peaks to obtain a new clean Avenue training set. Best viewed in color.

Qualitative evaluations : To measure the efficiency of the features vectors to discriminate

a normal and an abnormal frame, we visualise the distribution of feature vectors extracted from test samples on each dataset. For the visualisation to be possible, we do it for each of the CGAN separately. Figures 4.12, 4.13, 4.14 and 4.15 show the distribution of test vectors infer for respectively Avenue, Ped1, Ped2 and ShanghaiTech dataset. The axes of each figures denote the value of PSNR before normalization. Because we use PSNR based feature vector, by definition normal and abnormal samples are respectively associated to high and low values.

Ped1 and Ped2 datasets contain only grayscale images. Thus, CGAN-3 that is based on RGB images is reduced to gray level content and we observe a straight line distribution. All other points plots respect the dimension given in table 4.1.

For all plots we observe that two clusters appear in the distributions. Unfortunately, it is clear that both of these clusters are not linearly separable in this raw space : normal and abnormal points distribution partly occupied the same part of the space. Figure 4.16 is another illustration of this non-linearly separation problem for each CGAN.

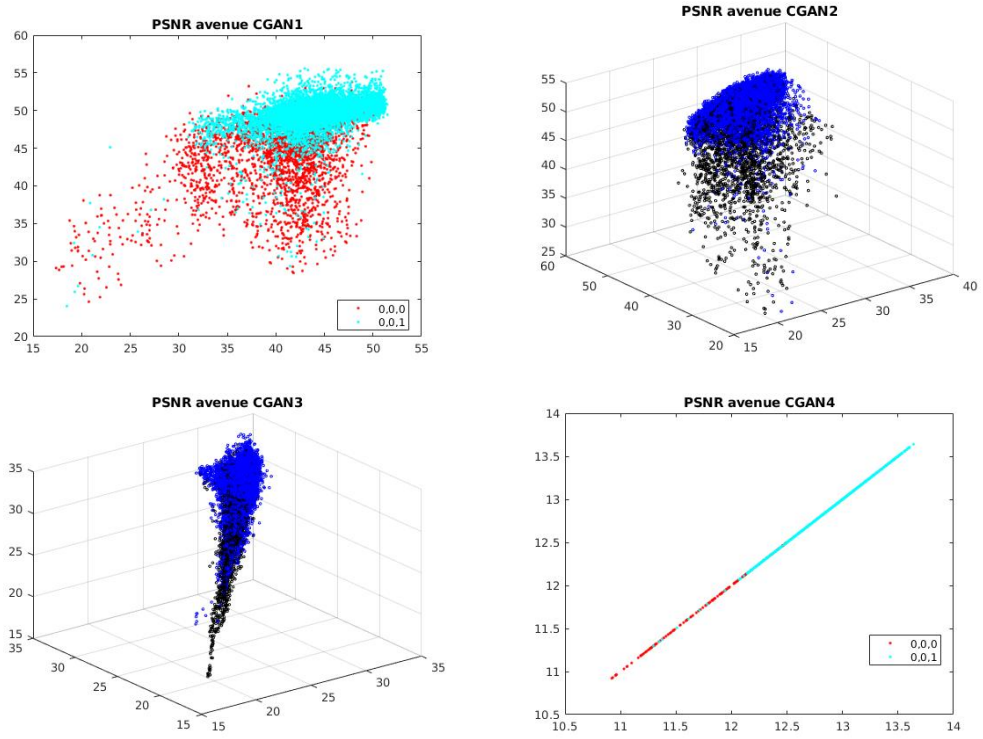


Figure 4.12 – Distribution of normal and abnormal samples in Avenue dataset. The axes of each figure denote the value of PSNR before normalization. For CGAN-2 and CGAN-3, normal points are blue and abnormal points are black. For CGAN-1 and CGAN4, normal = light blue and abnormal = red.

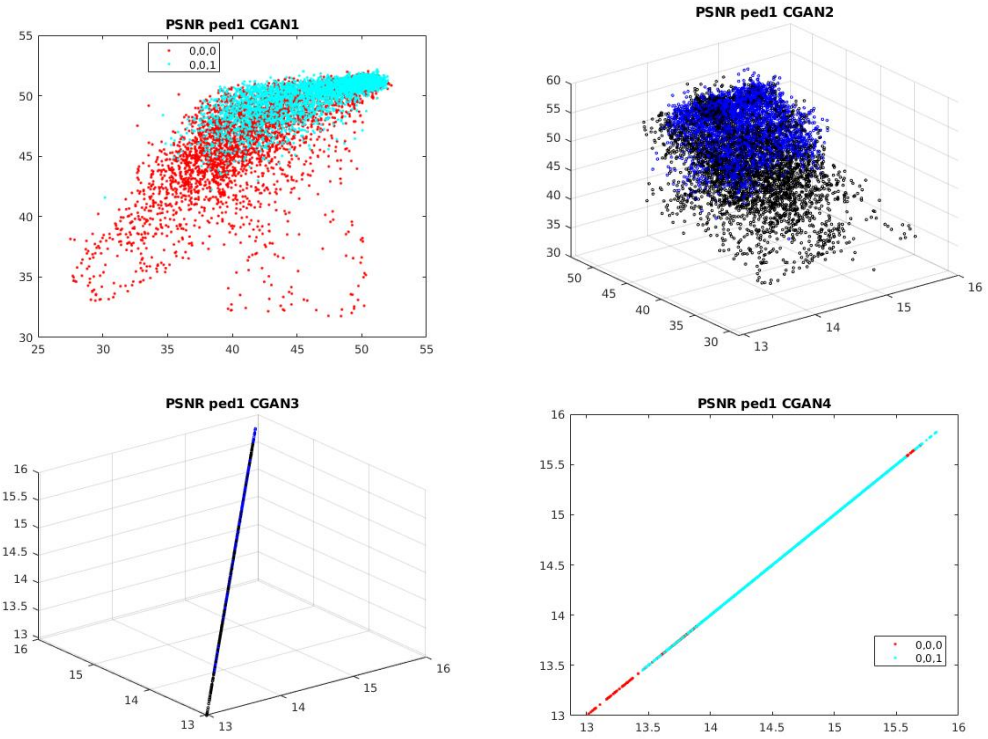


FIGURE 4.13 – Distribution of normal and abnormal samples in Ped1 dataset. The axes of each figure denote the value of PSNR before normalization. For CGAN-2 and CGAN-3, normal points are blue and abnormal points are black. For CGAN-1 and CGAN4, normal = light blue and abnormal = red.

The multi-channel CGANs backbone defines a feature space in which the images content is projected and in which abnormality can be detected. Generally, the SOTA proposed two natural approaches for assigning abnormal labels for image samples as described in works applying future prediction approach [53, 72, 52, 36, 101]. On the one hand, an abnormal score is computed by thresholding the error map [53, 72, 101]. On the other hand, authors propose to use learning model to classify the input into normal and abnormal classes [36, 52].

Table 4.2 shows the performance for both kind of methods. It clearly appears that the methods based on learning classifiers yield better results even on challenging dataset such as Avenue and ShanghaiTech.

In our work, we propose to train a Support Vector Machine to separate the normal and abnormal features vectors. From this point, we leave the unsupervised learning to adopt a semi-supervised strategy. The next section describes this part of our pipeline and the evaluation results we obtained for all the selected datasets.

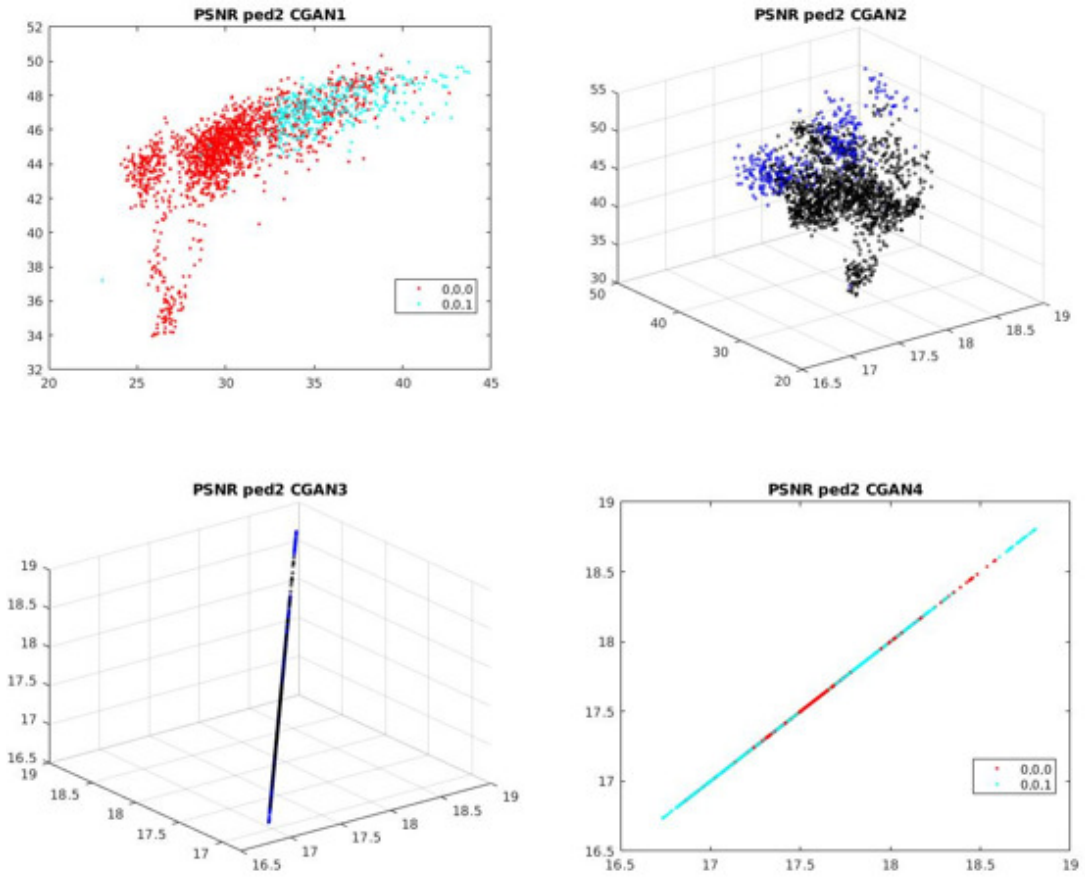


FIGURE 4.14 – Distribution of normal and abnormal samples in Ped2 dataset. The axes of each figures denote the value of PSNR before normalization. For CGAN-2 and CGAN-3, normal points are blue and abnormal points are black. For CGAN-1 and CGAN4, normal = light blue and abnormal = red.

4.2 Abnormality detection by Support Vector Machine

In this second part of the work, we describe how we can perform abnormality detection from the PSNR based feature map inferred from our multi-CGANs backbone. The analysis of the distribution proposed in the previous section clearly shows that both class cannot be linearly separated. Thus, we proposed to use a kernel based technique to discriminate the two classes. In this work, we propose to use Support Vector Machine (SVM) by following a supervised strategy. A binary SVM is used as the final layer to design the supervised classifier. It takes the feature maps provided by our network backbone as input after a PSNR transformation. In this section we show that a SVM is able to accurately detect the abnormal frames of a video. We go further and our experimental results will show how the accuracy is improved with regards to the size of

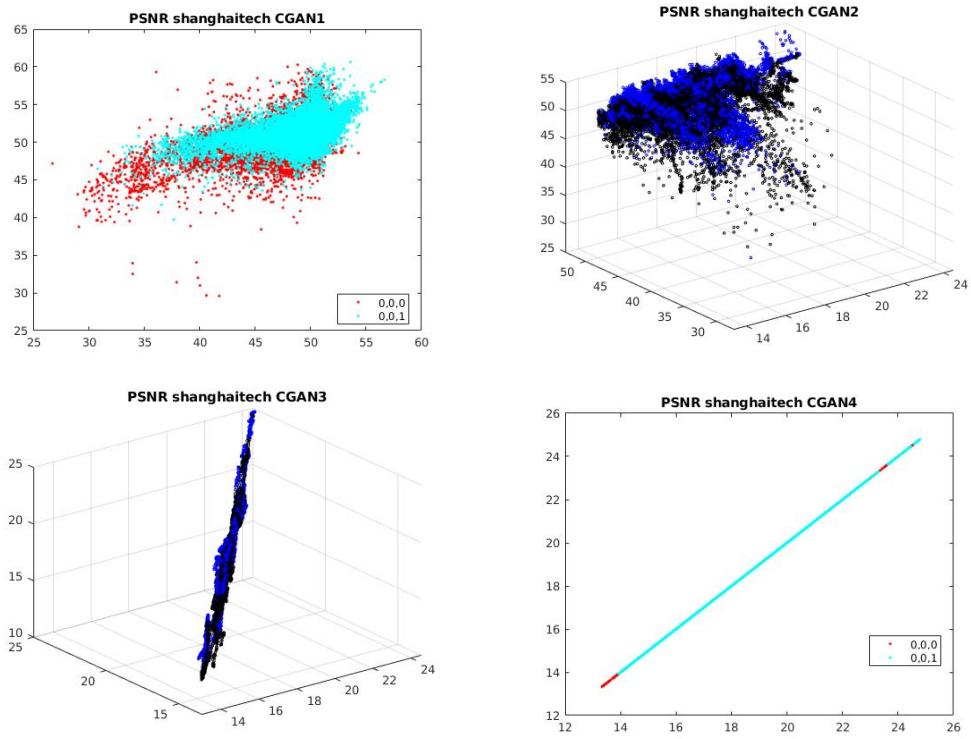


FIGURE 4.15 – Distribution of normal and abnormal samples in ShanghaiTech dataset. The axes of each figure denote the value of PSNR before normalization. For CGAN-2 and CGAN-3, normal points are blue and abnormal points are black. For CGAN-1 and CGAN4, normal = light blue and abnormal = red.

the dataset used to train the SVM. We finally propose a framework to localise the objects of the scene that generate the detected anomaly.

The next two sub-sections are respectively dedicated to the description of this SVM layer (section 4.2.1) and of the abnormal object localisation ((section 4.2.2)). In the last sub-section (section 4.2.3) are presented the quantitative evaluation of the SVM based abnormality detection and the quantitative evaluation of abnormal object localisation task.

4.2.1 SVM based frame-level anomaly detection

In [36], Ionescu et al. "...believe that including any form of supervision is an important step towards obtaining better performance...". In this paper, the authors propose to train a supervised one-versus-rest classifier on feature maps representing different kinds of normality defined by different clusters obtained on the training samples. On the contrary, instead of training SVMs with only normal samples from training dataset, Liu *et al.* proposed an interesting alternative supervised scenario [52]. Classical semi-supervised video anomaly detection assumes that only

	Methods	Avenue	Ped1	Ped2	SHT
Thresholding methods	Liu *et al.* [53]	0.85	**0.83**	0.95	0.73
	Nguyen *et al.* [72]	0.87		0.96	
	Vu *et al.* [101]	0.72	0.82	**0.99**	
Learning classifiers methods	Ionescu *et al.* [36]	0.90		0.98	**0.85**
	Liu *et al.* [52]	**0.93**			0.77

Tableau 4.2 – Comparison of Frame level AUC on 4 datasets between simple thresholding inference models and complex learning inference models. All reported methods constructed their features space by future prediction networks.

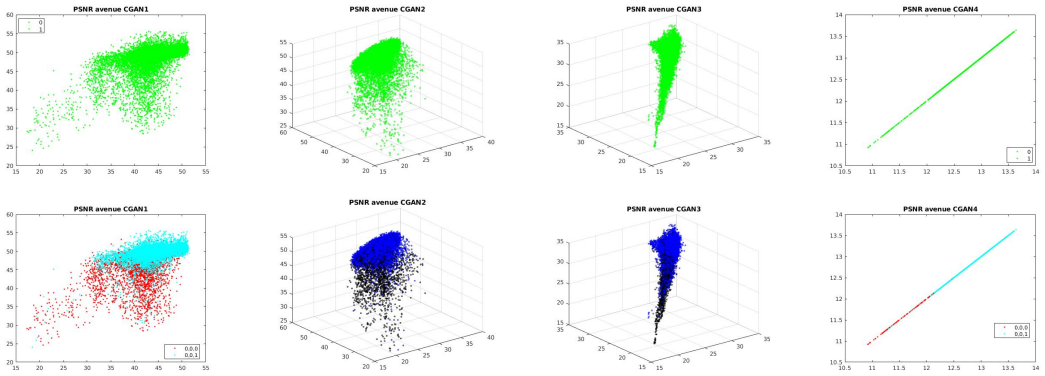


Figure 4.16 – Distribution of testing samples in Avenue dataset corresponding to our CGANs feature space. The first row illustrates the distribution of all samples by green point. The second row separates the distribution of abnormal and normal samples by different color : normal = blue ; abnormal = black or red.

normal data are available in the training set because of the rare and unbounded nature of anomalies. It is obvious that these infrequently observed abnormal events can actually help detecting identical or similar abnormal events when taken into account during he training. This is a line of thinking that motivates us to study open-set supervised anomaly detection with only a few types of abnormal observed events and many available normal events.

In this work, we use the feature maps provided by PSNR values computed on unsupervised multi-channel pix2pix CGANs output to train a supervised binary SVM classifier. As described previously, the training parts of the used datasets contain only normal samples. Because training the SVM requires negative and positive samples, we split each test set into two parts : a first part for learning and a second part for testing the performance of the trained model. The test part of the used dataset contains both abnormal and normal samples. During the evaluation, we analyze how the performance is modified regarding the size of the dataset used for training i.e. regarding the used test set ratio. We aim at defining a trade-off between the performance

improvement and the amount of work required for the annotation task.

In detail, by extracting a part of the test set from one dataset, we have two typical classes for each frame : one frame can be normal or abnormal. Each frame is represented by its features vector i.e. the PSNR feature vector extracted from the output of the multi-CGANs backbone. The full size of features vector is 9-dimensional corresponding to the 9-streams of CGANs when all the streams are considered. During the experiments, we analyse the performance by using each stream separately and by combining them. In practice, the feature vector can be from 1-dimensional to 9-dimensional vector depending on which combinations are evaluated. Before training the binary SVM classifier, all features vectors are normalised.

Support Vector Machine description

Theoretically, the SVM based binary classification algorithm allows to categorise an unseen object into two separate groups. This classification is based on the properties of the objects and of a set of known samples, which are already classified and that has been used to construct the SVM model. The properties are defined by a n-dimensional features vector, and the model is achieved by creating a linear partition of the feature space into the two classes. An unseen sample is classified by calculating the position of the features vector relatively to the linear model. SVM is not only restricted to linear the discriminative problem. Thanks to a technique known as the kernel trick, it is possible to estimate various types of non-linear decision boundaries. This trick performs a mapping of the original feature space to a higher-dimensional space, in which the problem of binary classification becomes near a linearly separable one. However, even if a kernel is applied, the boundary is always defined in the original space by the training set of feature vectors.

In a n-dimensional space the SVM algorithm constructs an optimal hyperplane that discriminates the samples from two classes, and the optimal hyperplane equation is given by :

$$\boldsymbol{x\beta + b = 0} \tag{4.14}$$

where :

— $\boldsymbol{x}$ is a sample vector

— $\boldsymbol{\beta}$ contains the coefficients that define an orthogonal vector to the hyperplane.

— $\boldsymbol{b}$ is the bias term

This hyperplane splits the feature space into two parts for which the sign of the function $\boldsymbol{f(x) = x\beta + b}$ is negative or positive. Once the model is obtained, classifying a new sample consists in determining the part of the space where its feature vector is located.

The SVM model, defined by parameters $\boldsymbol{\beta}$ and $\boldsymbol{b}$, is estimated by applying a training algorithm on a set of positive and negative training samples $(\vec{x_i}, y_i)$ where $\vec{x_i}$ is the feature vector of the i^{th} sample. y_i is equal to $\mathbf{1}$ or $\mathbf{-1}$ with regards to the class to which the sample belongs.

