Contribution à la Perception d'Environnement pour la Smart Mobilité

Redouane Khemmar

Normandie Université

Habilitation à Diriger des Recherches

Ecole Doctorale MIIS – ED 590

Spécialité : Automatique, Signal, Productique, Robotique

Contribution à la Perception d'Environnement pour la Smart Mobilité

Présentée et soutenue par
Radouane KHEMMAR

HDR soutenue publiquement le 28/11/2022 devant le jury composé de		
M. Fawzi NASHASHIBI	Directeur de Recherche INRIA Centre de Paris	Président du Jury
M. Dominique GRUYER	Directeur de Recherche Université Gustave Eiffel – IFSTTAR – Département COSYS	Rapporteur
M. Fabrice MÉRIAUDEAU	Professeur des Universités Université de Bourgogne (ImVia)	Rapporteur
Mme Samia BOUCHAFA-BRUNEAU	Professeur des Universités Université d'Evry Val d'Essonne/Université Paris-Saclay (IBISC)	Rapporteur
M. Gareth HOWELLS	Professeur des Universités University of Kent (School of Computing)	Examinateur
M. Adel GHAZEL	Professeur des Universités ESIGELEC (IRSEEM)	Examinateur
M. Paul HONEINE	Professeur des Universités Université de Rouen Normandie (LITIS)	Examinateur
M. Abdelaziz BENSRHAIR	Professeur des Universités INSA Rouen Normandie (LITIS)	Examinateur – Garant

Institut de Recherche en Systèmes Electroniques Embarqués (IRSEEM) – UR 4353

UNIVERSITÉ DE ROUEN NORMANDIE

MIIS

ESIGELEC irseem

Liste des Abréviations, Sigles et Acronymes

ADAS	Advanced Driver Assistance System
YOLO	You Only Look Once
KITTI	Karlsruhe Institute of Technology and Toyota technological Institute at chicago
IA	Intelligence Artificielle
SSS	Small Sample Size
ORL	Olivetti Research Lab
LBP	Local Binary Patterns
CNN	Convolutional Neural Network
SORT	Simple Online Realtime Tracking
FKE	Filtre de Kalman Etendu
LSTM	Long Short-Term Memory
STAF	Spatio-Temporal Attention Framework for Understanding Road Agents Behaviors
RailSem19	A Dataset for Semantic Rail Scene Understanding
INRIA	Institut National de Recherche en Sciences et Technologies du Numérique
SLAM	Simultaneous Localisation and Mapping
NUScenes	NuTonomy Scenes
SGD	Stochastic Gradient Descent
RoI	Region Of Interest
DRN	Deep Regression Networks
FPS	Frame Per Seconde
IoU	Intersection over Union
M-TL	Multi-Task Learning
GTAV	Grand Theft Auto V
RCSF	Réseau de Capteurs Sans Fils
BESM	Bureau d'Etudes Systèmes Mécatroniques
AES	Automotive Embedded Systems
CEIS	Connected Embedded Intelligent Systems
SEI	Système Embarqués et Instrumentation
IIS	Instrumentation, Informatique et Systèmes
ATER	Attaché Temporaire d'Enseignement et de Recherche

CIES	Centre d'Initiation à l'Enseignement Supérieur
IRSEEM	Institut de Recherche en Systèmes Electroniques EMbarqués
APP	Apprentissage Par Projets
SAM	Système A Microprocesseurs
IMU	Inertial Measurement Unit
SURF	Speeded Up Robust Features
SPA	Shortest Path Algorithm
SVC	Space Variant Convolution
ESRORAD	Esigelec and Segula technologies ROad and RAilway Dataset
AST	Adaptive Super Twisting
LLP	Lazy Learning Paradigm
HOG	Histogram of Oriented Gradient
DPM	Deep Part Model
SPP	Sparsity Preserving Projection
BAT 3D	3D Bounding Box Annotation Tool
LPP	Local Preserving Projection
NPE	Neighborhood Preserving Embedding
NN	Nearest Neighbor
CS	Compressed Sensing
PTZ	Pan Tilt Z Camera
ROAD	ROad event Awareness Dataset for Autonomous Driving
RBA	Road Behavior Understanding
MOT	Multi-Object Tracking
ITS	Intelligent Transportation System
NDC	Normalized Device Coordinate
JAAD	Joint Attention in Autonomous Driving
HDD	Honda Research Institute Driving Dataset
GCN	Graph Convolutional Networks
mAP	mean Average Precision
RNN	Recurrent Neural Network
NMS	Non Maximum Suppression
CA	Class Accuracy
ICP	Inception Conditional Probability
RPN	Regional Purpose Network
GAP	Global Average Pooling
GCN	Graph Convolutional Networks
FCN	Fully Convolutional Networks
FC	Fully Connected

Table des matières

Introduction

Contexte. Les véhicules autonomes au sens large (robots mobiles, voitures, drones, trains, etc.) sont de plus en plus présents dans notre quotidien, ouvrant de nouvelles perspectives en matière de smart mobilité. D'une manière générale, un système de navigation autonome doit comprendre 3 fonctions essentielles : perception, décision et action. Un véhicule doit percevoir son environnement, le comprendre, prendre des décisions d'actions pour les exécuter par la suite. Donc plus le système est capable de percevoir son environnement, plus il prendra de meilleures décisions lui permettant, *in fine*, de déclencher des actions de haute qualité répondant aux différentes exigences de sécurité, de confort et d'énergie. La perception d'environnement représente donc un enjeu majeur de la smart mobilité.

Cette perception doit garantir non seulement une bonne détection d'objets, mais aussi une bonne interprétation de la scène. La détection d'objets, leur localisation ainsi que leur tracking temps-réel sont des tâches très importantes pour la perception d'environnement. La détection d'objet vise à trouver tous les objets d'intérêts dans l'image en identifiant leurs catégories et en localisant leurs positions. Dans la smart mobilité multimodale routière et ferroviaire, les objets d'intérêt à détecter sont les véhicules, les piétons, les bus, les cyclistes, les arbres, etc. Les informations relatives aux positions des objets, combinées aux images de profondeurs permettent d'estimer la distance et les localiser donc dans l'espace par rapport au véhicule. La prédiction des trajectoires de ces objets peut ensuite être calculée à l'aide des algorithmes de suivi d'objets (filtre de Kalman par exemple) afin d'assurer un tracking temps-réel.

Au cours des dernières décennies, plus précisément depuis 2012, le deep learning est devenu un outil très puissant en raison de sa capacité à traiter de grandes quantités de données. L'apparition de nombreuses méthodes basées sur l'apprentissage profond a conduit à des progrès significatifs. L'intérêt d'utiliser de plus en plus de couches cachées et intermédiaires entièrement connectées a dépassé les techniques traditionnelles de la vision par ordinateur, en particulier dans la reconnaissance des formes, la détection d'objets et la classification. Avec l'essor de ces approches et l'émergence des réseaux de neurones convolutifs (Convolutional Neural Network - CNN), ainsi le besoin d'améliorer la perception d'environnement, a conduit au développement de nouvelles méthodes de détection d'objets, estimation de la

profondeur, tracking, segmentation sémantique, etc. Cela est fait en utilisant soit un capteur de profondeur de type LiDAR, soit des caméras.

Problématique. Bien que la détection d'objets soit un domaine très recherché et un facteur clé d'une bonne perception d'environnement, elle demeure le problème le plus difficile dans la vision par ordinateur (forme, éclairage, occlusion, etc.). Malgré le recours aux technologies de l'Intelligence Artificielle (IA) pour la smart mobilité est devenu plus qu'indispensable, peu de méthodes se concentrent sur l'aspect temps-réel, essentiel pour les applications réelles et ce, en raison des coûts de calcul élevés. La question de la vitesse représente donc un véritable défi pour un système embarqué. Sous l'impulsion du big data, les modèles CNN peuvent être formé efficacement à l'aide d'un grand nombre de données d'annotation. Ces dernières années, les performances des algorithmes de détection d'objets par deep learning se sont améliorées. Cependant, ces dernières présentent des lacunes évidentes dans les scènes complexes, en partie à cause du manque de données étiquetées.

Par exemple dans la smart mobilité routière et ferroviaire, même si les objets d'intérêt à détecter sont les mêmes (véhicules, piétons, cyclistes, etc.), le secteur routier est largement couvert par la littérature scientifique, c'est moins le cas pour le ferroviaire (train autonome) ou encore la santé (fauteuil roulant intelligent). Ceci est en partie dû à l'absence de jeux de données spécifiques avec des données vérité terrain dédiées à la détection d'objets. Cela représente, en plus, un autre défi dans la mesure où les approches de détection d'objets, d'estimation de distance et de tracking nécessitent une phase d'entrainement avec un jeu de données incluant une vérité terrain. Dans le cas du secteur ferroviaire, les datasets publics dédiés aux véhicules autonomes ne fonctionnement pas car les images ont été prises depuis un point de vue de la route, différent de celui du rail. Le même constat pour le secteur de santé où les fauteuils roulants se déplacent sur des trottoirs.

En plus des deux contraintes évoquées qui sont la précision et la vitesse, les approches de perception d'environnement doivent prendre en compte l'aspect énergétique dans le sens où ces algorithmes fonctionnent souvent sur des systèmes embarqués où la contrainte d'énergie demeure élevée. Il est donc essentiel que nos ADAS assurent, en plus de la sécurité, une gestion optimale de l'énergie à bord.

Contribution du HDR. Mes travaux de recherche sont concernés par cette problématique de perception d'environnement pour la smart mobilité. Ce mémoire a comme objectif de présenter, d'une manière synthétique, mes travaux de recherche, mes contributions scientifiques effectuées durant ces dernières années ainsi que mes perspectives de recherche. Mes travaux de recherche sur la vision par ordinateur ont commencé en tant que Master Européen à l'Université de Poitiers où j'avais travaillé sur la compression d'images fixes pour le codage

source. Par la suite, et en tant que doctorant à l'Université de Strasbourg, mes travaux de recherche ont porté sur la reconstruction d'objets 3D pour l'inspection dimensionnelle et le contrôle qualité. J'ai été ATER pendant 2 ans au sein de la même Université où j'ai continué à travailler sur la reconstruction d'objets 3D par contours actifs. J'ai par la suite effectué plusieurs postes en industrie en tant que chef de projet R&D avant d'intégrer l'ESIGELEC, en octobre 2009, en tant qu'enseignant chercheur au sein du pôle Instrumentation, Informatique et Systèmes (IIS) de l'IRSEEM. Depuis, j'ai poursuivi mes travaux de recherche sur la perception d'environnement pour la smart mobilité principalement mais aussi avec élargissement à plusieurs activités connexes propres à la recherche par projets. Mes travaux de recherche autour de la smart mobilité se répartissent donc sur 2 axes :

— Perception d'environnement pour la smart mobilité multimodale routière et ferroviaire
— Perception d'environnement pour la robotique mobile et la santé

L'objectif est d'atteindre un niveau d'analyse et de compréhension de scènes complexes permettant d'assurer une smart mobilité de très haut niveau de sécurité, de confort et de consommation d'énergie optimale. Associer la vision aux approches basées deep learning permet d'avoir des solutions robustes d'analyse et de compréhension de scènes complexes. Cela fait appel à deux briques essentielles et complémentaires :

1. Système fusion multicapteurs : permettant d'enrichir davantage la perception avec des données hétérogènes (caméra, LiDAR, RADAR, GPS/IMU). Cela renforce la décision qui à son tour permettra de déclencher des actions de très hautes précisions.
2. Perception d'environnement basée IA : permettant l'exploitation des données collectées pour une meilleure prédiction des situations dans une scène.

Il s'agit donc du développement d'une plateforme générique ouverte pour expérimenter et valider des concepts technologiques et scientifiques de partenaires académiques et industriels. Le domaine d'application est principalement la smart mobilité par IA allant de la robotique mobile aux véhicules autonomes. La figure 1 illustre l'architecture globale de mes travaux de recherche touchant à la smart mobilité par IA.

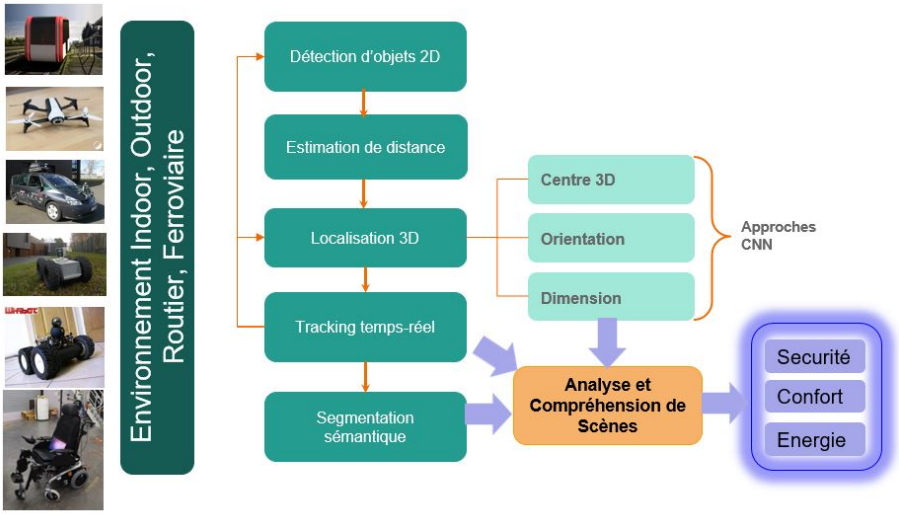

FIGURE 1 – Architecture globale de mes travaux de recherche sur la smart mobilité par IA.

Organisation du mémoire. Ce mémoire, rédigé en vue de l'obtention de l'Habilitation à Diriger des Recherches (HDR), présente une synthèse de mes activités d'enseignement et de recherche menées à l'ESIGELEC et au sein de l'IRSEEM. Il est structuré autour de 7 chapitres.

Dans le **Chapitre 1**, une synthèse de mes activités pédagogiques, administratives et scientifiques est présentée où je résume l'ensemble de mon parcours en tant qu'enseignant chercheur au sein de l'ESIGELEC ainsi qu'à son Institut de Recherche en Systèmes Electroniques EMbarqués (IRSEEM, UR 4353). Dans ce chapitre, je présente mon parcours, mes activités d'enseignement, mes responsabilités pédagogiques, mes encadrements, mes responsabilités de recherche, ainsi qu'une liste de mes publications scientifiques. Dans les chapitres suivants, une description de mes contributions au travers mes activités de recherche et de mes co-encadrements menés depuis 2009, sera présenté.

Le **Chapitre 2** est consacré à la détection d'objets 2D, leur localisation ainsi que leur tracking temps-réel. Il comprend 2 grandes parties. Dans un premier temps, nous présentons un nouveau système basé sur l'apprentissage profond de bout en bout pour la détection multi-objets, l'estimation de la profondeur, la localisation et le suivi d'objets dans un environnement routier. Pour la détection d'objets, une nouvelle approche basée YOLOv3 a été développée. Pour la localisation d'objet, nous présentons deux approches différentes basées Monodepth2 et MADNet. Enfin, nous introduisons une nouvelle méthode de suivi d'objets basée sur une version améliorée de l'approche SORT. Un filtre de Kalman étendu est développé pour améliorer l'estimation de la position des objets pour un meilleur tracking.

Dans la deuxième partie, et suite à certaines limites des algorithmes d'apprentissage

profond, nous commençons par une évaluation des méthodes d'estimation de la profondeur supervisée et non supervisée (BTS et Monodepth2) sur les deux jeux de données à grande échelle, KITTI et NuScenes. Nous présentons par la suite notre nouveau protocole d'évaluation de la profondeur mieux adapté aux applications de conduite autonome. Il comprend une méthode basée sur des plages de distance et un protocole pour évaluer les prédictions de la profondeur des objets. Dans l'ensemble, les résultats obtenus prouvent la robustesse de nos modèles d'apprentissage pour la smart mobilité routière. Ces travaux ont été réalisés dans le cadre de la thèse d'Antoine Mauri, du postdoc de Rim Trabelsi et de plusieurs stagiaires Master 2.

Dans le **Chapitre 3**, nous présentons une percée significative dans le domaine de la perception via une nouvelle approche de détection et tracking temps-réel d'objets 3D par apprentissage profond. Ce chapitre comprend trois grandes parties. Tout d'abord, nous présentons notre nouveau dataset ESRORAD (Esigelec and Segula ROad and RAilway Dataset) hybride et multimodale dédié à la smart mobilité routière et ferroviaire. Afin de combler le manque de jeux de données ferroviaires avec vérité terrain permettant la détection d'objets 3D, nous nous sommes intéressés à la conception de notre propre dataset. Ce dernier comprend deux ensemble de données, un virtuel appelé GTAV (Grand Theft Auto V) et l'autre réel développé à partir de scènes réelles routières et ferroviaires acquises dans deux villes Normandes : Rouen et Le Havre. Ce nouveau jeu de données nous permettra à la fois d'entraîner, mais aussi de valider nos approches de détection d'objets 3D, ce qui est une première mondiale. Ces travaux ont été réalisés dans le cadre de la thèse d'Antoine Mauri et du partenariat stratégique avec SEGULA Technologies.

Dans la deuxième partie, nous nous concentrons sur le développement de notre nouvelle approche de détection et tracking temps-réel d'objets 3D par IA. Nous avons créé deux méthodes de détection d'objets 3D différentes offrant l'avantage de détecter l'objet et sa localisation via un seul réseau. La détection d'objets comprendra désormais non seulement la position de l'objet ainsi que sa classe, mais aussi sa distance, son centre 3D, sa dimension et son orientation. Nous présentons un nouveau réseau de neurones convolutifs (CNN) monoculaire de détection d'objets 3D temps-réel basé sur YOLOv5 respectant les contraintes temps-réel de l'embarqué (légèreté, vitesse et précision). La troisième partie est dédiée au tracking temps-réel d'objets 3D. Afin de prévenir les risques de collisions, tant sur la route que sur le rail, la détection et la localisation d'objets ne suffisent pas. Nous présentons notre approche de suivi d'objets basée sur notre modèle de détection d'objet 3D couplé au filtre de Kalman pour prédire les positions futures des objets dans les images suivantes. Nous avons validé l'ensemble de nos approches sur notre propre dataset ESRORAD mais aussi sur KITTI. Ces travaux ont été menés dans le cadre de la thèse d'Antoine Mauri.

Le **Chapitre 4** est dédié à la détection et tracking temps-réel d'objets en vue de l'analyse et compréhension de scènes routières complexes par deep learning. Il est structuré autour de deux parties. La première partie évoquera la problématique de la détection de piétons et comment nous l'avons solutionné via une nouvelle approche basée sur l'Histogram Oriented Gradient (HOG). Nous présenterons aussi une étude comparative entre deux approches de détection d'objets, SSD et YOLOv3 ainsi que l'estimation de distance des objets par deep learning. Afin de valider ces approches, nous les avons appliqués sur deux scénarios différents, détection de piétons dans une scène routière et tracking temps-réel de personnes par un drone. La deuxième partie évoquera notre nouvelle approche d'analyse et de compréhension de haut niveau des activités des agents routiers (piétons, vélos, véhicules). Nous avons développé un nouveau réseau profond spatio-temporel appelé STAF (Spatio-Temporal Attention Framework for Understanding Road Agents Behaviors) basé sur un mécanisme d'attention qui exploite des caractéristiques visuelles clés au fil du temps. Des expériences ont été menées sur deux scénarios différents sur deux datasets dédiés à la compréhension de comportement d'agents routiers. Ces travaux ont été menés dans le cadre des travaux de postdoc de Rim Trabelsi ainsi que plusieurs stagiaires Masters 2 et ingénieurs sur les projets de recherche RIN M2NUM [1], RIN M2SiNUM [2] et SEGULA.

Le **Chapitre 5** présente une contribution significative de perception d'environnement pour la robotique mobile dédiée à la santé. Il comprend 4 parties. Dans un premier temps, nous présentons nos travaux de recherche sur la vision omnidirectionnelle appliquée à la détection et reconnaissance de visages en vue des applications de contrôles d'accès biométriques. Ces travaux ont été réalisés dans le cadre du projet NOBA (NOmad Biometric Authentication) par le postdoctorant Jin-xin Liu ainsi que plusieurs stagiaires Master 2. Dans un second temps, nous nous concentrons sur nos approches de détection, localisation et suivi d'objets pour un robot mobile de type fauteuil roulant intelligent. Cela comprendra des approches de détection d'objets, d'estimation de la profondeur, de tracking d'objets et de localisation par SLAM. Nous présentons aussi nos travaux de navigation autonome du fauteuil basée sur la vision omnidirectionnelle. La troisième partie quant à elle sera dédiée à la perception d'environnement outdoor du fauteuil roulant et sera consacrée à la présentation de notre nouveau jeu de données dédié à la segmentation sémantique temporelle pour la navigation du fauteuil roulant sur les trottoirs. Ce jeu de données est unique car il permet d'entrainer et valider nos modèles CNN sur des environnements outdoor particuliers comme les trottoirs. Enfin, la quatrième et dernière partie sera dédiée à la segmentation sémantique rapide pour

1. M2NUM : Modélisation Mathématique et Simulation NUMérique.
2. M2SiNUM : Modélisation Mathématique avancée et Simulations NUMériques pour l'innovation.

la perception supervisée des trottoirs. L'ensemble de ces travaux a été validé via le développement d'une plateforme (TRL7) de fauteuil roulant intelligent. Ces travaux ont été menés dans le cadre des projets de recherche COALAS et ADAPT par les ingénieurs de recherche Yassine Nasri, Louis Lecrosnier, Edgard Petit, Aristide Laignel et Vishnu Pradeep ainsi que plusieurs stagiaires Master 2 et ingénieurs.

Dans le **Chapitre 6**, nous présentons nos travaux de recherche sur les ADAS de dernière génération à vocation non seulement de renforcer la sécurité mais aussi et surtout d'optimiser la consommation d'énergie à bord de véhicules autonomes. Ce chapitre est structuré autour de 4 parties : (*i*) ADAS pour l'optimisation de la consommation d'énergie, (*ii*) optimisation de l'estimation de la consommation d'énergie par machine learning, (*iii*) diagnostic et détection de défauts à distance et (*iv*) tracking temps-réel de véhicules. Dans un premier temps, nous présentons notre système de perception par fusion multicapteurs dédié à l'optimisation de la gestion de l'énergie pour un véhicule autonome tenant compte des propriétés environnementales et topologiques de l'itinéraire. Dans un second temps, nous présentons une nouvelle approche d'amélioration de l'estimation de la consommation d'énergie par machine learning ainsi que notre nouveau ADAS basé V2G. La plateforme a été validée en simulation et dans des environnements intérieurs (banc d'essai de véhicule) et extérieur (trafic réel).

Ces travaux ont été réalisé dans le cadre des projets VIRTUOSE[3] et SAVEMORE[4] par plusieurs stagiaires Master 2 et ingénieurs. La troisième partie est différente car elle ne concerne pas le véhicule autonome mais traite la gestion optimisée de l'énergie et sera dédiée au diagnostic et détection de défauts à distance de champs éoliens via les réseaux de capteurs sans fils. Nous présentons une approche d'estimation des paramètres d'un champ éolien par Filtre de Kalman Etendu (EKF). Les Réseaux de Capteurs Sans Fil (RCSF) et l'Internet des Objets (IoT) ont été présenté comme des solutions pour la détection et le diagnostic de défauts de champs éoliens. Ces travaux ont été effectué dans le cadre de la thèse de Lavinius Ioan GLIGA. La quatrième partie abordera, quant à elle, la problématique de tracking temps-réel de véhicules par RFID dans des environnements de la chaine logistique portuaire. Ces travaux ont été effectué dans le cadre du projet RORO-MAX[5] par des stagiaires Master 2.

Le **Chapitre 7** représente une conclusion générale de mes activités de recherche avec un bilan synthétique et ouvre des perspectives sur les axes de travaux de recherche que je souhaite poursuivre dans les années à venir. Dans un souci de synthèse et de cohérence, certains de mes travaux de recherche n'ont pas été présenté dans ce mémoire d'HDR.

3. VéhIcule électRique intelligenT à prolongateUr d'AutOnomie et Sources d'Energie multiples.
4. The Smart Autonomous Vehicle for Urban Mobility using Renewable energy.
5. Roll-On Roll-Off.

Chapitre 1

Synthèse des Activités Pédagogiques, Administratives et Scientifiques

Sommaire

Ce premier chapitre synthétise mes activités pédagogiques, administratives et scientifiques. Dans un premier temps, je présenterai ma formation ainsi que mon parcours professionnel. Mes activités d'enseignement et de recherche seront par la suite présentées avant de terminer par ma production scientifique durant ces dernières années.

1.1 Curriculum Vitae

1.1.1 Situation administrative

Je suis actuellement Enseignant-Chercheur à l'ESIGELEC Rouen (Ecole d'ingénieurs-es généralistes). J'exerce mes activités pédagogiques au sein du département Systèmes Embarqués et Instrumentation (SEI) de l'ESIGELEC dans lequel je suis :

— Responsable Master Automotive and Embedded Systems (AES).
— Responsable Master Connected and Embedded Intelligent Systems (CEIS)
— Responsable académique des étudiants en échange

Mes activités de recherche se déroulent à l'IRSEEM au sein du pôle Instrumentation, Informatique et Systèmes (IIS).

1.1.2 Parcours professionnel : une vue synthétique

Depuis 2016	**Responsable du Master** of Science - MSc. in Automotive Embedded Systems (MSc. AES). Dual Degree : ESIGELEC-Manipal School of Information Science (India) & Muthoot University (India).
Depuis 2015	**Responsable du Master** of Science - MSc. in Connected and Embedded Intelligent Systems (MSc. CEIS). Dual Degree : ESIGELEC-Manipal School of Information Science (India).
Depuis 2015	**Conseiller Pédagogique** des étudiants internationaux en échange.
Depuis 10/2009	**Enseignant Chercheur** à l'ESIGELEC/IRSEEM Rouen : enseignement en systèmes embarqués, co-encadrement (thèses) et encadrement (stages, ingénieurs de recherche et postdoctorants), tutorat apprentissage, coordination de projets de recherche, réalisations de contrats industriels.
2008-2009	**Chef de Projet** Recherche et Développement (R&D) : Analyse et traitement de données. Groupe Alten. Strasbourg.
2008	**Chef de Projet** R&D : Smart mobilité ferroviaire. Groupe THALES, Vélizy - Paris.
2007-2008	**PostDoc (Chef de Projet)** R&D : Traitement d'images pour la Gestion Electronique des Documents (GED), dématérialisation, classification et indexation d'images. Groupe Jouve, Lens.
2006-2007	**Enseignant Chercheur** ATER à temps complet : Attaché Temporaire d'Enseignement et de Recherche (ATER) à l'Université de Strasbourg. Cycles Licence L1, L2 et L3 en Mathématique et Informatique, L3 en Bioinformatique, Master 2 IISA.
2005-2006	**Enseignant Chercheur** ATER à temps partiel : ATER à l'Université de Strasbourg. Cycles Licence - L2 Mathématique, L2 Informatique.
2002-2005	**Doctorant** en Traitement d'Images et Vision par Ordinateur : Equipe Modèle, Image et Vision (MIV), Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection, (LSIIT). Université de Strasbourg.
2002-2005	**Moniteur** CIES Alsace : Centre d'Initiation à l'Enseignement Supérieur (CIES). Doctorant Moniteur à l'Université de Strasbourg. Cycle Licence - L1 Mathématique.

1.1.3 Formation : une description qualitative

2002-2005 **Docteur** en Traitement d'Images et Vision par Ordinateur de l'Université de Strasbourg : Extraction contrôlée d'indices images et automatisation de la reconstruction 3D. Application à la mesure dimensionnelle par vision par ordinateur. Laboratoire LSIIT, Télécom Physique Strasbourg, Université de Strasbourg.

2002-2005 **Moniteur** au CIES Alsace (Université de Strasbourg) : Formation d'Initiation à l'Enseignement Supérieur (CIES).

2001-2002 **DEA (Master Européen Recherche)** à l'Université de Poitiers : Traitement de l'Information, Informatique, Images et Automatique (T3IA). Option traitement d'images, Laboratoire Signal, Image et Communication (SIC), SP2MI. Université de Poitiers.

1998-2001 **DEA** en Contrôle Industriel à l'Université de Batna (Algérie). Option contrôle industriel et systèmes embarqués.

1997 **Ingénieur** en Electronique à l'Université de Batna (Algérie). Option Contrôle Industriel.

1.2 Activités d'enseignement

1.2.1 Synthèse des enseignements

Globalement, trois types d'enseignement ont été entrepris en adéquation avec mon parcours professionnel :

— 2009-2022 : Enseignant Chercheur - ESIGELEC (Université Rouen Normandie)
— 2005-2007 : ATER Université de Strasbourg
— 2002-2005 : Moniteur CIES Alsace - Université de Strasbourg

Je présente dans cette section une synthèse des enseignements effectués à l'ESIGELEC depuis 2009. L'ensemble de mes enseignements s'effectue au sein du département SEI de l'ESIGELEC. Ces enseignements sont à hauteur de 350 heures équivalent TD par an et portent principalement sur la robotique mobile, les systèmes embarqués, la mécatronique et la vision par ordinateur.

Répartition des enseignements par catégories

La figure 1.1 synthétise mes enseignements regroupés en 4 grandes familles : systèmes embarqués, robotique mobile, mécatronique et vision par ordinateur.

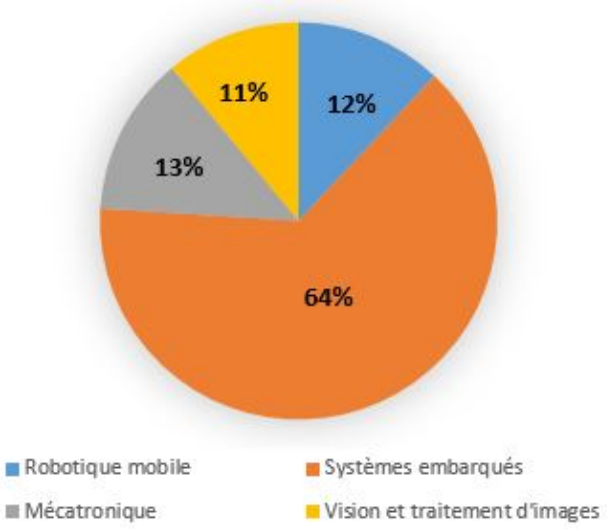

FIGURE 1.1 – Répartition des enseignements par disciplines.

Répartition annuelle de la charge pédagogique

La répartition annuelle des enseignements liés à ma charge pédagogique ainsi que leurs natures (Cours, CM, APP, TD, TP) sont présentées dans la figure 1.2.

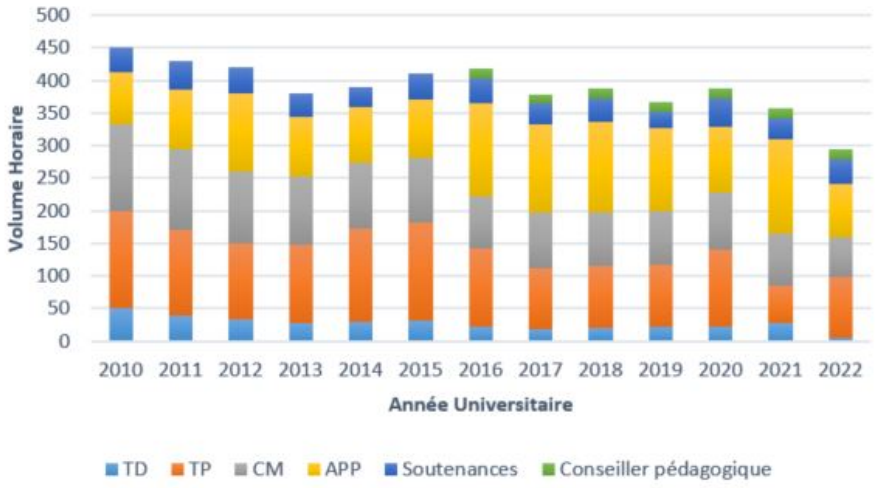

FIGURE 1.2 – Evolution de la charge pédagogique par activités.

Détails des enseignements

Je présente dans cette section la liste des enseignements que je dispense en fonction de leur intitulé, public concerné, type d'enseignement, nombre d'heures équivalent TD par an ainsi

que la langue d'enseignement. Les modules pour lesquelles je suis (ou j'ai été) responsable sont indiqués en caractères gras :

- **Bureau d'Etudes Systèmes Mécatronique (BESM)**, Ingénieur 3ème année, (CM, APP), 42h, français.
- **Robotique Mobile & Perception (RM&P)**, Ingénieur 3ème année, (CM, APP), 30h, français.
- Systèmes à Microprocesseurs (SAM), Ingénieur 2ème année, (CM, TD, TP, APP), 60h, français.
- Langage C, Ingénieur 2ème année, (CM, TD, TP), 20h, français.
- **Approches Mécatroniques (AM)**, Ingénieur 2ème année, (CM, APP), 40h, français.
- **Instrumentation et Systèmes**, Ingénieur 2ème année, (CM, TD, TP), 30h, français.
- Architecture des Ordinateurs, Ingénieur 1ère année, (CM, APP), 20h, français.
- **Découverte des Systèmes Embarqués**, Ingénieur 1ère année, (CM, TP), 7h, français.
- **Introduction of Embedded Systems**, Ingénieur 1ère année, (CM, TP), 7h, anglais.
- **Mobile Robotics and Perception**, Master MSc. AES, (CM, APP), 30h, anglais.
- **MSc. AES Project**, Master MSc. AES, (CM, APP), 40h, anglais.
- **MSc. CEI Project**, Master MSc. CEI, (CM, APP), 40h, anglais.
- **Bibliographical Studies**, Master MSc. AES, (CM, APP), 14h, anglais.

A noter aussi qu'en qualité de mes fonctions d'ATER effectuées de 2005 à 2007, j'ai pu dispenser plusieurs enseignements à l'Université de Strasbourg :

- Programmation fonctionnelle en CAML, Licence L1, (TD, TP), 27h, français.
- Algorithmie et programmation C, Licence L1, (TD, TP), 10h, français.
- Système d'exploitation Linux, Licence L1, (TD, TP), 10h, français.

1.2.2 Responsabilités pédagogiques

Au sein du département SEI, et en plus de mes activités d'enseignement, j'assure aussi plusieurs responsabilités pédagogiques en liens directs avec mes activités d'enseignement et celles du département.