The training algorithm finds and sets up the maximum margin length while keeping samples in the positive ($y = 1$) or negative ($y = -1$) part of the space. The optimal hyperplane is separating the two sets of samples such as it is the farthest from any training observations. The smallest perpendicular distance to a training observation from the hyperplane is known as the margin. The optimal hyperplane depends on the margin defined by a subset of the positive and negative training samples known as *support vectors*. It appears clearly that the optimal hyperplane and thus the classification accuracy depends on the support vectors.

The SVM training is based on an optimization and relies on a primal and a dual formalization of the SVM problem.

— For separable samples, the objective function is defined by solving $\mathbf{arg\,min}\|\beta\|$ with respect to the β and b such that $y_j f(x_j) \geq 1$, for all $j = 1, .., n$.

— For non-separable samples, the objective function is defined by solving $\mathbf{arg\,min}(0.5\|\beta\|^2 + C\sum\theta_j)$ with respect to the β, b, and θ_j such that $y_j f(x_j) \geq 1 - \theta_j$ and $\theta_j \geq 0$ for all $j = 1, .., n$ and for a positive scalar box constraint C. The algorithm applies slack variables (θ_j) to penalize the objective function for samples that cross the margin boundary for their class. $\theta_j = 0$ for samples that do not cross the margin boundary for their class, otherwise $\theta_j \geq 0$.

The algorithm applies the Lagrange multipliers to optimize the objective function, which presents n coefficients $\alpha_1, .., \alpha_n$. For both separable and non-separable cases, the dual formalization for a linear SVM is as follows :

— For separable samples,

$$L = \mathbf{arg\,min}(0.5\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\alpha_k y_j y_k x_j x_k - \sum_{j=1}^{n}\alpha_j) \tag{4.15}$$

w.r.t. $\alpha_1, .., \alpha_n$ subject to $\sum\alpha_j y_j = 0$, $\alpha_j \geq 0$ for all $j = 1, .., n$ and Karush-Kuhn-Tucker (KKT) Complementarity Conditions.

— For non-separable samples, the objective function is similar to separable classes, except for the additional condition $0 \leq \alpha_j \leq C$ for all $j = 1, .., n$.

The final resulting score function is :

$$\tilde{f}(x) = \sum_{j=1}^{n}\tilde{\alpha}_j y_j x x_j + \tilde{b} \tag{4.16}$$

where $\tilde{b}$ is the estimate of the bias and $\tilde{\alpha}_j$ is the j^{th} estimate of the vector $\tilde{\alpha}$ for $j = 1, ..., n$. Written this way, the score function is free of the estimate of β as a result of the primal formalization.

When classes discrimination requires the non-linear boundary, SVM is looking for an optimal separating hyperplane but in a transformed predictor space obtained by applying a kernel trick.

This extension to the non-linear case results in increasing the feature space dimension through the use of functions known as kernels ϕ. For nonlinear SVM, the algorithm constructs a Gram matrix using the rows of the predictor data X. The dual formalization replaces the inner product of the observations in X with corresponding elements of the resulting Gram matrix. The Gram matrix of a set of n vectors $x_1, .., x_n; x_j \in R^p$ is an $n - by - n$ matrix with element (j, k) defined as $G(x_j, x_k) =< \phi(x_j), \phi(x_k) >$, an inner product of the transformed predictors using the kernel function ϕ. Various kernel functions exist like the following :

- Gaussian kernel : $G(x_j, x_k) = exp(-\frac{\|x_j - x_k\|^2}{2\sigma^2})$
- Radial basis function for $\gamma > 0$: $G(x_j, x_k) = exp(-\gamma\|x_j - x_k\|^2)$
- Linear kernel : $G(x_j, x_k) = x_j x_k + c$
- Polynomial kernel for integer $q > 0$: $G(x_j, x_k) = (\alpha x_j x_k + c)^q$.
- Sigmoid kernel : $G(x_j, y_k) = \tanh(\alpha x_j x_k + c)$

In this case, the dual formalization for nonlinear SVM is :

$$L = \arg\min(0.5 \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k G(x_j, x_k) - \sum_{j=1}^{n} \alpha_j) \tag{4.17}$$

w.r.t. $\alpha_1, .., \alpha_n$ subject to $\sum \alpha_j y_j = 0$, $0 \leq \alpha_j \leq C$ for all $j = 1, .., n$ and Karush-Kuhn-Tucker (KKT) Complementarity Conditions. $G(x_j, x_k)$ are elements of Gram matrix. By this way, the final resulting score function is :

$$\tilde{f}(x) = \sum_{j=1}^{n} \tilde{\alpha}_j y_j G(x, x_j) + \tilde{b} \tag{4.18}$$

The KKT complementarity conditions are optimization constraints required for optimal nonlinear programming solutions. In SVM, the KKT complementarity conditions are :

$$\begin{cases} \alpha_j [y_j f(x_j) - 1 + \theta_j] = 0 \\ \theta_j (C - \alpha_j) = 0 \end{cases} \tag{4.19}$$

for all $j = 1, .., n$ where $f(x_j) = \phi(x_j)\beta + b$ with $\phi(x_j)$ is a kernel function and θ_j is a slack variable. If the classes are fully distinguishable, then $\theta_j = 0$ for all $j = 1, .., n$.

As it will be presented in the next section about evaluation results, we have evaluated different kernels to reach the best results. To enhance performance of SVM classifiers, we also apply $k - fold$ cross-validation strategy. This strategy proposed to divide the training dataset into k subsets and to apply the following procedure for each of the k subsets :

- A model is trained using $k - 1$ of the folds as training data ;
- The model performance is evaluated on the remaining subset (*i.e.* the one that has not been used for training)

The performance measure reported by k-fold cross-validation is then the average of the values

computed in the loop. In our experiment, we set $\boldsymbol{k} = \mathbf{5}$.

4.2.2 Abnormality object localisation

The objective of this last task is to localise the object related to the abnormal event detected at the frame-level. For each abnormal frame, we run a fast detector to compute bounding boxes (BB) for each object in the sequence. Each BB is considered as a new input and a PSNR score is computed for each one by taking the normal sum of 9 PSNR scores corresponding to 9 channels. To decide if one BB is or not one object that is producing the anomaly, we assume that heat region appears at the position of the BB in the error map. Thus, the BBs yielding the minimum values for PSNR scores or values smaller than a threshold refer to the abnormality objects.

After calculating PSNR score for each CGAN stream to encode feature vectors, we learn binary SVM classifier using the Classification Learner Toolbox on Matlab.

To reduce the computational complexity in practical experiments, we do not re-pass each BB throughout the whole framework but we apply directly each BB to the error maps. The error map of each frame is the subtraction of the generated images (optical flow and gray or color images) and the corresponding target frames as illustrated in Figure 4.17.

We compare our method with the state-of-the-art object-centric model (Figure 2.17) proposed by Ionescu *et al*. [36]. Ionescu *et al*. propose to extract object-centric features at the first step and then use those features for the rest of their pipeline. Their work helps them in neutralising the effect of background information which is usually stable throughout the sequence and highlights the distinctions around the objects. In a trade-off, their method increases the complexity of feature space and time consumption for every task. In our side, we apply the object detector after inferring the CGAN streams and after classifying normal/abnormal frames by SVM. Obviously, normal frames appear more often than abnormal ones. Hence, we avoid object detection task in normal frames and thus significantly reduce time processing. Second, we apply a simple PSNR score for both frame-level and object localization model instead of applying K-means clustering to generating the normal classes for unsupervised SVM as their method [36]. By reusing the error maps at frame-level classification stage, we accelerate our process by directly calculating PSNR score of bounding boxes on the error maps. Obviously, our methods look more natural because we apply global features for the global task (*i.e*. frame-level classification) and object-centric metrics for abnormality object localization.

Generally, this framework applies a hard-decision strategy to localize the abnormal objects. It leads to the issue that the final performance highly depends on the accuracy of the bounding box detector. The missing detections or confusing detections can affect the final decision. Another problem is how to pre-define the number of abnormal objects in cases of hard-decision. In popular datasets, this number should be one or two objects. Then we could search for two

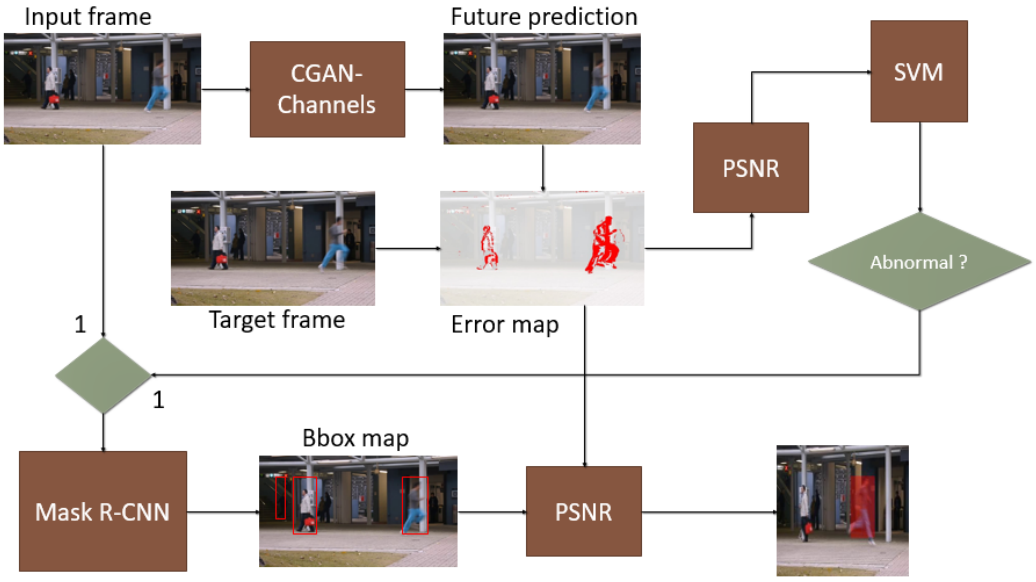


FIGURE 4.17 – The pipeline of abnormality object localisation framework. After the frame-level anomaly detection stage, we obtain the labels for each frame. For each abnormal frame, we run a fast detector to compute bounding boxes (BB) for each object in the sequence and we directly apply each BB to the error maps. PSNR scores are computed for each one by taking the normal sum of 9 PSNR scores and BBs yielding minimum values for PSNR scores refer to the abnormality objects.

candidates BBox obtaining minimum PSNR scores. It is based on the fact that the key object in our real-world scenario is abnormal one, so the false positive decision is more acceptable than false negative one. We could also go beyond this limitation by searching for a soft threshold of PSNR score.

During evaluation, we apply the Mask R-CNN [29] as the object detector. This model is based on a Resnet101 architecture trained on Imagenet dataset. All steps are implemented on Matlab with Nvidia GeForce GTX 1080.

4.2.3 Evaluation results

Evaluation metrics

As in the literature, we use frame-level Area-Under-Curve (AUC) metric as main measurement for quantitative evaluation and fair comparison with state-of-the-art methods. The type of AUC that we applied is AUC of Receiver Operating Characteristic (ROC) curve. AUC of ROC curve is an evaluation measurement for the classification problems at various threshold settings. ROC curve is a probability curve and AUC represents the degree or measure of the classifier

performance. AUC introduces how much the model is capable of separating between classes. The higher the AUC, the better the model at predicting negative sample as negative and positive sample as positive. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where TPR is on the y-axis and FPR is on the x-axis (Figure 4.18).

$$TPR = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{4.20}$$

$$FPR = \frac{True\ Negative\ (TN)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{4.21}$$

A strong classification model has its AUC near to 1 which means it has a good measure of separability (Figure 4.18). A weak model has its AUC near to 0 which means it has the worst measure of separability. In case of $AUC = 0$, the model is predicting all positive samples as negative and vice versa. Particularly, when $AUC = 0.5$, it means the model has not the ability of distinguishing between class or the model is achieving the performance at randomize level. For the EER, the best system achieves the lowest EER. For quantitative evaluation of abnormality localization, we use the same bounding box AUC and EER metrics reported in [101]. If the intersection between a detected box and the ground-truth box is smaller than 40% of the area of ground-truth box, the detected box is removed.

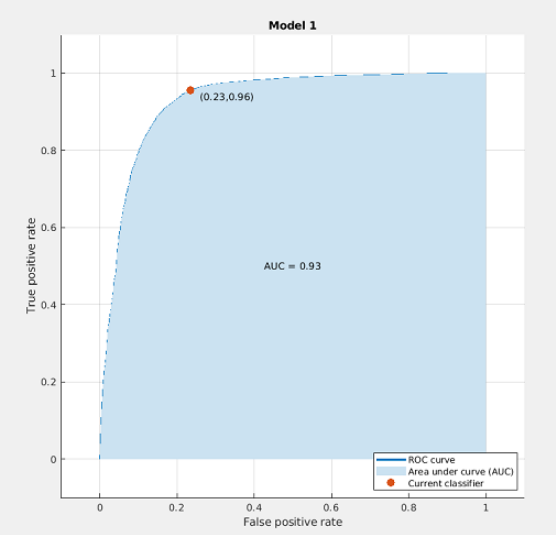


Figure 4.18 – An illustration of AUC-ROC. This performance belong to a good model based on AUC metric. The ROC curve is rapidly asymptotic to 1 while FPR is still small.

Besides, we propose some alternative evaluation metrics that are more adaptive to object-

level. We find that the bounding box condition above is not good enough for evaluating the object-centric systems because it tends to report us how many frames are true positive w.r.t. the size of the intersection, so it is still frame-level. It cannot directly illustrate the bounding box performance. Hence, we define three other metrics as following :

$$IOU_{50}\ Rate = \frac{Number\ of\ B_{50}}{Number\ of\ detected\ boxes} \tag{4.22}$$

$$IOU_{75}\ Rate = \frac{Number\ of\ B_{75}}{Number\ of\ detected\ boxes} \tag{4.23}$$

$$mIOU = \frac{\sum IOU}{Number\ of\ detected\ boxes} \tag{4.24}$$

where B_{50} and B_{75} denote the boxes that have the IOU with ground-truth boxes greater than 50% and 75%. For $mIOU$, there are two possible ways to treat the true negative frames (*i.e.* detected box = ground-truth box = 0). One way, we remove true negative frames and report only true positive cases. Other way, we set IOU = 1 for each true negative frame.

Evaluation of SVM based abnormality detection

In this section, we present the evaluation of the supervised part of our solution. We recall that the proposed solution is based on a multi CGANs backbone whose outputs are used to decide if the input frame is containing an anomaly or not. The network learning is based on a semi-supervised strategy *i.e.* the network weights and the PSNR based decision is trained with only normal image sequences.

To train the supervised SVM of our architecture, the training set has to contain anomaly samples. Only the test parts of the used datasets contain the required positive samples. To be fair during evaluation, we split the test set of each of the datasets into two subsets respectively called SVM-train and SVM-test set for respectively the training and the inference phase. To explore the effect of the supervised scenario on the performance, SVM-train is built by extracting from 10% up to 80% of the samples of the test set, the SVM-test being the remaining samples.

As proposed at the end of the section 4.2.1, we optimise the classifier with regards to the used kernel function *i.e.* we run the optimisation process on the SVM learner with the following kernel functions : polynomial, gaussian and linear functions. Thus, the SVM-train is splitted into a train and a validation set for 5-fold cross-validation. Similarly to multi-CGANs backbone evaluation, for ShanghaiTech dataset, we run the optimisation process only once when SVM-train equal 80%. Then we apply optimised parameters for the rest of the evaluation (*i.e.* from 10% to 80%).

As for multi-CGANs backbone, the effectiveness of the supervised layer is measured by the value of AUC. By sequentially increasing the SVM-train set size from 10% of the original test set size up to 80%, we report the evolution of performance on Figure 4.19. We clearly observe that the larger the size of the SVM-train set we create, the better the performance we obtain. We compare our best results for each benchmark with recent state-of-the-art methods in Table 4.5. We split those methods in two groups. The first group contains the fully unsupervised methods without adding extra abnormal samples from original test set to training set. We notice that while the method of Ionescu *et al*. [36] is unsupervised, they calculated abnormals scores by supervised SVM strategy. The second group contains the semi-supervised and supervised methods that insert abnormal samples into training set. Generally, it is difficult to compare the methods in different scenarios, especially methods from the second group, because the number of testing samples is not the same. Our solution gets promising results on all of the 4 datasets. We achieve a moderate 2% improvement on Ped1 dataset while producing a competitive result on Ped2 and Avenue datasets. Particularly, it outperforms by 9% compared with state-of-the-art method [36] on the challenging ShanghaiTech dataset. Considering the effect of the supervised scenario, we show that from 50% of the original test set size, we surpass state-of-the-art performance on Avenue, Ped1 and ShanghaiTech.

Figures 4.20,4.21,4.22 and 4.23 show the minimum classification error curve during the optimization process of the SVM parameters for each dataset. The optimization process is time consuming and depends on the size of the dataset, from 30 minutes for Pedestrian dataset to 6 hours for ShanghaiTech. As we previously mentioned, optimisation if done once. Based on those figures, we confirm the necessity and effective performance of SVM optimisation process. Thanks to the optimisation, we sharply reduce the minimum classification errors after a certain number of iterations. We also achieve best results for each dataset using various set of parameters.

After having the optimised model, we draw the ROC curve and calculate AUC performance. Results are illustrated in Figure 4.24,4.27,4.30 and 4.33 for all of the four datasets. The shape of the curve expresses the good performances we get. All of the ROC curves are rapidly asymptotic to 1 while False Positive Rate is still very small.

Then we present the confusion matrix in Figure 4.25,4.28,4.31 and 4.34. We find that most of the wrong decisions are False Positive while False Negative Rate is much lower. It can explain the high performance on AUC metric of our models. In real word transportation scenario, if we consider the abnormal activities are the dangerous activities then False positive is acceptable and we must reduce the False Negative Rate. It means that our models are very adaptive to real-world applications.

Besides, we draw the samples distributions for each class in Figure 4.26,4.29,4.32 and 4.35.

Almost all normal samples situate in higher values than abnormal ones. This distribution corresponds to the fact that PSNR techniques produce high scores if predicted images tend to be similar to the source images.

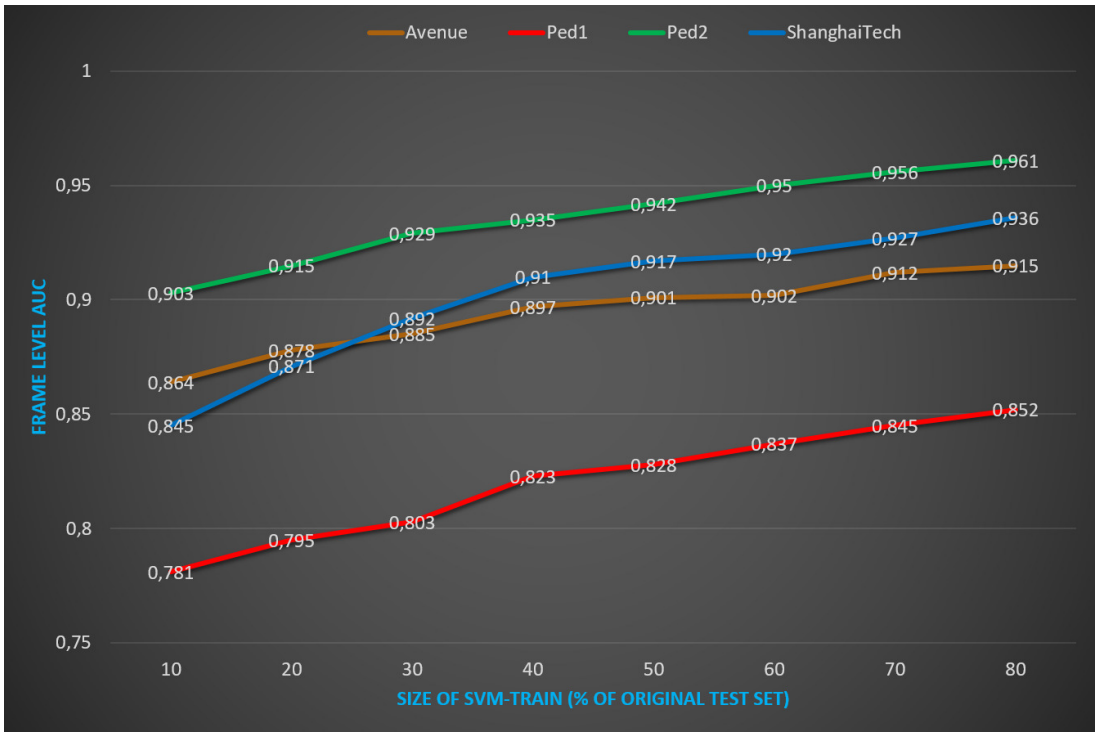


FIGURE 4.19 – Evolution of AUC performance according to the size of the SVM-train set. The vertical axis shows the AUC performances and the horizontal axis shows the size of the SVM-train set corresponding to how many samples of original test set are taken. By sequentially increasing the SVM-train set size from 10% of the original test set size up to 80%, we report the evolution of performance. Obviously, the larger the size of the SVM-train set we create, the better performance we obtain. Considering the effect of the supervised scenario, we show that from 50% of the original test set size, we surpass state-of-the-art performance on Avenue, Ped1 and ShanghaiTech. Best viewed in color.

MSE and PSNR comparison : We do the first experiments to evaluate the performance of feature encoding using PSNR and MSE. We choose Avenue benchmark as the reference dataset because it involves color images with acceptable dataset size (Pedestrian dataset have only grayscale frames and ShanghaiTech is too large). Results are illustrated in Table 4.3. The large margin about 10% on all streams and combination shows that PSNR technique is significantly better for frame level anomaly detection task.

Next, we apply PSNR technique for the other benchmarks. We go further by investigating

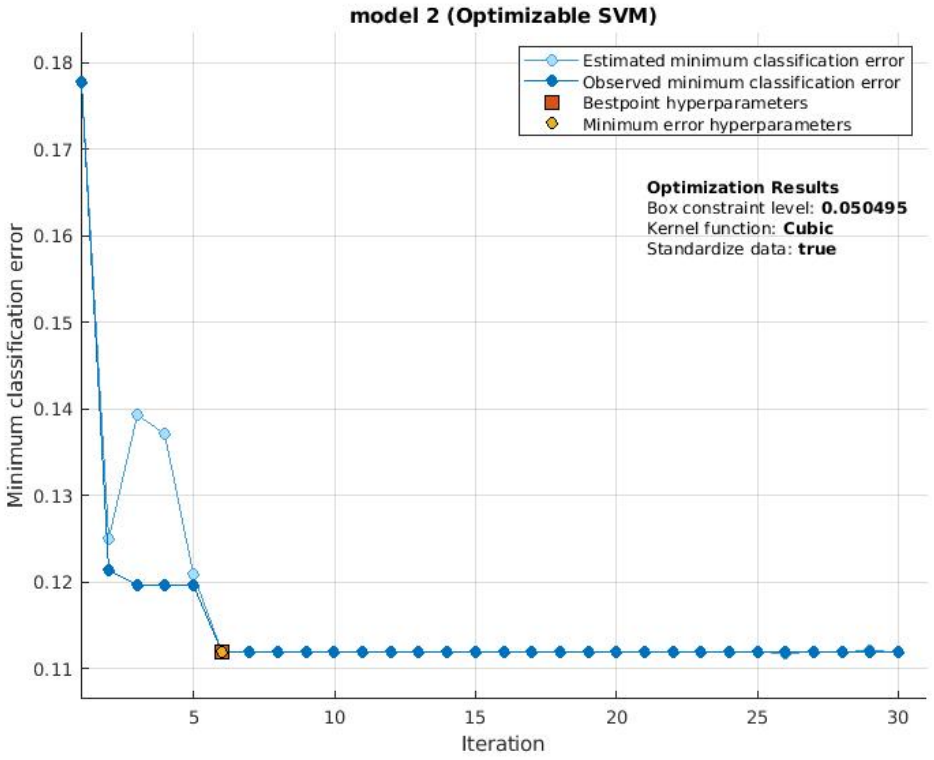


FIGURE 4.20 – Illustration of optimization process when we train SVM binary classifier on Avenue dataset. The minimum classification error is sharply decreased after 6 iterations then becomes stable. We achieve optimized parameters at box constraint level 0.05 with the cubic kernel.

various combinations of our 4 streams. Multi-stream is combined by concatenating PSNR features of each stream to produce a new vector. Results are shown in Table 4.4. On two grayscale datasets Ped1 and Ped2, CGAN-3 and CGAN-4 produce almost the same performance. The longer temporal horizon of prediction of CGAN-4 brings us a small improvement 1∼2% with respect to CGAN-3. On the other RGB datasets, CGAN-3 surpasses CGAN-4 with a huge difference of 10%. Generally, CGAN-2 achieves best performance among the 4 streams. It shows that learning both appearance and motion evolution can help us generating better features. Obviously, the combination of CGAN-1 (flow) and CGAN-4 (grayscale) produces similar or slightly better performance than CGAN-2 (combines flow with grayscale). This result also proves that taking into account longer temporal horizon of prediction, as CGAN-4, improves performance. Besides, when several streams are combined, better performance is achieved. A four streams combination always produces the best results : that is a strong experimental proof of the relevance of our proposed idea about multi-channel framework.

CGAN Stream	MSE	PSNR
CGAN-1	0.72	0.82
CGAN-2	0.69	0.81
CGAN-3	0.73	0.79
CGAN-4	0.62	0.67
CGAN-(1+2+3+4)	0.81	0.90

Tableau 4.3 – Frame-level AUC performance comparison between two methods of prediction error encoding : PSNR and MSE. Results are reported on Avenue dataset. 80% samples of test set are used for training SVM.

CGAN Stream	Avenue	Ped1	Ped2	SHT
CGAN-1	0.82	0.75	0.89	0.68
CGAN-2	0.81	0.78	0.92	0.78
CGAN-3	0.79	0.64	0.75	0.73
CGAN-4	0.67	0.65	0.77	0.62
CGAN-(1+2)	0.83	0.80	0.93	0.80
CGAN-(3+4)	0.86	0.70	0.77	0.81
CGAN-(1+4)	0.86	0.79	0.93	0.76
CGAN-(1+2+3)	0.86	0.83	0.95	0.87
CGAN-(1+2+4)	0.83	0.83	0.93	0.83
CGAN-(1+2+3+4)	0.92	0.85	0.96	0.94

Tableau 4.4 – Frame-level AUC performance on all 4 benchmarks using PSNR encoding. 80% samples of test set are used for training SVM. SHT = ShanghaiTech

	Methods	Avenue	Ped1	Ped2	SHT
Unsupervised methods	Luo *et al.* [62]	0.77		0.88	
	Nguyen *et al.* [72]	0.87		0.96	
	Hinami *et al.* [33]	0.89		0.92	
	Vu *et al.* [101]	0.72	0.82	**0.99**	
	Ouyang *et al.* [75]	0.89		0.97	0.81
	Ionescu *et al.* [36]	0.90		0.98	0.85
Semi and supervised methods	IVC with OS [52]	0.83			0.56
	IVC with OS & FL [52]	0.83			0.50
	IVC with OS & FL &2streams [52]	0.81			0.50
	TripleLoss + OCSVM [52]	0.80			0.50
	Hasan *et al.* [27]	0.80	0.75	0.85	0.61
	Luo *et al.* [63]	0.82		0.92	0.68
	Ionescu *et al.* [38]	0.81	0.68	0.82	
	Liu *et al.* [53]	0.85	0.83	0.95	0.73
	Liu *et al.* [52]	**0.93**			0.77
	Ours	0.92	**0.85**	0.96	**0.94**

Tableau 4.5 – Comparison of Frame level AUC on the 4 datasets between ours solution and recent state-of-the-art methods that reported their performance on at least 2 similar datasets. We split those methods in two groups. The upper group contains the fully unsupervised methods without adding extra abnormal samples from original test set to training set. The second group contains the semi-supervised and supervised methods that insert abnormal samples into training set. Generally, it is difficult to compare the methods in different scenarios, especially in the second group, because the number of testing samples is not the same. In our case, all the output channels of our model are used. 80% samples of test set are used for training SVM. SHT = ShanghaiTech

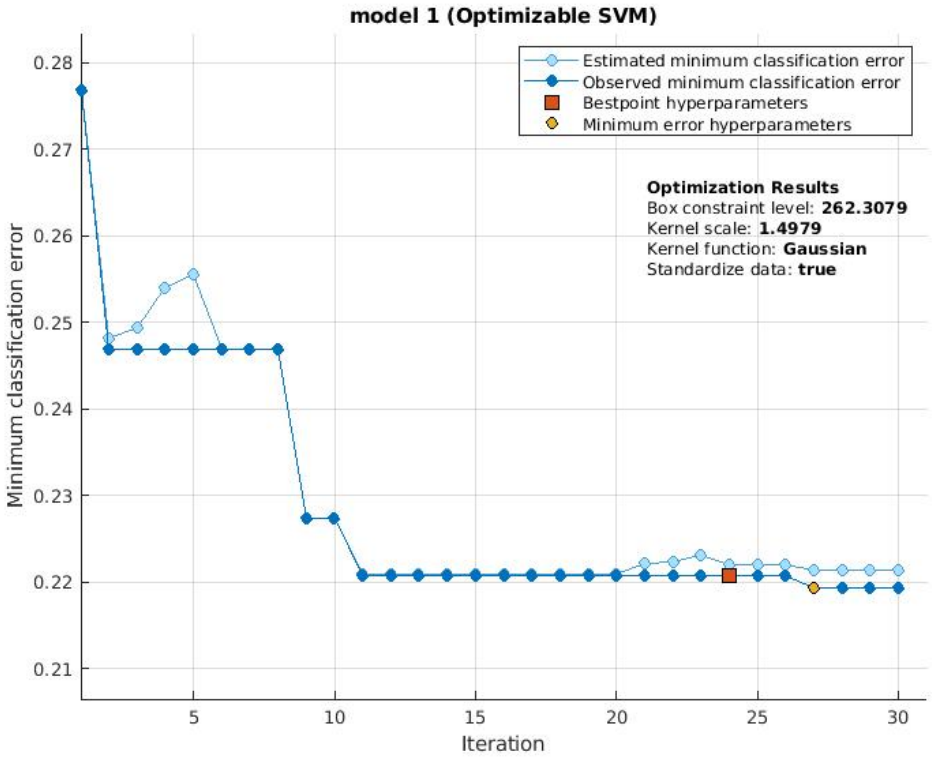


FIGURE 4.21 – Illustration of optimization process when we train SVM binary classifier on Ped1 dataset. The minimum classification error is sharply decreased after 11 iterations then becomes stable. We achieve optimized parameters at box constraint level 262.3 with Gaussian kernel.

Evaluation of abnormality objects localization

Quantitative evaluation : We evaluate the quantitative performance of our abnormality object localization models on Avenue dataset. We use AUC and ERR metrics at pixel-level. The obtained results are reported in Table 4.6. Most of the state-of-the-art researches [53, 72, 36] have not reported quantitative performance for abnormality localization task. These authors only propose a qualitative analysis to show that the error maps are relevant *i.e.* heat scores are located where abnormality objects are localised. Hence, we compare our quantitative results with recent methods of Vu *et al.* [102, 101] that applied the same evaluation metrics.

We achieve significant improvements on both metrics. We also report our performance on IOU_{50} *Rate*, IOU_{75} *Rate* and $mIOU$ in Table 4.7. A half of true positive boxes are acceptable boxes with IOU greater than 0.5. The similar rate of IOU_{50} and IOU_{75} shows that most of the acceptable boxes (*i.e.* B_{50}) are detected with high accuracy (*i.e.* B_{75}). For $mIOU$, there is a large

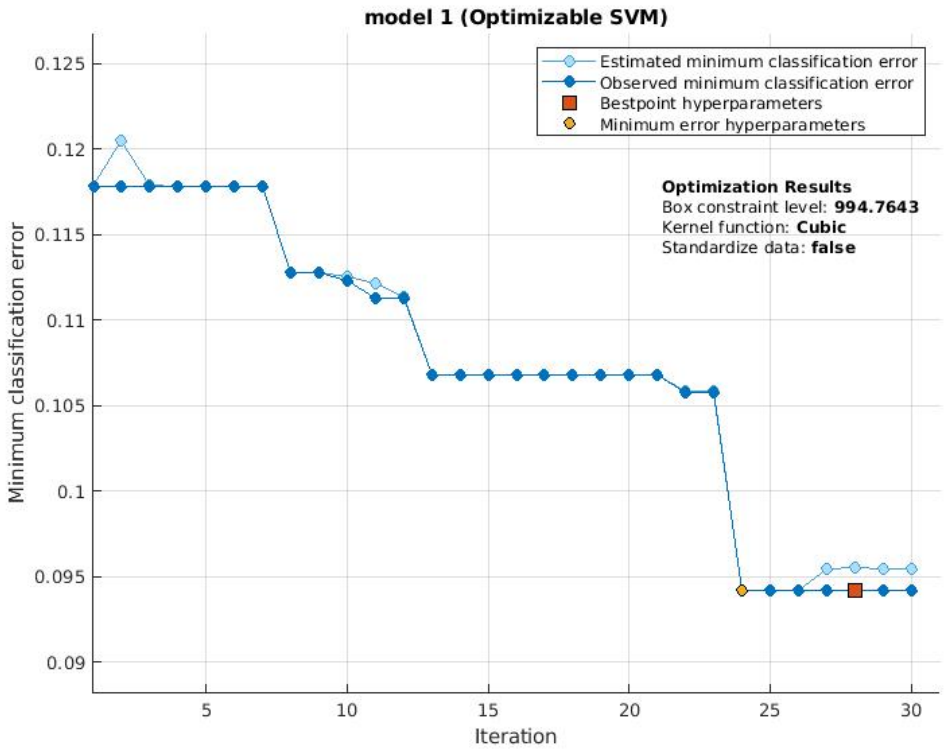


FIGURE 4.22 – Illustration of optimization process when we train SVM binary classifier on Ped2 dataset. The minimum classification error is sharply decreased after 24 iterations then becomes stable. We achieve optimized parameters at box constraint level 994.76 with the cubic kernel.

margin between both performances because there are more true negative samples than true positive samples.

Single abnormal object localization : Our method achieves promising performance when a single object is generating the anomaly. We investigate the bounding box for which the PSNR score is minimum. Figure 4.36 shows the qualitative results of single abnormal object localization on Avenue dataset. Each row represents a type of abnormality : running, dancing, throwing object, moving to wrong direction. We can exactly detect the abnormal object in every frame. Due to the inaccuracy of Mask R-CNN, the bounding boxes are not perfect in some cases, *e.g.* 4^{th} image of the first and the second row, and the 2^{nd} image of the third row.

Figure 4.37 shows the qualitative results of single abnormal object localization on Pedestrian datasets. Those datasets are more challenging than Avenue dataset because the camera view is quite far, most of the objects are small and overlapped, and only grayscale images are avai-

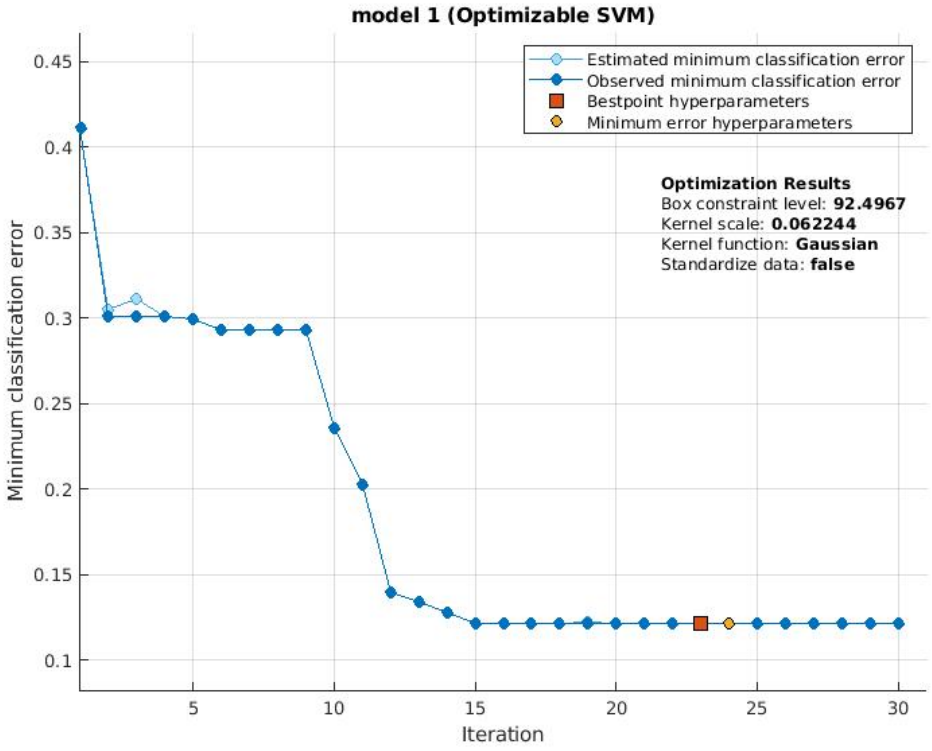


FIGURE 4.23 – Illustration of optimization process when we train SVM binary classifier on ShanghaiTech dataset. The minimum classification error is sharply decreased after 15 iterations then becomes stable. We achieve optimized parameters at box constraint level 92.50 with the Gaussian kernel.

lable. There are two typical types of abnormality in Pedestrian datasets : wrong vehicles (car, bicycle,*etc*.) moving near pedestrians and in wrong direction. Generally, the abnormal vehicles are easier to detect than the wrong moving direction. While all abnormal vehicle are localized in the first, second and fourth row, there are some false positive errors in the third row where the abnormal person are crossing the road. For Pedestrian dataset, Mask R-CNN detector provides the same detection/classification errors on bounding boxes that already appends in Avenue dataset.

Figure 4.38 illustrates the qualitative results of single abnormal object localization on ShanghaiTech dataset. There are two types of abnormality illustrated in this figure : jumping (first row) and wrong vehicles (all other rows). All abnormal objects are detected, but there are still some flaws due to detection/classification errors of Mask R-CNN, especially in the last row : instead of localising all the persons with a bicycle, our algorithm reports only the umbrella as

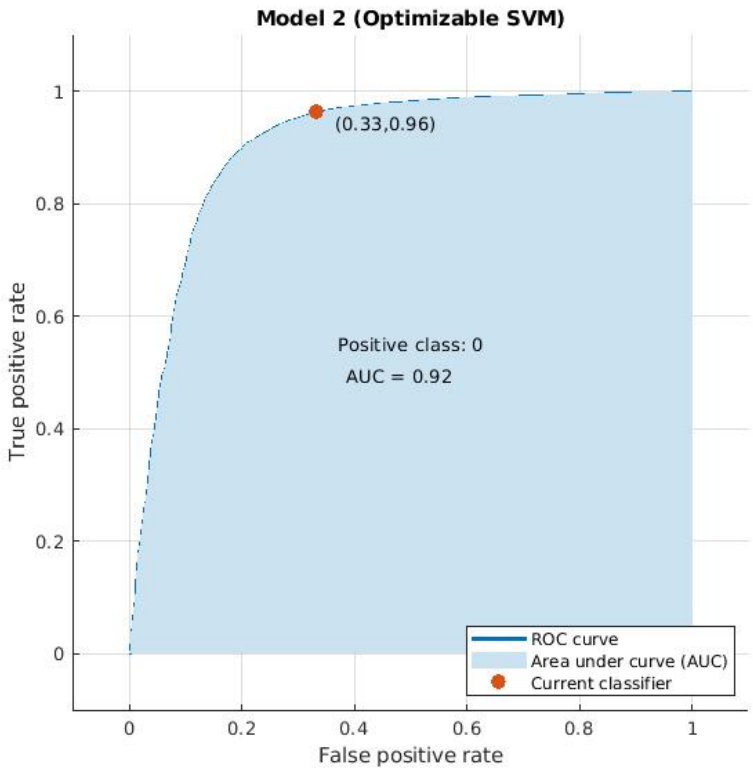


Figure 4.24 – Illustration of AUC performance achieved at optimized SVM model on Avenue dataset. The ROC is rapidly asymptotic to 1 while False positive rate is still very small. It means that our model achieves good performance.

the abnormality object.