Responsable Master MSc. AES

Le MSc. AES (Automotive Embedded Systems) est un Master of Sciences sur les systèmes embarqués pour l'automobile. MSc. AES est un Master en commun à double diplôme entre l'ESIGELEC et deux Universités indiennes, Manipal et Muthoot. Les étudiants font un semestre en Inde dans l'une des deux universités partenaires, un semestre à l'ESIGELEC (Rouen) et par la suite effectueront deux semestres sous forme de deux stages : premier stage de 4 à 6 mois pour valider le diplôme français (3 semestres sont nécessaires pour l'obtention

du diplôme ESIGELEC) et un deuxième stage de 4 à 6 mois pour valider le diplôme indien (4 semestres sont requis pour le diplôme de l'Université de Manipal). Les étudiants peuvent aussi effectuer un seul stage de 8 à 10 mois afin de valider leur double diplôme. Il est ouvert uniquement aux étudiants des Universités de Manipal et Muthoot disposants d'un Master 1 en électronique, génie-électrique, télécommunications et informatique.

Responsable Master MSc. CEIS

Le MSc. CEIS (Connected Embedded Intelligent Systems) est un Master of Sciences sur les systèmes embarqués intelligents et connectés. Le CEIS est un Master en commun à double diplôme entre l'ESIGELEC et l'Université de Manipal (Inde). Il était initialement appelé Control and Embedded Instrumentation (CEI). Il date depuis 2009 et a nécessité, en collaboration avec nos partenaires industriels ainsi que notre partenaire académique indien, une adaptation de son contenu pédagogique afin de répondre aux besoins des industriels. J'ai donc effectué ce travail de revalorisation du contenu pédagogique en lien direct avec la direction des relations internationales. Le nouveau programme a bénéficié d'une labellisation par la CGE (Conférence des Grandes Ecoles). Comme le Master AES, les étudiants font un semestre en Inde, un semestre à l'ESIGELEC (Rouen) et par la suite effectueront deux stages : premier stage de 4 à 6 mois pour valider le diplôme français et un deuxième stage de 4 à 6 mois pour valider le diplôme indien. Il est ouvert uniquement aux étudiants de l'Université de Manipal disposant d'un Master 1 en électronique, génie-électrique, télécommunications et informatique.

A noter que mes principales missions pour les deux Masters MSc. AES et CEIS sont donc :

- Planification des enseignements en relation avec le service de scolarité (direction des formations) et la direction des relations internationales.
- Préparation de l'arrivée des étudiants en France en étroite collaboration avec la direction des relations internationales.
- Faire le suivi des enseignements dispensés et réalisation des points réguliers avec les étudiants dans une démarche qualité d'amélioration continue.
- Recherche et contact d'intervenants (académiques et industriels, nationaux et internationaux) pour l'élaboration de la planification des enseignements.
- Organisation d'un comité de pilotage (Copil), tous les deux ans, afin d'améliorer le contenu pédagogique au regard des exigences et recommandations des partenaires académiques et industriels.
- Réalisation de la veille sur l'évolution technologique en relation avec les programmes des deux Masters et en cohérence avec les enseignements du département SEI.
- Accompagnement des étudiants dans leur démarche de recherche de stages auprès de

nos partenaires académiques et industriels nationaux et internationaux.
— Planification, en étroite collaboration avec la responsable des stages Masters, des sessions de soutenances (4 par an) des stages.
— Promotion des deux Masters auprès de nos partenaires industriels.
— Interaction avec nos partenaires Indiens (Universités de Manipal et de Muthoot) ainsi qu'avec notre bureau ESIGELEC en Inde sur le suivi des deux programmes : transfert des relevés de notes, choix d'électifs, conventions de stages ainsi que les évaluations des soutenances.

Conseiller pédagogique des étudiants internationaux en échange

Je suis conseiller pédagogique des étudiants internationaux en échange qui souhaitent effectuer une partie de leur cursus universitaire à l'ESIGELEC. Nous avons 40 partenaires dans le monde avec lesquels nous avons des échanges pour effectuer des mobilités à l'ESIGELEC d'un ou deux semestres. Mes principales missions pour cette responsabilité sont donc réparties en trois phases :

1. Avant la mobilité
   — Conseiller les étudiants pour leurs choix de programmes d'études à l'ESIGELEC (cycle Master ou Ingénieur).
   — Elaborer le contrat pédagogique mentionnant la liste des cours qu'ils doivent suivre pendant leur mobilité à l'ESIGELEC.
   — Valider le cursus pédagogique auprès de la scolarité et la direction des relations internationales.
2. Pendant la mobilité
   — Accompagner les étudiants sur le fonctionnement pédagogique à l'ESIGELEC.
   — Modifier si besoin le contrat pédagogique des étudiants.
3. Après la mobilité
   — Vérifier que les bulletins de notes correspondent bien au contrat pédagogique.
   — Valider le contrat pédagogique avant communication à l'Université d'origine de l'étudiant.

1.2.3 Valorisation de l'enseignement - pédagogie active

Mon statut d'enseignant chercheur m'a permis de choisir tous les enseignements que j'ai dispensé. Jusqu'à maintenant, j'ai mis en place de nouvelles formations, j'ai assuré des cours, j'ai conçu des formations en pédagogie active, j'ai encadré des travaux pratiques et/ou dirigés. J'ai pu enseigner tous les niveaux : Licence, Master 1 et 2 et Ingénieur.

J'ai contribué (et je contribue toujours) à la mise en place de la pédagogie active au sein de l'ESIGELEC. La pédagogie active a été introduite à l'ESIGELEC au début des années 2000 mais elle a pris sa forme d'Apprentissage Par Problème (APP) en 2012 via la forte collaboration de l'ESIGELEC avec notre partenaire Belge : Université de Louvain.

Confronter les étudiants à des problèmes concrets émanant du monde industriel est une méthode très efficace qui a également pour objectifs de responsabiliser les étudiants, développer leur autonomie, développer le travail d'équipe et les rendre acteurs dans le processus d'apprentissage. Cela passe par plusieurs déclinaisons de la pédagogie active. Tout d'abord via les projets ingénieurs qui concernent les étudiants de 2ème (via les projets semestre 8) et 3ème années ingénieurs (via les projets semestre 9). Ces projets peuvent être effectués en collaboration avec nos partenaires industriels, mais aussi en impliquant nos étudiants sur nos projets de recherche effectués au sein de notre laboratoire de recherche IRSEEM. Les étudiants, par groupe de 6, travaillent donc en mode projet avec des objectifs clairs et des livrables fixés par des cahiers de charges précis. Chaque groupe dispose d'un chef de projet assurant l'interface entre les étudiants et l'équipe d'encadrement constituée d'un commanditaire (client : industriel, laboratoire de recherche), d'un binôme de suivi et d'un instructeur. J'ai contribué dans ce dispositif selon les projets (souvent commanditaire et instructeur à la fois) en tant que :

— Commanditaire : proposer aux étudiants un projet d'une partie de mes travaux de recherche touchant à l'étude et la réalisation d'une plateforme donnée.
— Binôme d'encadrement : assurer en binôme, avec un collègue, le suivi en agile de la conduite du projet (état d'avancement, planning prévisionnel, livrables, budget, etc.).
— Instructeur : assurer le suivi technique du projet.

La pédagogie active en APP a fait aussi l'objet de la refonte de plusieurs modules. Mon département SEI a été pilote de la mise en place de l'APP. Plusieurs modules ont ainsi basculé en mode APP. Le module de tronc commun de 2ème année ingénieur (semestre 7), Systèmes à Microprocesseurs (SAM), était le module pilote via lequel nous avons mis en œuvre la première expérience d'APP au département. J'ai contribué d'une manière très significative à sa refonte. Le module APP comprend trois grandes parties :

— Langage C : les étudiants apprennent la programmation C embarqué.
— APP (Apprentissage Par Problèmes) : durant cette partie, les étudiants travaillent en groupe sur différentes thématiques liées à la programmation de microcontrôleurs à savoir : les entrées/sorties, les interruptions, l'ADC et le timer.
— APP (Apprentissage Par Projets) : les étudiants sont complètement en autonomie en mode projet où ils travaillent sur le développement d'une application de robotique mobile sur une plateforme que nous avons développé à l'ESIGELEC. Ils disposent

d'un livret pédagogique en guise de cahier de charges leur permettant de gérer leur projet en autonomie. Les séances sont divisées en deux types : tutorées (avec tuteur) et non tutorées (autonomie totale sans tuteur). Cette partie est validée via un concours à la fin en guise de démonstration de la solution.

Dans cette démarche, j'ai pu contribuer à l'élaboration du livret pédagogique respectant les règles de la pédagogie active, au développement de la plateforme (robot mobile avec microcontrôleur Texas Instruments MSP430), à la mise à jour annuelle du cahier de charges, au suivi des étudiants, à la préparation et au déroulement du concours, ainsi qu'au système d'évaluation.

1.2.4 Activités pédagogiques annexes

Dans le cadre de ma fonction d'enseignant chercheur, et particulièrement sur l'aspect enseignement, je suis amené à effectuer plusieurs activités annexes liées à la pédagogie : 1. tutorat d'apprentissage, 2. encadrement des projets ingénieurs, 3. encadrement des stages ingénieurs et 4. suivi et visites en entreprise des stagiaires de fin d'études ingénieurs.

Tutorat apprentissage

Chaque année, et en tant que tuteur pédagogique, j'assure le suivi des étudiants en apprentissage à l'ESIGELEC. Ce suivi s'effectue sous forme de visites d'apprentis en entreprise en interaction directe avec les tuteurs industriels. En plus, plusieurs réunions de suivis sont effectuées avec l'apprenti pour s'assurer du bon déroulement de sa mission d'apprentissage. Au-delà de l'aspect pédagogique, cela m'a permis aussi de développer des collaborations avec des partenaires industriels sur des projets pédagogiques et/ou recherches.

Encadrement des projets ingénieurs

Dans le cadre de la formation des étudiants ingénieurs, je suis amené à encadrer des projets ingénieurs en troisième année. Ces projets émanent de mes travaux de recherche ou de nos partenaires industriels. Les étudiants travaillent en équipe de 6 et développent leurs projets en 5 mois (octobre à février).

Encadrement des stages ingénieurs

Au sein du département SEI, je suis amené à encadrer des stages ingénieurs ayant comme objectif d'alimenter les travaux pédagogiques du département : développement de nouvelles maquettes pédagogiques ou refonte de certains contenus pédagogiques.

Suivi et visites des stages ingénieurs

En plus de mes encadrements des stagiaires, j'assure le suivi des stages de nos étudiants en entreprise. Ce suivi s'effectue par des points réguliers avec le stagiaire ainsi que son tuteur en entreprise et fait l'objet d'une visite en entreprise un mois avant la fin du stage.

1.3 Activités de recherche : encadrement

1.3.1 Co-encadrement de travaux de recherche : Thèses

Depuis que j'ai intégré l'ESIGELEC en tant qu'enseignant chercheur, j'ai pu co-encadrer et monté 5 Thèses (dont une soutenue), 7 post-doctorants et ingénieurs de recherche, 8 stages Master 2, 19 stages fin d'études ingénieurs, ainsi que de nombreux autres stages (1ère et 2ème année ingénieur). Ci-dessous la synthèse de mes co-encadrements de thèses :

Thèse 1 (soutenue)

Thèse 1	**Description**
Sujet	Diagnostic et détection de défauts à distance d'un parc éolien.
Type de Thèse	Thèse cotutelle entre l'ESIGELEC et l'Université Politehnica (Bucarest).
Doctorant	L. Gliga [1].
Directeur de Thèse	Dr/HDR. H. Chafouk (ESIGELEC) et Pr. D. Popescu (Univ. Polytehnica de Bucarest).
Dates/Ecole doctorale	2016-2019 (soutenue le 19/11/2019). MIIS (ED 590).
Taux d'Encadrement et Jury	**30%**. Dan Stefanou (Univ. Politehnica, Président), Radu Emil Precup (Politehnica Univ. of Timisoara, Rapporteur), Cristina Stoica-Maniu (CentraleSupélec, Rapporteur), Moussa Boukhnifer (Univ. de Lorraine, Examinateur), Jérôme Bosche (UPJV, Examinateur), Dumitru Popescu, Houcine Chafouk, R. Khemmar.
Contributions et publications associées	Collecte et traitement de données via LoRa. Filtre de Kalman Étendu (FKE) pour estimer les paramètres en cas d'anomalies [2].

Thèse 2 (en cours)

Thèse 2	**Description**
Sujet	Détection et tracking temps-réel d'objets 3D par deep learning. Application à la smart mobilité routière et ferroviaire.
Type de Thèse	Thèse CIFRE entre l'ESIGELEC et SEGULA Technologies (Paris).
Doctorant	A. Mauri [3].
Directeur de Thèse	Pr. R. Boutteau (Université Rouen Normandie).
Dates/Ecole doctorale	2019-2022 (soutenance prévue le 14/11/2022), MIIS (ED 590).
Taux d'Encadrement	**30%**.
Contributions	Détection d'objets 3D et tracking temps-réel par deep learning. Application à la smart mobilité routière et ferroviaire (train autonome) [4], [5], [6], [7], [8], [9], [10] et [11].

Thèse 3 (arrêtée)

Thèse 3	**Description**
Sujet	Analyse de scènes complexes par deep learning multitâches. Application à la smart mobilité.
Type de Thèse	Thèse cotutelle entre l'ESIGELEC et SUP'COM (Université de Carthage à Tunis).
Doctorante	S. Chebbi [12].
Directeur de Thèse	Pr. F. Abdelkefi (SUP'COM Tunis) et Dr/HDR. A. Cabani (ESIGELEC).
Dates/Ecole doctorale	2020-2023 (Thèse arrêtée le 01/09/2021 pour des raisons personnelles), MIIS (ED 590).
Taux d'Encadrement	**30%**.
Contributions	Analyse et compréhension de scènes complexes par deep learning multitâches. Application à la smart mobilité.

Thèse 4 (en cours)

Thèse 4	Description
Sujet	Architectures d'estimation hybride basées sur l'apprentissage en ligne pour le suivi de véhicules intelligents.
Type de Thèse	Thèse sur un projet de recherche ANR (Projet ArtIsmo [1]).
Doctorant	H. BESSAFA [13].
Directeur de Thèse	Dr/HDR. A. Zemouche (CRAN - Université de Lorraine).
Dates/Ecole doctorale	14/10/2021-30/09/2024, IAEM-Lorraine (ED 77).
Taux d'Encadrement	**30%**.
Contributions	Estimation des paramètres d'un véhicule autonome par deep learning. Application à la smart mobilité routière.

Thèse 5 (en cours)

Thèse 5	Description
Sujet	Contrôle/commande robuste et perception d'environnement par intelligence artificielle pour véhicules autonomes.
Type de Thèse	Thèse CIFRE entre l'ESIGELEC et SEGULA Technologies (Paris).
Doctorant	A. Evain [14].
Directeur de Thèse	Dr/HDR. S. Ahmedali (IBISC - Université d'Évry Val d'Essonne [2]).
Dates/Ecole doctorale	14/03/2022 - 31/04/2025. MIIS (ED 590).
Taux d'Encadrement	**40%**.
Contributions	Analyse et compréhension de scènes routière et ferroviaire par deep learning multitâches. Détection et tracking temps-réel d'objets 3D. Contrôle/commande robuste d'un train monorail autonome.

1. Algorithmes d'Estimation Intelligents pour la Smart Mobilité.
2. Université d'Évry Val d'Essonne / Université Paris-Saclay.

1.3.2 Encadrement de Post-doctorants et d'Ingénieurs de recherche

Nom	Année	Projet	Sujet
A. Laignel (IR)	2021-2022	ADAPT	Asservissement visuel d'un fauteuil roulant intelligent.
E. Petit (IR)	2022	ADAPT	Asservissement visuel d'un fauteuil roulant intelligent.
V. Pradeep (IR)	2021-2022	ADAPT	Segmentation sémantique rapide pour un fauteuil roulant intelligent.
L. Lecrosnier (IR)	2019-2021	ADAPT	Perception de scènes : détection, localisation et tracking d'objets pour un fauteuil roulant intelligent.
R. Trabelsi (PostDoc)	2019-2020	M2SiNUM	Analyse et compréhension de comportement d'agents dans les scènes routières par approches spatio-temporelles.
Y. Nassri (IR)	2013-2014	COALAS	Navigation autonome d'un fauteuil roulant intelligent sous ROS par vision omnidirectionnelle.
Ji-Xin LIU (PostDoc)	2012-2013	NOBA	Détection et reconnaissance de visage par fusion de capteurs omnidirectionnel et caméra PTZ.

En plus, trois postdoctorants sont prévus pour la période septembre à décembre 2022 sur trois projets de recherche (ANR ArtIsmo, RIN AntiHpert [3] et Carnot ESP CETRIA [4]).

1.3.3 Encadrement de stages Master 2

TARIKERE SRINIVAS Shaswath (2021-2022, Stage M2 MSc. AES) : Intégration des algorithmes de perception d'environnement pour un fauteuil roulant sur une Jetson XAVIER (Projet ADAPT).

Gajula SREENIVASULU Jayadeep (2021-2022, Stage M2 MSc. AES) : Labellisation d'un dataset multimodale routier/ferroviaire pour la smart mobilité (Thèse CIFRE SEGULA).

3. Opérateur 4.0 et anticipation dynamique de ses perturbations dans les ateliers de production.
4. Carte d'énergie temps-réel basée-IA pour l'écomobilité.

Charuvil Kumaran Aravind (2021, Stage M2 MSc. AES) : Instrumentation d'un fauteuil roulant intelligent et validation de la perception d'environnement (Projet ADAPT).

Varghese George Renjue (2019-2020, Stage M2 MSc. AES) : Instrumentation d'un fauteuil roulant intelligent (Projet ADAPT).

Alias Paul (2019-2020, Stage M2 MSc. CEI) : ADAS d'estimation d'énergie pour un fauteuil roulant intelligent (Projet ADAPT).

Pokala Adithya (2019, Stage M2 MSc. CEI) : Développement d'un vSLAM pour un fauteuil roulant intelligent (Projet ADAPT).

Kadapanatham Shashank Rao (2019, Stage M2 MSc. SEE) : Deep learning-based stress and fatigue detection for smart wheelchair (Projet ADAPT).

Raj Aditya (2011, Stage M2 Ranchi University) : Détection et reconnaissance de visage par fusion de capteurs catadioptric et caméra PTZ (Projet NOBA).

1.3.4 Encadrement de stages Ingénieurs

François Garnier (2022, Stage Projet Fin d'Etudes Ingénieur (PFE)) : Détection et tracking d'objets 3D par deep learning. Application à la smart mobilité routière et ferroviaire (Thèse CIFRE SEGULA).

KOUNOUHO Messmer Foinel Mahougnon (2022, Stage PFE) : Intégration et validation d'algorithmes de détection d'objets 3D par deep learning. Application à la smart mobilité routière et ferroviaire (Thèse CIFRE SEGULA).

Roland Mouissi (2021-2022, Stage PFE) : Intégration d'algorithmes de perception d'environnement pour un fauteuil roulant intelligent (Projet ADAPT).

Edgard Petit (2021-2022, Stage PFE) : Planification de trajectoire pour un fauteuil roulant intelligent (Projet ADAPT).

Grassi Marcos (2021-2022, Stage PFE) : Segmentation sémantique de séquences d'images par deep learning pour le suivi de trottoirs d'un fauteuil roulant intelligent (Projet ADAPT).

Laignel Aristide (2021, Stage PFE) : Segmentation sémantique pour le suivi de trottoirs et asservissement d'un fauteuil roulant intelligent (Projet ADAPT).

Dulompont Camille (2021, Stage PFE) : Développement d'un nouveau dataset routier/ferroviaire pour la détection d'objets 3D (Thèse CIFRE SEGULA).

Kefi Naceur (2020, Stage PFE) : Détection et tracking d'objets par deep learning pour un fauteuil roulant intelligent (Projet ADAPT).

Benmoumen Tahar (2020, Stage PFE) : Développement d'une caméra stéréoscopique pour la détection d'objets. Application à la smart mobilité routière et ferroviaire (Thèse CIFRE SEGULA).

Mauri Antoine (2019, Stage PFE) : Détection, localisation et tracking d'objets par deep learning. Application à l'analyse de scènes routières et ferroviaires (Projet SEGULA).

Latour Rodolphe (2019, Stage PFE) : Détection et suivi d'objets par deep learning. Application à la navigation autonome d'un fauteuil roulant intelligent (Projet ADAPT).

Da Silva Moura Ronaldo (2019, Stage PFE) : Modélisation énergétique d'un fauteuil roulant intelligent (Projet ADAPT).

Atahouet Amphani (2018, Stage PFE) : Détection et tracking d'objets par deep learning. Application à l'analyse de scènes routières et ferroviaires (Projet SEGULA).

Zhihao Chen (2018, Stage PFE) : Détection de piétons par deep learning. Application à la navigation autonome (Projet M2NUM).

Delong Li (2017, Stage PFE) : Détection et reconnaissance contrôlée d'objets (obstacles, piétons). Application à la navigation autonome (Projet M2NUM).

Gouveia Mathias (2017, Stage PFE) : Asservissement visuel d'un drone. Application au tracking temps-réel de personnes (Projet SEGULA).

Mbourou Diyeki Gael (2016, Stage PFE) : Développement d'une plateforme pédagogique pour la robotique mobile et perception (Projet ESIGELEC).

Nanou Adler Abdias KEDOTE (2016, Stage PFE) : Asservissement visuel d'un robot mobile de type wifibot (Projet ESIGELEC).

Bonardi Fabien (2013, Stage PFE) : Détection et reconnaissance de visage par vision omnidirectionnelle pour le contrôle d'accès biométrique (Projet NOBA).

1.3.5 Encadrement d'autres stages

Henry Victor (2021, Stage Technicien ESIGELEC) : Test et validation d'un robot mobile (Master DJI) pour des applications de perception d'environnement par IA.

Bonnard Mathéo (2021, Stage exécutant ESIGELEC) : Test et validation de drones dédiés au tracking temps-réel.

Nya Laetitia (2016, Stage Technicien ESIGELEC) : Navigation autonome d'un robot mobile de type wifibot.

Delong Li (2016, Stage Technicien ESIGELEC) : Développement d'une station météo intelligente.

Freddy Donald Tioguim Kana (2015, Stage Technicien ESIGELEC) : Modèle prédictif de gestion optimisée d'énergie d'un véhicule autonome (projet VIRTUOSE).

Lanoe Maxime, Ricardo Romain, Clouet Benjamin, Ruault Florian (2014, Stage Initiation à la Recherche CESI) : Modélisation d'un réseau électrique intelligent exploitant les capacités de stockage d'énergie de véhicules électriques (projet SAVEMORE).

Chuard Rémi, Labbes Guillaume, Cavalier Pierre, Delamare Alex (2014, Stage Initiation à la Recherche CESI) : Modélisation de la consommation d'énergie pour les ADAS de véhicules électriques (Projet VIRTUOSE).

Clybouw Alexis, Lopez Lucs, Février Romain (2013, Stage Initiation à la Recherche CESI) : Modélisation V2G. Application à l'utilisation d'un véhicule électrique autonome comme support de stockage d'électricité (projet SAVEMORE).

1.4 Activités de recherche : rayonnement scientifique

1.4.1 Montage, coordination et réalisation de projets

L'IRSEEM représente la division recherche de l'ESIGELEC. Nous développons une recherche partenariale (telle que définie par l'Agence Nationale de la Recherche - ANR [5]) et une activité de transfert à destination de la filière automobile régionale et nationale, aéronautique, électronique, télécommunications et énergie autour d'un thème fédérateur pour ces industries : les systèmes embarqués et la smart mobilité. Les thématiques de recherche de l'IRSEEM sont définies en liaison étroite avec les partenaires industriels et académiques du laboratoire et évoluent selon les besoins économiques, environnementaux et sociétaux. Un Service d'Ingénierie en Recherche et Développement (SIRD) qui joue le rôle d'une équipe transversale assurant la diffusion et la valorisation des travaux de recherche (transfert technologique) auprès des entreprises, PME et grands groupes. C'est dans ce cadre que j'effectue ma recherche au sein du pôle IIS touchant aux développement des systèmes de vision embarqués dédiés à la smart mobilité.

Ci-dessous la liste des principaux projets régionaux, nationaux et internationaux auxquels j'ai contribué, monté ou coordonné :

5. https ://anr.fr

Projet	Dispositif de financement (Budget IRSEEM)	Partenaires	Rôle dans le projet	Dates Début/Fin
CETRIA : Carte d'énergie temps-réel basée IA pour l'écomobilité	Carnot ESP (96.9k€, 80%)	CERTAM, IRSEEM	Porteur du projet	2022-2024
AntiHpert : Opérateur 4.0 et anticipation dynamique des perturbations dans les ateliers de production	RIN Recherche (125.5k€, 33%)	CESI, IRSEEM, Univ. Le Havre	Participation scientifique	2021-2025
ArtIsmo : Algorithme d'estimation intelligent pour la smart mobilité	ANR (284.4k€, 50%)	IRSEEM, CRAN, Autres [6]	Participation scientifique	2020-2024
ADAPT : Assistive Devices for empowering disAbled People through robotic Technologies	Interreg IVA (1,06M€, 11%)	IRSEEM, INSA Rennes, Autres [7].	Porteur du projet	2016-2022
M2SiNUM : Modélisation Mathématique avancée et Simulations Numériques pour l'innovation dans l'environnement et la santé	RIN et FEDER (58.3k€, 11%)	LITIS, IRSEEM, CORIA, GREYC, Autres [8].	Participation scientifique	2018-2021

6. ArtIsmo : IRSEEM, CRAN (Univ. de Larraine), IBISC (Univ. d'Évry - Univ. Paris-Saclay), LISEC (Minnesota), FAAR Industry (Paris).

7. ADAPT : IRSEEM, INSA Rennes, UPJV Picardie, Univ. of Kent, HEC, Kent Surrey Sussex Academic Health Science Network, Univ. of Kent, CCC Univ., Kent NHS Foundation.

8. M2SiNUM : IRSEEM, LMI, LMNO, LMRS, LITIS, LMN, CORIA, LMAH, GREYC, DITM.

Projet	**Dispositif de financement (Budget IRSEEM)**	**Partenaires**	**Rôle dans le projet**	**Dates Début/Fin**
M2NUM : Modélisation Mathématique. Applications et simulations NUMériques pour les énergies renouvelables, l'écomobilité, l'imagerie et la physique	RIN + FEDER (58.3k€, 15%)	LITIS, IRSEEM, CEREMA, CORIA, LMAH, LMRS.	Participation scientifique	2014-2017
ARGOS : Développement d'un robot mobile de surveillance de plateformes offshore	Challenge Total/ANR (500k€, 100%)	IRSEEM et Sominex.	Porteur du projet	2015-2018
COALAS : Cognitive Assisted Living Ambient System	INTERREG IVA (400k€, 24%)	IRSEEM, UPJV, Autres [9]	Porteur du projet	2012-2015
VIRTUOSE : Développement d'un Véhicule Electrique avec Prolongateur d'Autonomie	APE Région Normandie et FEDER (994k€, 83%)	IRSEEM , Lab. CEVAA Rouen.	Porteur du projet	2012-2015
SAVEMORE : Smart Autonomous VEhicle for Urban MObility using Renewable Energy	INTERREG IVA + GRR EEM (258k€, 45%)	IRSEEM, Univ. of Kent, LITIS	Porteur du projet	2013-2016

9. COALAS : IRSEEM, Univ. of Kent, Univ. of Essex, NHS Kent.

Projet	Dispositif de financement (Budget IRSEEM)	Partenaires	Rôle dans le projet	Dates Début/Fin
RORO MAX : Traçabilité de véhicules dans la chaine logistique portuaire	Région (83.7k€, 10%)	IRSEEM, GPMH, Univ. Rouen, ISEL, CRITT.	Participation scientifique	2011-2014
NOBA : NOmad Biometric Authentication	INTERREG IVA (1Mk€, 55%)	IRSEEM, University of Kent (UK).	Porteur du projet	2010-2013
ENEVATE : European Network of Electric Vehicles and Transferring	INTERREG IVA (207k€, 10%)	IRSEEM, Univ. of Cardiff, Autres [10].	Participation scientifique	2010-2013

1.4.2 Valorisation de la recherche par des contrats

Dans le cadre du développement des activités de recherche de l'IRSEEM via les contrats industriels, j'ai participé activement à la mise en place des conventions de partenariat avec des partenaires industriels. J'ai donc pu développer certaines activités représentées dans le tableau ci-dessous :

10. ENEVATE : IRSEEM, Univ. of Cardiff, ATC (Pays-bas), IRSEEM, Campus Automobile de Francorchamps (Belgique), Flemish Institute for Technical Research (Belgique), Bayern Innovativ, Inno Germany, Forschungszentrum Jülich (Allemagne), European Automotive Strategy Network (UK), National Renewable Energy Centre (UK), Future Transport Systems (UK), Electricity Supply Board (Pays-bas), Pôle Véhicule du Futur (France).

Partenaires	Projet	Role	Dates
Faurecia	ADAS pour véhicule autonome	Développement d'un ADAS "Park assist" et d'un système de fusion de caméras	2022-2023
SEGULA	Détection et tracking d'objets 3D pour la smart mobilité routière et ferroviaire	Développement d'algorithmes avec essais de validation et co-encadrement de deux doctorants ainsi que plusieurs stagiaires	2016-2022
FAAR Industrie	Intégration et validation d'algorithmes pour véhicule autonome	Participation scientifique et technique	2020-2024

J'ai aussi réalisé plusieurs expertises avec des partenaires industriels en "hors contrat" via des montages de projets non aboutis : Durisotti (France), Michelin (Suisse), Insoftdev (Romanie), HeatSoft (Rouen), etc.

1.4.3 Comité de lecture et reviewing

Comité de relecture dans des revues Scientifiques

J'ai effectué et continue à reviewer des articles de revues scientifiques :
— MDPI Sensors : Multidisciplinary Digital Publishing Institute Journal of Sensors
— MDPI Journal of Applied Sciences
— MDPI Journal of Imaging
— JEI : Journal of Electronic Imaging
— EAAI : Engineering Applications of Artificial Intelligence (Elsevier)
— IEICE : Fundamentals of Electronics, Communications and Computer Sciences
— London Journals Press
— International Journal of Vehicle Information and Communication Systems (IJVICS)

Comité de relecture dans des Conférences

J'ai aussi contribué comme relecteur dans les conférences suivantes :
— WSCG : Computer Graphics, Visualization and Computer Vision
— ITSC : IEEE Conference on Intelligent Transportation Systems
— EST : International Conference on Emerging Security Technologies

1.4.4 Participation à des réseaux de recherche

Je participe régulièrement à des webinaires recherche : CHIST-ERA, Horizon Europe et INTERREG Mer du Nord. Je participe aussi aux journées du GDR ISIS et du GDR Robotique ainsi qu'aux séminaires de la filière automobile : Nextmove [11], pôle TES [12], pôle véhicule du futur [13] et IFSTTAR [14].

1.4.5 Collaboration et partenariats

J'ai pu développer plusieurs partenariats et collaborations à plusieurs niveaux : international, national et régional. Au niveau international :

- University of Kent (UK) avec Konstantinos Sirlantzis et Gareth Howells via la réalisation de plusieurs projets de recherche de type INTERREG (NOBA, SAVEMORE, COALAS, ADAPT). C'est une collaboration très solide concrétisée par plus de 10 ans de travaux communs. Nous avons aussi collaboré dans la préparation de plusieurs éditions de la conférence EST ainsi que dans le montage de plusieurs projets (CHIST ERA 2018, 2019, 2020 et 2021).
- University College of London (UCL - UK) avec Tom Carlson via la réalisation du projet ADAPT (2017-2022).
- Canterbury Christ Church University (CCCU - UK) avec Eleni Hatzidimitriadou via le projet ADAPT (2017-2022).
- NHS (UK) avec Mathieu Pepper via la réalisation de deux projets de recherche COALAS (2013-2016) et ADAPT (2016-2022).
- University of Essex avec via le projet de recherche INTERREG COALAS (2013-2016).
- University of Leads avec Evangelos via le montage de projet de recherche (CHIST ERA 2021).
- University of Lituania (EDI) avec Kaspars Ozols via le montage de projets de recherche (CHIST ERA 2021).
- INSOFTDEV avec Ionut Rascanu via le montage de projet de recherche (CHIST ERA 2021).
- Université of Romania Bukarest avec Popesco via le co-encadrement de la Thèse en cotutelle de Lavinus Gliga (2016-2019).
- Polytechnique de Montréal (Canada) avec Jérôme le Ny via une collaboration sur le montage de plusieurs projets H2020 CHIST-ERA (2018 et 2019).

11. `https://nextmove.fr/`
12. `https://www.pole-tes.com/`
13. `https://www.vehiculedufutur.com/`
14. Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux.

— LISEC Minnesota, avec Rajesh Rajamani via le projet de recherche ArtIsmo (projet ANR : 2021-2025).
— Ecole Supérieure des Télécommunications de Tunis (SUP'COM), Tunisie. Cette collaboration s'est concrétisée par l'encadrement de plusieurs stages Master 2 et du co-encadrement de le Thèse de Samar Chebbi sous la direction de Fatma Abdelkefi.

Au niveau national :

— UPJV (Université Picardie Jules Vernes) avec Guillaume Caron et Fabio Morbidi du laboratoire MIS via la réalisation de plusieurs projets de recherche INTERREG (COALAS, ADAPT).
— INSA Rennes avec Marie Babel via la réalisation du projet de recherche INTERREG ADAPT (2016-2022).
— Pôle St Hélier (Rennes) avec Axel Petit, Philippe Gallien via la réalisation du projet de recherche ADAPT (2016-2022).
— IBISC (Université d'Evry) avec Dalil Ichalal via la réalisation du projet de recherche ANR ArtIsmo (2021-2025).
— CRAN (Université de Lorraine) avec Ali Zemouche via la réalisation du projet de recherche ANR ArtIsmo (2021-2025).
— NEOMA Paris avec Ouail Oulmakki via le montage de projets INTERREG MDN (2021, 2022).
— FAAR Industrie avec Randolph Toom et Emmanuel d'Arfeuille via le montage de plusieurs projets de recherche et la réalisation du projet de recherche ANR ArtIsmo (2021-2025).
— SEGULA Technologies avec Madjid Haddad et Sébastien Breteche via l'encadrement de plusieurs stages Master 2 et le co-encadrement de deux Thèses CIFRE : Antoine Mauri (2019-2022) et Alexandre Evain (2022-2025). Cela représente une collaboration solide marquée par plusieurs années de partenariat sur le train autonome.