Generally, we can exactly detect and localise the abnormal object in various types of abnormality : running, jumping, dancing, throwing object, moving to wrong direction, abnormal vehicles, *etc*. Interestingly, our method performs pretty well when occlusion occurs and in overlapped scenarios. In contrast, there are also some flaws related to the size and position of bounding boxes. Sometimes, they are too large or too small. In several other cases, the boxes only belong to parts of abnormal object, *e.g*. the umbrella existing in figure 4.38. Most of the imperfect cases come from the fact that our anomaly object localisation depends on the accuracy of the Mask R-CNN detector. Sometimes, overlapped objects are detected by Mask R-CNN at the same location, *e.g*. bicycle, person and umbrella in ShanghaiTech samples. When one of those objects achieves minimum PSNR score, our algorithm chooses only this box but not the total location.

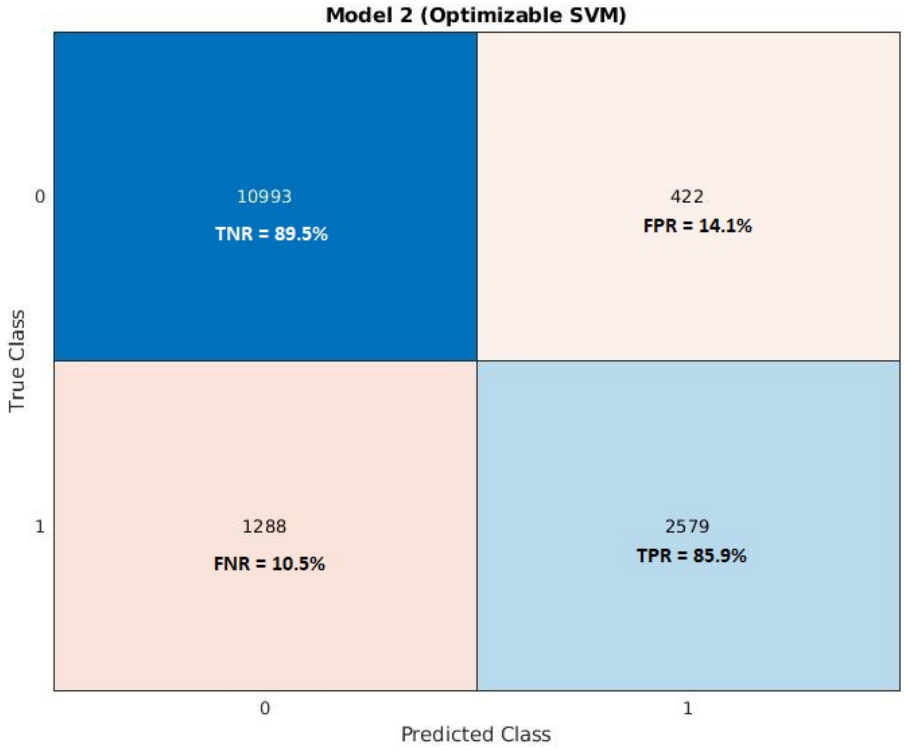


FIGURE 4.25 – Illustration of confusion matrix achieved with optimized SVM model on Avenue dataset. We find that most of the wrong decisions are False Positive while False Negative Rate is much lower. It can explain the high performance on AUC metric of our models. In real word transportation scenario, if we consider the abnormal activities are the dangerous activities then the False positive is acceptable and we must reduce the False Negative Rate. It means that our model is very adaptive to real-world applications.

Multiple abnormal objects localization : In the case of existing multiple abnormal objects, the maximum number of objects to detect is 2. It means that we report the minimum and second minimum PSNR objects. Figure 4.39 illustrates the qualitative results of multiple abnormal objects localizations. Generally, our methods achieve moderate performance. We successfully detect the first object (*i.e.* the minimum PSNR score) while the second object are sometimes false positives. Those errors often happens when the false positive object are near the main abnormal object. We can explain this problem by investigating the motion features and the inaccuracy of Mask R-CNN.

On the one hand, the optical flow estimation is not perfect. The motion errors are accumulated throughout our pipeline, from flow estimation to flow reconstruction. Obviously, the difference of optical flow is always an important part of the PSNR score. Combining with the

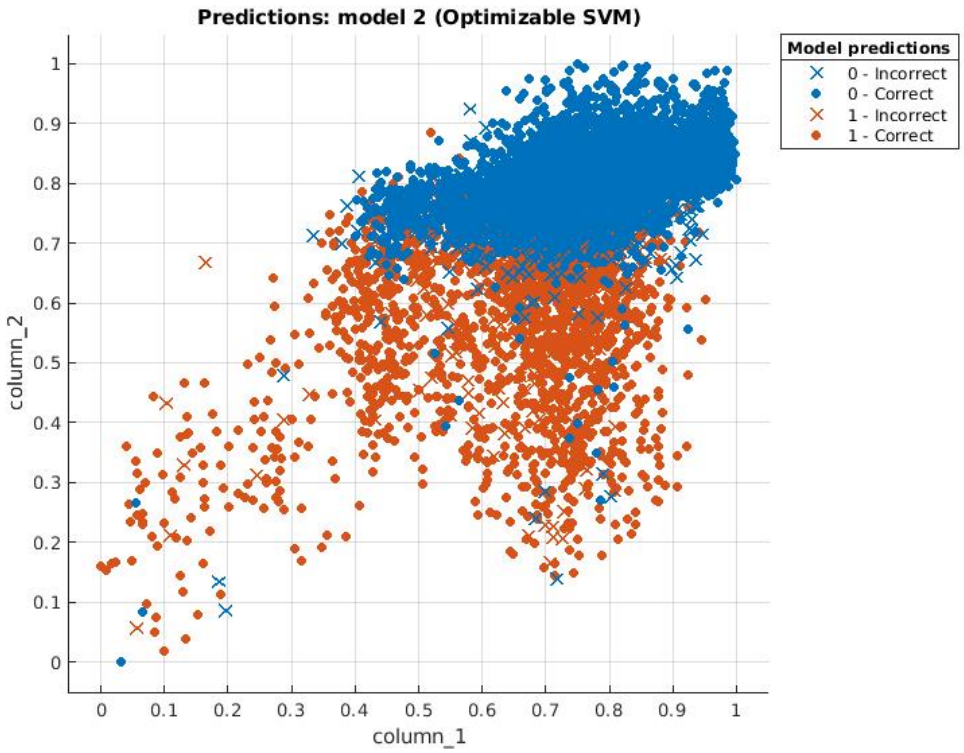


FIGURE 4.26 – Illustration of sample distribution achieved at optimized SVM model on Avenue dataset. Almost all of the normal samples situate in higher values than abnormal ones. This distribution corresponds to the fact that PSNR techniques produce high scores if predicted images tend to be similar to the source images.

existing drawback of Mask R-CNN (*i.e.* bounding boxes are too large or too small), the impact of motion score can be spread out and causes the effect on neighbour objects. By this way, the boxes near the first object tend to obtain lower PSNR score than the second real abnormality object.

On the other hand, overlapped objects are detected by Mask R-CNN at a same location, *e.g.* bicycle and person in pedestrian samples. Obviously, both objects usually obtain lower PSNR score than the rest. Therefore, our algorithm chooses the two objects at the same location instead of searching for a new position.

Figure 4.40 shows us several examples selected from Avenue, Ped1, Ped2 and ShanghaiTech to demonstrate true positive and failure cases of anomaly object localization. Obviously, we can strongly detect various types of abnormal events in the crowed and cluttered background, for different camera angles and involving not only single but also multiple objects. Failure cases

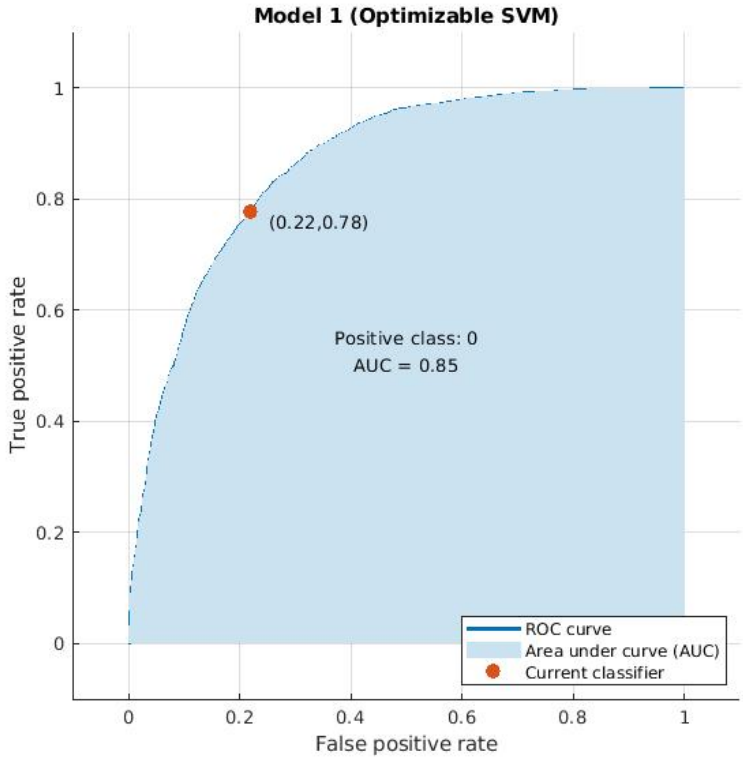


FIGURE 4.27 – Illustration of ped1 performance achieved with optimized SVM model on Avenue dataset. The ROC is rapidly asymptotic to 1 while False positive rate is still very small. It means that our model achieves good performance.

appear when abnormality objects are too close to the other normal boxes : the PSNR score of those boxes is affected by the features of the near abnormality object. Other common errors are directly caused by false negative abnormality of the frame-level detection. Generally, those problems come from two reasons. On the one side, SVM cannot classify frames as abnormal and produces false negatives : for those frames the object detector is not applied. On the other side, some frames are true negatives (*i.e.* SVM is correct) but the PSNR score is not relevant enough for localising the abnormality objects. A promising solution is proposing an adaptive thresholding rather than the hard-decision strategy.

4.3 Conclusions and discussions

In this chapter, a CGAN based SVM anomaly detector is proposed. The multi-CGANs backbone predicts the future information : it is based on four CGANs taking various types of appearance and motion information as input and producing prediction of similar information as

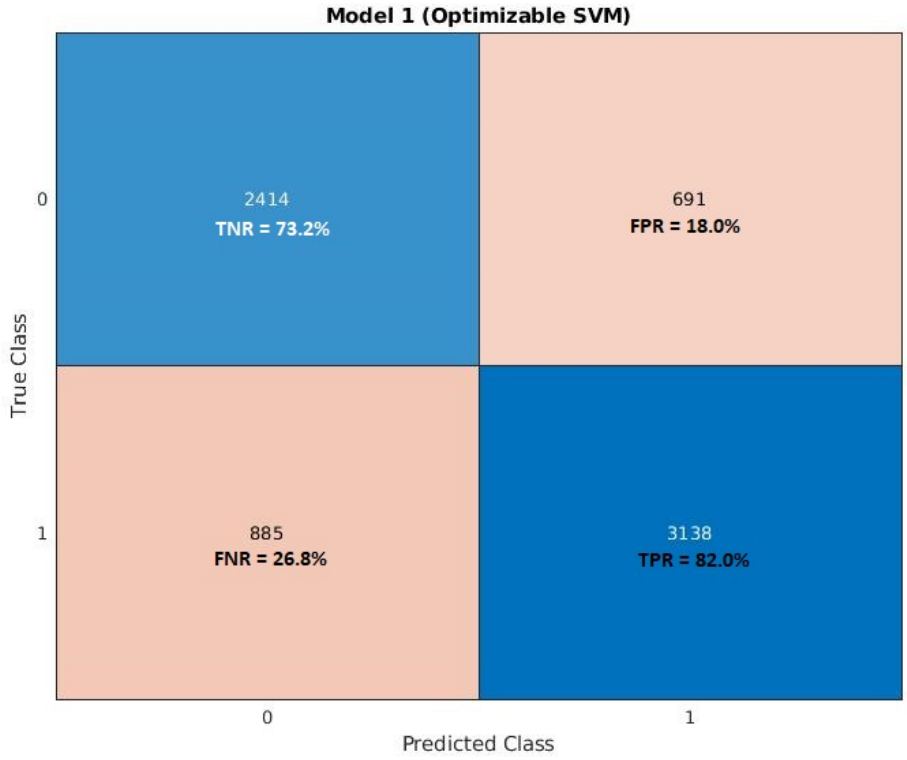


FIGURE 4.28 – Illustration of confusion matrix achieved with optimized SVM model on ped1 dataset. We find that almost all of the wrong decisions are False Positive while False Negative Rate is much lower. It can explain the high performance on AUC metric of our models. In real word transportation scenario, if we consider the abnormal activities are the dangerous activities then the False positive is acceptable and we must reduce the False Negative Rate. It means that our models is very adaptive to real-world applications.

output. These CGANs combined to PSNR metric provide a relevant feature space to discriminate normal and abnormal events. The discriminative model is based on a Support Vector Machine trained following a classical supervised strategy. While the binary SVM is applied for frame-level anomaly detection, we propose an anomaly object localisation based on a Mask R-CNN detector.

Our evaluations highlight our good achievements for anomaly detection on 4 reference datasets. Our experiments show that we achieve very promising results although we have not optimised all of the steps. For example, Full-Flow [13] and Lucas-Kanade [61] could be replaced by recent SOTA methods to reach better performance.

Strengths : Our anomaly detection framework is very flexible. It can be easily extended by adding, removing or replacing any of the CGAN streams and inference methods. About

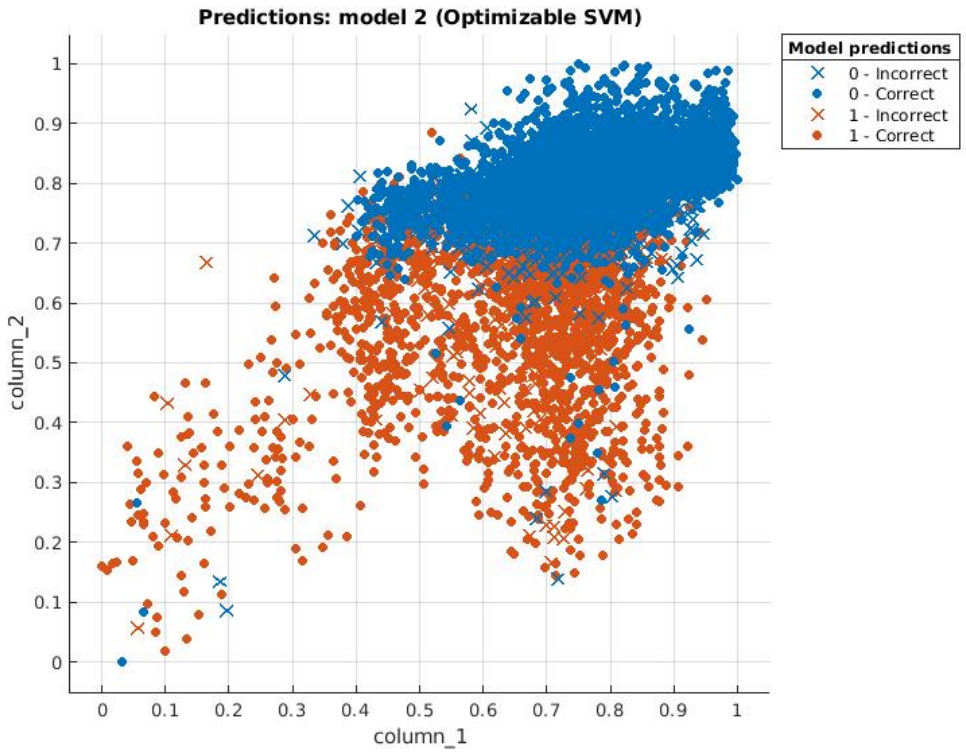


FIGURE 4.29 – Illustration of sample distribution achieved with optimized SVM model on ped1 dataset. Almost all of the normal samples situate in higher values than abnormal ones. This distribution corresponds to the fact that PSNR techniques produce high scores if predicted images tend to be similar to the source images.

computational complexity, our CGANs stage has the same complexity as the SOTA. The inference stage has moderate complexity : higher than [53] using simple threshold but more simple than [36] using very high-dimension features classified by K-means with multi-class SVM.

Limitations : On the one hand, the mix of unsupervised CGAN streams with supervised SVM helps us obtain impressive performances on off-line cases where all test data are prepared and well pre-processed. For online works, we will get troubles when brand-new abnormal actions suddenly occur without existing in the previous learned databases. On the other hand, the anomaly object localisation is significantly affected by the detector we use (Figure 4.41). Sometimes, bounding boxes is imperfectly generated regarding its size and its labels. We can go beyond this limitation by not applying a high-level detector such as Mask R-CNN but only a low-level region proposal network. Moreover, the proposed framework is not an end-to-end network structure at the moment. An end-to-end approach will be more globally optimal by avoiding several optimisation process done locally.

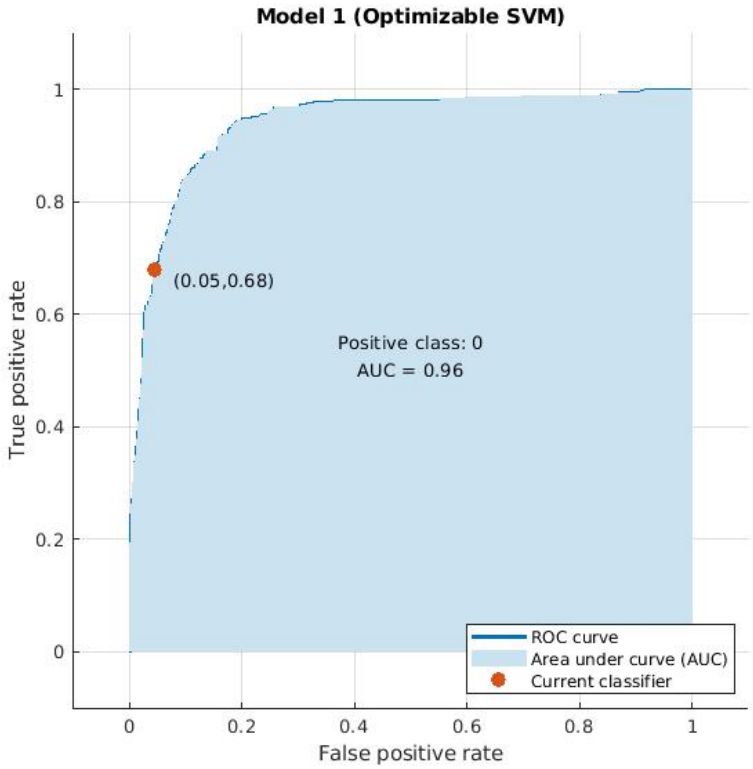


FIGURE 4.30 – Illustration of optimization performance achieved with optimized SVM model on ped2 dataset. The ROC is rapidly asymptotic to 1 while False positive rate is still very small. It means that our model achieves good performance.

Methods	AUC	EER
OC-SVM [102]	33.16	47.55
GMM [102]	43.06	43.13
Multilevel Representations [101]	52.82	38.83
Ours	**74.43**	**30.21**

TABLEAU 4.6 – Pixel-level AUC and EER performance of abnormal object localization on Avenue dataset.

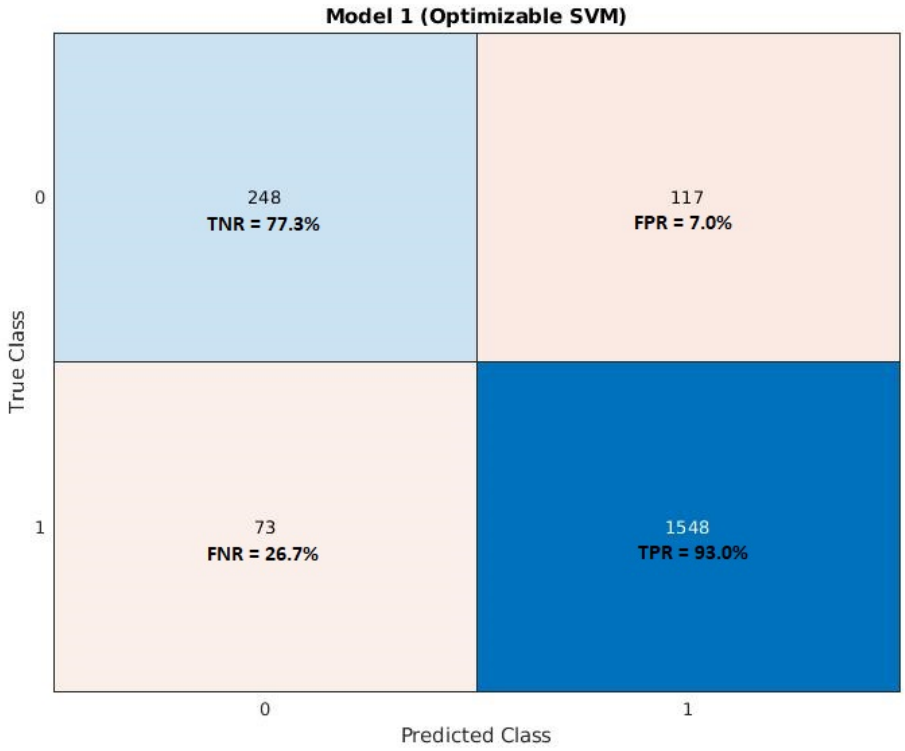


FIGURE 4.31 – Illustration of confusion matrix achieved with optimized SVM model on ped2 dataset. We find that almost all of the wrong decisions are False Positive while False Negative Rate is much lower. It can explain the high performance on AUC metric of our models. In real word transportation scenario, if we consider the abnormal activities are the dangerous activities then the False positive is acceptable and we must reduce the False Negative Rate. It means that our models is very adaptive to real-world applications.

Metric	Performance
IOU_{50} *Rate*	0.49
IOU_{75} *Rate*	0.41
$mIOU$	0.42
$mIOU + TN$	0.86

TABLEAU 4.7 – Pixel-level performance of abnormal object localization on Avenue dataset. We report mIOU in both cases : i, $mIOU$ denotes the performance without true negative samples and ii, $mIOU + TN$ denotes the performance with true negative samples *i.e.* for each true negative box, we set IOU = 1.

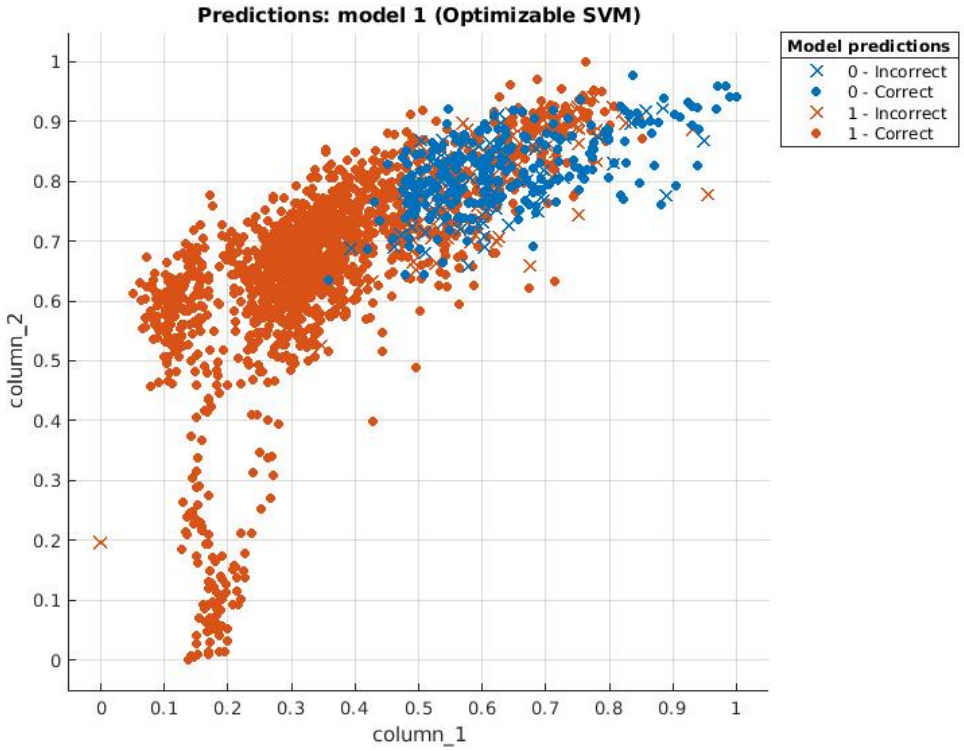


FIGURE 4.32 – Illustration of sample distribution achieved with optimized SVM model on ped2 dataset. Almost all of the normal samples situate in higher values than abnormal ones. This distribution corresponds to the fact that PSNR techniques produce high scores if predicted images tend to similar source images.

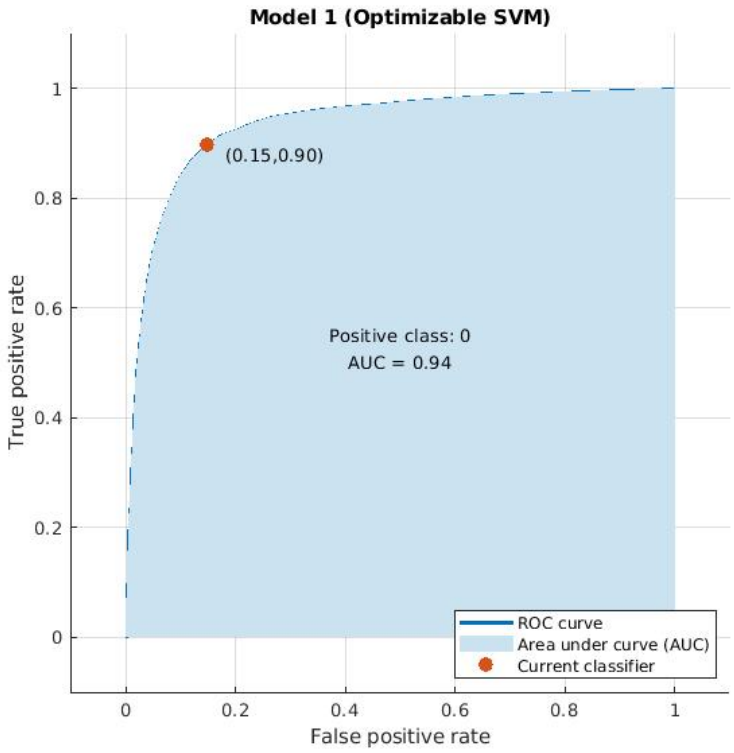


FIGURE 4.33 – Illustration of optimization performance achieved with optimized SVM model on ShanghaiTech dataset. The ROC is rapidly asymptotic to 1 while False positive rate is still very small. It means that our model achieves good performance.

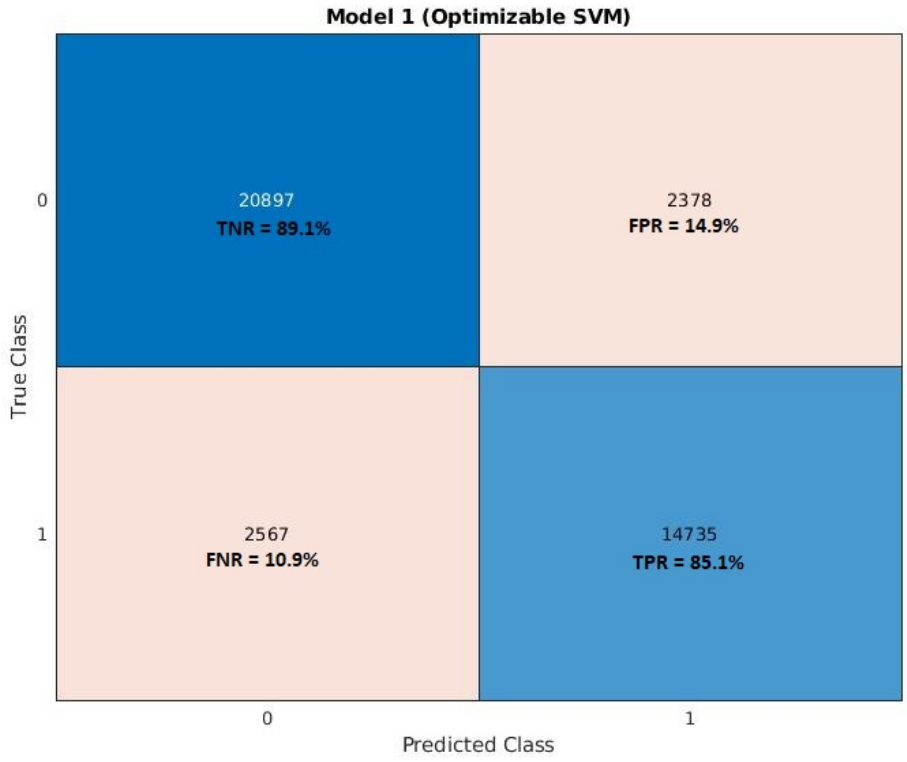


Figure 4.34 – Illustration of confusion matrix achieved with optimized SVM model on ShanghaiTech dataset. We find that almost all of the wrong decisions are False Positive while False Negative Rate is much lower. It can explain the high performance on AUC metric of our models. In real word transportation scenario, if we consider the abnormal activities are the dangerous activities then the False positive is acceptable and we must reduce the False Negative Rate. It means that our models is very adaptive to real-world applications.

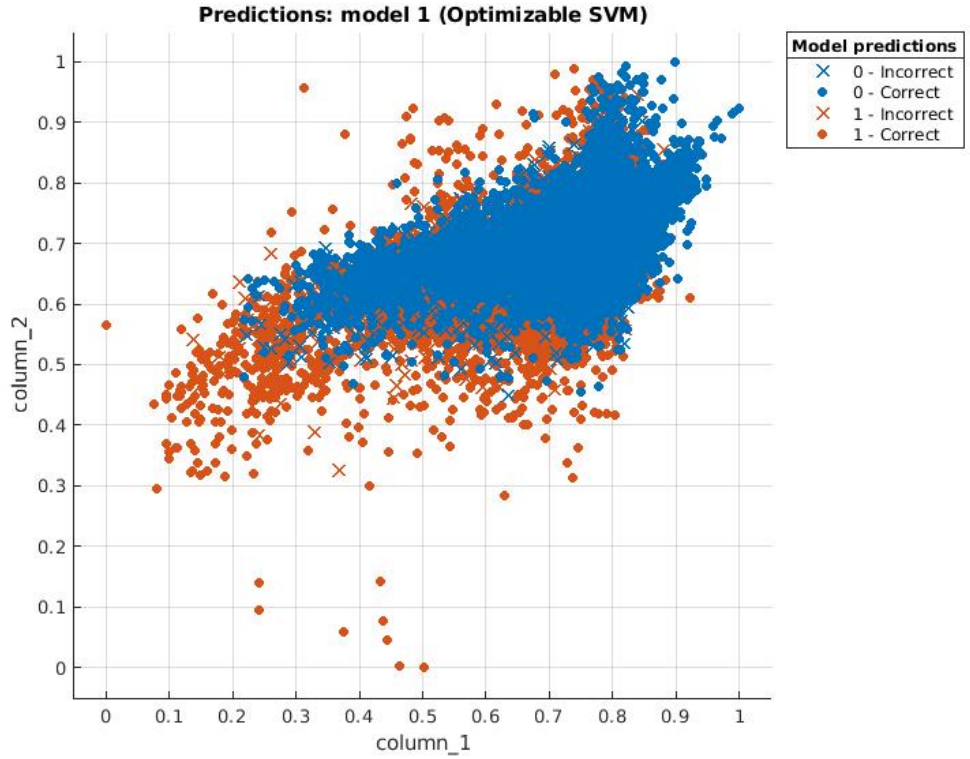


FIGURE 4.35 – Illustration of sample distribution achieved with optimized SVM model on ShanghaiTech dataset. Almost all of the normal samples situate in higher values than abnormal ones. This distribution corresponds to the fact that PSNR techniques produce high scores if predicted images tend to be similar to the source images.

Figure 4.36 – Qualitative results of single abnormal object localization on Avenue dataset. Each row represents a type of abnormality. From the top row to the bottom row : running, dancing, throwing object, moving to wrong direction. We can exactly detect the abnormal object in every frame. Due to the drawback of Mask R-CNN, the bounding boxes are not perfect in some cases, *e.g.* 4^{th} image of the first and second rows, 2^{nd} image of the third row.

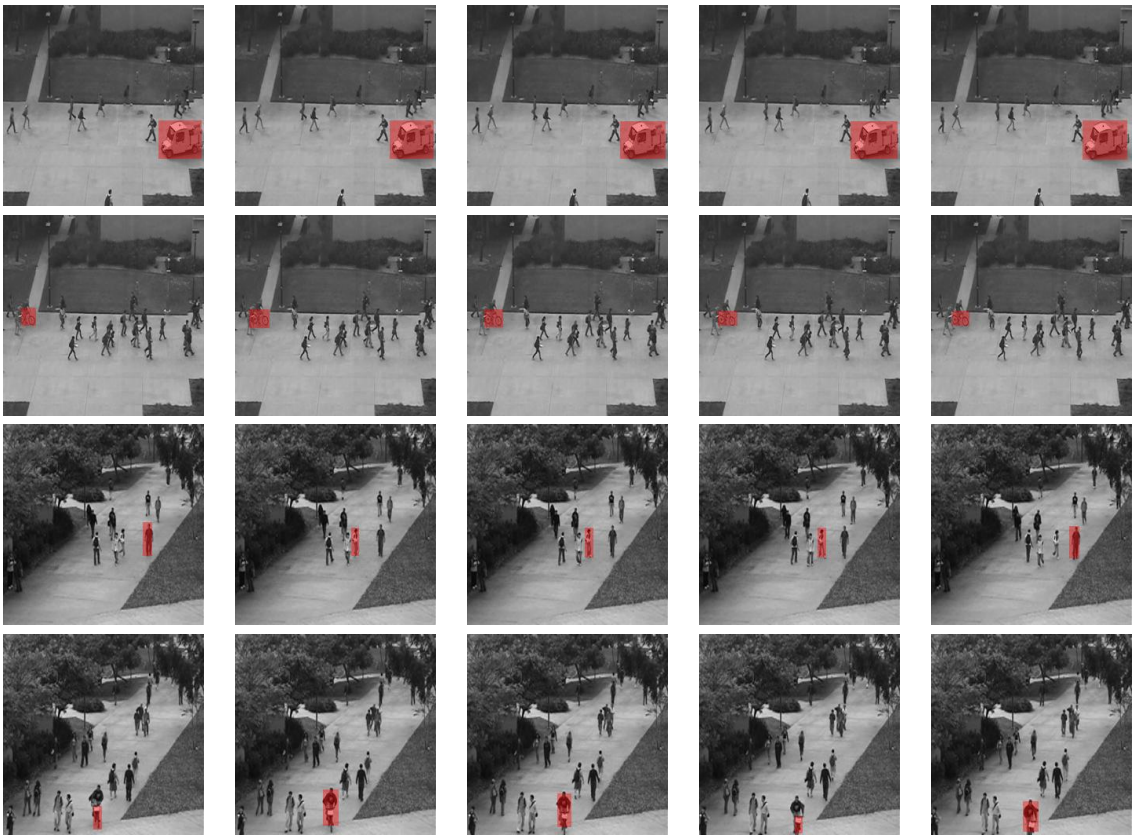


FIGURE 4.37 – Qualitative results of single abnormal object localization on Ped1 and Ped2 datasets. There are two typical types of abnormality in Pedestrian datasets : wrong vehicles (car, bicycle,*etc*.) and moving to wrong direction. Generally, the abnormal vehicles are easier (to detect) than the wrong moving direction. While all abnormal vehicles are localized in the first, second and fourth row, there are some false positive errors in the third row where the abnormal persons are crossing the road. The same drawback of Mask R-CNN bounding boxes still happens as in Avenue dataset.

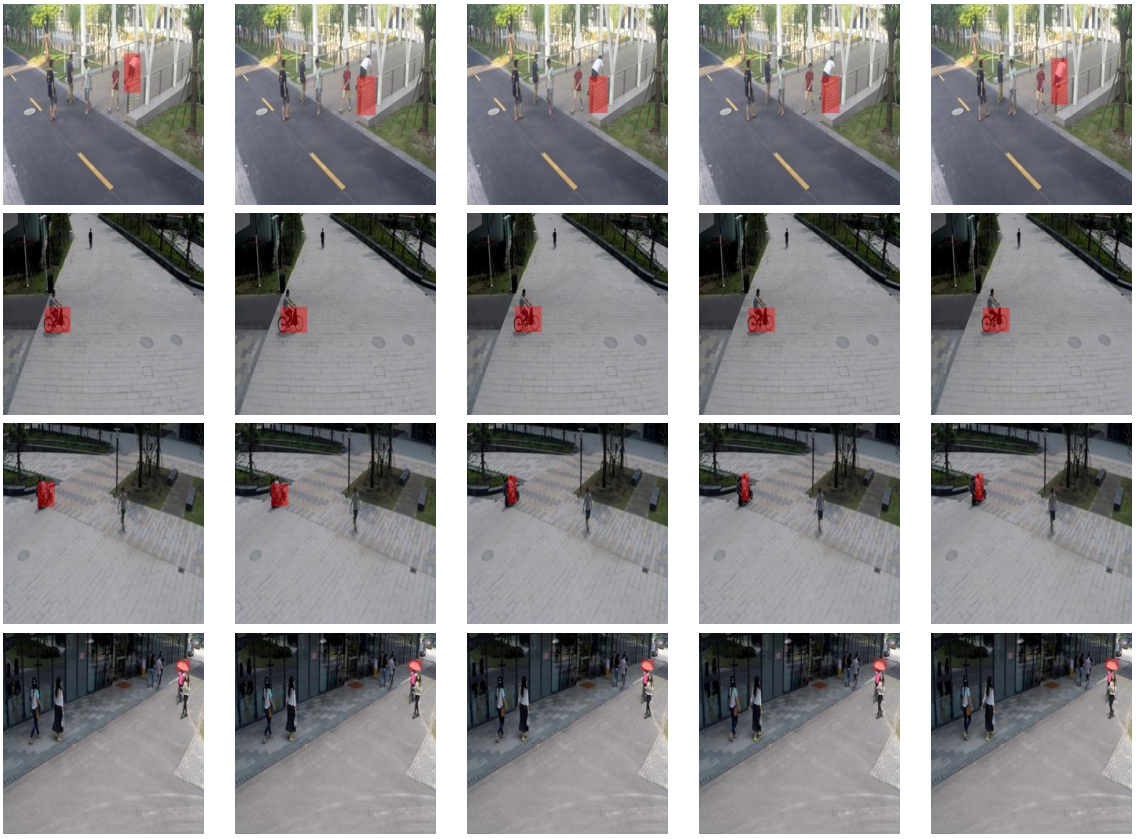


Figure 4.38 – Qualitative results of single abnormal object localization on ShanghaiTech dataset. There are two types of abnormality illustrated in this figure : jumping (first row) and wrong vehicles (all other rows). All abnormal objects are detected, but there are still some flaws due to the drawback of Mask R-CNN, especially in the last row. Instead of localizing all the persons with a bicycle, our algorithm chooses only the umbrella.