Au niveau régional :

— LITIS (INSA Rouen) avec Samia Ainouz, Nicolas Forcadel, Rémi Boutteau et Stéphane Mousset. Cette collaboration dure depuis plusieurs années sous plusieurs formats : réalisation de projets de recherche (INTERREG SAVEMORE, RIN M2NUM, RIN M2SiNUM), montage de plusieurs projets de recherche, co-encadrement de Thèse CIFRE d'Antoine Mauri avec Rémi Boutteau, etc.
— CEREMA Rouen avec Peggy Subirats via le montage de plusieurs projets de recherche.
— CERTAM Rouen avec Benjamin Sibille via le projet de recherche Carnot ESP CETRIA (2022-2024).
— Université Rouen Normandie : avec plusieurs laboratoires de recherche : LMI INSA Rouen, LMRS, LMN, CORIA. Ceci via les projets de recherche RIN M2NUM et

M2SINUM.
— GREYC (Université Caen Normandie) avec Lyes Khoukhi via le montage de projets RIN émergents (2021-2022).
— ISEL (Université Le Havre Normandie) avec Adnane Yassine et Ibrahima Diarrassouba via le projet de recherche RIN AntiHPert (2021-2024).
— LMAH (Université Le Havre Normandie) via le projet de recherche RIN M2SiNUM.
— LMNO (Université de Caen Normandie) via les projets de recherche RIN M2SiNUM et M2NUM.
— CESI Rouen avec M'hammed Sahnoun, via le projet de recherche RIN Recherche AntiHpert (2021-2024).
— CHU Rouen avec Lucie Ménard et Deborah Laval via le projet ADAPT (2016-2022).

1.4.6 Valorisation de la recherche : l'ATER

Ce travail a permis le développement d'un système de reconstruction 3D dédié à l'évaluation dimensionnelle de pièces manufacturées. L'avantage majeur de notre approche reste sa modularité et sa flexibilité. Cependant, plusieurs améliorations quant à la précision de la reconstruction 3D obtenue, et à la qualité de l'évaluation finale, sont nécessaires. Une réflexion a ainsi été menée sur les extensions possibles du travail réalisé.
L'ajustement des paramètres fixés *a priori* dans le graphe de contrôle (Situation Graph Trees) a permis au système de planification de s'adapter, au mieux, aux conditions effectives d'acquisition et donc d'améliorer l'évaluation dimensionnelle. Nous avons ainsi étudié la possibilité de contrôler une deuxième tête de mesure, stéréoscopique par exemple, afin d'enrichir le système et compléter ainsi les primitives 3D surfaciques avec de nouvelles primitives 3D contours. Le SGT développé est capable de contrôler plusieurs têtes de mesure à la fois. Chaque tête de mesure est utilisée indépendamment afin de générer des données 3D spécifiques, relatives soit aux contours (vision stéréoscopique), soit aux surfaces (lumière structurée). Toutefois, la coopération active de ces deux approches permettait une reconstruction 3D plus complète d'objets, caractérisées par de grandes surfaces plutôt uniformes et non texturées.

1.4.7 Valorisation de la recherche : le postdoc et la R&D industriel

J'ai effectué un postdoc au sein du pôle R&D du groupe Jouve en Gestion Electronique des Documents (GED). Mes travaux de recherche portaient sur la classification automatique des documents pour une meilleure gestion en dématérialisation. Les approches utilisées se basaient sur la classification par SVM avec des techniques d'apprentissage supervisée. Par la suite, je voulais faire de la R&D en entreprise et acquérir une vraie expérience industrielle,

c'est pour cela que j'ai intégré, en qualité de chef de projet R&D, le groupe Alten/THALES en charge de la smart mobilité ferroviaire. J'ai acquis une forte expérience professionnelle en R&D académiques et industrielles durant ces longues dernières années. Je dispose aussi d'une forte expérience professionnelle en gestion et conduite de projets scientifiques et industriels touchant à l'informatique et aux systèmes embarqués. J'ai donc pu développer plusieurs compétences dans divers domaines liés à la smart mobilité.

1.5 Liste des publications

1.5.1 Articles dans des revues internationales

[1] Decoux Benoit, **Khemmar Redouane**, Ragot Nicolas, Venon Arthur, Grassi-Pampuch Marcos, Mauri Antoine, Lecrosnier Louis, and Pradeep Vishnu. A Dataset for Temporal Semantic Segmentation Dedicated to Smart Mobility of Wheelchairs on Sidewalks. Journal of Imaging, MDPI, 2022, 8(8), pp.216, DOI : 10.3390/jimaging8080216, (JCR IF 3.57) [11].

[2] Pradeep Vishnu, **Redouane Khemmar**, Lecrosnier Louis, Duchemin Yann, Rossi Romain, Decoux Benoit. Self-Supervised Sidewalk Perception Using Fast Video Semantic Segmentation for Robotic Wheelchairs in Smart Mobility. Sensors, MDPI, 2022, 22(14), pp.5241, DOI : 10.3390/s22145241, (JCR IF 3.57) [15].

[3] **Redouane Khemmar**, Antoine Mauri, Camille Dulompont, Jayadeep Gajula, Vincent Vauchey, Madjid Haddad, and Rémi Boutteau. Road and Railway Smart Mobility : A High-definition Ground Truth Hybrid Dataset. Sensors, MDPI, 22(10), pp.3922, DOI : 10.3390/s22103922, (JCR IF 3.57) [10].

[4] Rim Trabelsi, **Redouane Khemmar**, Benoit Decoux, Jean-Yves Ertaud, Rémi Boutteau. STAF : Spatio-Temporal Attention Framework for Understanding Road Agents Behaviors. IEEE Access, 2022, 22 (10), pp.55794-55804, DOI : 10.1109/ACCESS.2022.3176861, (JCR IF 3.36) [16].

[5] Rim Trabelsi, **Redouane Khemmar**, Benoit Decoux, Jean-Yves Ertaud, Rémi Boutteau. Recent Advances in Vision-Based On-Road Behaviors Understanding : A Critical Survey. Sensors, MDPI, 2022, 22(7), pp.2654, DOI : 10.3390/s22072654, (JCR IF 3.57) [17].

[6] A. Mauri, **Redouane Khemmar**, B. Decoux, M. Haddad, Rémi Boutteau. Lightweight convolutional neural network for real-time 3D object detection in road and railway environments. Journal of Real-Time Image Processing (JRTIP), Springer Verlag, 19, 499–516 (2022), DOI : 10.1007/s11554-022-01202-6, (JCR IF 3.36) [4].

[7] Louis Lecrosnier, **Redouane Khemmar**, Nicolas Ragot, Romain Rossi, Jean Yves Ertaud et al. Object Detection, Localisation, and Tracking-based Deep Learning for Smart Wheelchair. AMSE, IIETA. Modelling, Measurement and Control C, Vol. 82, No. 1-4, December, 2021, pp. 1-5, DOI : $10.18280/\text{mmc}_c.821 - 401$, [18].

[8] Antoine Mauri, **Redouane Khemmar**, Benoit Decoux, Madjid Haddad, Rémi Boutteau. Real-Time 3D Multi-Object Detection and Localization Based on Deep Learning for Road and Railway Smart Mobility. Journal of Imaging, MDPI, 2021, 7(8), pp.145, DOI : 10.3390/jimaging7080145, (JCR IF 2.45) [5].

[9] Louis Lecrosnier, **Redouane Khemmar**, Nicolas Ragot, Benoit Decoux, Romain Rossi et al. Deep Learning-Based Object Detection, Localisation and Tracking for Smart Wheelchair Healthcare Mobility. International Journal of Environmental Research and Public Health (IJERPH), MDPI, 2021, 18(1), pp.91, DOI : 10.3390/ijerph18010091, (JCR IF 3.57) [19].

[10] Adnane Cabani, Peiwen Zhang, **Redouane Khemmar**, Jin Xu. Enhancement of energy consumption estimation for electric vehicles by using machine learning. IAES International Journal of Artificial Intelligence (IJ-AI), University of Leicester, United Kingdom, 2021, 10(1), pp.215, DOI : 10.11591/ijai.v10.i1.pp215-223, (JCR IF 2.36 Q2) [20].

[11] Lavinius Ioan Gliga, **Redouane Khemmar**, Houcine Chafouk, Dumitru Popescu. A Survey of Wireless Communication Technologies for an IoT-connected Wind Farm. Wireless Personal Communications, Springer Verlag, 122, 2253-2272 (2022), DOI : 10.1007/s11277-021-08991-2, (JCR IF 1.67) [2].

[12] Antoine Mauri, **Redouane Khemmar**, Benoit Decoux, Nicolas Ragot, Romain Rossi et al. Deep Learning for Real-Time 3D Multi-Object Detection, Localisation, and Tracking : Application to Smart Mobility, Sensors, MDPI, 2020, 20(2), pp.532, DOI : 10.3390/s20020532, (JCR IF 3.57) [6].

[13] L. Ménard, A. Petit, É. Leblong, M. Stein, E. Hatzidimitriadou, **Redouane Khemmar** et al. Novel Robotic Assistive Technologies : Choosing Appropriate Training for Healthcare Professionals. Modelling, measurement and control C, AMSE, 2020, 81 (1-4), pp.43-48, DOI : $10.18280/\mathrm{mmc}_c.811-40$, [21].

[14] **Redouane Khemmar**, Li. Delong, B. Decoux. Real-Time Pedestrian Detection-based Faster HOG/DPM and Deep Learning Approaches. International Journal of Computer Applications (IJCA), Foundation of Computer Science, 2020, 176 (42), pp.34-38, DOI : 10.5120/ijca2020920539, (JCR IF 1.07) [22].

[15] Adnane Cabani, **Redouane Khemmar**, Jean Yves Ertaud, Romain Rossi, Xavier Savatier. ADAS multi-sensor fusion system-based security and energy optimisation for an electric vehicle. International Journal of Vehicle Autonomous Systems (IJVAS), Inderscience, 2019, 14(4), pp.345-366, DOI : 10.1504/IJVAS.2019.102445, (JCR IF 1.18) [23].

[16] Yassine Nasri, Vincent Vauchey, **Redouane Khemmar**, Nicolas Ragot, Konstantinos Sirlantzis et al. ROS-based Autonomous Navigation Wheelchair using Omnidirectional Sensor. International Journal of Computer Applications (IJCA), Foundation of Computer Science, 2016, 133(6), pp.12-17, DOI : 10.5120/ijca2016907533, (JCR IF 1.07) [24].

[17] **Redouane Khemmar**, F. Bouzbouz, N. Ragot, X. Savatier. The Application of RFID Technology in a Port. International Journal of Computer Applications (IJCA), Foundation of Computer Science, 2014, 86(7), pp.41-50, DOI : 10.5120/15001-3249, (JCR IF 1.07) [25].

[18] **Redouane Khemmar**, Jean-Yves Ertaud, Xavier Savatier. Face Detection & Recognition based on Fusion of Omnidirectional & PTZ Vision Sensors and Heteregenous Database. International Journal of Computer Applications (IJCA), Foundation of Computer Science, 2013, 61. (JCR IF 1.07) [26].

1.5.2 Communications avec actes dans des congrès internationaux

[19] **Redouane Khemmar**, Antoine Mauri, Benoit Decoux, Madjid Haddad, Rémi Boutteau et al. Environment Perception-based Deep Learning. Application to Road and Railway Smart Mobility. V-ASET2021. 5th Edition of Applied Science, Engineering and Technology Virtual., Dec 2021, Teams, United Kingdom [7].

[20] Antoine Mauri, **Redouane Khemmar**, Benoit Decoux, Tahar Ben Moumen, Madjid Haddad et al. A Comparative Study of Deep Learning-based Depth Estimation Approaches : Application to Smart Mobility 8th International Conference on Smart Computing and Communications (ICSCC 2021), Jul 2021, Kochi, India [8].

[21] Antoine Mauri, **Redouane Khemmar**, Rémi Boutteau, Benoit Decoux, Jean Yves Ertaud et al. A new Evaluation Approach for Deep Learning-based Monocular Depth Estimation Methods. The 23rd IEEE International Conference on Intelligent Transportation Systems (ITS), Sep 2020, Rhodes (virtual conference), Greece [9].

[22] Rim Trabelsi, **Redouane Khemmar**, Rémi Boutteau, Benoit Decoux, Jean Yves Ertaud. Toward Comprehensive Road Agents Behavior Understanding. CSTI 2020 : 1er Colloque Francophone des Systèmes de Transports Intelligents, Nov 2020, Tunis, Tunisia [27].

[23] **Redouane Khemmar**, Matthias Gouveia, Benoit Decoux, Jean Yves Ertaud. Real-Time Pedestrian and Object Detection and Tracking-based Deep Learning. Application to Drone Visual Tracking. WSCG'2019 - 27. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2019, May 2019, Plzen, Czech Republic [28].

[24] **Redouane Khemmar**, Houcine Chafouk. Diagnostic et Détection de Défauts à distance des Eoliennes A l'aide de l'Internet des Objets (IoT). COFMER'03 (Energie solaire, Energie éolienne, Biomasse & Bioénergie, Efficacité énergétique & Stockage d'énergie), Apr 2019, Marrakech, Maroc [29].

[25] Zhihao Chen, **Redouane Khemmar**, Benoit Decoux, Amphani Atahouet, Jean-Yves Ertaud. Real-Time Object Detection, Tracking, and Distance and Motion Estimation based on Deep Learning : Application to Smart Mobility. Eighth International Conference on Emerging Security Technologies (EST), Jul 2019, Colchester, United Kingdom [30].

[26] Nicolas Ragot, **Redouane Khemmar**, Adithya Pokala, Romain Rossi, Jean-Yves Ertaud. Benchmark of Visual SLAM Algorithms : ORB-SLAM2 vs RTAB-Map. Eighth International Conference on Emerging Security Technologies (EST), Jul 2019, Colchester, United Kingdom [31].

[27] Adnane Cabani, **Redouane Khemmar**, Jean Yves Ertaud, Joseph Mouzna. Intelligent navigation system-based optimization of the energy consumption. 2015 IEEE Intelligent Vehicles Symposium (IV), Jun 2015, Seoul, South Korea. pp.785-789, [32].

[28] **Redouane Khemmar**, Jean Yves Ertaud, Xavier Savatier, Kostas Sirlantzis. V2G-based Smart Autonomous Vehicle for Urban Mobility using Renewable Energy The 4th International Conference on Smart Systems, Devices and Technologies (SMART2015), Jun 2015, Bruxelles, Belgium [33].

[29] **Redouane Khemmar**, Fadoua Bouzbouz, Nicolas Ragot, Xavier Savatier. Vehicle tracking based RFID technology and ITS application on automotive supply chain. 10th ITS European Congress, Helsinki, Finland 16–19 June 2014, Jun 2014, Helsinki, Finland [34].

[30] **Redouane Khemmar**, Aditya Raj, Jean Yves Eratud, Xavier Savatier. Face Detection and Recognition under Heterogeneous Database Based on Fusion of Catadioptric and PTZ Vision Sensors. Proceedings of the 8th International Conference on Computer Recognition Systems. CORES 2013, 226, Springer International Publishing, pp.171-185, Advances in Intelligent Systems and Computing [35].

[31] Ji-Xin Liu Ji-Xin Liu, **Redouane Khemmar**, J. Ertaud, X. Savatier. Compressed Sensing Face Recognition Method in Heterogeneous Database with Small Sample Size Problem. 2012 Eighth International Conference on Signal-Image Technology & Internet-Based Systems (SITIS 2012), Nov 2012, Naples, France. pp.80-84, [36].

[32] Ernest Hirsch, Alex Lallement, **Redouane Khemmar**. Steps Towards an Intelligent Self-Reasoning System for the Automated Vision-Based Evaluation of Manufactured Parts Workshop on Applications of Computer Vision, in conjunction with ECCV, May 12, Graz, Austria, 2006, May 2006, Graz, Austria [37].

1.5.3 Communications avec actes dans des congrès nationaux

[33] Antoine Mauri, **Redouane Khemmar**, Benoit Decoux, Jean-Yves Ertaud, Madjid Haddad et al. Une nouvelle approche pour l'évaluation des méthodes monoculaires d'estimation de la profondeur basées sur l'apprentissage profond. ORASIS 2021, Centre National de la Recherche Scientifique [CNRS], Sep 2021, Saint Ferréol, France [38].

[34] Jean-Yves Ertaud, Nicolas Ragot, Ronaldo da Silva Moura, Paul Alias, Marie Babel, **Redouane Khemmar** et al. Modélisation énergétique d'un fauteuil roulant électrique pour prédire la faisabilité d'effectuer un trajet. Handicap 2020, Nov 2020, Paris, France [39].

[35] **Redouane Khemmar**, Louis Lecrosnier, Romain Rossi, Jean Yves Ertaud, France Rouvray et al. Detection, Localisation et Tracking d'Objets basé Deep Learning pour un Fauteuil Roulant Intelligent. Handicap 2020. Technologies pour l'autonomie et l'inclusion, Nov 2020, Paris, France [40].

[36] Nicolas Ragot, **Redouane Khemmar**, Adithya Pokala, Romain Rossi, Benoit Decoux et al. Trajectography estimation for a smart powered wheelchair ORB-SLAM vs RTAB-MAP. A preprint M2SC (modélisation systémique de systèmes cyber-physiques), Jun 2019, Poitiers, France [41].

[37] Aditya Pokala, **Redouane Khemmar**, Nicolas Ragot, Romain Rossi, Jean Yves Ertaud et al. Estimation de trajectoire pour un fauteuil roulant électrique Modélisation systémique de systèmes cyber-physiques, Jun 2019, Poitiers, France [42].

[38] **Redouane Khemmar**, Fabien Bonardi, Jean Yves Ertaud, Xavier Savatier. Biometrie authentication platform-based multisensor fusion. 2017 Seventh International Conference on Emerging Security Technologies (EST), Sep 2017, Canterbury, France [43].

[39] René Emmanuel Datondji, Nicolas Ragot, Yassine Nasri, **Redouane Khemmar**, Rémi Boutteau. Odométrie visuelle par vision omnidirectionnelle pour la navigation autonome d'une chaise roulante motorisée. Journées francophones des jeunes chercheurs en vision par ordinateur, Jun 2015, Amiens, France [44].

[40] **Redouane Khemmar**, Fadoua Bouzbouz, Nicolas Ragot, Xavier Savatier. LA RFID au service du terminal roulier du GPMH. ATEC ITS France, Jan 2014, Paris, France [45].

[41] **Redouane Khemmar**, Nicolas Ragot, Fadoua Bouzbouz, Jean-Yves Ertaud, Anne-Marie Kokosy et al. Enhancing the Autonomy of Disabled Persons : Assistive Technologies Directed by User Feedback. 2013 Fourth International Conference on Emerging Security Technologies (EST), Sep 2013, Cambridge, France. pp.71-74, [46].

1.5.4 Chapitre de livre

[42] **Redouane Khemmar**, Rim Trabelsi, Benoit Decoux, Rémi Boutteau. Development in Vision-Based On-Road Behavior Understanding. Scholarly Community Encyclopedia. April, 18th. 2022. `https://encyclopedia.pub/entry/21840`, accessed on 21 April 2022 [47].

[43] **Redouane Khemmar**, Houcine Chafouk. Diagnostic et Détection de Défauts à distance des Eoliennes A l'aide de l'Internet des Objets (IoT). 2021, (accepté, en cours de publication) [48].

[44] **Redouane Khemmar**, Alex Lallement, Ernest Hirsch. Design of an Intelligent Self-Reasoning System Using a Situation Graph Tree for the Automated Vision-Based 3D Reconstruction of Manufactured Parts. 2007, [49].

1.5.5 Mémoires

[45] **Redouane Khemmar**. Extraction contrôlée d'indices images et automatisation de la reconstruction 3D : Application à la mesure dimensionnelle par vision par ordinateur. Traitement du signal et de l'image [eess.SP]. Thèse de Doctorat en Vision par Ordinateur de l'Université de Strasbourg. 2005, [50].

[46] **Redouane Khemmar**. Compression d'images fixes par critères d'informations. [Rapport de recherche] Laboratoire Signal, Image et Communication (SIC). Mémoire de Master 2 Recherche (DEA Recherche), Université de Poitiers. 2002, [51].

[47] **Redouane Khemmar**. Développement d'une carte 2D d'environnement statique pour un robot mobile autonome. [Rapport de recherche] Mémoire de projet de fin d'études Ingénieur, Université de Batna (Algérie), 2000, [52].

Chapitre 2

Détection, Localisation et Tracking d'Objets 2D par Apprentissage Profond

Sommaire

2.1 Introduction

Dans les tâches essentielles de la vision par ordinateur, nous avons assisté à des avancées significatives dans la détection, la localisation et le suivi d'objets. Cependant, il n'existe actuellement aucune méthode pour détecter, localiser et suivre des objets dans des environnements routiers tenant compte des contraintes temps-réel. Nous avons donc initié, depuis des années, des travaux de détection et de suivi multi-objets par apprentissage profond appliqués à la smart mobilité multimodale routière et ferroviaire. Nous avons développé un détecteur efficace basé sur YOLOv3 [53] que nous avions adapté à notre contexte. Par la suite, pour localiser avec succès les objets détectés, nous avons développé une méthode adaptative visant à extraire des informations 3D (cartes de profondeur). Pour ce faire, une étude comparative a été réalisée prenant en compte deux approches : Monodepth2 [54, 55] (vision monoculaire) et MADNEt [56] (vision stéréoscopique). Ces approches ont ensuite été évaluées sur des jeux de données de profondeur afin de discerner la solution temps-réel la plus performante. Par la suite, nous avons exploité les informations issues de la détection et localisation d'objets pour améliorer le tracking basé sur l'approche SORT [57]. Nous avons introduit un filtre de Kalman étendu (FKE) [58] pour mieux estimer la position des objets. Des expériences approfondies menées sur le jeu de données KITTI [59] prouvent que nos approches [60] surpassent ceux de l'état de l'art. Nos travaux portaient donc sur le développement de 2 axes : détection et localisation temps-réel d'objets présents dans la scène et prédiction de leur comportement.

Nous avons opté pour un système monocapteur basé uniquement vision par caméra afin de simplifier la solution et la rendre plus accessible. Pour cela, plusieurs défis sont à relever comme par exemple la localisation fiable d'objets 3D sans faire appel aux capteurs de profondeurs (RADAR ou LiDAR), détection précise d'objets, ou encore l'allègement de l'approche développée pour qu'elle soit intégrée dans un système embarqué temps-réel. La figure 2.1 illustre l'architecture de la solution développée.

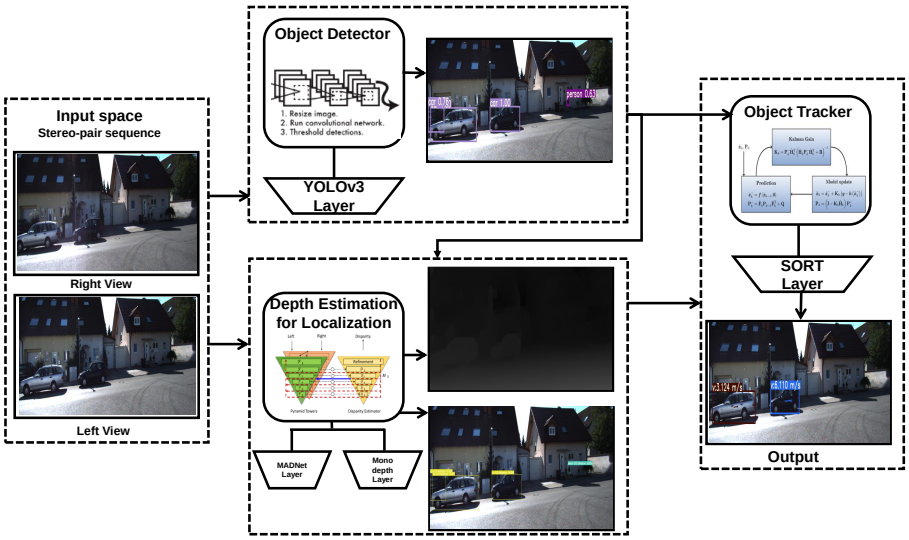

FIGURE 2.1 – Vue d'ensemble du système proposé, composé de trois éléments principaux : détection, estimation de la distance pour la localisation et suivi d'objets.

2.2 Etat de l'art

2.2.1 Détection d'objets

La détection d'objets représente un problème clé de la vision par ordinateur avec deux grands défis : la détection d'objets dans les images et l'estimation de leur position. Au cours des dernières années, plusieurs méthodes basées sur les réseaux de neurones convolutifs (CNN) ont été proposées pour surmonter ces difficultés. Les méthodes basées sur les CNN peuvent être divisées en deux catégories principales : (*i*) les méthodes à une étape, qui fournissent une estimation de la position et des classes en une seule étape, et (*ii*) les méthodes à deux étapes, qui détectent d'abord les régions des images où un objet pourrait être présent, puis appliquent un classifieur à ces régions. Parmi les méthodes de pointe à une étape, nous pouvons citer Single-Shot Detector (SSD) [61] et You Only Look Once (YOLO) [62]. Elles fournissent en sortie la probabilité de chaque classe possible (non pas pour l'ensemble de l'image mais pour un ensemble de positions régulièrement espacées et pour différentes échelles) et rapports d'aspect de rectangles appelés boîtes. Quant à la catégorie à deux étages, le RCNN (Region-proposal CNN) [63] signale des résultats exceptionnels sur de nombreux benchmarks ainsi que ses versions améliorées [64, 65]. Ces méthodes partagent des performances similaires du point de vue de la précision moyenne (mAP), un critère qui quantifie la qualité de la détection (proportion de détection correcte) en fonction du rappel, qui est la proportion d'objets détectés. L'inconvénient mineur de ces réseaux est l'utilisation de deux réseaux indépendants qui rend la prédiction lente sur des systèmes embarqués. En revanche, avec les architectures à une étape, la classification est faite sur des boîtes (fixes

en taille et nombre) dans des couches prédéfinies, puis la localisation d'objets détectés est ajustée par régression. Ces architectures sont généralement plus rapides que celles à deux étapes, avec des performances similaires [61]. Considérant que notre application est basée sur des données spatio-temporelles, la détection et localisation imprécises d'objets trop petits sont moins importantes que les performances temps-réel. Nous choisissons donc d'utiliser l'approche de détection d'objets à une seule étape YOLOv3 [53].

2.2.2 Estimation de la distance

Pour estimer la distance entre le véhicule et les objets de la scène, de nombreux capteurs sont disponibles, par exemple les capteurs laser-ultrasons [66] et les capteurs temps de vol TOF [67]. L'utilisation de ces capteurs avec des caméras rend le système plus complexe et plus coûteux. Contrairement à la plupart des travaux similaires et afin de réduire la complexité du calcul, nous étudions ici une caméra monoculaire pour l'estimation de distance. La littérature manque encore de jeux de données comprenant une carte de distance pour chaque instance. En plus, les CNN présentent un problème lié à leurs couches entièrement connectées (i.e. la classe de l'étiquette est discrète alors que sa distance est continue). Pour résoudre ce problème, il faut inclure une couche de régression et une autre de classification à la fin du réseau. De plus, en optant pour un schéma non supervisé, il pourrait être intéressant d'estimer la profondeur pour des images monoculaires, comme le modèle Monodepth [55] qui est entraîné principalement sur des images stéréo mais déduit les cartes de disparité des images monoculaires.

2.2.3 Tracking d'objets

Ces dernières années, le MOT (Multi-Object Tracking) basé sur le deep learning a atteint des performances de pointe en termes de qualité de suivi [68]. Par exemple, un détecteur d'objets comme Faster-RCNN [65] associé à un filtre de Kalman linéaire permet un bon compromis entre temps de traitement et qualité du suivi, comme le montre SORT (Simple Online and Realtime Tracking) [57]. Afin d'évaluer les performances du suivi, nous devons définir quelles sont les erreurs possibles. Un premier type d'erreur est un échec, c'est-à-dire un objet qui existe dans une séquence d'images mais qui n'est pas détecté dans une ou plusieurs images. Un deuxième type d'erreur est le faux positif, lorsqu'un objet détecté est associé par le "traqueur" à un objet et à une trajectoire de vérité terrain, mais ne correspond pas à un objet existant. Un troisième type d'erreur est la non-concordance, lorsque les objets détectés correspondent à des objets existants mais ne sont pas associés par le traqueur à des objets corrects. La somme de ces trois types d'erreurs, calculée sur le nombre total d'objets présents

dans toutes les images, définit le Multi Object Tracking Accuracy (MOTA) [69], qui est un critère couramment utilisé pour le tracking.

2.3 Architecture du système

Dans cette section, nous présentons une vue d'ensemble du système que nous avons conçu ainsi que les outils matériels et logiciels utilisés pour développer notre plateforme. Afin de concevoir une technique de détection d'objets efficace, il est nécessaire de trouver le bon compromis entre vitesse et précision. Même si les approches d'apprentissage profond donnent des bons résultats dans la plupart des contextes difficiles, le coût de calcul requis pour leur entraînement sur des jeux de données à grande échelle demeure élevé. Bien que certaines architectures profondes appliquées à différents domaines puissent fonctionner sur des CPU, pour la smart mobilité routière et ferroviaire, le recours au GPU est crucial.

2.3.1 Matériels informatique et systèmes d'acquisition

Pour effectuer l'entraînement des différents modèles d'apprentissage profond, nous avons fait appel au supercalculateur MYRIA [1] qui est utilisé par de nombreuses institutions de recherche et entreprises régionales pour des applications allant de la mécanique des fluides aux calculs deep learning. MYRIA est composé de 66 nœuds de calcul biprocesseurs Broadwell (28 cœurs à 2,4GHz, 128Go de RAM DDR4) dont 20 nœuds dédiés aux calculs deep learning, chacun d'entre eux étant équipé soit de 4 GPU Kepler K80 (12Go de VRAM par GPU), soit de 2 GPU Kepler P100 (12Go de VRAM par GPU) ou encore de 4 GPU Nvidia V100 (32Go). Chaque utilisateur dispose d'un espace de stockage de 50Go. Pour les tests, nous utilisons à la fois un serveur équipé de 2 GPU GTX1080Ti (11Go de VRAM chacun) et un ordinateur avec GPU GTX 1050 4GB (16Go de RAM) et CPU i5 8300h. Pour effectuer des tests en conditions réelles, nous utilisons des caméras Intel RealSense D435 fournissant des cartes de profondeur des distances d'objets. Ces caméras ont été utilisées dans [31] pour faire du SLAM (Simultaneous Localization and Mapping) visuel. Nous utilisons plusieurs librairies open-source de haut niveau disponibles dans le domaine du deep learning (e.g. TensorFlow, Pytorch, etc.).

2.3.2 Jeux de données

La deuxième contrainte la plus importante dans le domaine du deep learning est l'acquisition d'ensembles de données (dataset) à grande échelle (large scale) nécessaires à l'apprentissage et à l'évaluation des performances des algorithmes proposés par rapport à ceux

1. `https://www.criann.fr`

de l'état de l'art. En raison de l'importance de la conduite autonome, il existe de nombreux jeux de données dédiés au domaine routier tels que KITTI [59] qui fournit les images d'une caméra stéréoscopique, la profondeur de la scène mesurée par un LiDAR Velodyne ainsi que les annotations des véhicules et des piétons pour les objets détectés. Outre KITTI, de nombreux autres jeux de données sont disponibles, tels que CityScapes [70], Pascal VOC [71], MS-COCO [72], ImageNet [73] et OpenImages [74]. Pour évaluer et entrainer nos approches, nous utiliserons principalement KITTI et NuScenes.

2.4 Détection, localisation et tracking d'objets : évaluation des algorithmes

2.4.1 Détection d'objets

Critères d'évaluation. Les critères d'évaluation des méthodes de détection d'objets sont caractérisés par la métrique précision moyenne (mean Average Precision (mAP)) ainsi que le temps de réponse (Frame Per Seconde (FPS)). Pour être temps-réel, le temps de calcul doit être inférieur à $100ms$ ou supérieur à 10 FPS. La précision moyenne (mAP), qui quantifie la qualité de la détection (proportion de détection correcte) en fonction du Rappel (proportion des objets qui sont détectés). Nous définissons la Précision et le Rappel par les équations 2.1 et 2.2 :

$$Precision = \frac{VP}{VP + FP} \tag{2.1}$$

$$Rappel = \frac{VP}{VP + FN} \tag{2.2}$$

Où VP : Vrai Positif, FN : Faux Négatif et FP : Faux Positif. On considère qu'un objet est bien détecté si l'Intersection sur Union (IoU) entre la boite détectée et celle de la vérité terrain est supérieure à un seuil choisi. Pour calculer le mAP nous devons tout d'abord définir 11 valeurs de Rappel equidistantes,
$Rappel_i = \begin{bmatrix} 0.0 & 0.1 & ... & 0.9 & 1.0 \end{bmatrix}$. Le mAP se calcule ensuite via l'équation 2.3 :

$$mAP = \frac{1}{11} \sum_{Rappel_i} Precision(Rappel_i) \tag{2.3}$$

Evaluation et choix de la méthode. Afin d'évaluer les différentes approches de détection d'objets, nous avons utilisé comme critère la précision moyenne (mAP) ainsi que le temps de calcul. Dans un premier temps, nous avons effectué une comparaison qualitative des approches de détection d'objets du dataset MS-COCO [72] dont SSD [61] et YOLOv3 [53].

Tous les tests ont été effectués à l'aide d'un modèle pré-entraîné basé sur COCO, composé de 80 classes. Les résultats quantitatifs de cette évaluation peuvent être trouvés dans le tableau 2.1. Une comparaison qualitative des performances entre YOLOv3 et SSD est présentée dans la figure 2.2. Parmi les méthodes étudiées, nous pouvons citer Faster RCNN [65] et Mask RCNN [63]. Cette dernière est capable d'extraire un masque de l'objet détecté représentant ainsi une meilleure localisation de ce dernier. En terme de vitesse, et s'appuyant sur les tests des différentes implémentations de chaque méthode, le masque RCNN est au-dessous des 10 FPS requis pour un système temps-réel. YOLOv3 est rapide lorsqu'il est implémenté sous PyTorch et la meilleure implémentation de SSD correspond à celle originale sur Caffe. En terme de précision, l'évaluation des performances de détection demeure une tâche coûteuse en temps. Nous nous sommes donc appuyés sur les résultats obtenus dans l'article YOLOv3 [53] pour choisir l'algorithme optimal. YOLOv3 fonctionne mieux que SSD, ce qui est confirmé par les tests d'inférence que nous avons effectués. Au vu des résultats obtenus dans le tableau 2.1, la méthode la plus rapide est YOLOv3 et la plus précise est la méthode à deux étage FPN FRCNN. Nous pouvons néanmoins constater que YOLOv3 possède une précision très proche de celle-ci (-1.2 mAP) est surpasse de loin l'approche SSD. Nous avons donc fait le choix d'utiliser YOLOv3 pour la partie détection d'objets de notre algorithme.