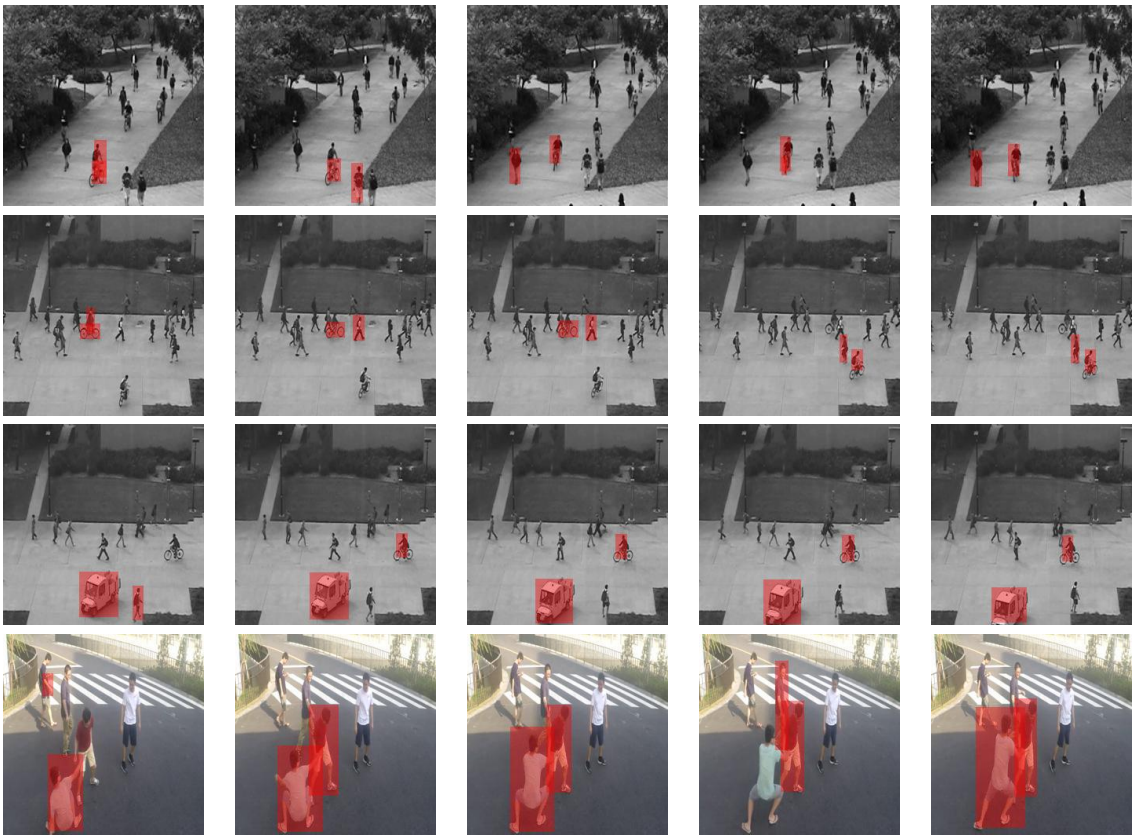


FIGURE 4.39 – Qualitative results of multiple abnormal objects localization on Pedestrian and ShanghaiTech datasets. We set the number of abnormal object to 2. While the first object (*i.e.* obtaining minimum PSNR score) is almost always true, the second object (*i.e.* obtaining the second minimum PSNR score) is false positive when its is near the first object. This drawback comes from the accumulated errors of the optical flow in estimating and generating steps. Besides, overlapped objects are detected by Mask R-CNN at the same location, *e.g.* bicycle and person in pedestrian samples. Obviously, both objects usually obtain lower PSNR score than the rest. Therefore, our algorithm chooses the two objects at the same location instead of searching for a new position.

Figure 4.40 – Qualitative results of abnormal objects localization on Avenue, Ped1, Ped2 and ShanghaiTech dataset. The left side contains correct localizations and the right side shows the failure cases with two main errors : false negative frames and detection of wrong objects. The true objects are marked by green boxes. Best viewed in color.

Figure 4.41 – Several imperfect cases of anomaly localisations affected by Mask R-CNN detector. Left : Too large bounding box ; Middle : Too small box ; Right : Wrong object. Best viewed in color.

Chapitre 5

Conclusion and Future works

This chapter summarises the essentials of my PhD thesis as well as presents some potential ideas for future works. The first section is dedicated to conclude the thesis. We briefly remind the general context and objectives of this project and synthesize our remarkable achievements for our proposed methods. The second section introduces what we plan to do to improve some limitations of our current works and to develop our frameworks to real-world applications.

5.1 Conclusion

This PhD thesis addresses two challenging problems in computer vision for the transportation applications : (1) vehicles and road users segmentation and tracking and (2) anomaly detection. By largely reviewing state-of-the-art methods for each task, we decide to follow the future prediction approaches through two possible scenarios. The first one deals with the classical hand-crafted generative methods based directly on optical flow estimation while the second one applies deep learning CGANs model to build a multi-channel generative framework that learned competitive features from both appearance and motion information.

For hand-crafted strategy, our purpose is to evaluate the capability of classical hand-crafted generative methods for improving segmentation and tracking. About the second strategy, we aim at making the most with the available deep generative model to improve SOTA performances obtained for various public datasets. The innovation relies on the flexibility of our model that is based on a combination of several conditional GAN, accepting various types of input information (*i.e.* RGB, grayscale and optical flow images). By producing features suitable for various classification techniques like Support Vector Machine, our model can be easily extended or lightened for further developments.

5.1.1 Improving segmentation and tracking by classical hand-crafted methods

Our first research is an initial work for evaluating the performance of the classical hand-crafted generative approach in future prediction and its capacity to improve segmentation and tracking of moving objects. By searching for a strong baseline among state-of-the-art methods, Mask R-CNN [29] has been chosen because of the promising performance for detection and instance segmentation of the objects. Tracking-by-detection approach has been adopted for tracking step. We focus on the impressive speeding-up performance of the IoU tracker that takes into account only the Intersection-Over-Union (IOU) between bounding boxes to match objects within consecutive frames. On the contrary, because this method does not use other visual information [7] and because of the failure detection of Mask R-CNN detectors, fragmented objects trajectories appear.

We propose to fill broken trajectories by predicting instance segmentation using Optical flow and then propose an enhanced tracker based on connecting fragmented trajectories by SURF feature [5].

Our solution first generates new detections or instance masks by translating backward and forward current information using optical flow vectors. We extend two methods proposed by [59] : Shift and Warp. The predicted/generated masks obtained by Warp method are then denoised by morphological operations. After this stage, the gaps of trajectories can be filled in. We use DAVIS dataset [77] with two popular classes of vehicles (*bus* and *car*) for evaluating the quantitative and qualitative performance of generating new detections. The quantitative results show that Shift-translation outperforms Warp-translation for long trajectories while Warp-translation integrated with morphological is more adaptive with short tracking. It also confirms the necessity of morphological operators for post-processing. The qualitative results show that our solution achieved stable performance with different types of flow estimation methods.
Generated segmentation and bounding boxes are matched along all the sequence by using SURF features. An IOU-matching algorithm is proposed to automatically fill fragmented parts of the trajectories. The entire process is appliued on DETRAC dataset [65]. We only realize a qualitative evaluation. At this stage of the work, quantitative evaluation was not required. The qualitative results show that our methods significantly improve the fragmented trajectories in particular sequences.

5.1.2 Anomaly detection by multi-channel deep generative neural networks

Although the results obtained for filling fragmented trajectories seems to be usable for easy tasks, for anomaly detection, the obtained performance is not sufficient. Obviously, there are important differences between anomaly detection and general action recognition problem. Firstly, the abnormal events rarely appear unlike normal event : it leads to an unbalanced scenario where

the number of samples in each class is significantly different. Secondly, the features of abnormal events usually do not necessarily follow any spatial or temporal relationship so that it raises the difficulty of pre-defining the structure or class of abnormal events. In order to tackle those challenges, we propose to train a generative model based on the observation of the apparent motion and appearance for the normal scenes. This model computes predicted images from past images. The abnormal situations are then detected by using error estimation between the generated output of the model and both appearance and optical flow provided by the sequence at the same time.

The proposed network is based on a multi pix-2-pix Conditional GAN (CGAN) [39] architecture with U-Net model. The Generator is an Encode-Decode network for generating future information based on appearance and optical flow images at previous time. The Discriminator tries to classify whether the generated samples are fake. We evaluate a four CGAN streams : each CGAN takes various types of appearance and motion information as input for which it produces a future prediction.

This multi-CGAN backbone provides a better feature space to project the specificities of normal and abnormal events. The SOTA proposed to infer the nature of event by unsupervised function. We decide to do that by using a supervised Support Vector Machine trained on the multi-outputs of the network encoded by Peak Signal-to Noise Ratio. The binary Support Vector Machine (SVM) is applied for frame-level anomaly detection. Finally, we jointly use a Mask R-CNN detector and the mutli-CGAN backbone to localize the obnormality object in each abnormal frame.

Our methods are largely evaluated on CUHK Avenue [58], USCD Pedestrians [66] and ShanghaiTech [53] datasets. Our experiment results demonstrate that PSNR features are better than mean square error maps computed by previous methods. For the strong performance objective, we achieve state-of-the-art frame-level AUC on Avenue, Ped1 and ShanghaiTech. Especially, for the most challenging Shanghaitech dataset, a supervised training model outperforms up to 9% the state-of-the-art on unsupervised strategy. We achieve those impressive results without optimization all steps in the pipeline, particularly optical flow extraction. We apply a very simple flow estimator pre-installed from OpenCV and classical Full-Flow [13] estimator. For the flexibility objective, we confirm the efficiency of integrating more channels because the performances of the unified network are significantly surpassing every single channel's result. Our model is also free to extend or lighten without any exigency due to the independent architecture of each channel. Each combination of various channels can be taken into account for different purposes. Our features are quite simple and suitable for many classification models in the inference phase.

5.2 Future works

We always keep our work in progress for improving both problems : vehicle segmentation-tracking and anomaly detection. There are many promising tasks to continue in perspectives : replacing classical optical flow estimators with recent state-of-the-art models, constructing a new dataset for anomaly detection, re-applying CGANs for the first problem, installing an end-to-end network for the second problem, *etc*. We discuss these prospects in the following parts.

Optimization of Optical flow extraction : Although we have evaluated LDOF [9] and FullFlow [13] for the first problem to confirm the independence of our methods regarding the type of flow, the second problem is still waiting for a better optical flow estimator. In both cases, all optical flow methods are classical and hand-crafted. Actually, there are many state-of-the-art models using deep learning techniques *e.g*. FlowNet-2 [35], PwC-Net [94], *etc*. Intuitively, the better the flow vectors are, the higher accuracy we could finally achieve. The first improvement could be to use one of these optical flow networks to infer the motion content for the multi-CGAN network. We will have to define if the output of the optical flow net is sufficiently accurate for anomaly detection.

Semi-supervised Dataset for Anomaly Detection : Actually, we evaluated our anomaly detection method for the unsupervised scenarios. Multi-CGAN backbone is trained with only normal samples while the SVM classifier is trained on a dataset containing both types obtained by splitting the test dataset. For unsupervised anomaly benchmark, this dataset is only used for the final evaluation. Even if we don't use the training part of the test dataset to blindly evaluate the model, to be completely fair we plan to build a new dataset adapted for semi-supervised scenarios. Moreover, one next step for evaluation will be to apply our methods on more datasets [93, 28, 48] that are more suitable for the supervised scenarios.

CGANs for Improving Segmentation and Tracking : A natural prospect for the first work is re-apply CGANs model of the second work for generating missing segmentations or bounding boxes in broken trajectories. There is always a trade-off here. On the one side, the combining of two strong deep learning models (*i.e*. Mask R-CNN and CGANs) can sharply increase qualitative and quantitative performance. On other side, we might consume much more time running for those deep networks. We have to deal with the balance between speed and accuracy.

From binary SVM to one-class SVM : For the second work, we also plan to evaluate our framework in the unsupervised scenario. Obviously, we could almost keep all parts of our model except the inference phase based on supervised binary SVM. By replacing binary SVM with one-class SVM, our classifiers will learn the boundary of normal samples and therefore be able to classify any points that lie outside the boundary as outliers or abnormal samples.

From hybrid model to end-to-end model : The ambitious goal for extending our second framework to the end-to-end model is replacing SVM inference phase by a sub neural network for detection/segmentation. To my personal knowledge, there have not existed successful models connecting CGANs to a discriminative segmentation/detection model yet. Recently, Nguyen *et al*. proposed a promising way to add a small brand of CNN models for classification into encode-decode model [71]. We might follow this approach in the future.