FIGURE 2.2 – Comparaison des performances : YOLOv3 (gauche) et SSD (droite).

Approche	mAP	Temps de calculs (ms)
SSD321	45.4	61
SSD513	50.4	125
R-FCN	51.9	85
FPN FRCN	**59.1**	172
YOLOv3-320	51.5	**22**
YOLOv3-416	55.3	29
YOLOv3-608	57.9	51

TABLE 2.1 – Evaluation des performances des méthodes de détection d'objets de l'état de l'art sur le jeu de données COCO.

2.4.2 Estimation de la profondeur

Métriques d'évaluation

Afin d'évaluer les méthodes les plus prometteuses, nous avons utilisé le temps de calcul, des résultats qualitatifs ainsi que l'écart quadratique moyen (RMSE) présenté dans l'équation 2.4 :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}\sum_{j}(g_{i,j} - p_{i,j})^2} \tag{2.4}$$

Avec i, j les coordonnées sur l'image, $g_{i,j}$ la vérité terrain, $p_{i,j}$ la prédiction de profondeur et N le nombre total de points. Cependant l'évaluation des algorithmes était une tâche difficile vue que chacune des méthodes évaluées avait ses propres caractéristiques (carte de profondeur, échelle, focale, baseline, etc.). Pour résoudre ce problème, nous avons testé chaque méthode sur une séquence du dataset KITTI [59], disposant d'une vérité terrain sur la profondeur garantie par un LiDAR. Nous avons ainsi pu transformer les cartes de disparités en cartes de profondeurs en utilisant l'échelle médiane entre l'inverse de la disparité prédite et la vérité terrain. D'une manière similaire, nous avons mis à l'échelle les cartes de profondeurs provenant de certaines méthodes en utilisant l'échelle médiane entre la carte de profondeur et la vérité terrain. L'équation 2.5 permet de transformer une carte de disparité en carte de profondeur non mise à l'échelle et l'équation 2.6 permet sa mise à l'échelle :

$$\widetilde{d_{i,j}} = \frac{1}{disp_{i,j}} \tag{2.5}$$

$$d_{i,j} = \widetilde{d_{i,j}} \times \frac{med(d_{gt})}{med(\widetilde{d})} \tag{2.6}$$

Avec i, j les coordonnées du point sur la carte, $disp$ la disparité, $\widetilde{d}$ la profondeur non mise à l'échelle et d la profondeur mise à l'échelle.

Evaluation de l'estimation de la profondeur

Notre objectif était de tester plusieurs algorithmes de deep learning d'estimation de distance. Pour cela, nous nous sommes concentrés sur les méthodes qui ne nécessitent pas de supervision lors de l'apprentissage et ce, à cause de la non disponibilité de la vérité terrain. Pour nos expériences, nous utilisons le jeu de données KITTI comprenant des cartes de profondeur avec vérité terrain. Afin d'identifier la méthode à utiliser, nous devons d'abord déterminer laquelle est la plus performante dans chaque catégorie : monoculaire et stéréoscopique. L'évaluation des algorithmes est problématique car chaque méthode produit sa propre carte de disparité rendant difficile l'obtention de la carte de profondeur. Pour sur-

monter ce problème, chaque méthode a été testée sur une séquence du dataset KITTI. Nous pouvons ainsi transformer la carte de disparité obtenue en une carte de profondeur en utilisant l'échelle médiane entre l'inverse de la disparité prédite et la vérité terrain. La précision de la méthode est ensuite estimée en calculant la différence quadratique moyenne entre la distance prédite par l'algorithme et la vérité terrain.

Approches monoculaires. Parmi les méthodes monoculaires testées, on trouve des approches telles que SfmLearner [75], MonoResMatch [76], Monodepth [55] et Monodepth2 [54]. À l'exception de [75] qui est entraîné sur des séquences d'images monoculaires, tous ces algorithmes nécessitent une paire d'images stéréoscopiques pour pouvoir effectuer l'entraînement. Cependant, les inférences sont réalisées sur des images monoculaires réduisant ainsi le coût dû à l'équipement. Des exemples de cartes de disparité obtenues peuvent être trouvés dans la figure 2.3. Les résultats de notre évaluation quant à eux, sont présentés dans le tableau 2.2.

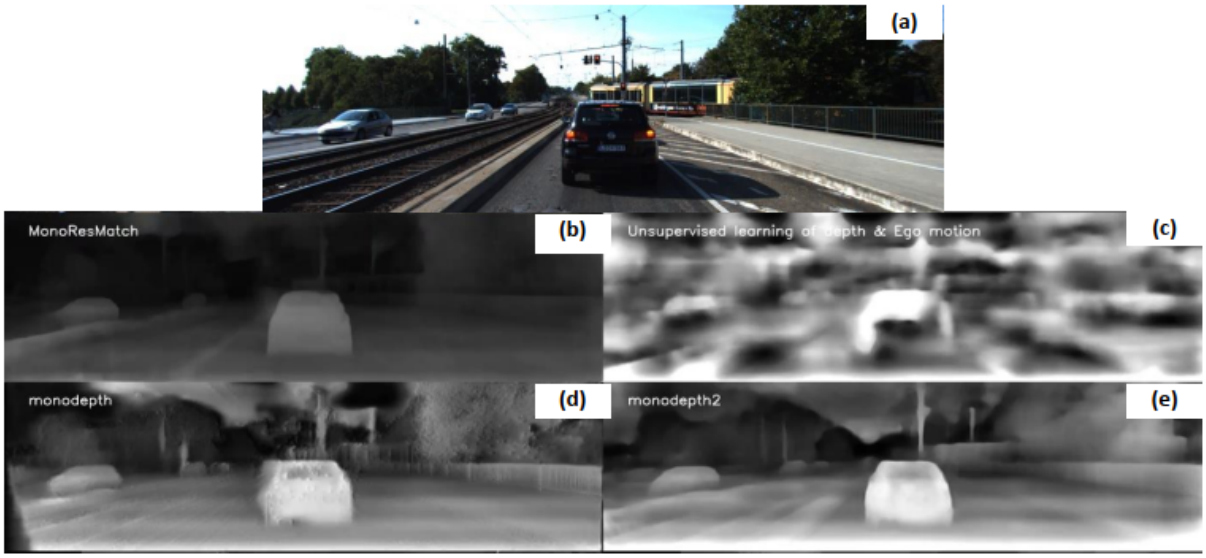

FIGURE 2.3 – Résultats des cartes de disparité obtenues par des approches monoculaires : (**a**) image originale, (**b**) MonoResMatch, (**c**) SfmLearner, (**d**) Monodepth et (**e**) Monodepth2.

Approches	**RMSE**	**Temps de calcul (ms)**
sfmLearner	16.530	50
Monodepth	6.225	200
MonoResMatch	5.831	1000
Monodepth2	5.709	50

TABLE 2.2 – Résultats expérimentaux sur le dataset routier KITTI. Les cartes de profondeurs provenant d'un capteur sont utilisées comme vérité terrain.

D'après les résultats du Tableau 2.2, il apparaît que Monodepth2 [54] est la méthode la

plus adaptée si une caméra monoculaire est utilisée. Celle-ci offre les meilleures performances en termes d'estimation de distance et de temps de calcul.

Approches stéréoscopiques. L'estimation de distance par caméra est souvent associée à l'utilisation de caméras stéréoscopiques calibrées. Ces méthodes classiques sont basées sur l'association des pixels entre les deux caméras pour obtenir une carte de disparité permettant de réaliser la carte de profondeur. Cependant, l'association ne fonctionne que si la scène filmée est suffisamment texturée. Cela peut être corrigé par des méthodes de post-traitement comme le filtre WLS [77], mais la qualité des résultats est très variable. C'est pourquoi les approches deep learning peuvent être une bonne alternative. Nous testons MADNet [56] qui a la particularité d'être adaptatif pouvant effectuer l'entraînement en même temps que l'inférence et ainsi pallier à un défaut majeur des algorithmes d'estimation de distance par deep learning : des performances dégradées sur des environnements inconnus. La figure 2.4 présente des exemples de cartes de disparité obtenues par ces méthodes. Les résultats quantitatifs sont présentés dans le tableau 2.3.

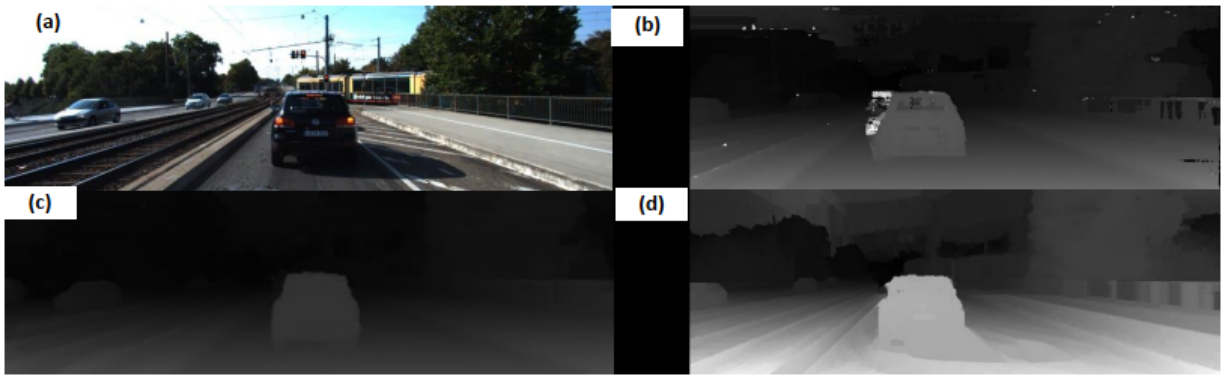

FIGURE 2.4 – Résultats des cartes de disparité obtenues par des approches stéréoscopiques : (**a**) image originale, (**b**) approche stéréo de base, (**c**) MADNet et (**d**) filtre stéréo WLS.

Approches	**RMSE**	**FPS**
Stereo-baseline	9.002	15
Stereo-WLS Filter	8.690	7
MADNet	4.648	7

TABLE 2.3 – Résultats expérimentaux des approches stéréoscopiques comparés à notre proposition basée sur MADNet.

Choix de l'approche. D'après notre étude d'évaluation des approches monoculaires (Monodepth2) et stéréoscopiques (MADNet), Monodepth2 est plutôt rapide (50 ms/image). Malgré un temps de calcul légèrement inférieur au seuil dit "temps-réel" (10 FPS), MADNet est

plus précise et offre des performances en estimation de distance supérieures à toutes les autres méthodes testées. Elle a également l'avantage de pouvoir s'adapter à de nouveaux environnements avec une stéréo permettant une estimation plus robuste que les méthodes monoculaires. Cette évaluation a permis aussi de mettre en évidence les performances supérieures des méthodes basées sur l'apprentissage profond vis-à-vis des approches "classique". Nous avons choisi la méthode MADNet car elle est robuste, temps-réel et peut s'adapter à de nouveaux environnements. Nous avons aussi fait le choix d'utiliser une caméra stéréoscopique avec MADNet car l'adaptation permet de compenser les problèmes liés au changement de domaine (lorsqu'une méthode perd en précision une fois appliquée sur des images différentes de celles utilisées pour l'entrainement et les tests).

2.4.3 Localisation d'objets

Grâce aux informations fournies par la détection d'objets et l'estimation de la profondeur, il était donc possible de localiser chaque objet dans un plan 3D du repère caméra. Notre problème se résumait donc à projeter les coordonnées en pixels d'un objet dans le repère caméra. Avec (o_x, o_y) les coordonnées du centre optique de la caméra, (s_x, s_y) la taille d'un pixel et f la distance focale. Les coordonnées Z peuvent être déterminées en utilisant les informations de la carte de profondeur (e.g. théorème 3D de Pythagore). Afin d'effectuer la localisation, nous combinons la détection d'objet et l'estimation de la profondeur dans un seul calcul parallèle (Figure 2.5) afin de vérifier si la contrainte temps-réel est réalisable.

FIGURE 2.5 – Résultats de la détection et localisation d'objets sur un échantillon du jeu de données KITTI.

2.4.4 Tracking d'objets

Tracking d'objets 2D. Plutôt que d'utiliser une approche classique de suivi visuel (Mean-shift, Camshift, etc.), nous avons plutôt choisi de tirer parti des informations que nous avions déjà acquises au travers de la détection et la localisation d'objets. Nous nous sommes donc

basés sur le Simple Online Real-Time Tracking (SORT) [57] qui offre de bonnes performances, un temps de calculs extrêmement faible et une architecture simple à modifier. Il utilise les boîtes englobantes fournies par un algorithme de détection pour initialiser les cibles qui sont ensuite suivies grâce à un filtre de Kalman [78] possédant un vecteur d'état X et un vecteur de mesure z, définis par :

$$X = (x \;\; y \;\; r \;\; a \;\; \dot{x} \;\; \dot{y} \;\; \dot{a})^t, \;\; z = (x \;\; y \;\; r \;\; a)^t \tag{2.7}$$

Où (x, y) sont les coordonnées de la boîte englobante, r son rapport (largeur sur hauteur) et a son aire. Le modèle de prédiction est basé sur une vitesse constante. L'association entre les cibles et les détections est faite en calculant l'IoU entre une détection et la position prédite par le filtre de Kalman. Si la détection obtient un IoU supérieur à un seuil prédéfini et a l'IoU le plus élevé parmi les autres détections, alors elle est associée à la cible et le vecteur d'état du filtre de Kalman correspondant à la cible est mis à jour. La performance de cet algorithme dépend directement de la qualité de la détection d'objets et le temps de calcul demeure inférieur à $1ms/frame$.

Tracking d'objets 3D. Compte tenu de notre objectif de suivi d'objets 3D, de nombreuses modifications ont été apportées à SORT. Ainsi, le filtre de Kalman a été remplacé par un Filtre de Kalman Etendu (FKE) :

$$X_s = (X \;\; Y \;\; Z \; a \;\; r \;\; \dot{X} \;\; \dot{Y} \;\; \dot{Z} \;\; \dot{a})^t, \;\; z = (x \;\; y \;\; d \;\; r \;\; a)^t \tag{2.8}$$

Avec (X, Y, Z) les coordonnées en 3D et d est la profondeur estimée par la détection. Afin d'effectuer des prédictions, nous utilisons le modèle à vitesse constante. Le vecteur de mesure z, composé des coordonnées mesurées avec la détection d'objet et l'estimation de la distance, nous permet de mettre à jour le vecteur d'état du filtre.

Résultats expérimentaux. Il est essentiel d'effectuer un bon ajustement des paramètres du filtre FKE pour espérer un suivi d'objets précis. La matrice de bruit de contrôle ajuste le comportement du filtre. Ainsi, le taux de confiance entre les mesures et le modèle de prédiction peut être ajusté. Pour filtrer un bruit de mesure important, il est préférable de réduire le taux de confiance des mesures, mais FKE aura plus de difficultés en cas de changement de trajectoire. Il s'agit donc de trouver le bon compromis entre le filtrage du bruit et le contrôle des mesures.

Nous avons également testé l'algorithme de tracking dans plusieurs environnements. Dans un premier temps, les tests ont été effectués en indoor (afin de vérifier la convergence des valeurs prédites). Par la suite, nous avons effectué des tests en outdoor dans le domaine

routier en utilisant le jeu de données KITTI [59]. Dans la figure 2.6, nous pouvons voir que le vélo à droite n'est plus détecté à partir de l'image $(t + 1)$. La prédiction du filtre de Kalman est alors utilisée pour estimer la position de cet objet. Si celui-ci n'est pas associé à une détection avant 3 images, l'algorithme supprime cette cible.

FIGURE 2.6 – Résultats du suivi d'objets dans un environnement routier (KITTI).

Notre méthode permet également de suivre une cible même si elle n'est pas associée à un blob de détection et ce, jusqu'à 3 images avant qu'elle ne soit supprimée. Nous pouvons voir dans la figure 2.7 les résultats quantitatifs et qualitatifs de notre méthode sur deux séquences différentes du jeu de données KITTI. Ces travaux de recherche ont fait l'objet de publications dans [79].

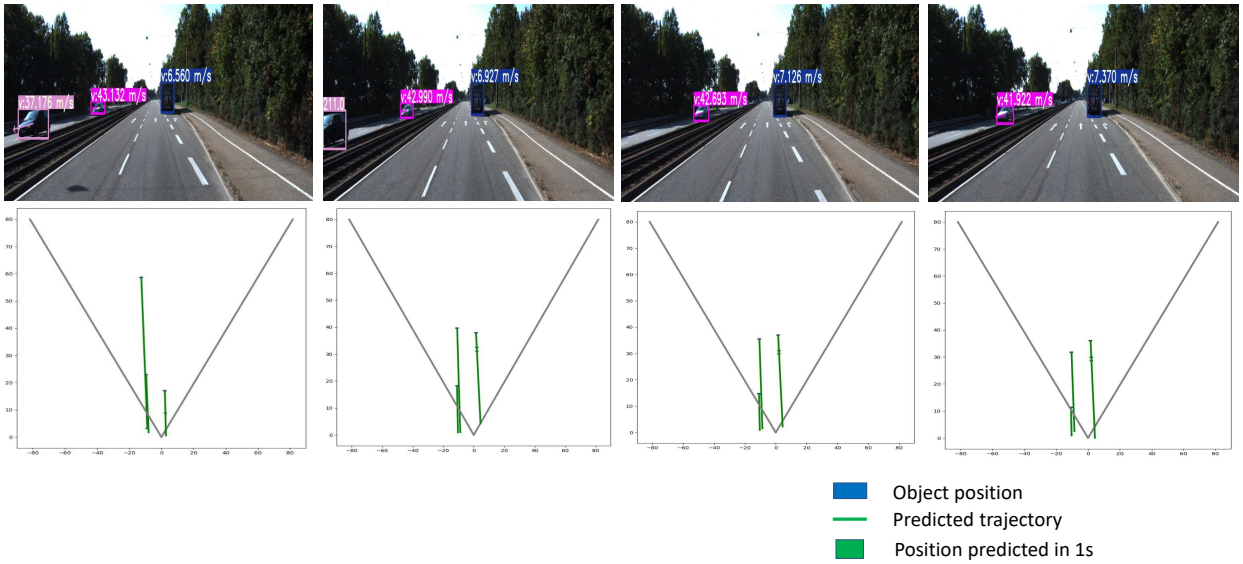

FIGURE 2.7 – Résultats du suivi d'objets sur une séquence KITTI. En haut : 4 images RGB provenant d'une séquence de route acquise à différents moments avec les boîtes de tracking correspondantes des objets en mouvement et leurs valeurs de vitesse. En bas : les cartes montrées permettent un aperçu simple mais complet des mouvements et des distances des objets suivis.

2.5 Evaluation de l'estimation de profondeur par apprentissage profond

2.5.1 Contexte & objectifs

L'estimation précise de la profondeur est nécessaire pour la perception d'environnement et peut augmenter considérablement la sécurité en estimant la distance des piétons et des véhicules. Dans ce contexte, nous avons déjà réalisé plusieurs travaux comme le suivi de personnes [80] ou la détection et le tracking d'objets, pour la smart mobilité routière [6, 30]. Dans cette section, nous aborderons les travaux que nous avons menés afin d'approfondir l'étude des méthodes d'estimation de profondeurs, indispensable pour la localisation d'objets. Dans la section précédente, nous avons déjà évalué exhaustivement les méthodes monoculaires pour l'estimation de profondeur, cependant, nous avions seulement évalué une seule méthode stéréoscopique pour l'estimation de profondeur basée deep learning. Une autre difficulté constatée est que le protocole d'évaluation, identique à celui de la littérature, que nous avions utilisé n'est pas réellement adapté pour la localisation d'objets.

En effet, l'évaluation des cartes de profondeurs fournies par les différentes méthodes effectuées sur la totalité de l'image ne fournissent pas d'information sur la qualité de la

prédiction de profondeur des pixels appartenant aux objets d'intérêts (voitures, piétons, etc.). De plus, le protocole d'évaluation actuel ne donne aucune information sur la dégradation de la précision de l'estimation de profondeur à mesure que la distance augmente. Ces informations sont vitales pour les applications de la conduite autonome. Cette section comprend deux contributions essentielles : (*i*) l'approfondissement de l'étude des méthodes stéréoscopiques basées deep learning pour l'estimation de profondeur et (*ii*) création d'un nouveau protocole d'évaluation mieux adapté aux environnements routiers ainsi qu'une évaluation comparative des méthodes les plus récentes sur deux gros jeux de données : KITTI [59] et NuScenes [81].

2.5.2 Etat de l'art

Estimation de profondeur monoculaire. Nous avons choisi, pour cette étude comparative, d'utiliser Monodepth2 qui avait été identifié dans les sections précédentes comme étant la méthode monoculaire la plus précise. Cependant, nous avons aussi identifié une nouvelle méthode, nommée BTS (Big-To-Small) [82] qui propose d'obtenir la profondeur à pleine résolution en fusionnant les sorties des différentes couches intermédiaires de la partie décodeur du réseau. Cette architecture est actuellement parmi les méthodes les plus performantes sur le benchmark d'estimation de la profondeur du jeu de données KITTI.

Estimation de profondeur stéréoscopique. Dans PSMNet [83], des travaux récents ont montré que l'estimation de la profondeur à partir d'un couple d'images stéréo peut être formulée comme une tâche d'apprentissage supervisé à résoudre avec des réseaux CNN. PSMNet est un réseau de correspondance stéréo pyramidale composé de deux modules principaux : le Spatial Pyramidal Pooling (SPP) et un CNN 3D. L'approche proposée crée ce qui est appelé le "volume coût" (*cost volume*) en envoyant les images stéréo d'entrée à deux pipelines de partage de poids composés d'un CNN pour le calcul des cartes de caractéristiques, d'un module SPP pour leur récolte et d'une couche de convolution pour leur fusion. Les sorties de ces deux pipelines sont ensuite combinées afin de former le volume coût. Dans [84] et afin d'améliorer l'approche proposée par PSMNet, une nouvelle méthode d'estimation de profondeur stéréoscopique a été proposée. Cette approche incorpore des améliorations pour l'architecture du "sablier empilé" présent dans le CNN 3D ainsi qu'un nouveau type de réseau stéréoscopique à corrélation par groupe (Group-Wise Correlation stereo Network ou GWC-Net). Cette approche comprend quatre modules : extraction des cartes de caractéristiques des images, construction de volumes de coûts, agrégation 3D et prédiction de la disparité. La contribution principale de cette approche vis-à-vis de PSMNet repose sur l'utilisation de la corrélation par groupe pour construire le volume de coût.

2.5.3 Métriques des erreurs utilisées dans l'évaluation approfondie

Afin d'évaluer les performances des approches d'estimation de profondeur, différentes erreurs statistiques sont utilisées : L'erreur relative (RE), l'erreur relative au carré (SRE), l'erreur quadratique moyenne (RMSE) et l'erreur quadratique moyenne logarithmique (logRMSE). Ces métriques donnent une évaluation globale de la performance d'une méthode sur l'ensemble de l'image testée. Nous notons par p la prédiction de profondeur, gt sa vérité terrain correspondante et N le nombre total de pixels de profondeur dans l'image. Enfin u et v correspondent aux coordonnées en pixels du point dans le repère image.

Erreur relative : RE (Relative Error) est détaillée dans l'équation (2.9) :

$$RE = \frac{1}{N} \sum_{u} \sum_{v} \frac{|gt_{u,v} - p_{u,v}|}{gt_{u,v}} \tag{2.9}$$

Erreur relative au carré : SRE (Squared Relative Error) est détaillée dans l'équation (2.10) :

$$SRE = \frac{1}{N} \sum_{u} \sum_{v} \frac{|gt_{u,v} - p_{u,v}|^2}{gt_{u,v}} \tag{2.10}$$

L'erreur quadratique moyenne : RMSE (Root Mean Squared Error) est donnée par l'équation (2.11) :

$$RMSE = \sqrt{\frac{1}{N} \sum_{u} \sum_{v} (gt_{u,v} - p_{u,v})^2} \tag{2.11}$$

Erreur quadratique moyenne logarithmique : logRMSE est donnée par l'équation (2.12) :

$$logRMSE = \sqrt{\frac{1}{N} \sum_{u} \sum_{v} (log(gt_{u,v}) - log(p_{u,v}))^2} \tag{2.12}$$

Pourcentage de pixels non conformes : BMP (Bad Matching Pixels) est donnée par l'équation (2.13), où C est le seuil de tolérance d'erreur :

$$[a]_{k=[1..3]} = \frac{1}{N} \sum_{u} \sum_{v} max(\frac{g_{u,v}}{p_{u,v}}, \frac{p_{u,v}}{g_{u,v}}) < C^k \tag{2.13}$$

Ces métriques donnent une évaluation statistique complète de la performance d'une méthode, mais nous les avons améliorés via le développement d'un nouveau protocole d'évaluation de la profondeur.

2.5.4 Résultats expérimentaux

Nous avons choisi d'explorer à la fois des méthodes stéréoscopiques (GWCNet, PSMNet) mais aussi monoculaires (BTS, Monodepth2) afin de comparer leur écart de précision. Ces approches ont été choisies car elles sont les plus précises de la littérature.

Évaluation des méthodes monoculaires. Nous avons utilisé les modèles pré-entraînés BTS et Monodepth2 pour l'évaluation sur le jeu de données KITTI. Les deux méthodes ont été évaluées sur la partie de test définie par Eigen [85]. BTS a été entraîné avec une résolution de 704×352 tandis que Monodepth2 a été entrainé en utilisant le mode monoculaire non supervisé sur la partie d'entraînement de KITTI définie dans les travaux de Zhou [75] avec une résolution de 1024×320. Les résultats sont présentés dans le tableau 2.4.

	RE	SRE	RMSE	logRMSE
Monodepth2	0.115	0.882	4.701	0.190
BTS	**0.060**	**0.249**	**2.798**	**0.096**

TABLE 2.4 – Évaluation de la profondeur monoculaire sur KITTI des méthodes Monodepth2 et BTS. SRE et RMSE sont exprimées en mètres.

Les résultats montrent que BTS est globalement plus performant. L'une des raisons de cette performance est probablement due à l'entrainement de BTS en mode supervisé alors que Monodepth2 a été entrainé en mode non supervisé. Monodepth2 exploite des séquences d'images pour l'entraînement non supervisé et utilise un modèle pour apprendre l'ego-motion de la caméra en mode supervision. Il en résulte un grand nombre d'approximations pendant l'apprentissage. De plus, le modèle BTS est beaucoup plus profond et plus lourd en termes de calculs que Monodepth2.

Évaluation des méthodes stéréoscopiques. Les résultats des méthodes stéréoscopiques sur KITTI montrent que GWCNet est significativement meilleur que PSMNet. L'évaluation quantitative (e.g. tableau 2.5) permet de déterminer la méthode la plus précise. Nous pouvons expliquer la différence de performance entre GWCNet et PSMNet par le fait que les fonctionnalités d'agrégation des caractéristiques rendent le réseau plus robuste face aux scènes routières difficiles (objets et régions sans texture, changements d'éclairage, etc.) qui sont plus présents dans GWCNet en raison de la corrélation de groupe fortement recommandée dans les réseaux stéréoscopiques.

	RE	SRE	RMSE	logRMSE
GWCNET	**0.018**	**0.048**	**0.981**	**0.042**
PSMNET	0.032	0.061	1.139	0.056

TABLE 2.5 – Évaluation de la profondeur stéréoscopique sur KITTI des méthodes GWCNet et PSMNet. SRE et RMSE sont exprimées en mètres.

Notre évaluation expérimentale a montré que BTS et GWCNet offrent de meilleures performances pour les images monoculaires et stéréoscopiques respectivement. Les approches stéréoscopiques ont une plus grande précision. Néanmoins, cette étude plus approfondie des méthodes d'estimation de profondeur ne permet pas de déterminer quantitativement la performance de ces approches pour l'estimation de profondeur d'objets. Il est aussi impossible, via cette évaluation, de mesurer la dégradation de la précision en fonction de la distance.

2.6 Nouveau protocole d'évaluation pour l'estimation de profondeur

Afin de pallier aux problèmes soulevés dans la section précédente, nous avons conçu un nouveau protocole d'évaluation de l'estimation de profondeur ayant comme objectif d'obtenir une évaluation cohérente avec notre tâche de localisation d'objets. Ce nouveau protocole comprend désormais deux parties : (i) évaluer la qualité de la prédiction de la profondeur en fonction de la distance afin de quantifier la dégradation de la précision lorsque la distance augmente, (ii) quantifier la précision de l'estimation de profondeur des points appartenant à un objet. Chacune des méthodes qui seront évaluées renvoie soit des disparités de distance soit la profondeur elle-même. Bien qu'il soit possible de trouver la profondeur à partir de la disparité, chacune des méthodes a été entraînée avec des résolutions d'images particulières. Nous avons donc opté pour une mise à l'échelle via les informations de vérité terrain. Cela nous a permis de prédire la profondeur pour chacune des méthodes sans faire appel à l'équation (2.1) (Section 2.4).

2.6.1 Évaluation de la profondeur selon l'objet

Alors que l'évaluation actuelle des estimations de la profondeur donne une évaluation complète de la performance globale d'une méthode donnée, elle est faite sur l'image globale et n'évalue pas la prédiction de la distance des objets. C'est pourquoi nous avons conçu un nouveau protocole d'évaluation qui nous permet de calculer l'erreur de prédiction de la profondeur pour les objets pertinents (personne, voiture, camion, etc.). Notre protocole d'évaluation est composé de 4 étapes :

- La carte de profondeur prédite est mise à l'échelle en utilisant une médiane (soit p la carte de profondeur prédite et g la carte de profondeur de la vérité terrain, la mise à l'échelle médiane est décrite comme suit : $p = p * \frac{med(gt)}{med(p)}$)
- Les masques d'objets sont générés en utilisant Mask-RCNN [86]
- Les masques générés sont ensuite utilisés pour segmenter les cartes de profondeur et les erreurs de profondeur sont calculées pour chaque masque dans l'image

— Enfin, la moyenne des erreurs est calculée pour chaque classe

Ce nouveau protocole d'évaluation permettra de mieux comprendre la façon dont une méthode donnée estime la distance d'objets. La figure 2.8 illustre un exemple d'application de notre protocole d'évaluation.

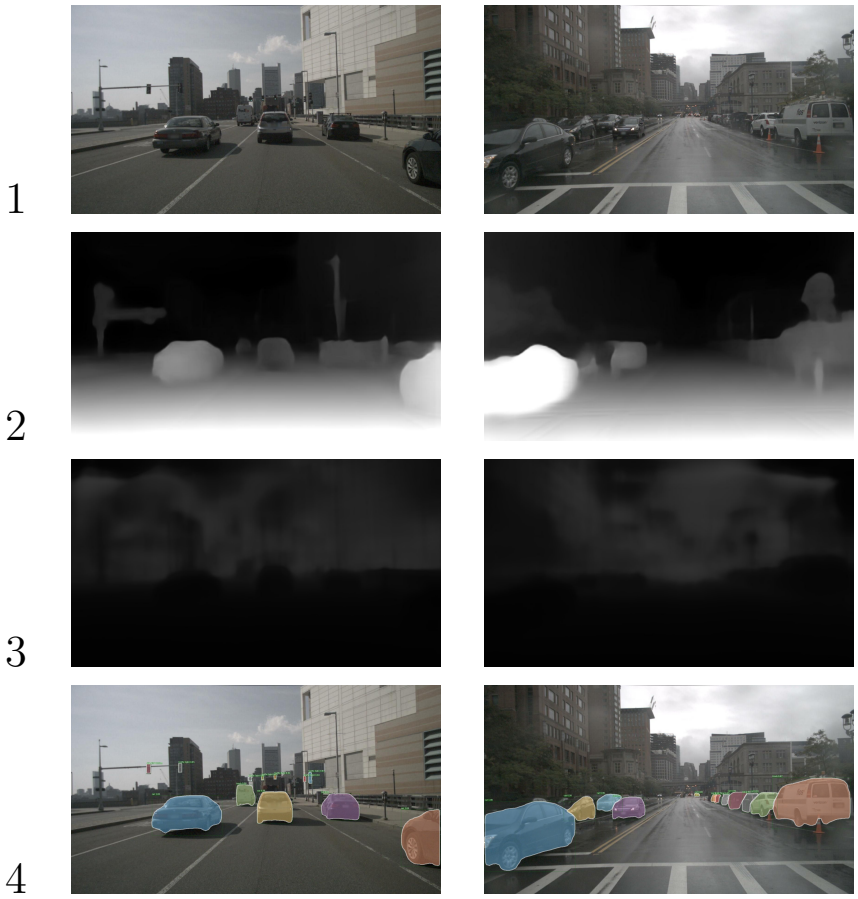

FIGURE 2.8 – Données d'entrée pour notre protocole d'évaluation de la distance selon la classe d'objets. 1. l'image d'entrée alimentant l'algorithme de prédiction de la profondeur, 2. la carte de disparité, 3. la carte de profondeur normalisée après mise à l'échelle médiane et 4. les masques d'objets issus de Mask-RCNN.

2.6.2 Évaluation de la profondeur sur des plages de distance

Un autre inconvénient du protocole classique d'évaluation de la profondeur est qu'il ne permet pas d'évaluer les performances d'une méthode selon la distance. Nous nous sommes inspirés des travaux de [87] qui décrivait un protocole d'évaluation sur des plages de distance pour des scènes intérieures. Nous avons adapté le protocole à des environnements extérieurs (scènes routières). Il comprend les étapes suivantes : 1. la carte de profondeur prédite est mise à l'échelle en utilisant une échelle médiane, 2. des plages de distance de $10m$ à $80m$ ont été créées, 3. chaque pixel est affecté à une plage de distance en fonction de sa valeur dans la

vérité terrain de la profondeur, 4. pour chaque plage de distance, les erreurs de profondeur sont calculées. Ce nouveau protocole permet d'évaluer la dégradation de l'estimation de la profondeur en fonction des distances.

2.6.3 Jeux de données : KITTI & NuScenes

Nous avons proposé une méthode basée sur des plages de distance permettant d'évaluer l'évolution de la précision de la profondeur sur la distance et un protocole pour évaluer les prédictions de la profondeur des objets en utilisant les masques d'objets générés par Mask-RCNN. Nous avons ensuite réalisé une évaluation d'une méthode non supervisée et d'une méthode supervisée avec Monodepth2 et BTS sur deux jeux de données à grande échelle pour les scènes routières, KITTI et NuScenes. Les résultats comparatifs de l'évaluation montrent que BTS a de meilleures performances que Monodepth2 sur tous les aspects. Nous avons également montré que les erreurs d'estimation de la profondeur des objets étaient significativement plus élevées que celles sur l'image entière pour les deux méthodes.

KITTI. Le modèle BTS a été entraîné avec des images provenant de l'entraînement d'Eigen [85] à une résolution de 704×352 et avec une vérité terrain dense. Les poids de Monodepth2 ont été entraînés en utilisant la vérité terrain monoculaire sur l'ensemble de Zhou [75], à une résolution de 1024×320. L'évaluation a été réalisée sur l'ensemble de test d'Eigen. Les résultats de notre évaluation de BTS et Monodepth2 sont présentés dans le tableau 2.6 (la valeur de C pour l'erreur de seuil définie dans l'équation (2.13) a été fixée à $1,25$).

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Classe de l'objet	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
Personne	0.314	**0.166**	5.721	**1.786**	8.430	**5.892**	0.326	**0.253**	0.601	**0.772**	0.829	**0.894**	0.920	**0.947**
Deux roues	0.131	**0.116**	0.517	**0.467**	2.810	**2.669**	0.172	**0.163**	0.829	**0.839**	**0.964**	0.962	0.993	**0.994**
Voiture	0.206	**0.137**	3.132	**1.491**	7.924	**6.052**	0.271	**0.223**	0.773	**0.838**	0.883	**0.922**	0.938	**0.955**
Camion	0.215	**0.122**	2.769	**0.826**	6.978	**4.523**	0.259	**0.177**	0.694	**0.854**	0.903	**0.969**	0.964	**0.985**

TABLE 2.6 – Évaluation de la distance selon l'objet sur KITTI : Monodepth2 (MD2) et BTS. Les erreurs de profondeur ont été calculées pour les classes d'objets ayant suffisamment d'instances dans l'ensemble de test. La RMSE est exprimée en mètres.

NuScenes. Nous avons entraîné les deux méthodes sur l'ensemble d'entraînement du jeu de données NuScenes. Pour BTS, nous avons réalisé un entraînement de 50 epochs avec une taille de batch de 20 et une résolution de 192×192, avec les données éparses de la supervision LiDAR. Nous avons entraîné Monodepth2 pendant 20 epochs avec une taille de batch de 12 et une résolution de 446×224. Les résultats de notre évaluation de BTS et de Monodepth2 sont présentés dans le tableau 2.7 (la valeur de C pour l'erreur de seuil définie dans l'équation (2.13) a été fixée à $1,25$).

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Classe de l'objet	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
Voiture	0.346	**0.218**	6.853	**2.144**	10.420	**6.862**	0.448	**0.278**	0.546	**0.708**	0.736	**0.880**	0.920	**0.949**
Personne	0.501	**0.384**	8.312	**3.910**	9.291	**7.858**	0.531	**0.449**	0.438	**0.492**	0.679	**0.717**	0.803	**0.839**
Bus	0.448	**0.228**	11.837	**2.226**	13.929	**7.811**	0.448	**0.274**	0.465	**0.644**	0.729	**0.891**	0.848	**0.958**
Camion	0.324	**0.218**	6.803	**2.091**	11.425	**7.263**	0.378	**0.260**	0.574	**0.674**	0.793	**0.902**	0.887	**0.964**
Moto	0.284	**0.245**	1.671	**1.430**	4.509	**3.917**	0.320	**0.288**	0.512	**0.658**	0.868	**0.869**	0.935	**0.946**

TABLE 2.7 – Évaluation de la distance selon la classe des objets sur NuScenes : Monodepth2 (MD2) et BTS. RMSE est exprimée en mètres.

2.6.4 Résultats expérimentaux

Nos résultats sur les deux jeux de données montrent que, globalement, BTS donne de meilleurs résultats que Monodepth2. Notre évaluation sur les plages de distance montre également que les deux méthodes, comme prévu, ont tendance à avoir une précision plus faible lorsque la distance augmente. Les erreurs de prédiction de l'estimation de la profondeur sont significativement plus élevées que les erreurs sur l'image globale. Cela peut s'expliquer par la grande variété de chaque classe d'objets qui rend l'apprentissage de la profondeur plus difficile, alors que l'environnement est moins variable, ce qui facilite l'apprentissage de la profondeur par ces méthodes. Nous constatons aussi que les erreurs peuvent être deux fois plus élevées que l'erreur globale pour les personnes et les voitures, et cela doit être pris en compte si ces méthodes sont utilisées dans le cadre du véhicule autonome. Enfin, les résultats obtenus sur le jeu de données NuScenes ont des erreurs plus élevées que ceux obtenus sur KITTI, ce qui peut s'expliquer par les différences dans l'entraînement entre les deux jeux de données, notamment que NuScenes possède des scènes plus complexes. En combinant nos deux protocoles d'évaluation, nous avons également calculé l'évolution de l'erreur pour des objets tels que les voitures sur des plages de distance (voir figure 2.9). Ces résultats comparatifs peuvent être utilisés pour évaluer l'adéquation d'une méthode d'estimation de la profondeur à un scénario particulier dans les environnements routiers. Par exemple, pour la conduite sur une route dans des conditions idéales, où nous supposons que le véhicule roule à 90km/h, la méthode doit être précise jusqu'à 60m (distance de sécurité recommandée entre deux véhicules à cette vitesse). Ces travaux de recherche ont fait l'objet de plusieurs publications dans [8, 9, 38] et [88].

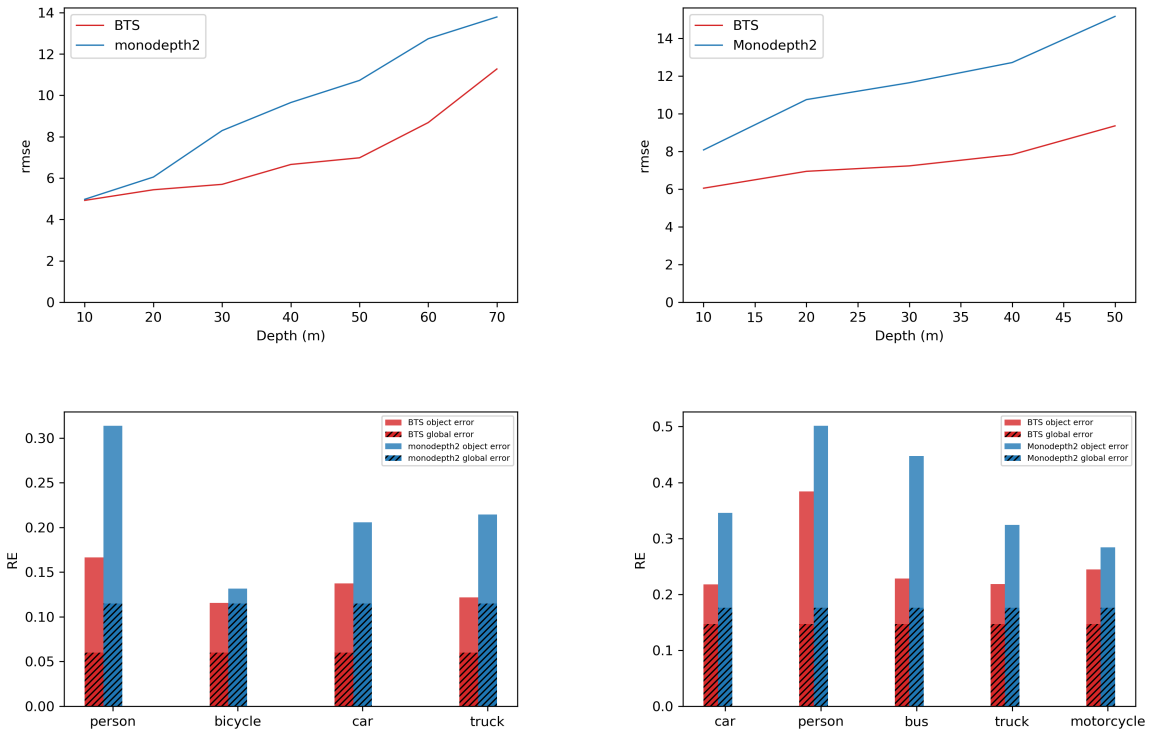

FIGURE 2.9 – RMSE (en mètres) pour la classe d'objets voiture selon la distance (haut gauche : KITTI, haut droite : NuScenes). Nos résultats d'erreur relative (RE) de BTS et Monodepth2 pour différentes classes d'objets comparés à l'erreur relative (RE) globale (hachuré) : bas gauche : KITTI, bas droite : NuScenes.

2.7 Conclusion

Nous avons présenté dans ce chapitre deux grandes parties : la détection, localisation et tracking d'objets 2D ainsi qu'un nouveau protocole d'évaluation de la distance d'objets. Pour la première partie, un système basé sur l'apprentissage profond de bout en bout pour la détection multi-objets, l'estimation de la profondeur, la localisation et le suivi d'objets dans un environnement routier réaliste a été présenté. Pour le module de détection d'objets, un détecteur efficace basé sur YOLOv3 a été proposé. Conjointement, l'estimation des cartes de profondeur a été introduite afin de rendre les informations de localisation d'objets détectés. À cette fin, nous utilisons deux types d'approches différentes, à savoir Monodepth2 et MADNet, pour concevoir le deuxième module de localisation d'objets. Enfin, nous avons introduit une nouvelle méthode de suivi d'objets basée sur une version améliorée de l'approche SORT. Un filtre de Kalman étendu (FKE) est développé pour améliorer l'estimation de la position des objets. Pour chaque étape, nous avons mis en place diverses expériences sur l'ensemble de données KITTI tout en comparant notre approche avec celles de l'état de l'art. Dans

l'ensemble, les résultats qualitatifs et quantitatifs rapportés prouvent que notre approche est suffisamment robuste pour relever les défis des scènes routières.

Dans la deuxième partie, nous avons présenté nos travaux visant à approfondir l'évaluation des méthodes d'apprentissage profond pour l'estimation de profondeur. Dans un premier temps, nous avons enrichi notre étude comparative en évaluant de nouvelles méthodes stéréoscopiques et monoculaires d'estimation de profondeur. Nous avons aussi conçu notre propre protocole dédié à l'évaluation de l'estimation de la profondeur ainsi que de mesurer les performances de celles-ci en fonction de la distance. Notre nouveau protocole d'évaluation a permis de mettre en avant la nécessité d'utiliser une méthode dédiée mêlant détection et localisation d'objets. Il a mis en avant les problèmes de précision des méthodes d'estimation de profondeur pour localiser les objets dans l'espace.

Les domaines routiers et ferroviaires sont assez similaires, ce qui permet aux algorithmes entrainés sur le domaine routier d'offrir de bonnes performances lors de l'inférence sur le domaine ferroviaire. Cependant, en raison de l'absence de jeux de données ferroviaires avec une vérité terrain suffisante pour entrainer et évaluer les différentes approches, à l'exception du jeu de données RailSem19 (sans vérité terrain) récemment publié [89], nous n'avons pas été en mesure de tester notre approche dans le domaine ferroviaire. Pour cela, nous proposons d'étendre la version actuelle de notre système d'acquisition de données en incluant un nouveau capteur stéréoscopique et un LiDAR afin de pouvoir collecter notre propre dataset multimodale routier et ferroviaire. C'est ce que nous allons présenter en partie dans le chapitre 3 ainsi que notre nouvelle approche de détection et tracking temps-réel d'objets 3D.

Chapitre 3

Détection et Tracking Temps-Réel d'Objets 3D par Apprentissage Profond

Sommaire

3.1 Introduction

Si le secteur routier est largement couvert par la littérature scientifique, c'est moins le cas pour la smart mobilité ferroviaire. Ceci est en partie dû à l'absence de jeux de données ferroviaires spécifiques avec vérité terrain dédiés à la détection d'objets 3D. Cela représente un véritable défi quant à l'entrainement des algorithmes de détection d'objets 3D. Dans ce chapitre, nous proposons de combler ces limites via deux grandes parties : (*i*) un nouveau jeu de données hybride et multimodale dédié à la détection 3D pour la smart mobilité routière et ferroviaire, (*ii*) une nouvelle méthode CNN légère (L-CNN) pour la détection et tracking temps-réel d'objets 3D. Tout d'abord, nous proposons un nouveau jeu de données comprenant des images prises du point de vue des voitures et des tramways. En prenant en compte à la fois les similarités entre le domaine routier et celui ferroviaire ainsi que la complexité inhérente de l'acquisition de données dans le domaine ferroviaire, nous avons fait le choix de nous orienter vers un jeu de données multimodal routier et ferroviaire. Ce nouveau dataset nous permettra à la fois d'entraîner et valider nos approches dans un environnement ferroviaire.

Ces dernières années, des méthodes basées sur les CNN ont été explorées pour détecter et localiser des objets dans l'espace 3D avec seulement des images. Malgré le nombre important d'approches proposées dans la littérature, dédiées à la détection d'objets 3D, nous constatons un déficit dans leur évaluation particulièrement dans des conditions environnementales réalistes. Nous proposons donc une nouvelle méthode dédiée à la détection et tracking temps-réel d'objets 3D combinant la détection et la localisation 3D et ce, afin d'éviter de devoir cumuler deux approches différentes et ainsi alléger le réseau. Elle est basée sur le détecteur d'objets 2D YOLOv5 [90]. Avec cette nouvelle approche, nous introduisons des boîtes d'ancrage 3D qui font de notre méthode la plus rapide, disponible pour la détection d'objets 3D tout en ayant une précision comparable aux méthodes de pointe.

3.2 Dataset hybride et multimodal pour la smart mobilité

3.2.1 Contexte & objectifs

L'enjeu principal pour le véhicule autonome est d'obtenir des données sur son environnement lui permettant d'adapter son comportement en fonction. Toute approche basée sur le deep learning nécessite un jeu de données de vérité terrain annoté pour effectuer l'apprentissage. La grande majorité est destinée à la conduite autonome en environnement routier et se concentre sur la détection d'objets et la segmentation de scènes. Nous pouvons d'ailleurs les diviser en deux grandes familles [91] : jeux de données avec et sans vérité terrain. Même

si dans les scènes routières et ferroviaires, les objets d'intérêt à détecter sont les mêmes : véhicules, piétons, cyclistes, etc., les datasets routiers ne peuvent pas être utilisés dans le domaine ferroviaire car les images sont prises du point de vue routier et non pas ferroviaire. A notre connaissance, il n'existe aucun dataset réel avec vérité terrain pour le domaine ferroviaire. Nous avons donc développé le "premier" dataset multimodal (routier et ferroviaire) et hybride (virtuel et réel) nous permettant l'entrainement et la validation de l'ensemble de nos modèles deep learning pour la détection d'objets 3D.

3.2.2 Etat de l'art

Dataset Monomodal

Collecter uniquement des images RGB sans vérité terrain est une tâche qui reste relativement raisonnable en temps et coût. Différents datasets monomodal ont été développés ces dernières années [70, 92, 93, 94], qui fournissent des annotations 2D et des étiquettes de segmentation de scènes pour les images RGB. CityScape [70] est un Benchmark de référence et un ensemble de données à grande échelle permettant d'entrainer des approches d'étiquetage sémantique. Il comprend également divers ensembles de séquences stéréo enregistrées dans les rues de 50 villes différentes, $5k$ images avec des annotations de haute qualité et $20k$ images supplémentaires avec des annotations grossières [70]. DriveSeg [95], développé par MIT et Toyota, montre la valeur des informations sur la dynamique temporelle avec 12 classes d'agents pour les tâches de segmentations routières. Nous pouvons également citer d'autres jeux de données dédiés à l'annotation de la détection des piétons tels que Pascal-VOC [71], INRIA dataset [96], TUD-MotionPairs [97] et NightOwls [98].

Dataset Multimodal

Datasets dédiés à la smart mobilité routière. Même si leur développement demeure long et coûteux, les ensembles de données multimodales représentent un avantage crucial dans l'utilisation des approches supervisées offrant une grande variété de vérité terrain provenant de différents capteurs. Le dataset routier le plus connu est sans doute KITTI [59] (Karlsruhe Institute of Technology and Toyota Technological Institute). Il a été le tout premier jeu de données multimodal. Les flux vidéo enregistrés pendant une durée de $6h$ comportent plus de $200k$ annotations d'objets 3D capturés en Allemagne dans différents scénarios et conditions de circulation. Les enregistrements ont été capturés et synchronisés à $10Hz$ grâce à un système d'acquisition équipé de 4 caméras haute résolution, d'un LiDAR Velodyne HDL-64E 3D ($100k$ points par image) et d'un système inertiel de navigation GPS/IMU 3D [59]. KITTI comprend des étiquettes de suivi d'objets 3D pour différentes classes d'objets (voitures, camions, tramways, piétons, cyclistes).

NuScenes [81] publié en mars 2019, est un jeu de données public à grande échelle pour la conduite autonome. NuScenes est plus récent que KITTI qui l'a d'ailleurs inspiré. Il comprend $1k$ scènes de conduite ($20s$ de durée chacune) à Boston et Singapour, villes connues pour leur trafic dense. Il contient 23 classes d'objets annotés avec des bounding boxes (boites englobantes) 3D précises à $2Hz$, ce qui constitue une bonne annotation pour la détection d'objets, l'estimation de la distance et le tracking. NuScenes comprend environ 1.4 million d'images caméra, $390k$ balayages LiDAR, 1.4 million de balayages RADAR et 1.4 million de bounding boxes d'objets. Il comprend un LiDAR, 5 RADAR, 6 caméras et une unité GPS/IMU [81]. ROAD [91] event Awareness Dataset for Autonomous Driving (ROAD), comprend 22 vidéos (8mn chacune) avec $122k$ de frames annotées, et $560k$ bounding boxes de détection avec $1,7M$ étiquettes individuelles. Le processus d'annotation suit une approche multi-label dans laquelle les agents de la route (véhicules, piétons, etc.), leur emplacement ainsi que les actions qu'ils effectuent sont identifiés [91].

Argoverse [99] comprend des annotations pour des images RGB, LiDAR/RADAR, et des bounding boxes 3D avec des informations de suivi pour 15 objets d'intérêt [99]. Lyft [100] utilise des caméras et LiDAR pour fournir des annotations 2D et des bounding boxes 3D pour 4 classes d'objets. Le jeu de données ouvert Waymo [101] est un nouveau dataset collecté via des LiDARs et des caméras à grande échelle. Il comprend 1150 scènes (chacune de $20s$). TITAN [102] est un dataset regroupant 700 clips vidéo bruts de 10 à 20 secondes étiquetés (avec odométrie) capturés dans des scènes de trafic urbain à Tokyo. Au total, TITAN comprend 10 heures d'enregistrements à 60 FPS, 50 étiquettes, capturés avec une camera GoProHero7 et une IMU qui enregistre des données odométriques synchronisées à $100Hz$. Enfin, il existe des jeux de données d'images virtuelles créés avec des simulateurs tels que CARLA [103] ou SYNTHIA [104]. Elles ont l'avantage d'être simples à acquérir et permettent donc d'avoir des données vérité terrain de masse. Cependant, ces jeux de données ont des graphismes datés, ce qui rend difficile leur utilisation en conditions réelles. Pour pallier à ce problème, des travaux récents ont porté sur l'acquisition d'un dataset via un jeu vidéo [105]. Par exemple le jeu vidéo GTAV [106] est utilisé pour acquérir des ensembles de données routiers similaires à KITTI ou CityScape.

Datasets dédiés à la smart mobilité ferroviaire. Bien que le nombre de datasets publics routiers ait augmenté au cours des dernières années, les applications ferroviaires ne sont pas autant développées. Il y a un manque notable de travaux concernant les trains/tramways autonomes. A notre connaissance, il n'existe que quelques jeux de données images dédiés à la smart mobilité ferroviaire [89, 107]. Cependant, ces derniers ne sont pas dédiés à la détec-

tion 3D et ne disposent pas de vérité terrain sur la profondeur ce qui limite leur utilisation dans des applications de smart mobilité réelles. Le dataset RailSem19 [89] est le premier jeu publié à proposer des données ferroviaires. Au final, plus de 8500 scènes, 350 heures de trafic ferroviaire et de tramway ont été collectées dans 38 pays différents et dans des conditions météorologiques variées. Parmi ses limites l'absence de vérité terrain et donc pas d'informations de profondeur des objets. Pour pallier à cette problématique, nous avons décidé de développer notre propre dataset hybride et multimodal dédié à la smart mobilité routière et ferroviaire. Il comprend deux ensembles de données : (*i*) dataset multimodal virtuel basé sur le jeux vidéo GTAV et (*ii*) dataset multimodal réel comprenant des scènes routières et ferroviaires collectées dans deux villes Normandes, Rouen et Le Havre.

3.3 Dataset virtuel multimodal routier et ferroviaire

3.3.1 Contexte & objectifs

Le principal avantage d'un jeu de données virtuel est la simplification de son acquisition et de son annotation. Cependant, cela se fait au détriment de la fidélité par rapport au monde réel. En effet, la plupart de ces jeux de données provenant de simulateurs (e.g. CARLA [103] ou encore SYNTHIA [104]) n'ont pas une fidélité graphique suffisante pour permettre aux méthodes entraînées sur ceux-ci d'offrir de bonnes performances. Pour surmonter ce problème, nous nous tournons vers les jeux vidéo comme Grand Theft Auto V (GTAV), comprenant des scènes routières et ferroviaires réalistes (i.e. passages à niveau, passages piétons, trafic intense, etc.). De nombreux travaux ont été réalisés dans ce sens pour construire des datasets routiers [105, 108]. Notre jeu de données ferroviaire a été acquis avec un pipeline d'acquisition de données basé sur les travaux de [106, 109]. Nous avons utilisé une version modifiée du code du jeu [110] permettant la conduite à un utilisateur. Cela a permis d'avoir un jeu de données hybride avec des scènes routières et ferroviaires.

3.3.2 Préparation de la vérité terrain

Méthode d'acquisition. Les données extraites via GTAV incluent celles des entités présentes dans l'environnement proche comme la classe de l'entité (personne, type de véhicule, etc.), leurs positions, les matrices de projections, les conditions météorologiques, la position de la caméra, etc. Nous avons modifié le jeu afin de l'adapter à nos besoins comme par exemple l'ajout d'une deuxième caméra stéréoscopique, l'usage d'un pilote automatique pour contrôler notre véhicule ou encore le pilotage de trains [110]. Toutes les données sont ensuite stockées dans une base de données SQL avec un post-traitement pour les adapter aux approches de détection d'objets 3D. Les données sélectionnées pour l'entrainement de

nos modèles sont structurées de la manière suivante : 1. classes des objets, 2. position et dimensions des boites englobantes 2D des objets présents sur l'image, 3. orientation des objets et 4. positions et dimensions des objets (exprimées en mètres).

Vérité terrain. Une fois les données sont collectées, il est nécessaire d'effectuer un traitement supplémentaire afin de créer une vérité terrain qui soit "exploitable" pour les approches d'apprentissage profond. En effet, le jeu GTAV ne permet pas d'obtenir les matrices intrinsèques et extrinsèques nécessaires à la reprojection du repère monde au repère image en pixels. Nous avons développé notre propre approche de transformation : coordonnées monde $->$ (matrice de vue) $->$ coordonnée caméra $->$ (matrice de projection) $->$ NDC (*Normalized Device Coordinate*) $->$ (redimensionnement) $->$ pixels dans l'image. En appliquant l'équation 3.1 :

$$\begin{pmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \\ 1 \end{pmatrix} = \begin{pmatrix} matrice\ de\ vue \\ (4 \times 4) \end{pmatrix} \begin{pmatrix} X_{monde} \\ Y_{monde} \\ Z_{monde} \\ 1 \end{pmatrix} \tag{3.1}$$

Nous obtenons les coordonnées dans le repère caméra à partir des coordonnées du repère monde. La matrice de vue correspond à une matrice extrinsèque entre le repère monde et le repère caméra. Les coordonnées en pixels peuvent ensuite être calculées via l'équation 3.2 :

$$\begin{pmatrix} x_{pixels} \cdot k \\ y_{pixels} \cdot k \\ k \end{pmatrix} = \begin{pmatrix} w & 0 & 0 & w \\ 0 & -h & 0 & h \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} matrice\ de\ proj \\ (4 \times 4) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \\ 1 \end{pmatrix} \tag{3.2}$$

Avec (w, h) la largeur et la hauteur de l'image en pixels. Afin de trouver l'orientation des objets selon l'axe Y (azimut), il est nécessaire de constituer la matrice extrinsèque entre l'objet et la caméra. À partir de cette matrice, nous pouvons extraire la matrice de rotation et donc par extension l'azimut. Pour restituer cette matrice, nous utilisons les équations 3.3 et 3.4 :

$$\begin{pmatrix} \mathbf{T}_{obj->cam} \\ (4 \times 4) \end{pmatrix} = \begin{pmatrix} matrice\ de\ vue \\ (4 \times 4) \end{pmatrix} \begin{pmatrix} \mathbf{T}_{monde->obj} \\ (4 \times 4) \end{pmatrix}^{-1} \tag{3.3}$$

$$\begin{pmatrix} \mathbf{T}_{obj->cam} \\ (4 \times 4) \end{pmatrix} = \begin{pmatrix} \mathbf{R} & & \mathbf{t} \\ (3 \times 3) & & (3 \times 1) \\ & & \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.4}$$

GTAV nous fournit directement le vecteur de rotation et de translation des objets dans le repère monde, ce qui nous permet de créer la matrice de transformation $\mathbf{T}_{monde->obj}$ entre le repère monde et le repère objet. $\mathbf{R}$ représente la matrice de rotation permettant l'obtention de l'azimut de l'objet dans le repère caméra et $\mathbf{t}$ est la position de l'objet dans ce repère.

3.3.3 Architecture du jeu de données virtuel

Notre dataset comporte $220k$ images stéréoscopiques (soit $110k$ échantillons) divisées en 72 séquences. Parmi ces séquences, 13 sont prises du point de vue tramway et 59 sont prises du point de vue voiture. Les environnements dans lesquels ont été menées les acquisitions varient entre urbain, campagne et autoroute. GTAV comprend 5 classes différentes (voiture, personne, camion, moto et vélo) avec une répartition 80% routiers et 20% ferroviaires. Le jeu comprend aussi des scènes de nuit, de jour, de pluie et de temps clair, offrant une grande variété de conditions de visibilité pour la smart mobilité routière et ferroviaire. La vérité terrain est générée automatiquement par l'API du jeu et comprend entre autres les boîtes englobantes 3D et 2D, la classe de l'objet, la carte de profondeur et la carte sémantique. Toutes ces annotations sont prises du point de vue d'une caméra stéréoscopique ($10Hz$). Cependant, pour combiner notre jeu de données virtuel avec le jeu de données KITTI, nous avons sélectionné uniquement les classes qui sont également présentes dans KITTI (voiture, personne et cycliste).

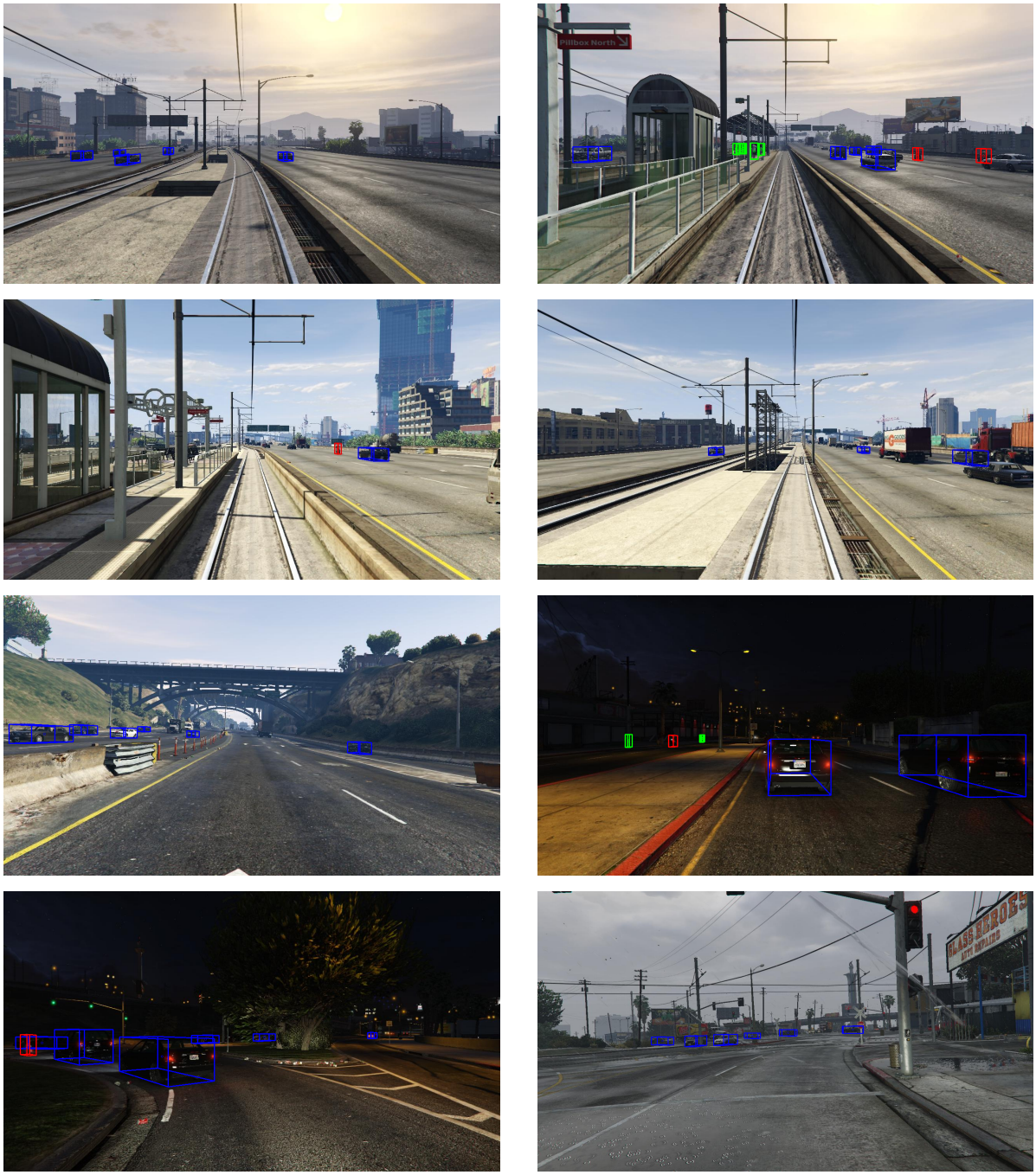

FIGURE 3.1 – Images avec vérités terrains en 3D. Notre jeu de données présente une variété d'environnements et de conditions (temps clair, pluie, jours, nuit, route, rails, etc.). Les vérités terrains présentées correspondent aux mêmes classes que KITTI : voiture, personne et vélo.

3.4 Dataset réel multimodal routier et ferroviaire

3.4.1 Contexte & objectifs

Afin de combler le manque de jeux de données d'images réelles pour le domaine ferroviaire, et pour étoffer davantage notre jeu de données virtuel avec des données réelles, nous avons entrepris l'acquisition et l'annotation de notre propre jeu de données réel. Ce nouveau dataset appelé ESRORAD (ESIGELEC and SEGULA technologies ROad and RAilway Dataset) [111], a pour objectif non seulement de tester et valider nos approches mais aussi de mettre à disposition des chercheurs et industriels un dataset ouvert leur permettant d'entrainer et valider leurs modèles. ESRORAD est le premier jeu de données comprenant des acquisitions, provenant d'environnements ferroviaires, dédiées à la détection d'objets 3D. Cela est particulièrement important car, malgré les nombreuses similarités entre les domaines routiers et ferroviaires, il nous est impossible actuellement de confirmer la validité d'une méthode de détection 3D sur le domaine ferroviaire et ce, à cause du manque de jeu de données dédiés.

Notre objectif est de développer un dataset hybride (virtuel et réel) et multimodal (routier et ferroviaire). Nous avons entrepris l'acquisition de $80k$ échantillons avec une création de vérité terrain progressive jusqu'à obtenir un jeu de données avec quelques milliers d'échantillons annotés. Ces derniers sont constitués d'acquisitions provenant d'une caméra stéréoscopique (images RGB) et de balayages provenant d'un LiDAR (nuage de points 3D). Ces données nous permettrons par la suite de créer une vérité terrain dédiée à la détection d'objets 3D.

3.4.2 Architecture du système d'acquisition

Les principales difficultés liées à ces travaux sont dues aux techniques requises comme la mise en place du système d'acquisition avec le calibrage des capteurs ou encore la synchronisation entre la caméra et le LiDAR. Afin de réaliser ces acquisitions dans les conditions de circulations les plus réalistes possibles, nous avons opté pour l'utilisation de notre véhicule de test IRSEEM (e.g. figure 3.2) comprenant les dispositifs suivants :

- Caméras stéréoscopiques Intel RealSense (L515)
- GPS AsteRx (septentrio)
- Unité de mesure inertielle (IMU) LANDYN (IXblue) composée d'un accéléromètre, un magnétomètre et un gyroscope
- Odomètre instrumenté sur la roue arrière droite
- LiDAR VLP16 de type Velodyne synchronisé avec le GPS
- Un système d'acquisition de données temps-réel RTMAPS (Intempora)

FIGURE 3.2 – Véhicule IRSEEM instrumenté pour la collecte de données. Le coffre sur le toit du véhicule comprend l'ensemble de l'instrumentation (Caméras, GPS/IMU, LiDAR, RADAR, Odomètre).

3.4.3 Calibrage et synchronisation du système d'acquisition

Calibrage du nouveau capteur. L'opération de calibrage de caméra correspond à déterminer la relation entre les coordonnées spatiales d'un point de l'espace avec le point associé dans l'image. Nous utilisons le modèle sténopé qui modélise une caméra par une projection perspective. La première étape du calibrage représente la transformation entre le repère monde et celui de la caméra. Cette transformation peut se décomposer en une rotation R et une translation T formant les paramètres extrinsèques de la caméra. Le calibrage des caméras est effectué via un motif damier permettant de déterminer les paramètres intrinsèques et de distorsion. Les paramètres extrinsèques sont déterminés en utilisant la fonction LiDAR de la caméra. L'objectif étant de superposer les scans LiDAR du véhicule avec les scans LiDAR de la caméra L515 pour obtenir la matrice de transformation entre les deux repères.

Afin d'obtenir un bon alignement des données intermodales entre le LiDAR et les caméras stéréoscopiques, l'exposition d'une caméra est déclenchée lorsque le LiDAR supérieur balaie le centre du champ visuel de la caméra. L'horodatage de l'image représente l'heure de déclenchement de l'exposition, et l'horodatage du balayage LiDAR représente l'heure à laquelle la rotation complète de l'image LiDAR actuelle est atteinte. Comme les caméras fonctionnent à $12Hz$ et le LiDAR à $20Hz$, il est essentiel de faire correspondre un scan LiDAR à une

image en identifiant les moments de capture.