Annexe A

BONUS EXPERIMENTAL RESULTS ILLUSTRATIONS

FIGURE A.1 – Larger illustration of Sequence 1a in Figure 3.14

FIGURE A.2 – Larger illustration of Sequence 1b in Figure 3.14

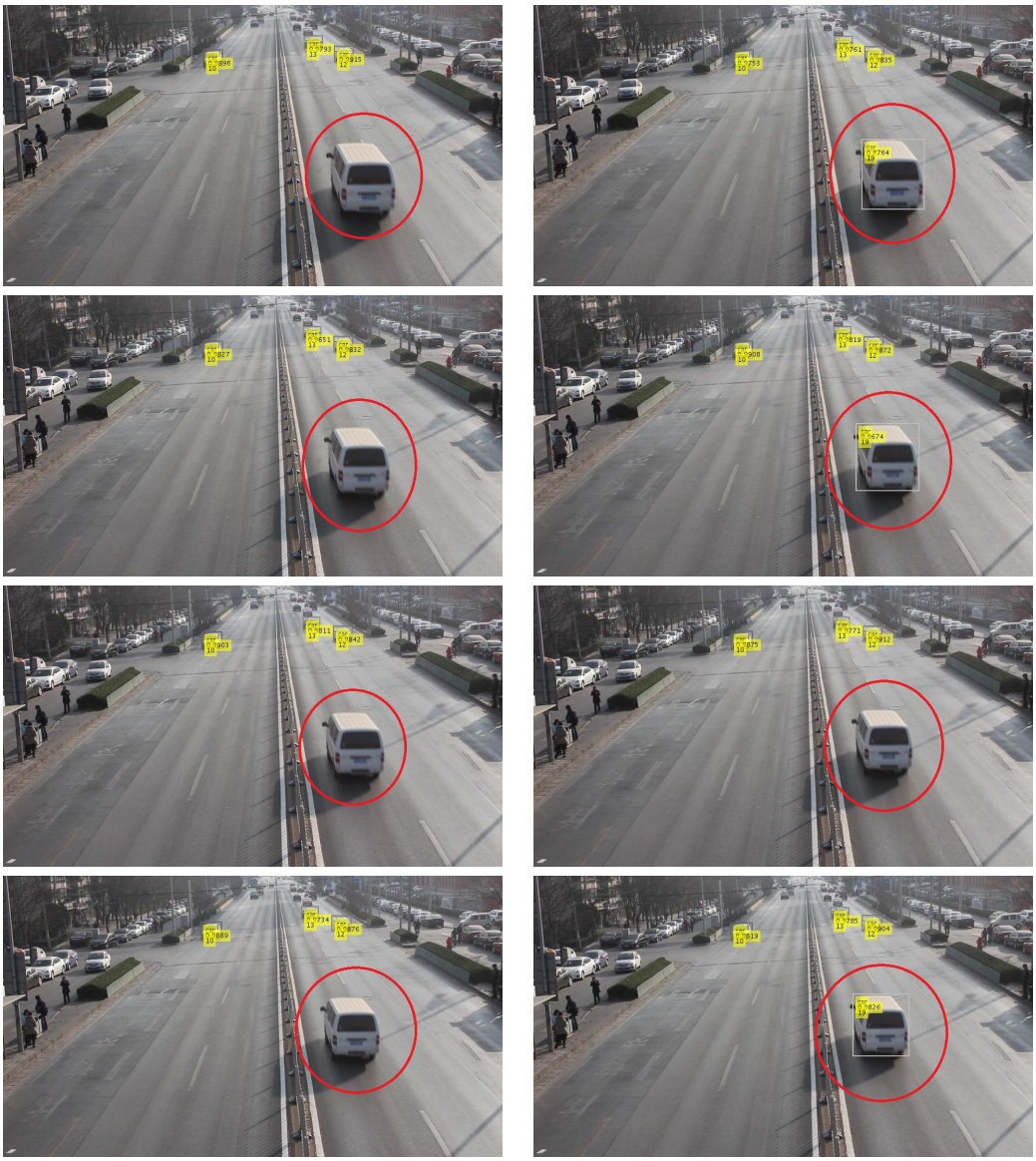


FIGURE A.3 – Larger illustration of Sequence 2a in Figure 3.14

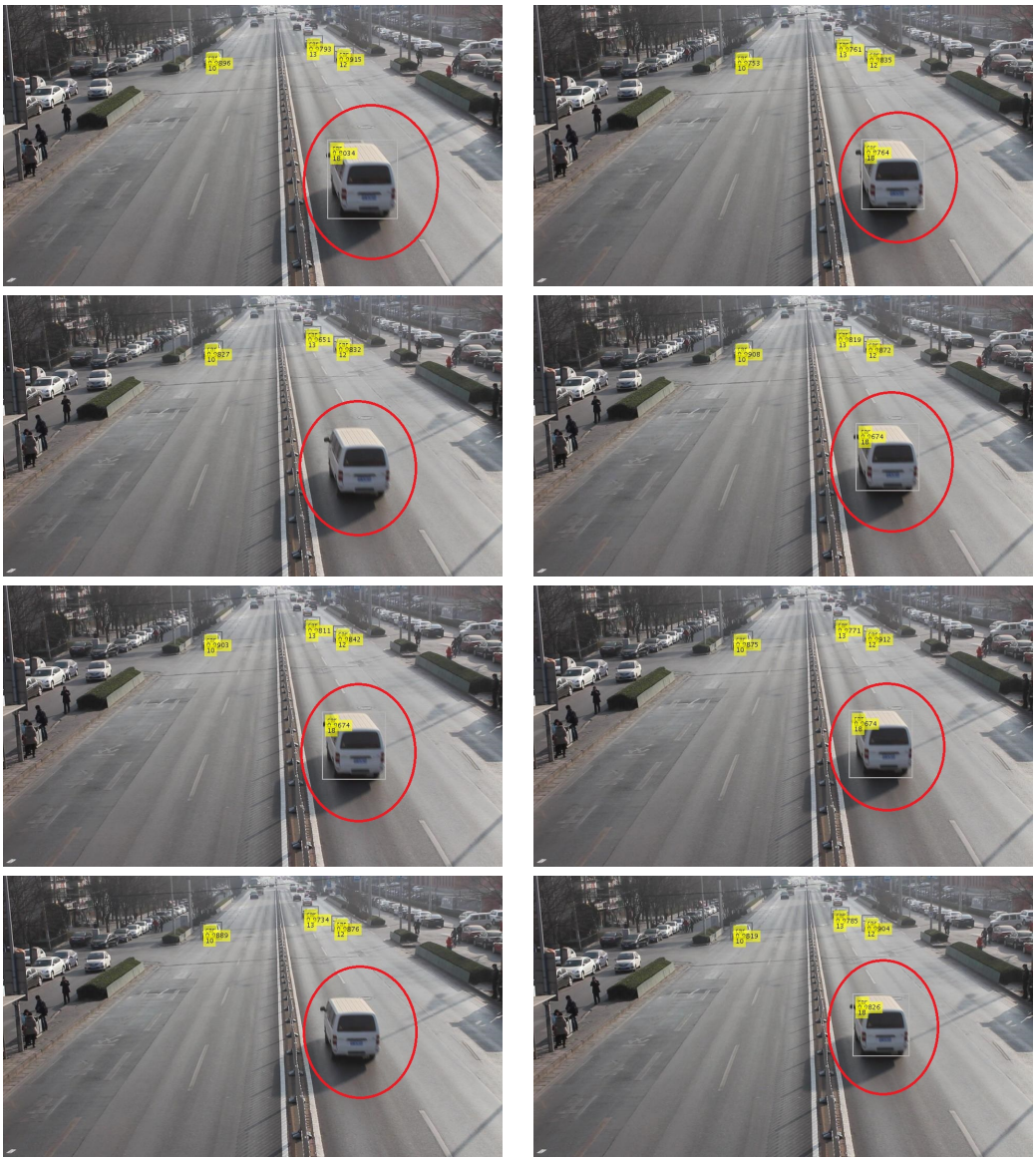


FIGURE A.4 – Larger illustration of Sequence 2b in Figure 3.14

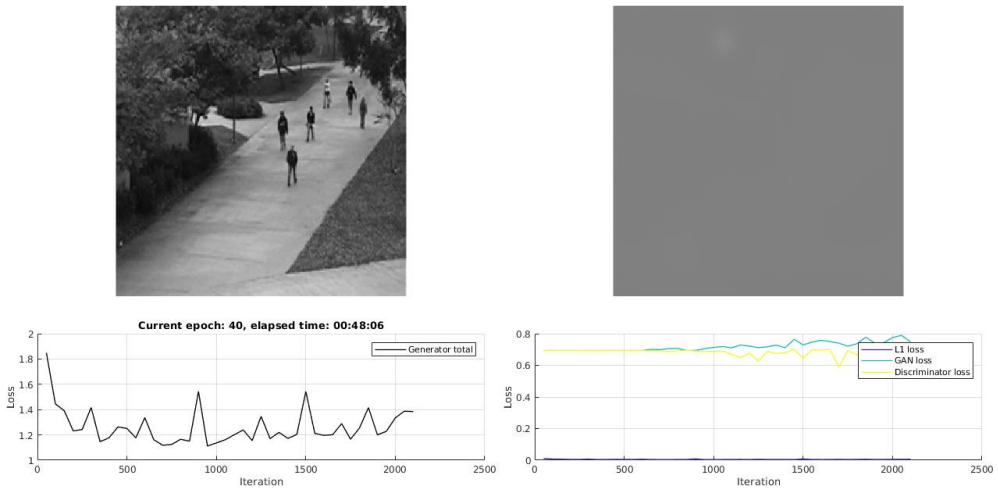


FIGURE A.5 – Illustration of training CGAN-1 on Ped1 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

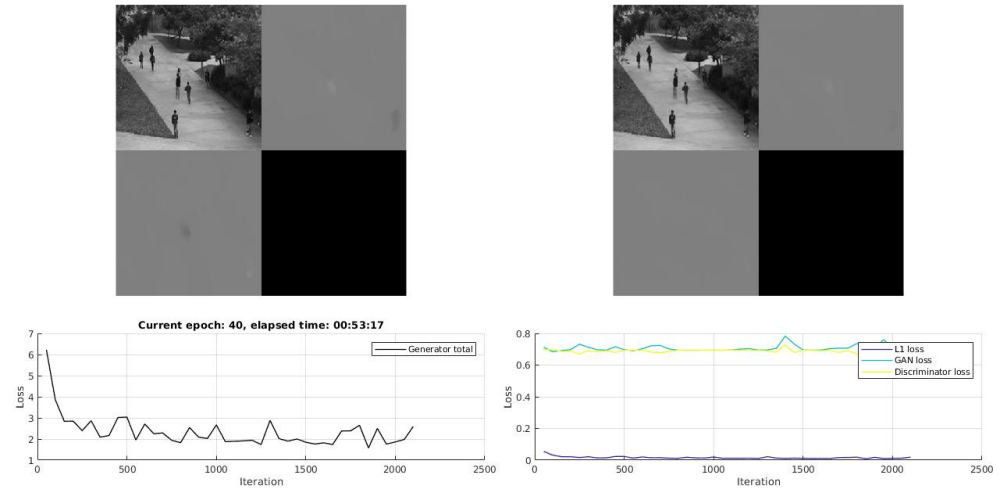


FIGURE A.6 – Illustration of training CGAN-2 on Ped1 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

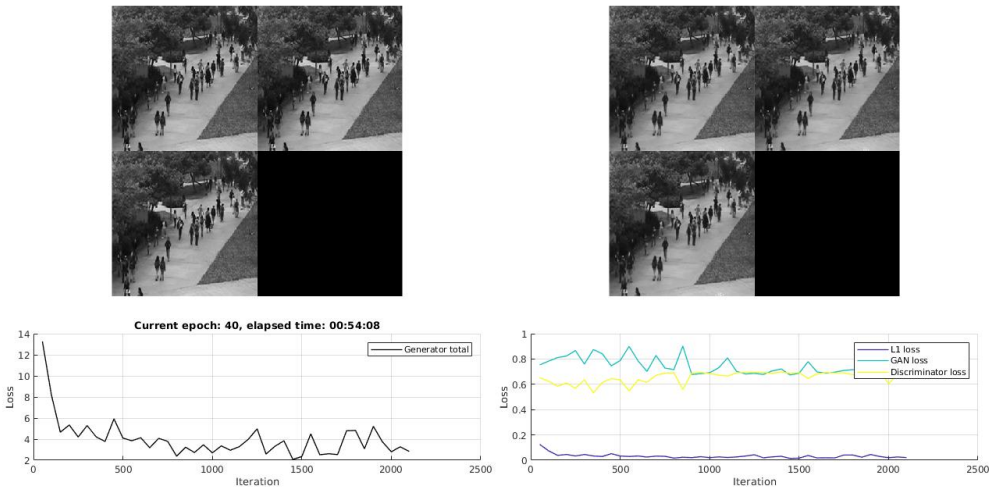


FIGURE A.7 – Illustration of training CGAN-3 on Ped1 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

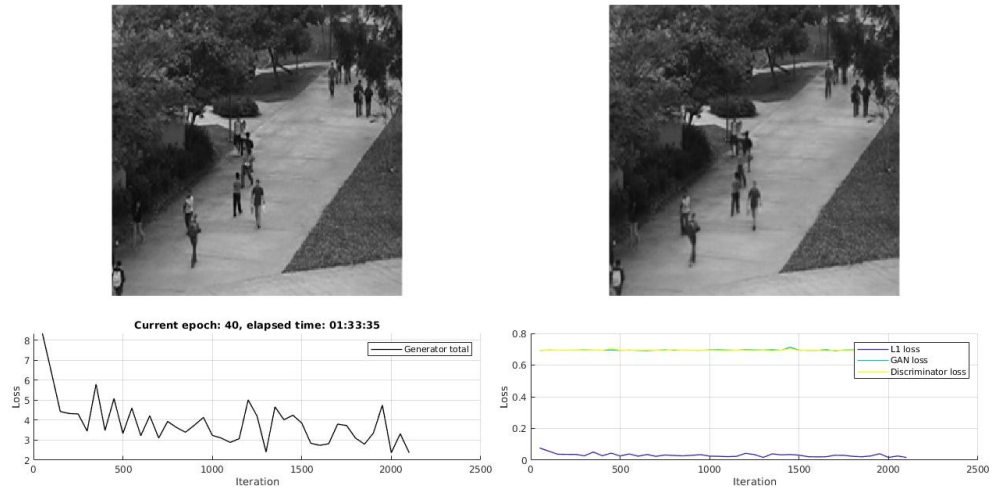


FIGURE A.8 – Illustration of training CGAN-4 on Ped1 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

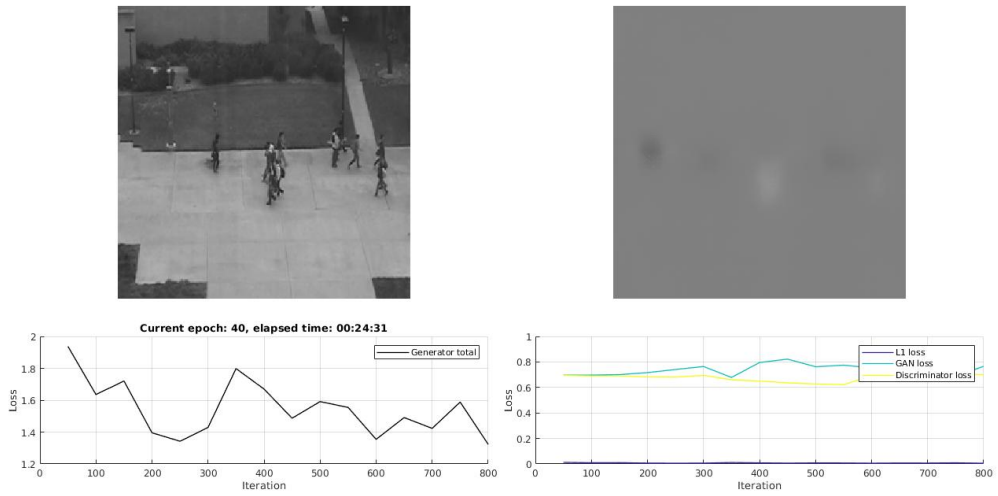


FIGURE A.9 – Illustration of training CGAN-1 on Ped2 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

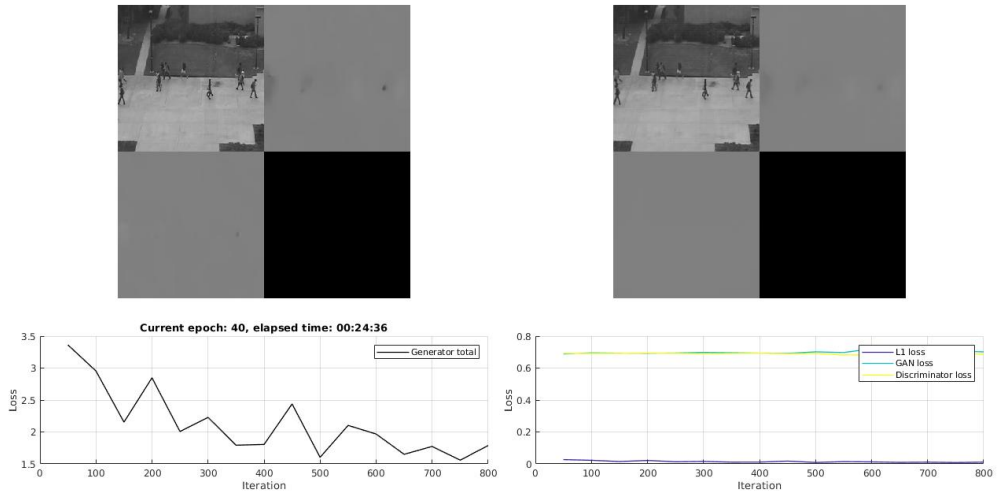


FIGURE A.10 – Illustration of training CGAN-2 on Ped2 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

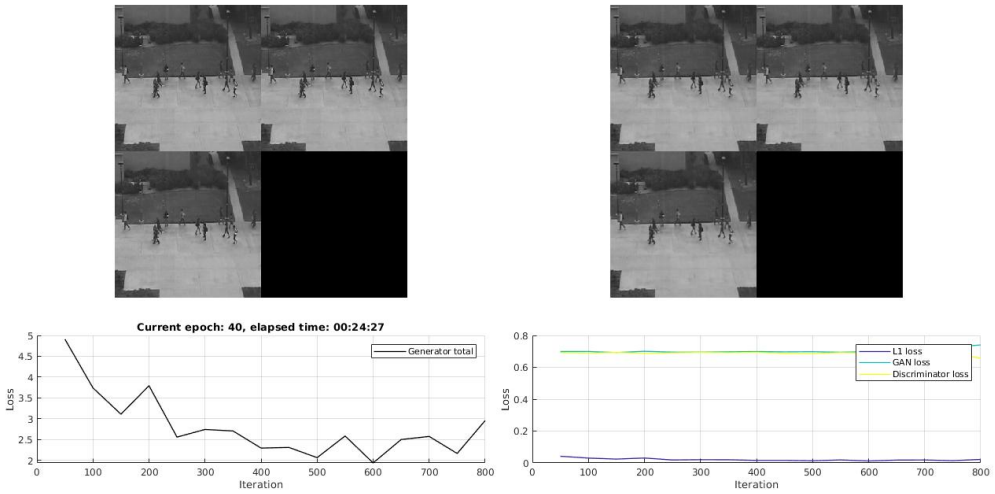


FIGURE A.11 – Illustration of training CGAN-3 on Ped2 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

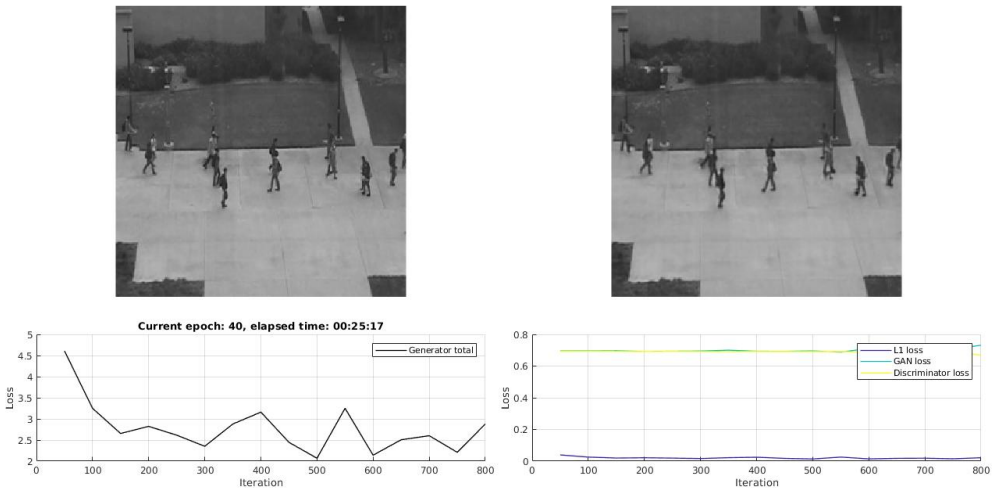


FIGURE A.12 – Illustration of training CGAN-4 on Ped2 dataset. We achieve good loss convergences at $E = 40$, $M = 128$ and $A = eLU$.

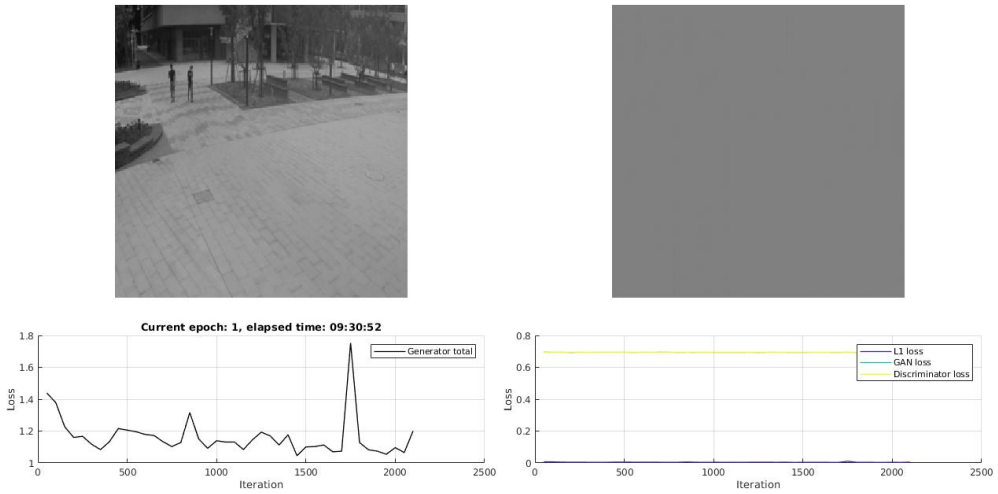


FIGURE A.13 – Illustration of training CGAN-1 on ShanghaiTech dataset. We achieve good loss convergences at $E = 1$, $M = 128$ and $A = eLU$.

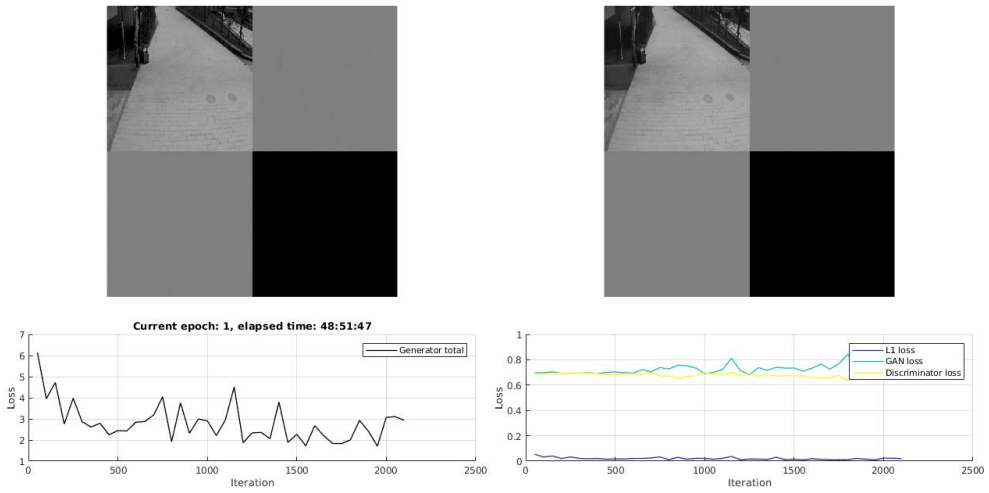


FIGURE A.14 – Illustration of training CGAN-2 on ShanghaiTech dataset. We achieve good loss convergences at $E = 1$, $M = 128$ and $A = eLU$.

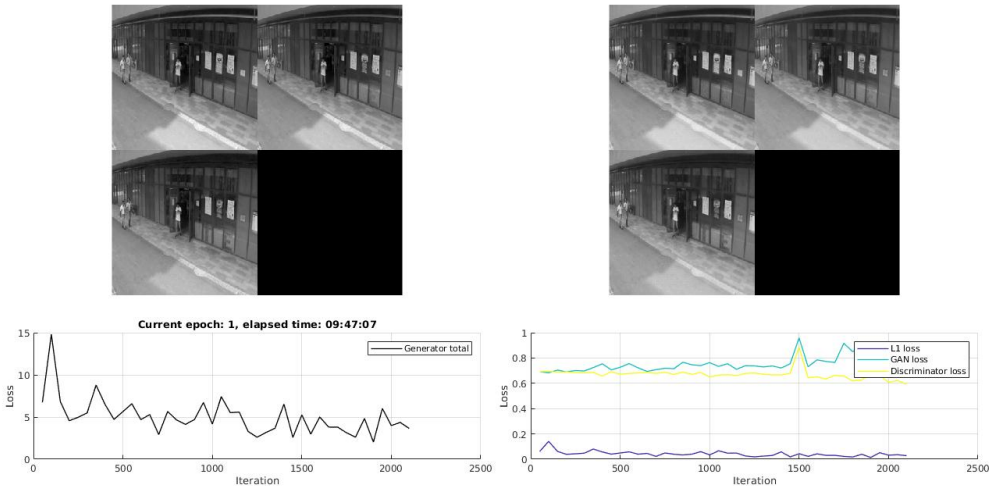


FIGURE A.15 – Illustration of training CGAN-3 on ShanghaiTech dataset. We achieve good loss convergences at $E = 1$, $M = 128$ and $A = eLU$.