Projection des points LiDAR sur l'image 2D. Afin de vérifier que les matrices de passage obtenues soient correctes, la matrice de transformation est appliquée sur les images et les nuages de points. Cela permet d'afficher chaque point LiDAR sur l'image (i.e. équation (3.5)). Soit (X_{LiDAR},Y_{LiDAR},Z_{LiDAR}) les coordonnées d'un point dans le repère LiDAR et $\mathbf{T}_{LiDAR->cam}$ la matrice extrinsèque entre le LiDAR et la caméra. Les coordonnées de ce même point dans le repère de la caméra (X_{cam},Y_{cam},Z_{cam}) sont obtenues via l'équation (3.5) :

$$\begin{pmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{LiDAR->cam} \\ (4 \times 4) \end{pmatrix} \begin{pmatrix} X_{LiDAR} \\ Y_{LiDAR} \\ Z_{LiDAR} \\ 1 \end{pmatrix} \tag{3.5}$$

Ce point doit ensuite être projeté grâce à la matrice intrinsèque de la caméra afin d'obtenir ses coordonnées en pixels. Avec (u,v) les coordonnées en pixels du point, (f_x, f_y) la focale horizontale et verticale de la caméra et enfin (c_x, c_y) le centre optique, la projection est détaillée dans l'équation (3.6) :

$$\begin{pmatrix} u \times Z_{cam} \\ v \times Z_{cam} \\ Z_{cam} \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_{cam} \\ Y_{cam} \\ Z_{cam} \end{pmatrix} \tag{3.6}$$

Synchronisation Caméras-LiDAR. Lors du précédent calibrage, la projection, en statique, des points LiDAR sur l'image 2D ne posait pas de problèmes vu que le véhicule était à l'arrêt, ainsi, même si le LiDAR et la caméra fonctionnaient à des fréquences différentes, la scène demeurait la même. En dynamique, ce décalage de déclanchement est plus problématique. En effet, pour réaliser la projection sur des scènes en mouvement, il fallait associer chaque scan LiDAR à l'image correspondante qui avait été capturée au même moment. Nous avons donc élaboré un "nouveau protocole" de synchronisation des données caméras-LiDAR.

3.4.4 Processus d'annotation des données

Choix des itinéraires et collecte des données. Afin de réaliser la collecte de données, nous avons planifié un protocole d'enregistrement prenant en compte deux agglomérations normandes : Rouen et Le Havre. Elles disposent d'un réseau de trafic routier et ferroviaire important comprenant non seulement des routes et de voies de tramways, mais aussi des zones dont la circulation des voitures sur la voie ferroviaire est autorisée. Cela nous a permis de faire des acquisitions de scènes routières et/ou ferroviaires avec un point de vue direct

sur les rails et ce, sans interruptions de la circulation du tramway. Nous avons pu collecter environ $100k$ images réparties entre Le Havre ($45k$ d'images) et Rouen ($55k$ d'images).

Processus d'annotation des données. L'annotation consistait en la documentation des images avec la vérité terrain. L'objectif était d'identifier, pour chaque image, nos trois classes d'objets (véhicule, piéton, cycliste) ainsi que la distance de chaque objet par rapport au véhicule. Toutes ces informations constituent la vérité terrain du jeu de données réel. Nous avons donc réalisé un benchmarking des outils d'annotation existants dans la littérature dont Supervisely [112] et Scale [113]. Nous avons utilisé 3D Bounding Box Annotation Tool (BAT 3D) [114], un nouveau système ouvert consacré à l'annotation d'ensembles de données 2D et 3D que nous avons modifié. L'annotation consistait d'afficher les nuages de points LiDAR et de placer manuellement des boîtes 3D correspondantes à la classe de l'objet identifié, puis d'afficher le résultat sur l'image couleur. La Figure 3.3 illustre quelques exemples d'annotation d'images ainsi que des résultats de prédiction de notre algorithme de détection d'objets entrainé sur le dataset virtuel GTAV. Ces travaux de recherche ont fait l'objet de plusieurs publications dans [5], [10] et [115].

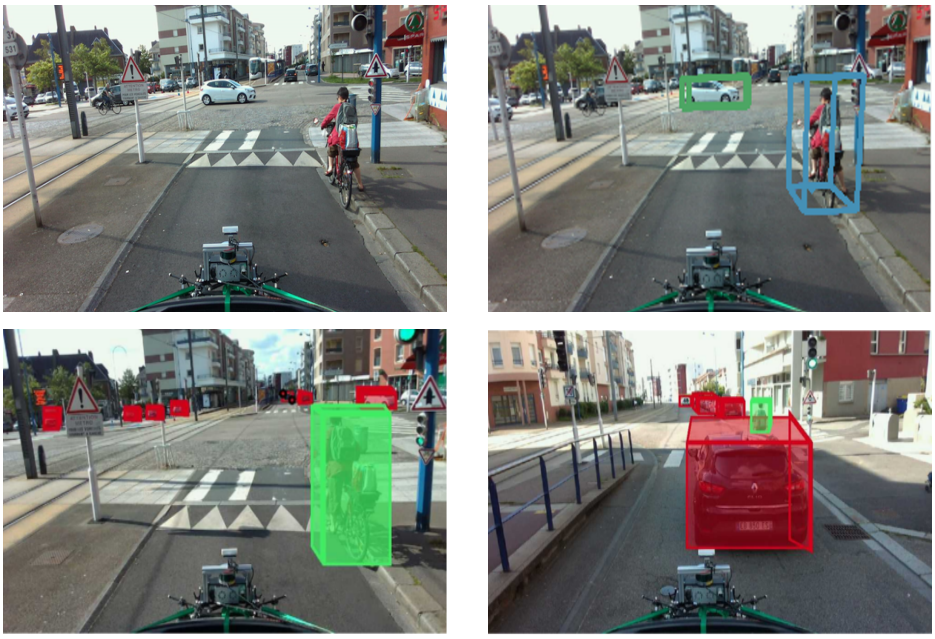

FIGURE 3.3 – Exemples d'annotations et de prédictions effectuées sur le dataset réel. Haut à gauche : image sans annotation, haut à droite : image avec annotation, les deux images en bas : prédictions de notre algorithme de détection d'objets entrainé sur GTAV.

3.4.5 Résultats expérimentaux

ESRORAD représente une nouveauté par rapport aux jeux de données de pointe et ce, pour diverses raisons. Il est actuellement le seul jeu de données, à notre connaissance, dédié non seulement à la smart mobilité routière mais aussi ferroviaire (train autonome). Il

combine des scènes routières (80%) et ferroviaires (20%) avec des données virtuelles ($220K$ images) et réelles ($100k$ images). Nous avons, à ce jour, annoté $2,5k$ images avec BAT-3D. La vérité terrain obtenue à l'issue du processus d'annotation comprend la classe de l'objet, sa position vis-à-vis du LiDAR (XYZ en mètres), ses dimensions (en mètres) et son orientation. Ces annotations fournies par BAT-3D sont relatives au données LiDAR, pour qu'elles soient exploitées par nos approches de détection d'objet 3D, il est nécessaire de les transformer dans les repères caméra/image. Nous pouvons obtenir la position de l'objet dans le repère caméra en utilisant la matrice extrinsèque, entre le LiDAR et la caméra, ainsi que l'équation (3.5). De la même façon, nous pouvons obtenir la boite englobante 2D de l'objet en re-projetant sa boite 3D dans le repère caméra via l'équation (3.6). L'azimut de l'objet relatif au repère caméra peut être calculé à partir de la matrice de transition entre le repère objet et le repère caméra. Cette matrice est obtenue par les équations (3.7) et (3.8) :

$$\begin{pmatrix} \mathbf{T}_{obj->cam} \\ (4 \times 4) \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{lidar->cam} \\ (4 \times 4) \end{pmatrix} \begin{pmatrix} \mathbf{T}_{obj->lidar} \\ (4 \times 4) \end{pmatrix} \tag{3.7}$$

$$\begin{pmatrix} \mathbf{T}_{obj->cam} \\ (4 \times 4) \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{t} \\ (3 \times 3) & (3 \times 1) \\ 0 \quad 0 & 0 \quad 1 \end{pmatrix} \tag{3.8}$$

Avec $\mathbf{R}$ la matrice de rotation entre l'objet et la caméra nous permettons de calculer l'azimut et $\mathbf{t}$ la matrice de translation. $\mathbf{T}_{lidar->cam}$ étant la matrice de transition entre le LiDAR et la caméra et $\mathbf{T}_{obj->lidar}$ celle entre l'objet et le LiDAR. La figure 3.4 illustre quelques exemples d'annotations d'images de notre jeu de données ESRORAD ainsi que des résultats de prédiction de notre algorithme léger de détection d'objets 3D entraîné sur le jeu de données NuScenes [81].

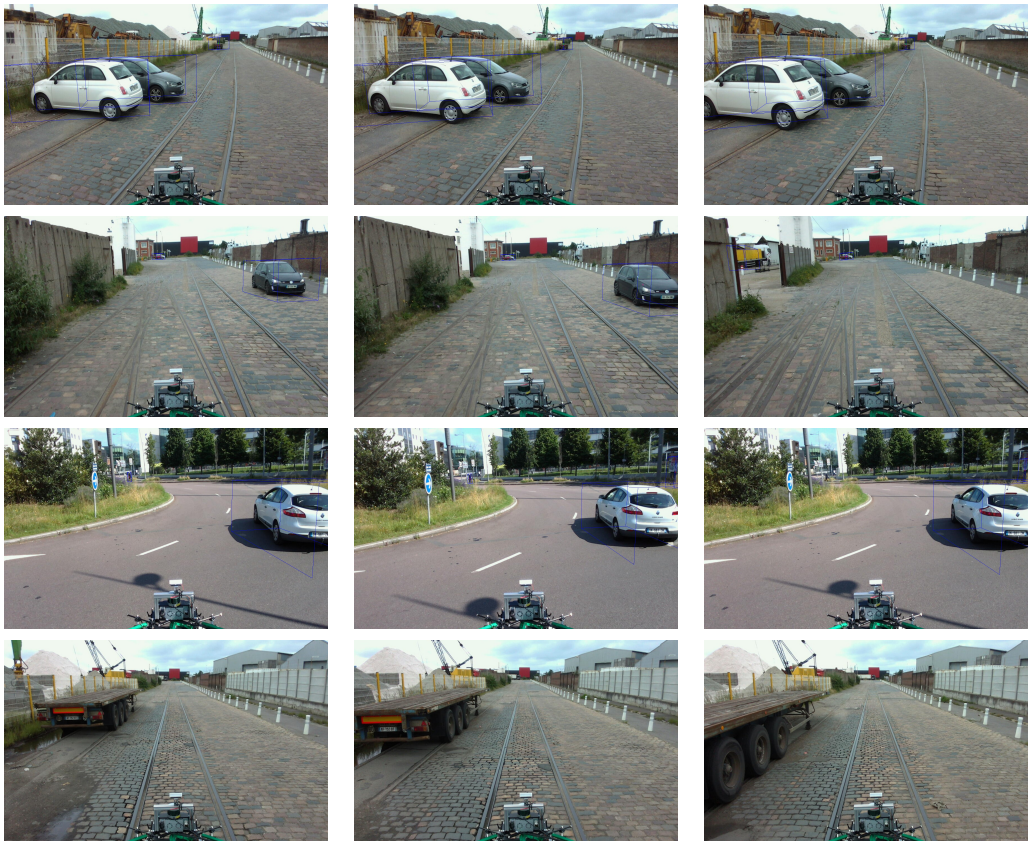

FIGURE 3.4 – Exemples d'annotations et de prédictions sur ESRORAD. Résultats obtenus par notre algorithme de détection d'objets 3D [115] léger et entraîné pendant 300 époques sur NuScenes [81].

3.5 Détection et tracking temps-réel d'objets 3D par apprentissage profond

3.5.1 Contexte & objectifs

Dans cette deuxième partie du chapitre, nous proposons une nouvelle méthode dédiée à la détection temps-réel d'objets 3D intégrant la détection et la localisation 3D et ce, afin d'éviter de devoir combiner les deux approches différentes et ainsi alléger le réseau. Ces dernières années, des méthodes basées sur les réseaux CNN ont été explorées pour détecter et localiser des objets dans l'espace 3D avec seulement des images comme entrées. Malgré leur nombre important, nous constatons un déficit quant à leur évaluation particulièrement dans le trafic ferroviaire. Notre nouvelle approche de détection temps-réel d'objets 3D comprend plusieurs prédictions : 1. position de l'objets dans l'image, 2. classe de l'objet (détection 2D), 3. distance de l'objet, 4. centre 3D de l'objet, 5. dimension de l'objet en 3D et 6. orientation de l'objet. Grâce à ces prédictions, il est possible de dessiner des boîtes englobantes 3D pour

chacun des objets détectés dans l'image et de les localiser dans le repère caméra. Il s'agit donc d'un nouvel algorithme de détection et tracking temps-réel d'objets 3D basé sur des méthodes de détection d'objets éprouvées (e.g. YOLOv5 [90]). Alors que les autres méthodes de l'état de l'art se concentrent principalement sur la précision de la détection 3D, notre approche est monoculaire et prend en compte deux critères importants, la précision et la vitesse.

Pour l'entrainement de notre approche, nous utiliserons notre nouveau jeu de données ESRO-RAD, présenté dans la section précédente, ainsi que les jeux de données routiers KITTI [59] et NuScenes [81]. Nous montrons que notre méthode est la plus rapide pour la détection temps-réel d'objets 3D, avec une précision proche de celle des méthodes les plus avancées sur KITTI. Nous démontrons également que le processus de pré-entraînement sur notre jeu de données virtuel GTAV améliore la précision sur les jeux de données réels tels que KITTI, permettant ainsi à notre méthode d'obtenir une précision encore meilleure que celles des approches de l'état de l'art.

3.5.2 Etat de l'art

Ces dernières années, de nombreuses méthodes de détection d'objets 3D basées sur l'apprentissage profond ont été proposées. Plusieurs d'entre elles sont des extensions d'approches de détection 2D, tirant ainsi parti de leurs très bonnes performances. Le traitement des paramètres 3D est ensuite ajouté à ces modèles, tandis que d'autres infèrent directement ces paramètres par un apprentissage de bout en bout. Dans [116], le problème est abordé en deux étapes. D'abord, des boîtes 3D candidates sont générées et notées en exploitant plusieurs caractéristiques, ensuite, les candidats ayant obtenu les meilleurs scores sont soumis à un détecteur d'objets modifié, Faster-RCNN [64], pour prédire les classes d'objets, les décalages des boîtes englobantes et l'orientation des objets. Dans [117], la détection 2D est effectuée par la proposition de régions multi-échelle MS-CNN [118], étendue pour régresser l'orientation et les dimensions des boîtes 3D. L'estimation de l'orientation utilise un principe MultiBin où le réseau prédit d'abord dans quel intervalle se trouve l'angle (classification), avant de prédire la déviation entre l'angle recherché et le centre de l'intervalle choisi (régression).

Dans DeepMANTA [119], les auteurs présentent un modèle comprenant la détection, la localisation, la caractérisation de la visibilité et l'estimation des dimensions 3D simultanées des véhicules. Il comprend deux étapes. La première étape produit des boîtes englobantes associées aux informations véhicules, tandis que la deuxième utilise ces sorties et un ensemble de données virtuelles de véhicules 3D pour récupérer les orientations et les localisations 3D. Dans [120], une carte de profondeur obtenue via un réseau entièrement connecté (FCN) est

fusionnée avec l'image d'entrée avant d'alimenter un réseau de proposition de région (RPN). Dans [108], un modèle CNN pour la détection et le suivi conjoints d'objets 3D est proposé. Il exploite un RCNN plus rapide pour prédire les boîtes englobantes 2D, qu'il associe par la suite à des informations 3D. Deux couches LSTM (Long Short-Term Memory) assurent le suivi à travers des séquences d'images, permettant un raffinement supplémentaire des positions des boîtes de délimitation 3D. Dans SMOKE [121], un CNN est entraîné de bout en bout en une seule étape au moyen d'une fonction de perte unifiée. Les paramètres des boîtes englobantes 3D sont régressés directement sans utiliser la détection 2D et ce, grâce à un réseau composé de deux branches de classification et de régression des paramètres 3D.

Dans MonoGRNet [122], la détection d'objets 3D est décomposée en quatre sous-tâches exécutées en parallèle puis combinées : détection d'objets 2D, estimation de la profondeur du centre de l'objet, estimation du centre 3D projeté et régression des coins locaux. M3D-RPN [123] exploite la convolution en fonction de la profondeur pour localiser les caractéristiques spécifiques et améliorer ainsi la compréhension de la scène 3D. GAM3D [124] utilise également des boîtes d'ancrage 3D/2D pour prédire la boîte de délimitation 3D d'objets et introduit un module de convolution, sensible à la profondeur, censé améliorer la précision du réseau. Dans [125], la détection 3D est traitée sous forme d'un problème de détection de points clés. Ainsi, un réseau pyramidal de caractéristiques de points clés (KFPN) est défini, fournissant le centre et les points de perspective des boîtes de délimitation 3D. Il permet un traitement temps-réel vu la nature sans ancrage de son détecteur ainsi le fait que l'épine dorsale du réseau soit basée sur une architecture légère (comme ResNet-18 [126] ou DLA-34 [127]).

La plupart de ces modèles sont destinés à améliorer les performances de la détection d'objets 3D en termes de précision. D'autres se concentrent davantage sur la vitesse, visant ainsi le temps-réel sur des GPU puissants. Depuis plusieurs années, nous nous concentrons sur des architectures CNN légères qui soient compatibles avec les contraintes temps-réel embarquées. Nous pensons qu'il est primordial d'aller plus loin dans cette voie pour faciliter le développement à grande échelle de solutions d'apprentissage profond pour la smart mobilité. Outre la vitesse, nous visons aussi le développement de modèles performants aussi bien en environnement routier que ferroviaire.

3.6 Détection d'objet 3D par approche multi-étages

3.6.1 Vue d'ensemble

Contrairement aux autres méthodes de détection 3D avec des images monoculaires, qui s'appuient sur un réseau RPN (Region Proposal Network) distinct, tel que Faster RCNN, notre méthode est basée sur le détecteur à étage unique YOLOv3 [53] pour effectuer les prédictions des boîtes de délimitation 2D. Elle partage la même base du réseau YOLOv3 afin d'extraire les caractéristiques de l'image réduisant ainsi le temps de calcul. Notre architecture permet de réduire considérablement la consommation de mémoire, tout en surpassant les autres méthodes de l'état de l'art en termes de vitesse et ce, au prix d'une précision inférieure. Notre méthode est entraînée de bout en bout, alors que d'autres modèles exigent que le détecteur 2D soit entraîné séparément, ce qui réduit le temps et le coût d'entraînement de notre approche. La figure 3.5 montre une vue d'ensemble de notre méthode de détection multi-objets 3D.

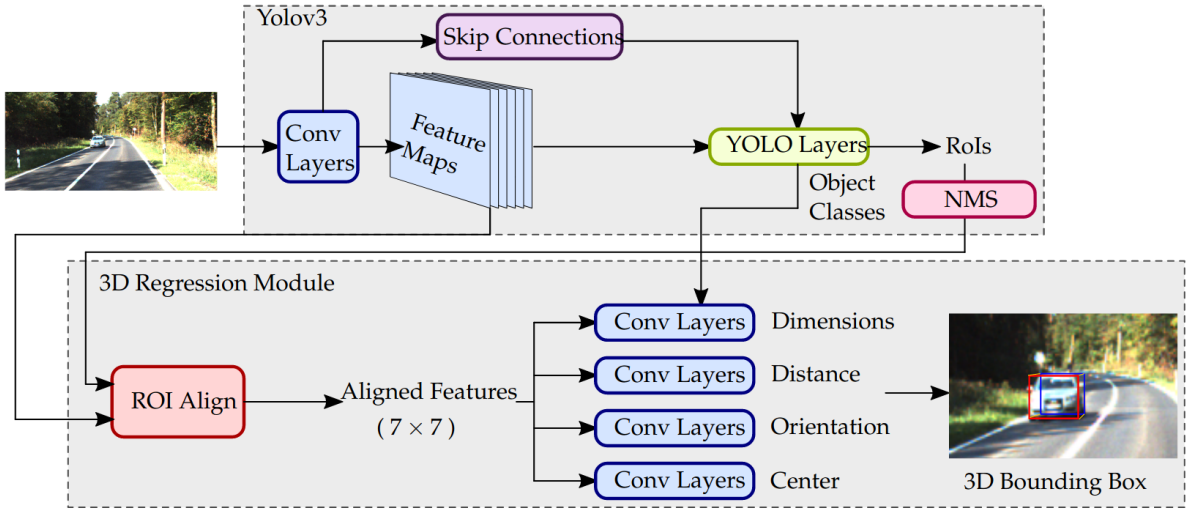

FIGURE 3.5 – Vue d'ensemble de notre méthode de détection multi-objets 3D. Une seule image RGB est utilisée comme entrée. Les caractéristiques convolutionnelles partagées sont ensuite extraites par la base du réseau. Nous extrayons ensuite les caractéristiques des RoIs en utilisant l'alignement des caractéristiques utilisé dans [128]. Les paramètres de la boîte de délimitation 3D sont prédits, et enfin la boîte de délimitation 3D est dessinée sur l'image.

3.6.2 Estimation de la boite englobante 3D

La boîte englobante 3D d'un objet peut être décrite par plusieurs éléments : la position de son centre 3D par rapport à la caméra $\mathbf{T} = [x\ y\ z]^\top$, ses dimensions $\mathbf{D} = [w, h, l]$ (avec $\mathbf{w}, \mathbf{h}, \mathbf{l}$ respectivement largeur, hauteur et longueur) et son orientation $\mathbf{R}(\phi, \theta, \psi)$ (avec ϕ, θ,

ψ respectivement les angles d'élévation (tangage), d'azimut (lacet) et de roulis). Étant donné $\mathbf{K}$ la matrice des paramètres intrinsèques de la caméra et
$\mathbf{X_o} = [x_o\ y_o\ z_o\ 1]^\top$ un point 3D dans le repère objet, la projection de ce point dans le plan image $\mathbf{x_{im}} = [u\ v\ 1]^\top$ est donnée par l'équation 3.9 :

$$\mathbf{x_{im}} = \mathbf{K}.\begin{bmatrix}\mathbf{R} & \mathbf{T}\end{bmatrix}.\mathbf{X_o}. \tag{3.9}$$

Où $[RT]$ représente la matrice extrinsèque avec R (matrice 3×3) la rotation et T (3×1) la translation. En considérant que le centre de la boîte englobante 3D est l'origine des coordonnées de l'objet, les coordonnées de la boîte englobante 3D sont : $\mathbf{X_1} = \begin{bmatrix}w/2 & h/2 & l/2\end{bmatrix}$, $\mathbf{X_2} = \begin{bmatrix}w/2 & -h/2 & l/2\end{bmatrix}$, ..., $\mathbf{X_8} = \begin{bmatrix}-w/2 & -h/2 & -l/2\end{bmatrix}$. Les coordonnées de la boîte englobante 3D dans l'image peuvent alors être obtenues en utilisant l'équation (3.9).

3.6.3 Paramètres prédits

Afin de mieux comparer les performances des approches multi-étapes avec celles des approches simples de détection d'objets 3D, nous prédisons des paramètres similaires à ceux présentés dans notre approche multi-étapes précédente [5].

Détection d'objets 2D. Pour effectuer la détection d'objets 2D, notre approche fait appel à YOLOv3 effectuant la prédiction des classes d'objets cls ainsi que les paramètres de la boite englobante $\mathbf{b}$ (position et dimension). Les prédictions de la boîte englobante sont ensuite utilisées comme RoIs pour l'alignement des caractéristiques du réseau (Feature Align) afin d'extraire celles de chaque RoI. Elles sont ensuite transmises au reste du réseau pour prédire les paramètres de la boîte englobante 3D.

Estimation de la distance de l'objet. Afin de déterminer le centre de la boîte englobante 3D, il est nécessaire de déterminer sa position sur l'axe Z des coordonnées de la caméra. Pour chaque RoI du détecteur d'objet, nous prédisons la distance du centre de l'objet $\tilde{z_o}$.

Prédiction du centre de l'objet. Nous supposons que le centre de la boîte englobante 3D est le centre 3D de l'objet. Pour prédire donc son centre il suffit de prédire sa position : $\mathbf{X_o} = [x_o\ y_o\ z_o\ 1]^\top$. Afin d'augmenter la précision de cette prédiction, nous cherchons à prédire le centre 3D projeté sur le plan image. Au lieu de prédire directement les coordonnées du centre de l'objet sur le plan image, nous utilisons les indices de la prédiction de la boîte englobante 2D et nous prédisons la position décalée en pixels, du centre de l'objet par rapport au centre de la boîte englobante 2D $\tilde{\mathbf{c}} = \begin{bmatrix}\tilde{cx} & \tilde{cy}\end{bmatrix}$. Nous réduisons donc la variance de la prédiction, ce qui facilite l'apprentissage de l'algorithme. Le centre de l'objet $\mathbf{X_o}$ peut ensuite être calculé en utilisant l'estimation de la distance de l'objet sur l'axe Z ainsi que la matrice de calibration inversée $\mathbf{K^{-1}}$ comme décrit dans l'équation (3.10) :

$$\begin{bmatrix} x_o \\ y_o \\ z_o \\ 1 \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} \widetilde{cx} \times z_o \\ \widetilde{cy} \times z_o \\ z_o \\ 1 \end{bmatrix} \tag{3.10}$$

Dimensions d'objets. Les dimensions d'objets ont une très faible variance au sein d'une même classe (voiture, camion, etc.). Par conséquent, au lieu de prédire directement les dimensions de l'objet, nous utilisons les dimensions moyennes de chaque classe d'objets comme une forte priorité pour l'estimation des dimensions.

Orientation de l'objet. L'orientation d'un véhicule est en général caractérisée par trois paramètres : tangage, roulis et lacet. Nous supposons que seul le lacet de l'objet (noté θ) importe pour la smart mobilité routière (tangage $\phi = 0$ et roulis $\psi = 0$). Or observé du point de vue caméra, un lacet peut conduire à plusieurs orientations (voir figure 3.6), ce qui ne permet pas de prédire directement l'angle θ. Nous prédisons donc l'angle observé α et récupérons l'orientation globale θ en utilisant l'équation (3.11) :

$$\theta = \alpha + arctan(\frac{x}{z}). \tag{3.11}$$

En suivant les travaux décrits dans [117], au lieu de considérer la prédiction de l'angle comme un problème de régression, nous adoptons une approche hybride classification/régression. Nous divisons les angles possibles en 2 intervalles et nous effectuons ensuite une tâche de classification pour prédire dans quel intervalle se trouve l'angle de l'objet. Ensuite, nous effectuons une régression de la différence entre le centre du "bin" et l'angle α.

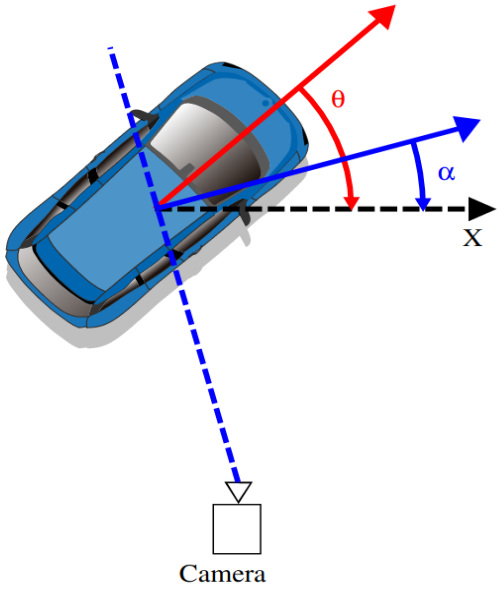

FIGURE 3.6 – Illustration du lacet θ de l'objet et de son orientation observée α. α est obtenue en calculant l'angle entre la normale (entre la caméra et le centre de l'objet) et l'axe X de la caméra.

3.6.4 Fonctions perte

Notre aproche de détection d'objets 3D fait appel à la fonction perte présentée dans l'équation (3.12) :

$$L = L_{yolo} + k_1 \times L_{center} + k_2 \times L_{distance} + k_3 \times L_{dim} + k_4 \times L_{orient}. \tag{3.12}$$

Nous utilisons la même perte que YOLOv3 pour la détection 2D et la prédiction de classe. Nous utilisons également $k_{i\in[1,...,4]}$ comme facteurs de pondération pour les pertes. Avec $\mathbf{c_k} = \begin{bmatrix} cx^k & cy^k \end{bmatrix}^T$ le centre de vérité terrain de l'objet k en pixels, $\mathbf{Rc_k} = \begin{bmatrix} Rcx_k & Rcy_k \end{bmatrix}^T$ le centre de la RoI pour l'objet k, N le nombre d'objets et $\tilde{\mathbf{O}}\mathbf{c_k}$ le décalage entre la boite englobante prédite par le détecteur d'objets et la position de l'objet à déterminer, la perte du centre s'écrit donc via l'équation (3.13) :

$$L_{center} = Moyenne(\frac{1}{N}\sum_{k=1}^{N}\left|\mathbf{c_k} - \mathbf{Rc_k} - \tilde{\mathbf{O}}\mathbf{c_k}\right|). \tag{3.13}$$

En supposant que z_k est la vérité terrain de la distance d'un objet k sur l'axe Z du repère caméra, N le nombre total d'objets, et $\tilde{z}_k$ la prédiction de distance de notre méthode, la perte

de distance est alors calculée en utilisant la perte $L1$ et est détaillée dans l'équation (3.14) :

$$L_{distance} = \frac{1}{N}\sum_{k=1}^{N} |z_k - \tilde{z}_k| \,. \tag{3.14}$$

Avec $\mathbf{d_k} = \begin{bmatrix} dx & dy & dz \end{bmatrix}^T$ les dimensions réelles d'un objet k en mètres, $\mathbf{dd}_k$ la dimension moyenne pour la classe d'objet k, N le nombre d'objets et $\tilde{d}_k$ la prédiction de notre méthode, la perte de dimensions est présentée dans l'équation (3.15) :

$$L_{dim} = Moyenne(\frac{1}{N}\sum_{k=1}^{N} \left|\mathbf{d_k} - \tilde{\mathbf{d}}_\mathbf{k} - \mathbf{dd_k}\right|). \tag{3.15}$$

Enfin, en supposant que α_k est l'angle observée de l'objet k, N le nombre total d'objets, et $\tilde{\alpha}_k$ la prédiction, nous utilisons la perte Smooth $L1$ comme perte d'orientation et la perte s'écrit selon l'équation (3.16) :

$$L_{orient} = \frac{1}{N}\sum_{k=1}^{N} e_k, \tag{3.16}$$

Où :

$$e_k = \begin{cases} 0.5(\alpha_k - \tilde{\alpha}_k)^2, & \text{si } |\alpha_k - \tilde{\alpha}_k| < 1 \\ |\alpha_k - \tilde{\alpha}_k| - 0.5, & \text{sinon} \end{cases}$$

3.6.5 Résultats expérimentaux

Entrainement sur KITTI. Nous avons entraîné notre méthode sur KITTI sur le même split d'entraînement que celui utilisé dans [117] contenant la moitié des échantillons et nous avons effectué l'évaluation sur l'autre moitié. KITTI offre plus de $7k$ images annotées pour l'entraînement avec des données sur les boîtes de délimitation 2D, la position de l'objet (XYZ), ses dimensions, l'angle de lacet de l'objet et l'orientation observée pour 3 classes d'objets différentes (voiture, vélo, personne). Nous avons transféré les poids d'un modèle YOLOv3 déjà entrainé sur COCO [72] (Transfert learning) sur notre nouveau modèle afin de réduire le temps d'entrainement et augmenter la précision.

Entrainement sur GTA. Nous avons utilisé notre dataset ESRORAD [111] pour entrainer et évaluer notre nouvelle méthode. L'entraînement a été réalisé sur un split contenant $107K$ images routières et ferroviaires. L'évaluation a été réalisée sur la partie comprenant $11K$ images et l'entraînement a été effectué sur 50 époques.

Evaluation de la Détection 2D. Pour évaluer la performance de la détection 2D, nous avons utilisé les métriques décrites par les auteurs de YOLOv3 sur chaque classe du jeu de données. Étant donné que la régression des paramètres de la boîte de délimitation 3D repose

sur des RoIs et des classes précises d'objets, l'évaluation de la précision du détecteur 2D s'avère donc une nécessité. Les mesures d'erreur sont la moyenne de la précision (AP), le rappel (R), la précision moyenne (mAP) et le score F1.

Evaluation de l'estimation de la distance. Nous avons utilisé la même évaluation que celle utilisée pour les méthodes d'estimation de la profondeur comme Monodepth2 [129] ou MADNet [56]. Les métriques utilisées sont les mêmes que celles décrites dans le Chapitre 2. Soient z_{gt} et z_{pd}, respectivement, la distance réelle et prédite de l'objet i, calculées à l'aide de l'équation (3.17), où $\delta = 1.25^k$:

$$\alpha_{k=[1...3]} = max(\frac{z_{gt}}{z_{pd}}, \frac{z_{pd}}{z_{gt}}) < \delta^k. \tag{3.17}$$

Evaluation de la dimension. L'évaluation de la prédiction de la dimension est effectuée à l'aide du score de dimension (DS) décrit dans [108]. Avec V_{pd} et V_{gt} le volume prédit et le volume de vérité terrain de l'objet, le DS est calculé en utilisant l'équation (3.18) :

$$DS = min(\frac{V_{pd}}{V_{gt}}, \frac{V_{gt}}{V_{pd}}). \tag{3.18}$$

Evaluation du centre de l'objet. Les prédictions du centre de l'objet sont évaluées à l'aide du score du centre (CS) comme décrit dans [108]. En supposant que x et y sont les coordonnées du centre projeté en pixels et w et h la largeur et la hauteur de la boîte de délimitation 2D, CS est calculé selon l'équation (3.19) :

$$CS = (2 + cos(\frac{x_{gt} - x_{pd}}{w_{pd}}) + cos(\frac{y_{gt} - y_{pd}}{h_{pd}}))/4. \tag{3.19}$$

Evaluation de l'orientation. Pour évaluer les prédictions d'orientation, nous utilisons le score d'orientation (OS) tel que décrit dans le benchmark KITTI. On pose α l'angle observé, L'OS est calculé en utilisant l'équation (3.20) :

$$OS = (1 + cos(\alpha_{gt} - \alpha_{pd}))/2. \tag{3.20}$$

Résultats. Notre méthode nécessite 2 étapes : (i) extraction des caractéristiques de l'image pour prédire les RoIs des objets et leurs classes, (ii) utilisation à la fois des caractéristiques extraites et les RoIs pour prédire les paramètres 3D des objets. Contrairement à l'approche [117] qui améliore la prédiction de l'orientation lors de la détection 3D, Notre approche prédit non seulement la détection d'objets 2D, mais aussi leurs distances, leur centres 3D, leurs orientations ainsi que leurs dimensions 3D. Toutes ces prédictions sont intégrées dans un réseau "tout-en-un". Les résultats quantitatifs de notre méthode sur KITTI et GTAV sont présentés dans la figure 3.7.