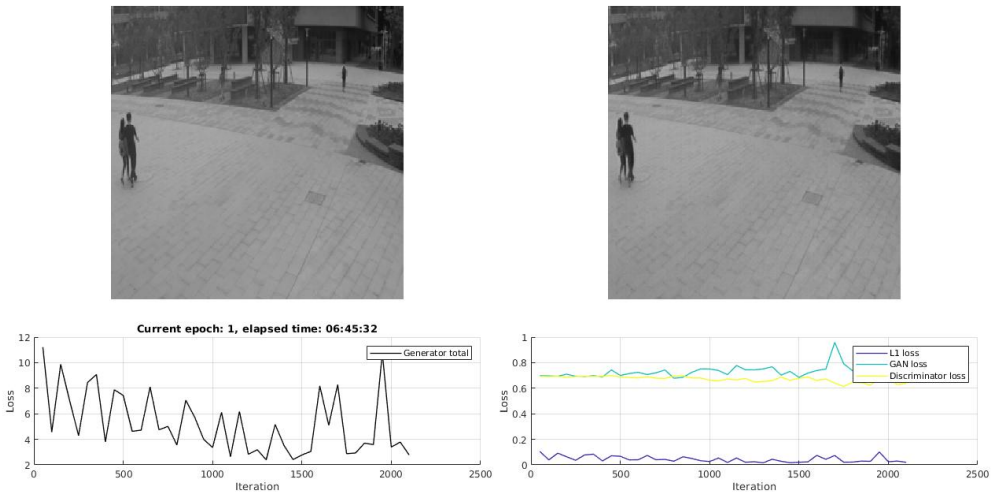


FIGURE A.16 – Illustration of training CGAN-4 on ShanghaiTech dataset. We achieve good loss convergences at $E = 1$, $M = 128$ and $A = eLU$.

Table des matières

ANOMALY DETECTION AND OBJECT TRACKING BY FUTURE PREDICTION USING GENERATIVE METHODS FOR TRANSPORTATION

Résumé

Actuellement, le traitement automatiquement des problèmes de transport devient un sujet actif. Dans le cadre de ce travail, nous visons à relever un défi spécifique dans ce domaine : la détection et le suivi des anomalies. Notre objectif est de construire un système flexible et efficace produisant des performances élevées sur diverses bases de données publiques. Le contexte de notre recherche est l'amélioration des approches précédentes pour obtenir de meilleurs résultats. Nous traitons deux scénarios conduisant à deux méthodes mentionnées dans les parties suivantes : (1) la segmentation et le suivi des véhicules et des piétons par des prédictions utilisant des méthodes génératives classiques basées sur des descripteurs a priori (hand crafted) et sur l'estimation des flux optiques ; (2) la détection des anomalies par des prédictions utilisant des systèmes génératifs multicanaux profonds et l'apprentissage supervisé.

Notre première recherche vise à l'évaluation des performances de l'approche générative classique pour les prévisions et la détermination des ses capacités à améliorer la segmentation et le suivi d'objets. Récemment, divers détecteurs d'apprentissage profond ont été proposés *e.g.* Mask R-CNN qui permettent une approche efficace du problème de suivi : le suivi par détection. A l'exception de tout autre information visuelle, ce type de tracker rapide ne prend en compte que l'intersection-sur-union (IOU) entre les boîtes de délimitation pour apparier les objets. Ainsi, l'absence d'informations visuelles du tracker IOU combinée avec les possibles défaillances des détecteurs créent des trajectoires fragmentées. Nous proposons alors un tracker amélioré basé sur la détection par suivi et sur l'estimation du flux optique. Notre solution génère de nouvelles détections ou segmentations basées sur une translation temporelle en avant et en arrière des résultats des détecteurs CNNs en utilisant les vecteurs de flot optique. Cette étape permet de combler une première partie des lacunes des trajectoires. Les résultats qualitatifs montrent alors que notre solution a obtenu des performances stables avec différentes méthodes d'estimation du flot optique. Les lacunes résiduelles au sein des trajectoires sont traitées en utilisant des caractéristiques SURF. La base de données DAVIS est utilisée pour évaluer la meilleure façon de générer de nouvelles détections. Enfin, le tracker résultant est testé sur la base de données DETRAC. Les résultats qualitatifs montrent que notre approche diminue très significativement la fragmentation des trajectoires. Pour les travaux futurs associés à ce tracker, nous prévoyons d'appliquer les réseaux CGAN développés dans le cadre de la seconde partie de notre travail afin de proposer un système compétitif de suivi d'objet basé prévision.

Malgré les résultats tangibles de cette première approche, les méthodes classiques présentent des limitations importantes concernant la détection d'anomalies qui est l'un de nos objectifs principaux. La fréquence plus faible des événements anormaux donne un scénario déséquilibré et leurs caractéristiques ne suivent généralement aucune relation spatiale ou temporelle. Face à ces défis, la plupart des méthodes de l'état-de-l'art se basent sur des réseaux prédictifs et utilisent les erreurs entre informations générées et réelles comme caractéristiques de détection. Inspirés par cette approche, d'une part, nous proposons un cadre multicanal flexible pour générer des caractéristiques multitypes au niveau image. D'autre part, nous étudions la possibilité d'améliorer les performances de détection par un apprentissage supervisé. Notre système est ainsi basé sur quatre GAN conditionnels (CGAN) prenant en entrée différents types d'informations d'apparence et de mouvement et produisant des informations de prédiction. Ces CGAN représentent la distinction entre événements normaux et anormaux. Ensuite, la différence entre les informations générées et les vérité-terrains est encodée par le pic du rapport signal / bruit (PSNR). Nous classons alors ces caractéristiques dans un contexte supervisé en construisant un petit ensemble d'entrainement à partir de quelques échantillons anormaux de l'ensemble de test original. C'est un Séparateur à Vaste Marge (SVM) qui est appliquée pour la détection des anomalies au niveau trame. Enfin, nous utilisons Mask R-CNN comme détecteur pour effectuer la localisation d'anomalies centrées objet. Notre solution est largement évaluée sur les bases de données Avenue, Ped1, Ped2 et ShanghaiTech. Nos résultats démontrent que les caractéristiques de PSNR combinées avec le SVM supervisé sont meilleures que les cartes d'erreurs calculées par les méthodes précédentes. En particulier, pour la base de données la plus difficile qu'est Shanghaitech, notre modèle surpasse jusqu'à 9% l'état-de-l'art des methodes non-supervisées. En perspective, nous prévoyons de construire une base de données pour la détection d'anomalies dans un cadre semi-supervisé, et d'intégrer un classifieur one-class SVM pour proposer un système "de bout en bout".

Mots clés : détection d'anomalies, apprentissage profond, modèle génératif, application au transport

---

CERI SN - IMT Lille Douai

– – – – –

Abstract

Today, automatic solving transportation problem becomes active subject. In our PhD project, we aim to address a specific challenge in this domain: anomaly detection and tracking. Our ultimate goal is constructing a flexible and effective framework producing high performance on various public datasets. The context of our research is applying and improving previous successful approaches to achieve better results. We deal with two scenarios leading to two methods mentioned in following parts: (1) vehicles and road users segmentation and tracking by future predictions using classical hand-crafted generative methods based on optical flow estimation; (2) anomaly detection by future predictions using multi-channels deep generative frameworks and supervised learning.

Our first research is evaluating the performance of the classical hand-crafted generative approach in future prediction and its capability for improving segmentation and tracking. Recently, there existed various strong deep learning detectors *e.g.* Mask R-CNN lead to an effective approach for tracking problem: tracking-by-detection. This very fast type of tracker considers only the Intersection-Over-Union (IOU) between bounding boxes to match objects without any other visual information. In contrast, the lack of visual information of IOU tracker combined with the failure detections of CNNs detectors create fragmented trajectories. We propose an enhanced tracker based on tracking by-detection and optical flow estimation in vehicle tracking scenarios. Our solution generates new detections or segmentations based on translating backward and forward results of CNNs detectors by optical flow vectors. This task can fill in the gaps of trajectories. The qualitative results show that our solution achieved stable performance with different types of flow estimation methods. Then we match generated results with fragmented trajectories by SURF features. DAVIS dataset is used for evaluating the best way to generate new detections. Finally, the entire process is tested on DETRAC dataset. The qualitative results show that our methods significantly improve the fragmented trajectories. For future work, we plan to apply CGANs streams of second work for the first task to propose a new competitive process of future prediction for segmentation and tracking.

Despite the moderate success of the first work, there is significant limitations of classical approaches to deal with our main task: anomaly detection. The lower frequency of abnormal events leads to an unbalanced scenario and the features of abnormal events usually do not follow any spatial or temporal relationship. It is also difficult to pre-define the structure or class of abnormal events. Facing to those challenge, most of state-of-the-art (SOTA) anomaly detection methods are based on apparent motion and appearance reconstruction networks and use error estimation between generated and real information as detection features. These approaches achieve promising results by only using normal samples for training steps. In this thesis, our contributions are two-fold. On the one hand, we propose a flexible multichannel framework to generate multi-type frame-level features. On the other hand, we study how it is possible to improve the detection performance by supervised learning. The multi-channel framework is based on four Conditional GANs (CGANs) taking various types of appearance and motion information as input and producing prediction information as output. These CGANs provide a better feature space to represent the distinction between normal and abnormal events. Then, the difference between those generative and ground-truth pieces of information is encoded by Peak Signal-to Noise Ratio (PSNR). We propose to classify those features in a classical supervised scenario by building a small training set with some abnormal samples of the original test set of the dataset. The binary Support Vector Machine (SVM) is applied for frame-level anomaly detection. Finally, we use Mask R-CNN as a detector to perform object-centric anomaly localization. Our solution is largely evaluated on Avenue, Ped1, Ped2 and ShanghaiTech datasets. Our experiment results demonstrate that PSNR features combined with supervised SVM are better than error maps computed by previous methods. We achieve SOTA performance for frame-level AUC on Avenue, Ped1 and ShanghaiTech. Especially, for the most challenging Shanghaitech dataset, a supervised training model outperforms up to 9% the SOTA on unsupervised strategy. Furthermore, we keep in progress several promising ways: building a new dataset for semi-supervised anomaly detection containing both normal and abnormal samples in its training set and applying one-class SVM to propose an end-to-end framework.

Keywords: anomaly detection, deep learning, generative model, transportation application

CERI SN - IMT Lille Douai

Bibliographie

[1] Z. Harchaoui A. Gaidon and C. Schmid. Temporal localization of actions with actoms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35, 2013.

[2] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet : A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12) :2481–2495, 2017.

[4] S. Bae and K. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225, 2014.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3) :346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia.

[6] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates : How important is recognition for 3d scene flow estimation in autonomous driving scenarios ? In *International Conference on Computer Vision (ICCV)*, 2017.

[7] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017*, Lecce, Italy, Aug. 2017.

[8] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1515–1522, 2009.

[9] Thomas Brox and Jitendra Malik. Large displacement optical flow : Descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 :500–513, 2011.

[10] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau. Deep manta : A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *CVPR*, 2017.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018.

[12] Liang-Chieh Chen, G. Papandreou, I. Kokkinos, Kevin Murphy, and A. Yuille. Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully

connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 :834–848, 2018.

[13] Q. Chen and V. Koltun. Full flow : Optical flow estimation by global optimization over regular grids. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4706–4714, 2016.

[14] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised Object Discovery and Localization in the Wild : Part-based Matching with Bottom-up Region Proposals. In *CVPR - IEEE Conference on Computer Vision & Pattern Recognition*, pages 1201–1210, Boston, United States, June 2015. IEEE.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] J.J. D. Oneata, J.J. Verbeek and C. Schmid. Efficient action localization with approximately normalized fisher vectors. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[17] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[18] C. Dicle, O. I. Camps, and M. Sznaier. The way they move : Tracking multiple targets with similar appearance. In *2013 IEEE International Conference on Computer Vision*, pages 2304–2311, 2013.

[19] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet : Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2758–2766. IEEE Computer Society, 2015.

[20] Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *NIPS*, 2016.

[21] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2flow : Motion hallucination from static images for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5947, 2017.

[22] Allison Del Giorno, J. Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *ECCV*, 2016.

[23] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society, 2015.

[24] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[26] S. Hamdi, S. Bouindour, K. Loukil, H. Snoussi, and M. Abid. Hybrid deep learning and hof for anomaly detection. In *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 575–580, 2019.

[27] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning temporal regularity in video sequences. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 733–742, 2016.

[28] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows : Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9) :1904–1916, 2015.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[32] João F. Henriques, Rui Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 2012.

[33] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3639–3647, 2017.

[34] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17 :185–203, 1981.

[35] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0 : Evolution of optical flow estimation with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1647–1655. IEEE Computer Society, 2017.

[36] Radu Tudor Ionescu, F. Khan, Mariana-Iuliana Georgescu, and L. Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. *CVPR*, pages 7834–7843, 2019.

[37] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *CVPR*, 2018.

[38] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2922, 2017.

[39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[40] V. Kantorov and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2014*, 2014.

[41] J. Kim and K. Grauman. Observe locally, infer globally : A space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928, June 2009.

[42] Alexander Klaser, Marcin Marszalek, and Cordelia Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Mark Everingham, Chris Needham, and Roberto Fraile, editors,

BMVC 2008 - 19th British Machine Vision Conference, pages 275 :1–10, Leeds, United Kingdom, Sept. 2008. British Machine Vision Association.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

[44] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB : a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[45] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Unsupervised Object Discovery and Tracking in Video Collections. In *ICCV 2015 - IEEE International Conference on Computer Vision*, pages 3173–3181, Santiago, Chile, Dec. 2015. IEEE.

[46] Ivan Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3) :107–123, Sept. 2005.

[47] S. Lee, H. G. Kim, and Y. M. Ro. Stan : Spatio-temporal adversarial networks for abnormal event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1323–1327, 2018.

[48] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li. The lv dataset : A realistic surveillance video dataset for abnormal event detection. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2017.

[49] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet : Multi-path refinement networks for high-resolution semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5168–5177. IEEE Computer Society, 2017.

[50] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO : common objects in context. *ECCV*, 2014.

[51] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD : single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 21–37. Springer, 2016.

[52] W. Liu, Weixin Luo, Zhengxin Li, P. Zhao, and Shenghua Gao. Margin learning embedded prediction for video anomaly detection with a few anomalies. In *IJCAI*, 2019.

[53] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - a new baseline. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2017.

[54] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[55] William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *ICLR*, 2016.

[56] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[57] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2) :91–110, Nov. 2004.

[58] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. *ICCV*, 2013.

[59] Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting Future Instance Segmentation by Forecasting Convolutional Features. In *ECCV 2018 - European Conference on Computer Vision*, volume 11213 of *Lecture Notes in Computer Science*, pages 593–608, Munich, Germany, Sept. 2018. Springer.

[60] B.D. Lucas and T.Kanade. An iterative image registration technique with an application to stereo vision. *Proceedings of Imaging Understanding Workshop*, pages 121–130, 1981.

[61] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.

[62] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, 2017.

[63] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349, 2017.

[64] J.Jia L.Xu and Y.Matsushita. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[65] Siwei Lyu, Ming-Ching Chang, Dawei Du, Longyin Wen, Honggang Qi, Yuezun Li, Yi Wei, Lipeng Ke, Tao Hu, Marco Del Coco, et al. Ua-detrac 2017 : Report of avss2017 & iwt4s challenge on advanced traffic monitoring. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–7. IEEE, 2017.

[66] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, June 2010.

[67] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[68] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *ICLR*, abs/1511.05440, 2015.

[69] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8) :873–889, Aug 2001.

[70] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[71] Trong-Nguyen Nguyen and Jean Meunier. Hybrid deep network for anomaly detection. *BMVC*, 2019.

[72] Trong Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. *ICCV*, 2019.

[73] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *ECCV*, 2017.

[74] Dan Oneata, Jérôme Revaud, Jakob Verbeek, and Cordelia Schmid. Spatio-Temporal Object Detection Proposals. In *ECCV - European Conference on Computer Vision*, volume 8691, pages 737–752, Zurich, Switzerland, Sept. 2014. Springer.

[75] Yuqi Ouyang and Victor Sanchez. Video anomaly detection by estimating likelihood of representations. *ICPR*, 2021.

[76] Nikos Papanikolopoulos. Unusual crowd activity dataset of university of minesota. `http://mha.cs.umn.edu/index.shtml`. Accessed : 2015.

[77] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[78] Justine Pinkney. Pix2pix, matlab, 2020.
[79] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
[80] B. Ramachandra, M. Jones, and R. R. Vatsavai. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[81] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581, 2017.
[82] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016.
[83] J. Redmon and A. Farhadi. Yolo9000 : Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.
[84] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1060–1069. JMLR.org, 2016.
[85] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc., 2015.
[86] Jérôme Revaud, Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Epicflow : Edge-preserving interpolation of correspondences for optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1164–1172. IEEE Computer Society, 2015.
[87] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net : Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.
[88] M. Sabokrou, Mohammad Khalooei, M. Fathy, and E. Adeli. Adversarially learned one-class classifier for novelty detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018.
[89] V. Saligrama, J. Konrad, and P. Jodoin. Video anomaly identification. *IEEE Signal Processing Magazine*, 27(5), 2010.
[90] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press.
[91] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
[92] K. Soomro, A. Zamir, and M. Shah. Ucf101 : A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, 2012.

[93] W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[94] D. Sun, X. Yang, M. Liu, and J. Kautz. Pwc-net : Cnns for optical flow using pyramid, warping, and cost volume. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[95] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015.

[96] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning Motion Patterns in Videos. In *CVPR - IEEE Conference on Computer Vision & Pattern Recognition*, Honolulu, United States, July 2017.

[97] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[98] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, abs/1706.08033, 2017.

[99] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3560–3569. PMLR, 2017.

[100] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 613–621, Red Hook, NY, USA, 2016. Curran Associates Inc.

[101] H. Vu, T. Nguyen, Trung Le, Wei Luo, and Dinh Q. Phung. Robust anomaly detection in videos using multilevel representations. In *AAAI*, 2019.

[102] Hung Vu, Tu Dinh Nguyen, Anthony Travers, Svetha Venkatesh, and Dinh Phung. Energy-based localized anomaly detection in video surveillance. In *Advances in Knowledge Discovery and Data Mining*, Lecture Notes in Computer Science, pages 641–653. Springer, 2017.

[103] Jacob Walker, Abhinav Gupta, and Martial Hebert. Dense optical flow prediction from a static image. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2443–2451, 2015.

[104] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1) :60–79, May 2013.

[105] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, Sydney, Australia, 2013.

[106] Jue Wang and Anoop Cherian. Gods : Generalized one-class discriminative subspaces for anomaly detection. *ICCV*, 2019.

[107] L. Wang, Y. Lu, H. Wang, Y. Zheng, H. Ye, and X. Xue. Evolving boxes for fast vehicle detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1135–1140, 2017.

[108] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4305–4314, 2015.

[109] T. Wang and H. Snoussi. Histograms of optical flow orientation for visual abnormal events detection. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 13–18, Sep. 2012.

[110] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow : Large displacement optical flow with deep matching. In *IEEE Intenational Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013.

[111] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1282–1289, 2014.

[112] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track : Online multi- object tracking by decision making. In *International Conference on Computer Vision (ICCV)*, pages 4705–4713, 2015.

[113] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers : Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2129–2137, 2016.

[114] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

[115] T. Zhang, H. Lu, and S. Z. Li. Learning semantic scene models by object classification and trajectory clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1947, June 2009.