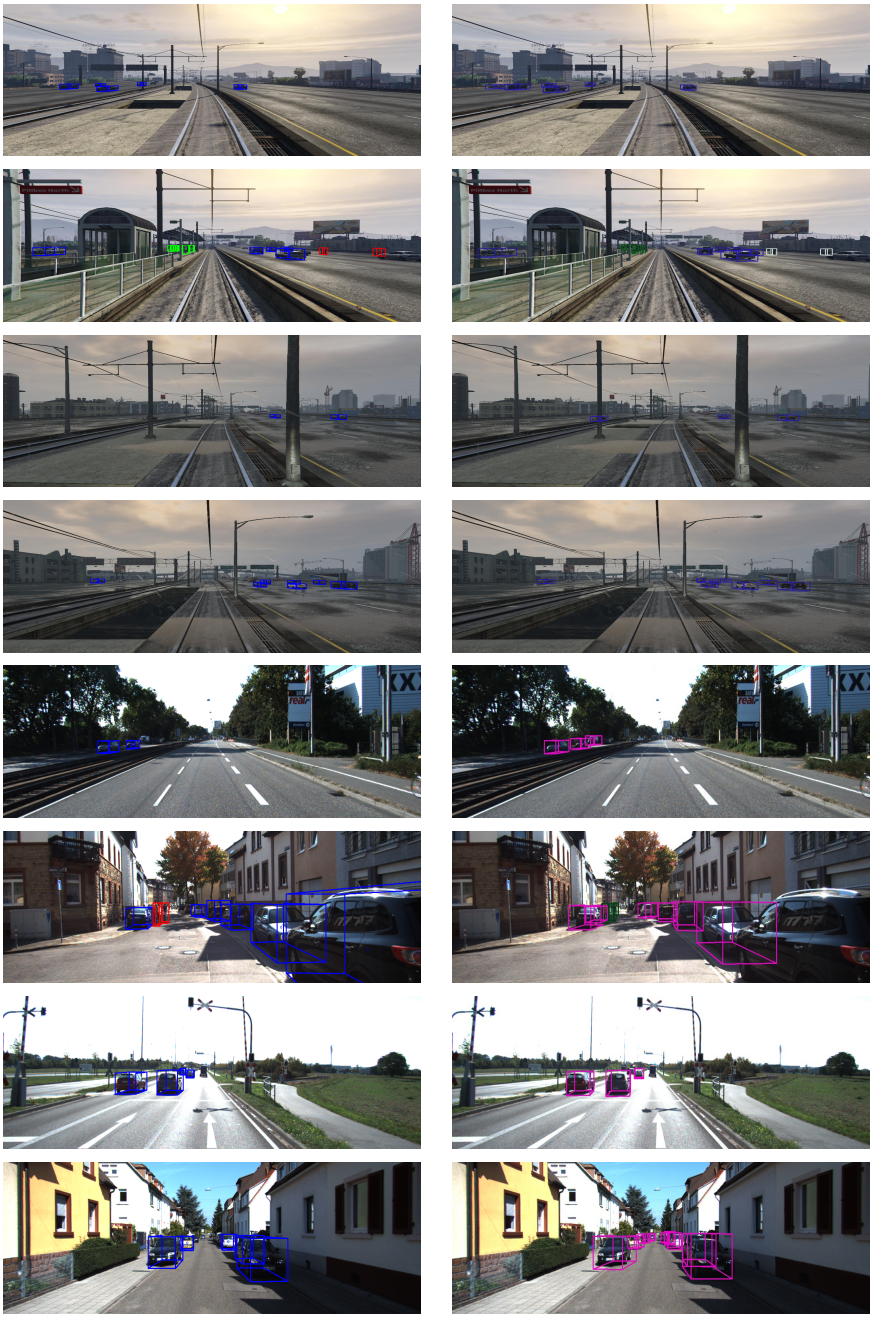

FIGURE 3.7 – Les résultats qualitatifs de notre méthode sur KITTI et GTAV. Les 4 lignes du haut : résultats obtenus sur GTAV, les 4 lignes du bas : résultats obtenus sur KITTI. Colonne de gauche : vérité terrain, colonne de droite : prédiction.

Les résultats obtenus montrent que l'architecture à réseau unique de notre approche nous permet de réduire considérablement le temps de calcul ainsi que l'empreinte mémoire, ce qui convient aux applications temps-réel embarquées. Bien qu'étant l'une des plus rapides, notre méthode de détection d'objets 3D n'atteint pas encore le niveau de précision des méthodes de pointe. Cela est dû au fait que la variation des RoIs pendant la phase de prédiction 2D entraîne une perte de précision lors de la prédiction des paramètres 3D. Il est donc nécessaire

de limiter autant que possible la prédiction directe des paramètres 3D. Par exemple, au lieu de prédire directement la position du centroïde de l'objet en mètres, nous prédisons le centre 3D projeté sur l'image 2D, permettant ainsi une meilleure prédiction du centre 3D. Notre réseau peut alors utiliser les informations liées à l'apparence de l'objet pour déduire la position du centre 3D de l'objet sur l'image. En combinant ces informations avec la prédiction de la distance de l'objet et la matrice de calibration, nous pouvons obtenir une prédiction du centre 3D de l'objet. Certes, cette approche permet d'améliorer la précision de la prédiction 3D de l'objet, mais elle n'atteint toujours pas la précision des données LiDAR. Cependant le prix de cette rapidité d'exécution est une perte en terme de précision assez importante. Il est donc devenu nécessaire de concevoir une seconde méthode pour la détection d'objets 3D qui ne souffre pas des problèmes liés à la variation des RoIs et de l'alignement des caractéristiques du réseau. C'est ce que nous allons présenter dans la section suivante.

3.7 Détection d'objets 3D par approche à étage unique

3.7.1 Vue d'ensemble

Nous avons jugé nécessaire de développer une nouvelle approche de détection d'objets 3D pour pallier aux problèmes liés à l'extraction des caractéristiques dans les RoIs. Pour ce faire, nous avons supprimé l'étape d'alignement des caractéristiques et le module de régression 3D afin de les remplacer par un réseau à architecture unique. Il va donc, dans le même module, prédire les paramètres 3D et la boîte de délimitation 2D grâce à une architecture à étape unique basée sur le détecteur 2D YOLOv5 [90]. La figure 3.8 illustre l'architecture de notre nouvelle approche.

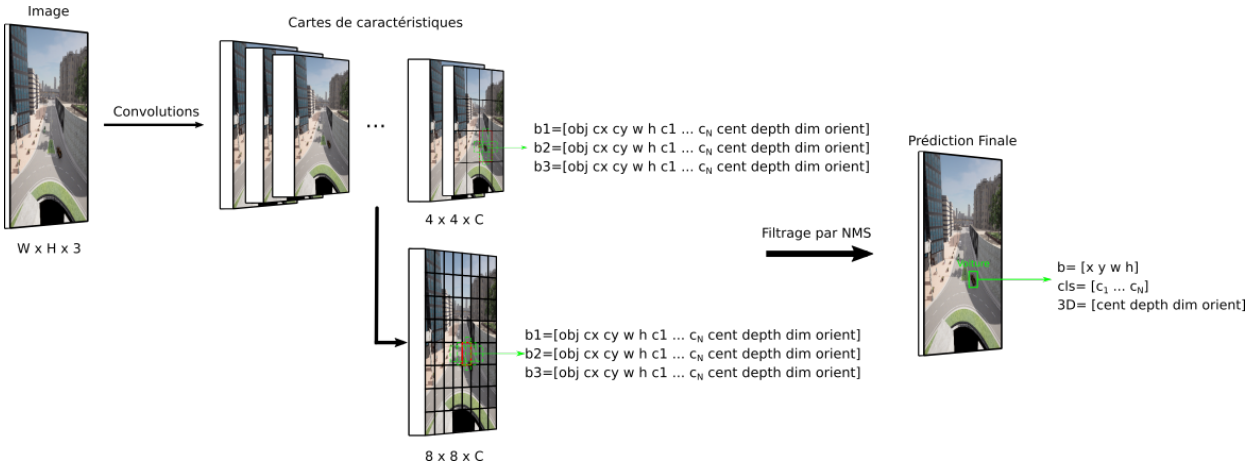

FIGURE 3.8 – Architecture de notre nouveau détecteur 3D basé sur YOLOv5. Nous avons modifié les "boîtes ancrées" pour prédire les paramètres 3D des objets. Une image de dimension ($W \times H \times 3$) est utilisée comme entrée du réseau. Des couches convolutionnelles successives transforment cette image en cartes de caractéristiques de différentes dimensions utilisées pour obtenir les "boîtes par défaut" (ou "boîtes ancrées"). Leurs prédictions pour chacune des cartes de caractéristiques sont combinées et filtrées à l'aide d'un filtre de Non-Maximum Suppression (NMS) pour obtenir la prédiction finale.

3.7.2 Détection d'objets en un seul étage

Notre méthode est un détecteur d'objets 3D en une étape basée sur YOLOv5. En effet, il s'agit d'un réseau convolutif léger qui offre une bonne précision, permettant une utilisation temps-réel. YOLOv5 offre un gain de performance significatif par rapport à YOLOv3 grâce à une architecture de réseau améliorée ainsi qu'à des innovations sur l'augmentation des données pendant l'entraînement. La conception à une étape de notre réseau permet de meilleurs temps de calcul par rapport aux méthodes à plusieurs étapes et notre réseau peut être entraîné de bout en bout. Notre approche utilise des boîtes d'ancrage hybrides pour régresser directement la boîte de délimitation 2D ainsi que les paramètres 3D. De cette façon, nous évitons le problème de la fluctuation de la précision lorsque les RoIs ne sont pas fixes. Cette nouvelle approche constitue la méthode la plus rapide pour la prédiction d'objets 3D et ce, sans sacrifier la précision. La figure 3.9 montre une vue d'ensemble de notre méthode de détection multi-objets 3D.

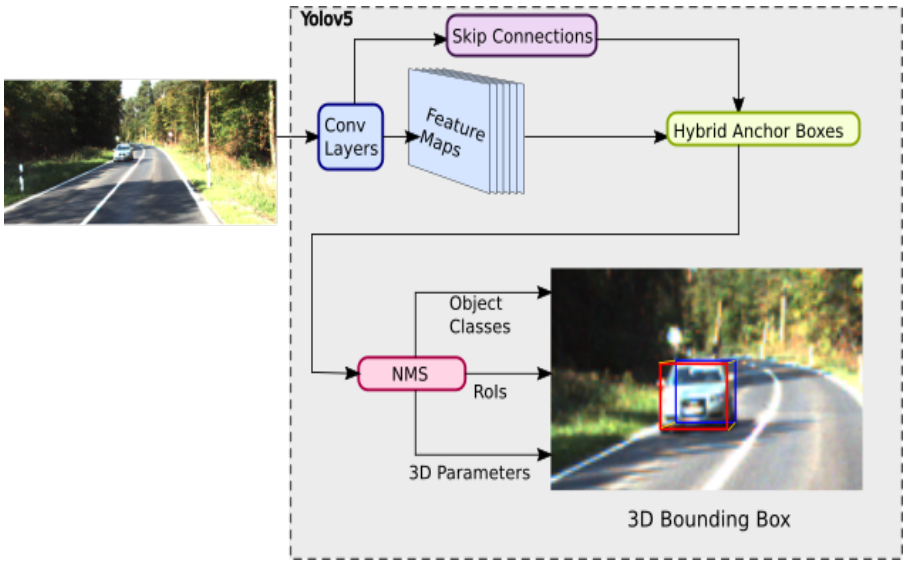

FIGURE 3.9 – Vue d'ensemble de notre méthode de détection multi-objets 3D. Nous utilisons nos boîtes d'ancrage hybrides pour prédire les boîtes de délimitation 2D et 3D. NMS est utilisé pour filtrer les prédictions. Parmi les paramètres 3D prédits : centre 3D projeté sur le plan image, distance, dimension et orientation de l'objet.

3.7.3 Paramètres prédits et boîtes d'ancrage hybride 2D/3D

Paramètres prédits et fonctions pertes. Les prédictions qui sont renvoyées par cette nouvelle méthode sont exactement les même que celles pour la méthode précédente : détection d'objets 2D, prédiction du centre projeté, distance, dimensions de l'objet et enfin son orientation. La perte utilisée lors de l'apprentissage de notre méthode est détaillée dans l'équation (3.21) :

$$L = L_{yolo} + k_1.L_{center} + k_2.L_{distance} + k_3.L_{dim} + k_4.L_{orient}. \tag{3.21}$$

Où L_{yolo} est la perte pour la détection 2D et la prédiction de classe, et L_{center}, $L_{distance}$, L_{dim} et L_{orient} sont les pertes pour le centre, la distance, les dimensions et l'orientation de l'objet 3D, respectivement. $k_{i\in[1,...,4]}$ sont les poids des différentes pertes. Pour la perte L_{yolo}, nous utilisons la même fonction que YOLOv5.

Boites d'ancrage hybride 2D/3D. Parmi les principales contributions de cette nouvelle approche est l'utilisation de nouvelles boites d'ancrages hybride afin de prédire à la fois les paramètres 2D (boite englobante, classe) et 3D (distance, centre, dimension et orientation). Ces nouvelles boites proviennent de YOLOv5 modifiée. Nous utilisons trois grilles de caractéristiques provenant de notre réseau avec des tailles différentes. Dans chaque grille, trois boîtes d'ancrage sont prédites. Ces boîtes contiennent les prédictions des classes d'objets, la position relative de la boîte englobante par rapport au centre de la grille, la probabilité

qu'un objet soit présent et enfin les paramètres de la boîte englobante 3D. L'équation (3.22) illustre la sortie de nos boîtes d'ancrage hybrides de notre détecteur d'objets 3D :

$$\mathbf{y} = \begin{bmatrix} \mathbf{b_{off}} & obj & \mathbf{cls} & \mathbf{c_{off}} & Z & \mathbf{D} & \mathbf{O} \end{bmatrix} \tag{3.22}$$

Où $\mathbf{y}$ est la valeur de sortie de l'une de nos nouvelles boîtes d'ancrage. Avec $\mathbf{b_{off}} = \begin{bmatrix} x & y & w & h \end{bmatrix}$ la prédiction du décalage de la boîte englobante 2D (avec x et y la position du centre de la boite sur les 2 axes de l'image et w et h la hauteur et la largeur de la boite en pixels. *obj* représente la prédiction de la présence de l'objet dans la boite, $\mathbf{cls} = \begin{bmatrix} cls_1 & ... & cls_N \end{bmatrix}$ le score pour les N classes cls, $\mathbf{c_{off}} = \begin{bmatrix} x_c & y_c \end{bmatrix}$ la prédiction du décalage entre le centre 3D projeté et le centre de la boîte englobante en pixels, Z est la position sur l'axe Z du centre 3D de l'objet en mètres, $\mathbf{D} = \begin{bmatrix} W_1 & H_1 & L_1 & ... & W_N & H_N & L_N \end{bmatrix}$ est le décalage entre les dimensions moyennes d'un objet selon sa classe N et la dimension de l'objet réelle prédite en mètres. La prédiction d'orientation
$\mathbf{O} = \begin{bmatrix} bin_1 & \overline{bin_1} & sin_1 & cos_1 & bin_2 & \overline{bin_2} & sin_2 & cos_2 \end{bmatrix}$, où bin_1 et bin_2 représentent la probabilité prédite de $\alpha \in \begin{bmatrix} -195 & 15 \end{bmatrix}$ et $\alpha \in \begin{bmatrix} -15 & 105 \end{bmatrix}$ respectivement avec α l'orientation observée de l'objet. $\overline{bin_k}$ représente la probabilité que l'angle ne se situe pas dans l'intervalle k. Enfin $sin_{i\in\{1,2\}}$ $cos_{i\in\{1,2\}}$ sont le sinus et le cosinus du décalage de l'angle observé par rapport au centre du i-ème bac.

3.7.4 Détails de l'entrainement

Taille du modèle. Comme proposé dans l'approche YOLOv5, notre méthode peut être déclinée en différentes versions avec une profondeur (nombre de couches) et une largeur (nombre de filtres dans chaque couche) différentes afin de faire varier la taille du réseau : Large, Medium et Small. Afin d'analyser les performances en précision et temps de calcul de chaque version, nous avons entraîné et évalué chacune d'entre elles sur les jeux de données KITTI et GTA.

Entraînement sur KITTI. Nous avons entraîné notre méthode sur KITTI sur la même répartition d'entraînement et de validation définie dans [130]. Pour accélérer le processus d'apprentissage et améliorer la précision, nous avons entraîné nos modèles avec les poids pré-entraînés de YOLOv5. Nous avons aussi utilisé les poids pré-entraînés de notre modèle sur GTAV. Nous démontrons que le pré-entraînement de notre modèle sur notre jeu de données améliore significativement la précision. De plus, comme dans [124], nous effectuons l'entraînement avec les images de gauche et de droite du split d'entraînement de KITTI, doublant ainsi le nombre d'images disponibles pour l'entraînement, ce qui améliore les performances de notre méthode.

Entraînement sur GTA. L'apprentissage de notre méthode a été effectué sur un échantillon contenant des images routières et ferroviaires ($107k$ images). L'évaluation a ensuite été réalisée sur la partie validation contenant 11630 images. L'apprentissage sur ce jeu de données a été effectué avec les couches de base de notre réseau provenant d'un modèle pré-entraîné de YOLOv5 sur le jeu de données COCO 2D [72] et ce, afin de réduire le temps d'apprentissage et d'améliorer la précision.

Augmentation des données. L'augmentation des données est très utile pour améliorer les performances des modèles CNN via la création de nouveaux échantillons d'images. Même si KITTI et GTA présentent une variété d'environnements et de conditions de trafic, le risque de surentraînement de nos modèles demeure toujours probable. Nous utilisons l'augmentation des données en mosaïque telle que définie dans YOLOv4 [131] qui consiste à mélanger 4 images d'entraînement améliorant ainsi la compréhension du contexte spatial des objets. Nous avons également appliqué des transformations de translation, de teinte, de saturation, de contraste ainsi que de mise à l'échelle aux images.

3.7.5 Résultats expérimentaux

Nos résultats (figure 3.10) montrent que notre approche, bien que moins complexe que les méthodes de l'état de l'art, a une précision comparable ou même supérieure à celles-ci. Notre modèle est également beaucoup plus rapide que toutes les autres approches testées, avec un temps de calcul de $11,2ms$ par image sur notre modèle "Large". Notre méthode est même capable d'atteindre une plus grande précision en utilisant des poids pré-entraînés sur GTAV, montrant ainsi le potentiel des jeux de données virtuels pour améliorer la détection d'objets 3D sur des jeux de données réels. Ainsi, notre modèle "Large" peut surpasser les méthodes de pointe dans des scénarios modérés et difficiles.

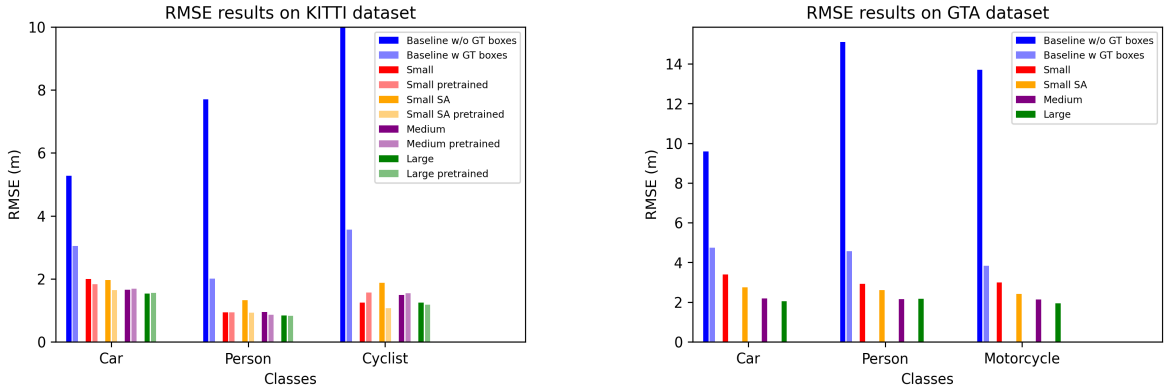

FIGURE 3.10 – RMSE de la profondeur obtenue par nos différents modèles par rapport à notre précédente méthode multi-étapes basée sur YOLOv3 [5] pour chaque classe sur les jeux de données KITTI et GTA. Notre nouvelle méthode améliore la précision sur KITTI de manière significative lorsque nous utilisons des poids pré-entraînés sur GTA.

Nous présentons également les différents modèles de notre approche actuelle : Small, SA Small, Medium et Large. Nous pouvons constater que la précision de ces modèles pour la détection 2D, dimension, centre et orientation ne bénéficie pas significativement des modèles plus lourds alors que l'estimation de la distance s'améliore en utilisant des modèles plus grands. Les résultats qualitatifs sur KITTI et GTA sont présentés dans la figure 3.11.

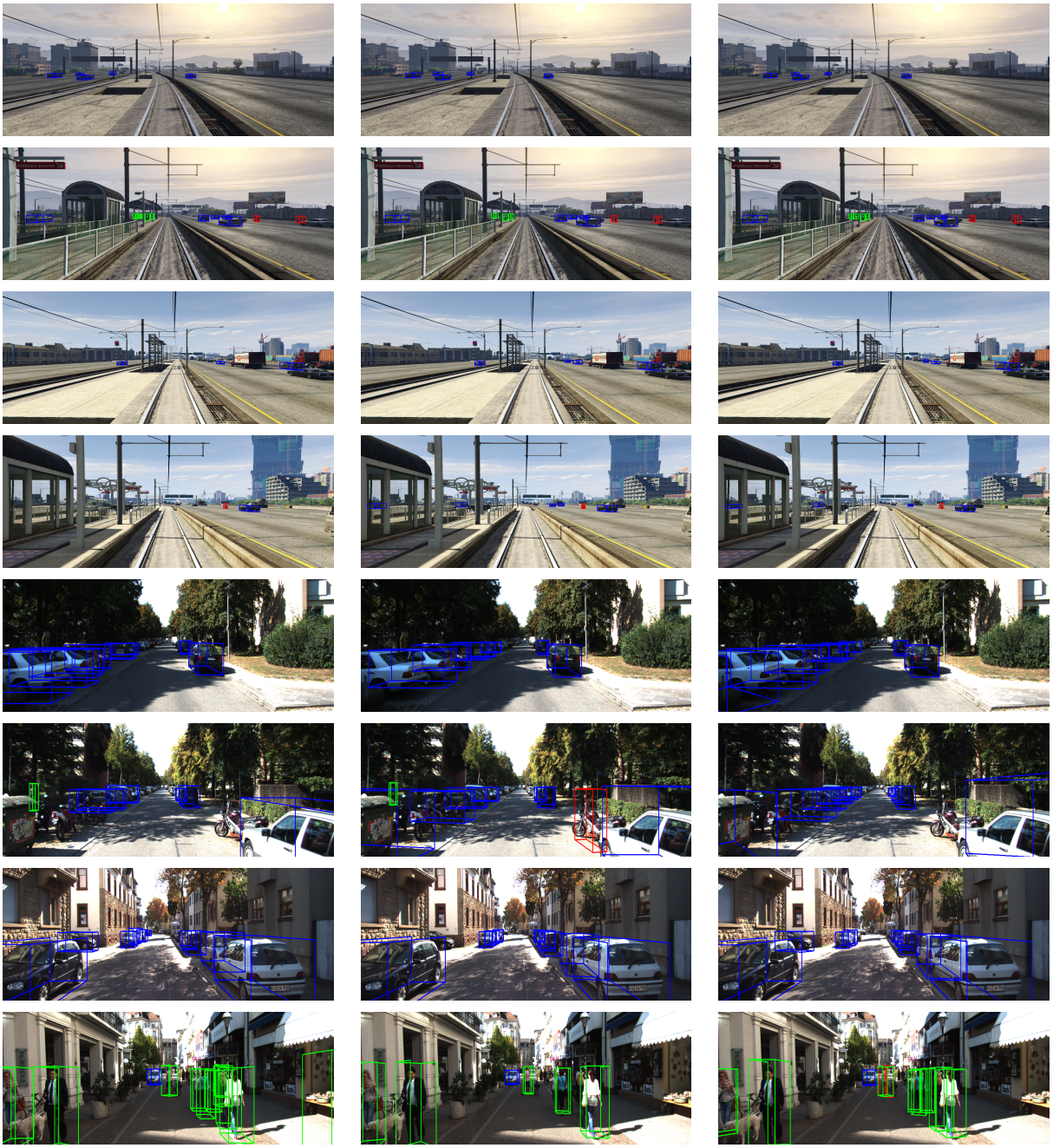

FIGURE 3.11 – Résultats qualitatifs de nos méthodes sur KITTI et GTA. Nous affichons les résultats de notre meilleur modèle (Large pré-entraîné sur GTAV) et de notre nouveau modèle léger pour les plateformes embarquées (Small SA pré-entraîné sur GTAV). Les 4 lignes du haut : résultats obtenus sur GTAV, les 4 lignes du bas : résultats obtenus sur KITTI. Colonne de gauche : vérité terrain, colonne du milieu : Small SA, colonne de droite : Large.

3.7.6 Création d'un modèle optimal

Un réseau qui a trop de paramètres aura un temps de calcul beaucoup plus élevé tout en ayant une plus grande tendance à se surentraîner, ce qui dégrade considérablement sa précision. En revanche, un modèle qui n'a pas assez de paramètres n'apprendra pas bien et aura une précision plus faible. Dans notre réseau, le nombre de paramètres définissant sa complexité dépend largement du nombre de couches et de filtres. La vitesse et la précision de notre approche dépendront également de la résolution des images d'entrée. Un travail a déjà été effectué pour obtenir un ensemble optimal (profondeur, largeur, résolution d'image) dans EfficientNet [132]. Les auteurs ont déterminé que la complexité du modèle, mesurée en Flops, était liée à la largeur (w), la profondeur (d) et la résolution de l'image (r) comme décrit dans l'équation 3.23 :

$$Flops \sim w^2 \times r^2 \times d \tag{3.23}$$

Afin de créer un modèle encore plus léger, nous avons choisi de mener une démarche similaire. Tout d'abord, nous avons défini la complexité de notre modèle avec l'objectif d'obtenir un temps de calcul de $33ms/image$ sur la Jetson TX2. La complexité choisie est de $\tilde{1}0$ GFlops. Nous sélectionnons ensuite toutes les combinaisons (largeur, profondeur, résolution) qui nous permettent d'obtenir cette même complexité. Enfin, nous entraînons chacun de ces modèles pendant 20 époques sur GTAV. Si nous définissons $w = 1$ et $d = 1$ comme la largeur et la hauteur du modèle "Large" et $r = 1$ pour définir une image de résolution 1280×720, le modèle optimal que nous avons trouvé a comme combinaison ($w = 0.4$, $d = 0.5$, $r = \frac{576}{1280} = 0.45$). Nous avons par la suite adopté la même procédure pour entraîner les autres modèles (Large, Medium, etc.). Nous avons d'abord entraîné YOLOv5 avec les mêmes dimensions de ce nouveau modèle sur le jeu de données COCO pendant 300 époques. Les poids pré-entraînés des premières couches sont ensuite transférés à notre modèle pour 15 époques d'entraînement sur GTAV. *In fine*, nous entraînons ce modèle sur les jeux de données KITTI et NuScenes. Ce nouveau modèle correspond au meilleur compromis entre faible temps de calcul et précision, comme l'illustre le tableau 3.1.

Base de donnée	Résolution	Détection 2D			Distance							Dimensions	Centre	Orientation	# Paramètres	temps/img
		AP	R	mAP@0.5	RE	SRE	RMSE	log RMSE	α_1	α_2	α_3	DS	CS	OS		
KITTI	608x192	0.642	0.518	0.469	0.056	0.133	2.15	0.081	0.985	0.998	0.999	0.871	0.953	0.863	6.0M	28ms
NuScenes	608x352	0.527	0.431	0.374	0.054	0.205	3.11	0.074	0.985	1	1	0.838	0.962	0.920	6.0M	28ms

TABLE 3.1 – Résultats quantitatifs de notre nouveau modèle optimal. Les modèles testés ont été entraînés sur KITTI et NuScenes.

3.8 Tracking d'objets 3D

Afin de prévenir les risques de collisions, tant sur la route que sur les rails, la détection et la localisation d'objets ne suffisent pas. En effet, il est nécessaire de prédire la position future des objets et les suivre en temps-réel. Dans le chapitre 2, nous avions déjà créé une méthode de suivi d'objets 3D afin de prédire leurs trajectoires. Les positions des objets étaient obtenues à partir d'un algorithme de détection 2D (YOLOv3) combiné à un second algorithme d'estimation de distance (MADNet). Les prédictions obtenues, bien que prometteuses, n'étaient pas au niveau de qualité de notre dernière approche de détection 3D. Notre objectif donc est de développer une nouvelle approche de tracking d'objets 3D rapide et précise. La qualité du suivi d'objet dépend largement de celle des prédictions de position. Plusieurs méthodes de suivi d'objets 3D utilisent déjà une approche similaire, combinant un détecteur 3D et un filtre de Kalman [108], mais aucune d'entre elles n'est compatible temps-réel. Nous appelons "tracklet" les instances d'objets qui sont suivis par notre algorithme. Notre approche de tracking d'objets 3D comprend donc trois parties : (*i*) association entre les détections et les tracklets, (*ii*) création de tracklets pour les détections non-associées et (*iii*) prédiction de la position des tracklets sur la frame suivante.

3.8.1 Association détection/tracklet

Afin de suivre avec précision les objets détectés, il est nécessaire d'associer les détections du temps t aux tracklets de $t-1$ dont la position a été prédite pour le temps t. Pour cela, nous calculons un score d'affinité entre la détection et le tracklet (i.e. un score permettant de caractériser la probabilité qu'un objet soit le même entre deux images consécutives) via l'équation (3.24) :

$$score = 0.5 \times Score_{coord} + 0.5 \times Score_{Feat} \tag{3.24}$$

Avec $Score_{coord}$ l'affinité des distances entre la position de détection et la position prédite du tracklet. $Score_{Feat}$ représente l'affinité des caractéristiques du réseau de la région où se trouve l'objet à t et celles des tracklets à $t-1$. Soit (X_{det}, Y_{det}, Z_{det}) et (X_{trk}, Y_{trk}, Z_{trk}) les positions du centre de l'objet détecté et la position prédite du tracklet, le score d'affinité des coordonnées $Score_{coord}$ est calculé en utilisant la distance euclidienne entre ces deux points 3D. En suivant les travaux de Deep SORT [57], nous utilisons une nouvelle métrique basée sur l'apparence des objets pour améliorer l'association entre les détections et les tracklets. Le calcul de cette métrique est possible car nous pouvons utiliser les couches intermédiaires de notre réseau ainsi que la fonction d'alignement des caractéristiques pour obtenir notre descripteur d'apparence d'objet. Soit $Feat_{det}$ et $Feat_{trk}$ les cartes de caractéristiques extraites à l'aide de la fonction d'alignement de caractéristiques sur les régions contenant la détection

à l'image t et le tracklet à $t-1$. Ces cartes se présentent sous la forme d'une matrice de dimensions (N_{filter},7,7) avec N_{filter} le nombre de filtres dans la couche réseau. Le score d'affinité des caractéristiques $Score_{Feat}$ est ensuite calculé en utilisant la distance euclidienne entre ces deux matrices (e.g. équation (3.25)) :

$$Score_{Feat} = exp\left(-\sqrt{\sum_{n}^{N_{filtre}}\sum_{k=1}^{7}\sum_{j=1}^{7}\frac{(\hat{f}_{n,k,j}-f_{n,k,j})^2}{500}}\right) \tag{3.25}$$

Où f et $\hat{f}$ les éléments des matrices $\mathbf{Feat}_{trk}$ et $\mathbf{Feat}_{det}$ respectivement. Pour chaque détection, nous calculons le score d'affinité *score* avec tous les tracklets. A la fin, nous pouvons créer une matrice contenant tous les scores entre les détections et les tracklets. Cette matrice de dimensions (N_{det},N_{trk}) représente le nombre de détections et de tracklets. L'association devient alors un problème d'optimisation combinatoire qui peut être résolu en utilisant l'algorithme hongrois [133] afin d'associer les détections les plus appropriées aux tracklets. Si aucun tracklet n'est associé à une détection, un nouveau tracklet est créé pour celle-ci. De la même manière, si aucune détection n'est associée à un tracklet, celui-ci est supprimé et le suivi est considéré donc comme terminé.

3.8.2 Prédiction

Le mouvement des objets suivis dans l'espace 3D est modélisé à l'aide d'un filtre de Kalman. Nous utilisons l'hypothèse d'une vélocité constante [134]. Ainsi, lorsqu'un objet détecté est initialisé dans notre approche et devient un tracklet, nous supposons que sa vitesse par rapport à la caméra est nulle (i.e. fréquence > 30Hz). Nous pouvons formuler ce modèle comme décrit dans l'équation (3.26) :

$$\begin{bmatrix} X_t \\ Y_t \\ Z_t \\ \dot{X}_t \\ \dot{Y}_t \\ \dot{Z}_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \\ \dot{X}_{t-1} \\ \dot{Y}_{t-1} \\ \dot{Z}_{t-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \Delta\dot{X} \\ \Delta\dot{Y} \\ \Delta\dot{Z} \end{bmatrix} \tag{3.26}$$

Avec (X,Y,Z) la position dans l'espace 3D et ($\dot{X}$,$\dot{Y}$,$\dot{Z}$) la différence de position entre le temps t et $t-1$ (vélocité). Enfin ($\Delta\dot{X}$, $\Delta\dot{Y}$,$\Delta\dot{Z}$) est la variation de la vitesse qui peut être modélisée par une distribution gaussienne $N(0,\sigma_v)$. Avec σ_v fixé en fonction du degré de notre certitude sur l'estimation *a priori* de la vélocité. Afin de modéliser le tracking avec le filtre

de Kalman, nous définissons d'abord le vecteur de mesure z_t donné par l'équation (3.27) :

$$\mathbf{z}_t = \mathbf{H} \begin{bmatrix} X_t \\ Y_t \\ Z_t \\ \dot{X}_t \\ \dot{Y}_t \\ \dot{Z}_t \end{bmatrix} + \gamma_t \tag{3.27}$$

Avec

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.28}$$

Où γ est l'erreur de mesure que nous supposons suivre une distribution gaussienne de moyenne nulle avec une covariance σ_γ. La phase de prédiction du filtre de Kalman s'écrit comme suit (équation 3.29 et 3.30) :

$$\begin{bmatrix} \bar{X} \\ \bar{Y} \\ \bar{Z} \\ \bar{\dot{X}} \\ \bar{\dot{Y}} \\ \bar{\dot{Z}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \\ \hat{\dot{X}} \\ \hat{\dot{Y}} \\ \hat{\dot{Z}} \end{bmatrix} \tag{3.29}$$

$$\bar{\Sigma}_t = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \hat{\Sigma}_{t-1} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_v & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_v & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_v \end{bmatrix} \tag{3.30}$$

Lorsqu'un objet est suivi pour la première fois et qu'un tracklet est créé, σ_v est initialisé à 0, les estimations de position initiale ($\hat{X}_{t=0}$,$\hat{Y}_{t=0}$, $\hat{Z}_{t=0}$) sont initialisées aux valeurs de position de la détection et les vitesses initiales ($\hat{\dot{X}}_{t=0}$,$\hat{\dot{Y}}_{t=0}$,$\hat{\dot{Z}}_{t=0}$) sont nulles. On donne à la matrice

de covariance $\hat{\Sigma}_{t=0}$ une valeur très élevée pour modéliser notre incertitude sur la première prédiction. Lorsqu'une détection est associée à un objet précédemment suivi, nous mettons à jour notre filtre de Kalman pour améliorer la prédiction.

3.8.3 Résultats expérimentaux

Dans cette section, nous présentons les résultats obtenus par notre nouvelle approche de tracking d'objets 3D. Les évaluations quantitatives sont en cours de réalisation. Les essais qualitatifs ont été réalisés à la fois sur des jeux de données routiers publics tels que KITTI et NuScenes mais aussi sur notre propre jeu de données ESRORAD comme le montre les figures 3.12 et 3.13.

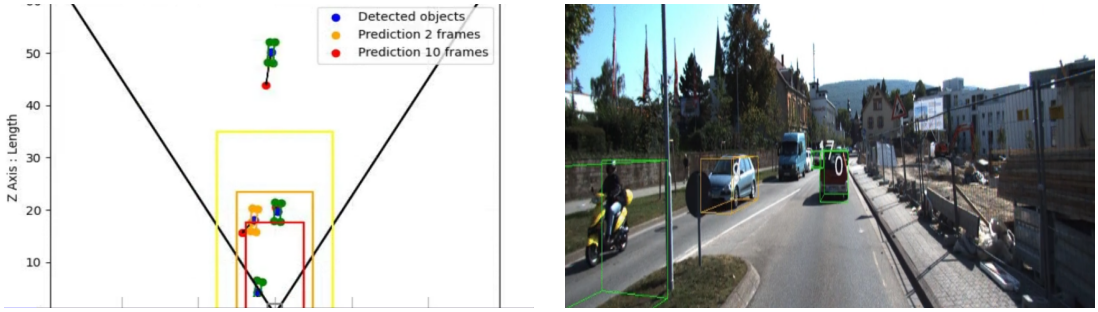

FIGURE 3.12 – Résultats de tracking 3D sur KITTI. A gauche sont tracées les trajectoires des objets suivis et à droite l'image avec les boîtes 3D des objets suivis dessinées.

FIGURE 3.13 – Résultats obtenus par notre méthode de suivi d'objets 3D. Les 2 lignes du haut : NuScenes, les 2 lignes du bas : ESRORAD. Le modèle "Large" entraîné sur NuScenes a été utilisé pour détecter les objets 3D. Dans la colonne de gauche sont présentés les résultats du suivi d'objet avec l'ID de l'objet au centre des boîtes. La lettre L indique qu'un objet est "perdu", aucune détection n'a été associée au tracklet mais l'algorithme continue à le suivre quelques images avant de l'effacer. Dans la colonne de droite sont présentés les résultats de notre détecteur d'objets 3D.

3.9 Conclusion

Nous avons présenté dans ce chapitre trois grandes parties : 1. le dataset ESRORAD pour la smart mobilité routière et ferroviaire, 2. une nouvelle approche de détection d'objets 3D et 3. une nouvelle méthode de tracking temps-réel d'objets 3D. ESRORAD est un jeu de données à grande échelle, le premier de son genre, qui peut être utilisé pour l'entraînement et la validation des modèles deep learning pour la smart mobilité routière et ferroviaire. La nature hybride de l'ensemble de données est obtenue en incluant des séquences synthétiques développées avec GTAV et des données réelles acquises dans les deux villes Normandes Rouen

et Le Havre. Les images sont annotées avec des boîtes de délimitation 3D montrant au moins trois classes d'objets différentes (personnes, voitures et vélos). L'évaluation comparative a été effectuée à l'aide de nos différents modèles CNN entraînés avec KITTI et NuScenes. ESRORAD a été validé et a donné des résultats prometteurs. Par conséquent, il représente un excellent choix pour l'entraînement et la validation des différents algorithmes de perception d'environnement. Nous fournissons un accès libre à notre jeu de données pour permettre aux chercheurs et aux entreprises de mener à bien leurs travaux de recherche. A moyen terme, ESRORAD vise à étendre les données ferroviaires en dehors des villes pour intégrer les paysages ruraux comme dans le cas des trains longue distance. La qualité des images et des vidéos sera améliorée grâce à un nouveau LiDAR haute résolution.

Dans la deuxième partie, nous avons présenté nos deux nouveaux algorithmes pour la détection d'objets 3D. Notre première approche est basée sur une architecture à deux étages avec deux modules, le premier étant le détecteur d'objets 2D temps-réel YOLOv3 et le second étage étant dédié à la prédiction des paramètres 3D. Si cette première méthode était prometteuse, notamment grâce à sa vitesse élevée, elle souffrait d'un manque de précision dû à la variation des RoIs des objets détectés. Nous avons donc présenté notre deuxième approche à un seul étage pour la prédiction simultanée des paramètres 3D et 2D. Elle présente également un temps de calcul bien meilleur avec une précision identique voire supérieure aux méthodes de l'état de l'art. Enfin, dans la troisième partie nous avons développé une nouvelle approche de tracking temps-réel d'objets 3D afin de prédire leurs trajectoires dans l'espace 3D et ainsi anticiper les risques de collisions. Des essais ont été réalisés pour valider nos approches en conditions réelles de trafic.

Chapitre 4

Analyse et Compréhension de Scènes par Deep Learning

Sommaire

4.1 Introduction

Ce chapitre est dédié à l'analyse et compréhension de scènes complexes routières et ferroviaires. Il comprend trois parties : 1. détection de piétons par les approches classiques de la vision par ordinateur, 2. détection d'objets par deep learning et 3. analyse et compréhension de scènes routières. Nous allons présenter une succession d'approches classiques et par deep learning de détection et tracking temps-réel d'objets et d'analyse et compréhension de scènes routières. Dans un premier temps, une nouvelle approche Faster-DPM de détection d'objets de type piétons sera présentée. Afin de tester la fiabilité de l'approche développée pour le tracking temps-réel d'objets, nous l'avons adaptée à l'asservissement visuel d'un drone. Par la suite, nous présenterons une contribution basée sur des approches deep learning dédiées à la détection, l'estimation de distance et le suivi d'objets pour la smart mobilité routière et ferroviaire. Une étude comparative entre les deux détecteurs d'objets, SSD et YOLOv3, sera présentée afin d'évaluer leurs performances. Nous présenterons aussi notre approche d'estimation de la distance basée sur l'algorithme monodepth. Nous avons fusionné les deux approches de détection et d'estimation de distance afin de partager les couches d'extraction de caractéristiques, ce qui permet d'améliorer leur efficacité.

Des expériences de validation des approches présentées ont été effectuées sur les jeux de données publics Cityscape et KITTI, mais aussi dans des conditions réelles de circulation dans le centre-ville de Rouen. Enfin, la troisième et dernière partie sera dédiée à notre nouvelle approche innovante basée LSTM pour l'analyse et compréhension du comportement des agents routiers. Des expériences menées sur deux jeux de données consacrés à la compréhension du comportement des agents routiers seront présentées. Combiner les différentes approches de détection, localisation et tracking temps-réel d'objets avec celles de l'analyse du comportement des agents routiers permettra, *in fine*, une meilleure interaction et prise de décision dans des environnements complexes.

4.2 Détection temps-réel de piétons par Faster-DPM

4.2.1 Contexte & objectifs

Dans le cadre du projet M2NUM sur la smart mobilité et particulièrement la détection de piétons ainsi que le partenariat industriel avec SEGULA sur le véhicule autonome (drone autonome), nous avons développé de nouvelles méthodes basées sur deux axes : (i) détection de piétons basée sur l'Histogram Oriented Gradient (HOG) [96] et le Deformable Part Model (DPM) amélioré, (ii) détection d'objets et/ou piétons basée deep learning. Afin de valider ces approches, nous les avons appliquées sur deux scénarios de smart mobilité : la détection

de piétons dans des scènes routières et le tracking temps-réel de personnes par un drone. Plusieurs tests effectués au laboratoire et dans des conditions réelles de trafic (centre-ville de Rouen) ont permis de valider la plateforme développée.

4.2.2 Détection de piétons par Faster-DPM

Les algorithmes HOG et DPM calculent d'abord les caractéristiques des images. Ils appliquent des classificateurs sur des bases de données d'images positives (avec piétons) et négatives (sans piétons). Nous avons réalisé toutes les expériences avec 6 jeux de données différents dédiés à la détection de piétons : ETH [135], INRIA [136], TUD Brussels [137], Caltech [138], Daimler [139] ainsi que notre propre jeu de données ESIGELEC dédié à la détection des piétons. Ils ont l'avantage d'être robuste et donnent des résultats relativement précis. Cependant, leur temps de calcul demeure élevé et rend donc leur utilisation difficile dans les applications de smart mobilité. Nous avons donc amélioré l'algorithme HOG/DPM pour qu'il soit compatible avec nos contraintes temps-réel tout en maintenant une précision élevée. Pour cela, nous avons modifié les 5 étapes de l'algorithme HOG (i.e. de la correction gamma en passant par le redimensionnement des images jusqu'au gradients négatifs et le vecteur de normalisation).

Notre nouvelle approche, appelée désormais "Faster-DPM", apportes-en plus d'une précision élevée, un gain en temps de calcul considérable. L'algorithme DPM est appliqué sur l'image entière et ceci à chaque itération de la séquence vidéo. Or le Faster-DPM est appliqué uniquement dans une région d'intérêt (RoI) dans laquelle la cible (ici le piéton) est située. Cela avait réduit considérablement le temps de calcul et avait permis également de mieux isoler l'objet du fond d'image. Dans un premier temps, le DPM est appliqué sur toute l'image pour localiser l'objet. Ensuite, et après avoir obtenu une RoI entourant l'objet à détecter, nous construisons une nouvelle boite englobante (RoI) adaptative intégrant une zone de tolérance. En cas de perte de l'objet à suivre, l'algorithme Faster-DPM est réappliqué, à nouveau, sur l'ensemble de l'image pour tenter de le retrouver.

Afin de mieux évaluer les performances de notre approche, nous avons effectué une comparaison de notre algorithme Faster-DPM avec un détecteur d'objet CNN basé SSD comme le montre la figure 4.1.

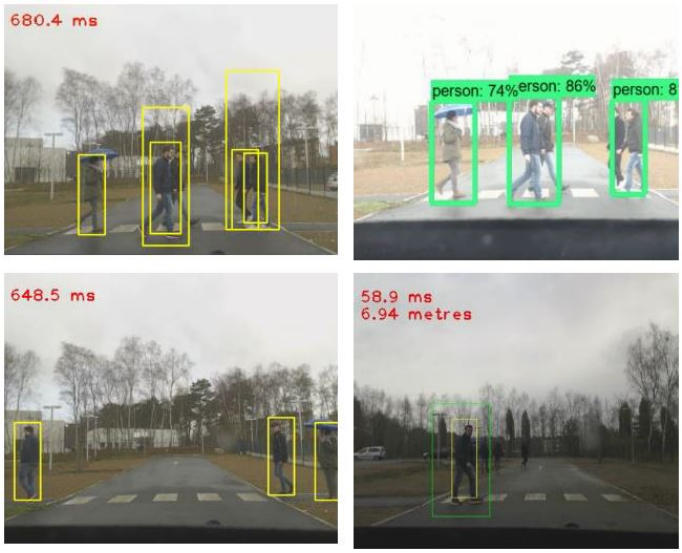

FIGURE 4.1 – Détection de piétons et estimation de distance par Faster-DPM et SSD. Image supérieure gauche, inférieure gauche et droite : détection de piétons et distance calculée par Faster-DPM avec RoI adaptative (rectangle vert). Image supérieure droite : détection de piétons basée SSD.

4.2.3 Asservissement visuel d'un drone par Faster-DPM

Nous avons aussi étendu l'application de notre nouvelle approche Faster-DPM en l'appliquant au tracking temps-réel de personnes par un drone. Afin d'asservir la position du drone par rapport à celle de la cible, plusieurs étapes ont été effectuées. Dans un premier temps, et une fois la cible est détectée par notre approche Faster-DPM, nous calculons le barycentre de la RoI qui devient le point à suivre en temps-réel par le drone. Par la suite, une estimation de la distance entre la cible et le drone est effectuée.

Enfin, nous avons mis en œuvre deux types de contrôleurs : PID et le contrôleur Adaptive Super Twisting (AST) [140]. La plateforme a été validée via différents scénarios en comparant les données de navigation mesurées (i.e. les deux contrôleurs PID et AST) à celles de vérité terrain fournies (i.e. trajectoires effectuées par la cible). Les essais de validation en indoor et outdoor ont montré, encore une fois, la robustesse de notre approche Faster-DPM. Ces travaux de recherche ont fait l'objet de publications dans [30] et [80].

4.3 Détection et tracking temps-réel d'objets par deep learning

4.3.1 Contexte & objectifs

Dans le cadre des travaux de recherche sur le projet M2NUM [1], nous avons développé un système de détection et tracking d'objets pour les applications de smart mobilité routière et ferroviaire. Nous avons adopté deux approches de détection d'objets : YOLOv3 et SSD. Nous avons par la suite effectué l'estimation de la distance d'objets basée sur l'algorithme Monodepth. En sortie, nous obtenons une carte de disparité que nous combinons avec la détection d'objet. Pour valider notre approche, nous avons testé deux modèles avec différents backbones dont VGG et ResNet utilisés avec Cityscape [70] et KITTI [59]. Comme dernière étape, nous avons développé une nouvelle méthode basée sur SSD pour analyser le comportement des piétons et des véhicules en suivant leurs mouvements (e.g. véhicule traversant de droite vers la gauche, piéton qui traverse, piéton à l'arrêt, etc.) et ce, même en cas de perte d'objets sur certaines images d'une séquence. L'ensemble des développements a été testé dans des conditions de trafic routier réel (centre-ville de Rouen) et ferroviaire (vidéos prises depuis le tramway de Rouen).

4.3.2 Détection d'objets par SSD et YOLOv3

Le modèle SSD a été entraîné sur le dataset PascalVOC [71], YOLOv3 quant à lui a été entraîné sur les deux jeux de données, PascalVOC et COCO [72]. Pour obtenir des résultats d'évaluation, nous avons entraîné deux modèles, l'un sur PascalVOC avec une classe "personne" et l'autre sur le jeu de données COCO. D'après les résultats obtenus, nous constatons que le modèle entraîné sur le jeu de données COCO avec 64115 images de personnes montre une amélioration en mAP de 5.6 par rapport aux résultats obtenus de l'entrainement sur PascalVOC avec $2k$ images de personnes. En plus, nous constatons que le nombre de classes n'influence pas de manière significative la vitesse d'inférence. Le modèle à une classe est seulement 0.4 FPS plus rapide que le modèle à 80 classes. Le temps d'inférence dépend principalement de la quantité de paramètres dans le réseau CNN, mais la suppression des classes n'a qu'un impact sur le nombre de paramètres de la couche entièrement connectée.

4.3.3 Estimation de la distance des objets par deep learning

Monodepth est une approche non supervisée basée CNN dédiée à l'estimation de la distance. Le modèle est entraîné sur des images stéréoscopiques mais testé en inférence sur

1. Modélisation Mathématique et simulation NUMérique

des images monoculaires et produit une carte de disparité. Nous avons testé deux modèles sur certaines images, avec différents backbones, VGG et ResNet, et sur différents jeux de données Cityscape et KITTI. Les valeurs de disparité représentent des distances relatives car les images d'entraînement et celles d'inférence sont prises par des caméras ayant des distances focales différentes. Pour obtenir la distance réelle, nous avons utilisé des images de calibration pour calculer le coefficient de baseline et la distance focale.

Ces cartes de disparité de sortie montrent l'influence que peut avoir le choix du dataset ainsi que du backbone. Le modèle entraîné sur Cityscape obtient de meilleurs résultats. Le choix du backbone influence aussi la qualité des résultats : VGG a une meilleure performance que ResNet. Nous combinons des résultats de la détection d'objets et de l'estimation de la profondeur : pour chaque objet détecté, l'histogramme des disparités estimées est calculé. Nous avons appliqué cette méthode sur des images réelles et synthétiques. Les résultats étaient difficiles à quantifier, car il n'y avait pas de jeux de données disponibles avec vérité terrain de distance. L'estimation de la distance donne de bons résultats lorsque les objets ne sont pas trop nombreux. Ce problème a été solutionné en appliquant une segmentation sémantique aux images, parallèlement au processus de détection d'objets, afin de n'utiliser que les pixels appartenant aux objets situés à l'intérieur des boîtes englobantes. La figure 4.2 montre quelques exemples de cartes de disparité et estimation de distance évaluées comme étant satisfaisantes. Ces travaux de recherche ont fait l'objet de publications dans [30] et [80].

FIGURE 4.2 – Cartes de disparité et estimation de distance. En haut : test de différents jeux de données : cartes de disparité résultantes de l'application de Monodepth sur une scène de trottoir, avec backbone VGG et entraînement sur deux datasets : Cityscape (milieu) et KITTI (droite). En bas : exemple d'estimation de distances. La disparité estimée est superposée à chaque objet détecté.

4.4 Analyse et compréhension de scènes par deep learning

4.4.1 Contexte & objectifs

L'analyse du comportement sur la route est une tâche essentielle pour la smart mobilité. Dans le cadre du projet M2SiNUM [2] dont l'objectif était la modélisation mathématique de problèmes issus de diverses applications pour l'environnement et le vivant, notre contribution menée par les travaux de la post-doctorante Rim Trabelsi, portait sur l'aide à la conduite autonome. Contrairement aux tâches de perception traditionnelles, nos travaux visaient à obtenir une compréhension de haut niveau des activités des agents routiers (i.e. véhicules, piétons, vélos, etc.). Nous avons développé un nouveau réseau profond spatio-temporel basé sur un mécanisme d'attention qui exploite des caractéristiques visuelles clés au fil du temps. Il s'agissait en fait d'une nouvelle approche appelée STAF (Spatio-Temporal Attention Framework for Understanding Road Agents Behaviors) à travers des couches de mémoire à long terme (LSTM) [141] qui utilise un mécanisme d'attention multi-têtes sur l'état passé des cellules afin de se concentrer sur les attributs qui sont pertinents dans le temps. Des expériences ont été menées sur deux scénarios différents sur des données provenant de Joint Attention in Autonomous Driving (JAAD) [142, 143] et Honda Research Institute Driving Dataset (HDD) [144], deux ensembles de données consacrés à la compréhension du comportement des agents routiers. L'évaluation des résultats obtenus prouvait que notre modèle STAF est plus performant que les algorithmes de pointe de la littérature, comme par exemple LSTM et ce, avec une précision moyenne (mAP) de 13%.

4.4.2 Etat de l'art

Approches spatio-temporelles

Le mécanisme d'attention a transformé le mode de fonctionnement des RNN/LSTM en leur permettant de se concentrer sur un certain segment d'une séquence spatio-temporelle donnée [145, 146]. Les modules d'attention standard sont employés dans les RNN standard entre les modules encodeur et décodeur où, pour un intervalle de temps donné, ils prennent en entrée les vecteurs de sortie de l'encodeur et les états correspondants du décodeur. La sortie du module d'attention est composée d'une séquence de "vecteurs de contexte" permettant au module suivant, et au décodeur, de se concentrer sur certaines parties de l'entrée lors de la prédiction de la sortie. L'une des propositions marquantes est le module "Transformer". Au lieu d'utiliser le mécanisme d'attention en conjonction avec les RNN, le module Transformer révolutionnaire introduit initialement dans [147] et réutilisé avec succès dans [148, 149],

2. Modélisation mathématique avancée et Simulations Numériques pour l'innovation dans l'environnement et la santé

prouve que cette nature séquentielle peut être capturée en utilisant uniquement le mécanisme d'attention et ce, sans aucune association avec d'autres machines. Dans [150], le mécanisme d'attention a été mis au-dessus des RNNs en introduisant le soi-disant "consciousness prior". Il consiste à formaliser la conscience de manière récurrente dans le temps, en retenant des fragments des représentations d'entrées dans la cellule récurrente comme s'il s'agissait d'une attention.

Les modules d'attention mentionnés précédemment et leurs extensions ont à peine été utilisés dans l'analyse vidéo avec une application pour la perception d'environnements routiers [151, 152, 153]. L'apprentissage spatio-temporel est pourtant bien maîtrisé dans des applications telles que la reconnaissance vidéo [154, 155] et le traitement [156]. Par exemple, [157] emploie le mécanisme d'attention par le biais de LSTM pour générer des représentations spatialement pondérées. Yang *et al.* [158] ont proposé d'utiliser l'attention pour détecter une région d'intérêt à chaque instant mais ils n'ont pas réussi à la localiser à chaque image. Des stratégies d'attention hiérarchique sont élaborées dans [159, 160] visant à extraire les poids de l'attention spatiale et leur vraisemblance conditionnée avec leur schéma d'attention temporelle correspondant.

L'attention spatio-temporel. L'attention temporelle a été largement utilisée dans des travaux récents sur le sous-titrage vidéo pour décider quelle(s) image(s) de la vidéo est (sont) importante(s) pour générer le mot suivant dans une légende. Cependant, ces systèmes transforment généralement les images vidéos brutes en caractéristiques CNN de haut niveau, ce qui marginalise des informations spatiales importantes pour le sous-titrage. Tu *et al.* [159] et Yu *et al.* [155] proposent des schémas d'attention hiérarchique qui conditionnent le mot du sous-titre actuel et les caractéristiques visuelles. Ils génèrent d'abord des poids d'attention spatiale, avant de générer le mot via les caractéristiques pondérées. Plus récemment, Aafaq *et al.* [161] utilisent l'ingénierie des caractéristiques spatio-temporelles pour améliorer les performances du sous-titrage. Dans [162], la saillance des objets est combinée au raisonnement par graphe temporel bidirectionnel. Cette approche est liée au modèle d'attention classée que nous avons proposé, mais notre formulation est beaucoup plus simple. Xu *et al.* [163] ont proposé un réseau FCN-LSTM de bout en bout considérant 4 actions discrètes dans l'apprentissage d'un modèle de conduite.

Compréhension du comportement sur la route

La compréhension de l'activité humaine joue un rôle important dans la réalisation de systèmes intelligents. Différents ensembles de données ont été proposés, comme le jeu de données vidéo étiquetées Cambridge-driving (CamVid) [164], AutoRate pour prédire l'inat-

tention du conducteur [165], Trajectory Prediction in Heterogeneous Traffic (TraPHic) [166] et Deep RObust Goal-Oriented trajectory prediction Network (DROGON) [167] pour remédier aux limites des travaux antérieurs. ROad [91] event Awareness Dataset for Autonomous Driving (ROAD) est un ensemble de données qui a été conçu pour tester les capacités de prise de conscience de la situation d'un robot-voiture. Le processus d'annotation suit une approche multi-label dans laquelle on reconnait les agents de la route (véhicules, piétons, etc.), leurs emplacements ainsi que leurs actions. La reconnaissance d'une action orientée vers un objectif est un problème de reconnaissance d'activité égocentrique. Le jeu de données Stanford-ECM [168] est similaire au notre via au moins les deux aspects suivants : (*i*) il définit des classes d'activités égocentriques pour les humains, similaires à nos classes orientées vers les objectifs pour les conducteurs et (*ii*) il fournit des vidéos égocentriques provenant d'un capteur portable pour l'apprentissage conjoint de la reconnaissance des activités et de la dépense énergétique, tandis que dans notre approche, nous fournissons des enregistrements multicapteurs d'un véhicule instrumenté pour l'apprentissage du comportement du conducteur.

Dans [169], Ramanishka présente l'ensemble des données de conduite du jeu de données HDD qui comprend une nouvelle méthodologie d'annotation avec une représentation à 4 couches. Un algorithme basé sur la détection du comportement du conducteur a été entraîné et testé pour le processus de validation. Rasouli [142] a présenté un nouveau jeu de données consacré à la compréhension des scènes de trafic qui combine la localisation des piétons, leur comportement ainsi que des données contextuelles. En plus de l'apprentissage des activités égocentriques, nous annotons également la manière dont les participants au trafic interagissent avec les conducteurs.

4.4.3 Compréhension du comportement des agents routiers

L'approche commune pour remédier au problème de prédiction du comportement des usagers de la route consiste à utiliser des facteurs dynamiques, tels que la trajectoire [170], la vitesse [171] ou le but final bien anticipé des piétons [155, 172]. Un autre indice comportemental est l'orientation de la tête du piéton afin de mesurer le niveau de conscience au moment de la traversée [158]. Ces études ont toutefois une portée limitée et ne prennent en compte que très peu d'éléments contextuels nécessaires pour prédire le comportement des piétons. Dans la pratique, il existe d'autres facteurs, en plus des facteurs spatio-temporels, qui peuvent influencer le comportement d'un piéton, notamment la structure "passage piétons" (e.g. signalisation, délimitation), les facteurs environnementaux (e.g. conditions météorologiques, visibilité) ou encore les caractéristiques individuelles des piétons (e.g. données démographiques).

Dans la figure 4.3, nous présentons notre description des composants d'analyse du comportement routier. Par agents, nous visons les différents acteurs du trafic routier : piétons, véhicules, vélos, motos, etc. Les travaux de l'état de l'art ont utilisé des approches communes d'apprentissage profond et de vision par ordinateur pour prédire les comportements des agents routiers : (i) des indices élémentaires comme la vélocité [173], les trajectoires [170] et (ii) souvent consacrés à un seul scénario comme la classification des manœuvres [174], la reconnaissance des états du conducteur [175] et les activités des piétons [176].

En vue d'une analyse complète de la scène routière, nous proposons une nouvelle approche d'apprentissage pour modéliser explicitement le comportement des agents routiers et leur interaction avec l'environnement. Notre contribution est triple :

1. Tout d'abord, nous proposons un réseau profond spatio-temporel basé sur un mécanisme d'attention qui exploite les caractéristiques visuelles clés au fil du temps.
2. Par la suite, en s'inspirant du modèle "Transformer" introduit dans [147] pour les tâches de traduction, nous proposons de réutiliser le mécanisme d'auto-attention à têtes multiples avec un réseau LSTM que nous appelions STAF pour Spatio-Temporal Attention Framework.
3. Enfin, pour prouver l'efficacité de notre modèle à traiter l'analyse des comportements routiers, nous réalisons des expériences approfondies sur les jeux de données JAAD et HDD dédiés à la compréhension des interactions entre les agents routiers, les comportements des conducteurs et le raisonnement causal.

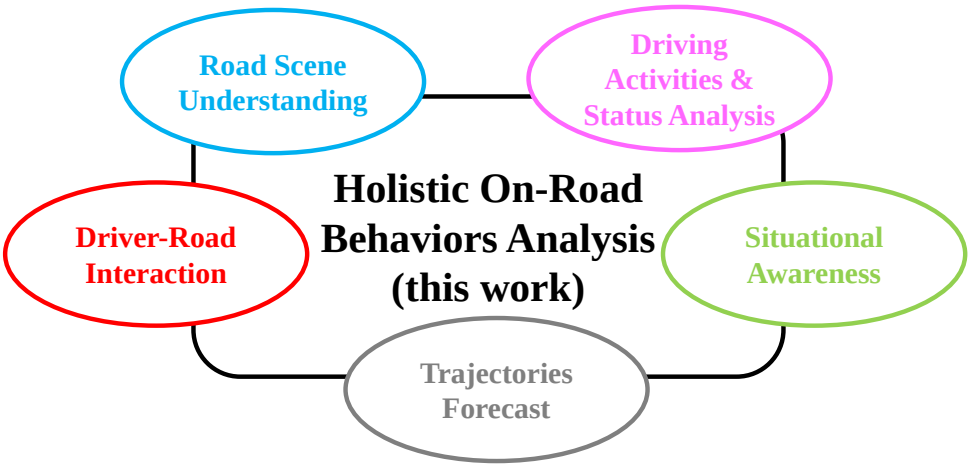

FIGURE 4.3 – Composants d'analyse du comportement sur route : conscience de la situation, interaction conducteur-route, analyse des activités de conduite et de l'état de la route, prévision de la trajectoire et compréhension de la scène routière, ainsi que des efforts holistiques qui permettent une compréhension générale des environnements routiers.

Ainsi, notre analyse du comportement routier comprend trois niveaux : couche supérieure, moyenne et inférieure (e.g. figure 4.4). À la couche la plus basse, les données telles que l'apparence, la profondeur et le mouvement sont collectées, prétraitées et étiquetées afin de préparer les paramètres pour les tâches des couches supérieures. À un niveau supérieur, plusieurs scénarios de compréhension de niveau intermédiaire peuvent se produire pour décrire et modéliser des données visuelles spatiales et/ou temporelles comme la détection et la reconnaissance d'agents routiers, la classification d'activités, l'identification des manœuvres, etc. Au plus haut niveau, un agrégat des résultats des couches inférieures est utilisé pour apprendre les comportements routiers globaux en incluant des paradigmes plus intuitifs comme la causalité. Dans ce travail, nous nous concentrons sur cinq tâches de haut niveau : 1. la connaissance de la situation, 2. l'interaction conducteur-route, 3. la compréhension de la scène routière, 4. la prévision de la trajectoire et 5. l'analyse de l'état de la conduite.

FIGURE 4.4 – Illustration des niveaux ascendants de l'analyse du comportement sur la route.

4.4.4 L'approche STAF

Vue d'ensemble. Dans la figure 4.5, nous présentons l'architecture de notre modèle STAF comprenant les étapes suivantes :

1. L'extraction des caractéristiques CNN à partir des images d'entrées.
2. L'épine dorsale du modèle STAF composée de 4 couches principales :
   — SA-LSTM (Spatial Attention-LSTM) layer.
   — Couche TAP (Temporal Average Pooling).
   — TA-LSTM (Temporal Attention-LSTM) layer.
   — Couche FC (Fully Connected).
3. Couche de perte d'entropie croisée pour mesurer la performance de la classification.

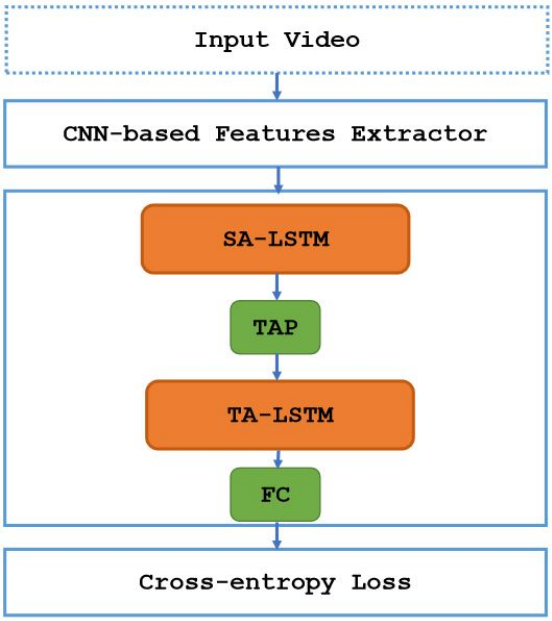

FIGURE 4.5 – Architecture du modèle STAF proposé. Après avoir extrait les caractéristiques CNN de l'image d'entrée, nous réalisons d'abord un LSTM spatial via SA-LSTM, puis une étape de pooling via TAP est réalisée, nous appliquons ensuite un LSTM temporel via TA-LSTM avant une étape entièrement connectée via la couche FC. Toutes les étapes sont suivies d'une couche de perte d'entropie croisée pour l'évaluation de la classification.

Modélisation de la compréhension du comportement. Inspiré par le mécanisme d'attention [147], nous proposons ici un réseau LSTM d'attention à têtes multiples qui permet au modèle de traiter de manière coopérative des données provenant de plusieurs sous-espaces de représentations différentes. Contrairement à l'auto-attention standard, l'état cellulaire c_{t-1} peut interroger ses propres k entrées passées avec une fenêtre d'attention de la taille de k (e.g. figure 4.6). Pour faciliter la notation, désignons par a_t le temps supplémentaire de concaténation des états c comme présenté dans l'équation 4.1 :

$$a_t = [c_{t-1}, c_{t-2}, \cdots, c_{t-k}] \tag{4.1}$$

L'indexation de ces états k dans le temps définit le codage positionnel (PE), utilisé à l'origine dans [147], qui utilise différentes sinusoïdes pour coder les positions, comme le montre l'équation (4.2)) :

$$\begin{aligned} PE(i, 2j) &= sin(\frac{i}{10000^{\frac{2j}{d_{model}}}}) \\ PE(i, 2j+1) &= cos(\frac{i}{10000^{\frac{2j}{d_{model}}}}) \end{aligned} \tag{4.2}$$

Où $i \in \{1 \cdots L\}$, L est la longueur de la séquence d'entrée, $j = \left[1, \cdots, \frac{d_{model}}{2}\right]$ et d_{model} est

la dimension de la sortie de chaque couche. Dans notre modèle STAF, et contrairement à ce qui a été proposé dans la littérature, nous n'avons pas besoin d'une phase aléatoire (formule proposée dans l'équation 4.7), puisqu'elle bénéficie d'une taille de fenêtre fixe qui permet d'omettre la dépendance de la fonction d'encodage au nombre variable de caractéristiques. Les états des cellules fenêtrées c_t sont concaténés aux caractéristiques au lieu de les ajouter comme dans [147].

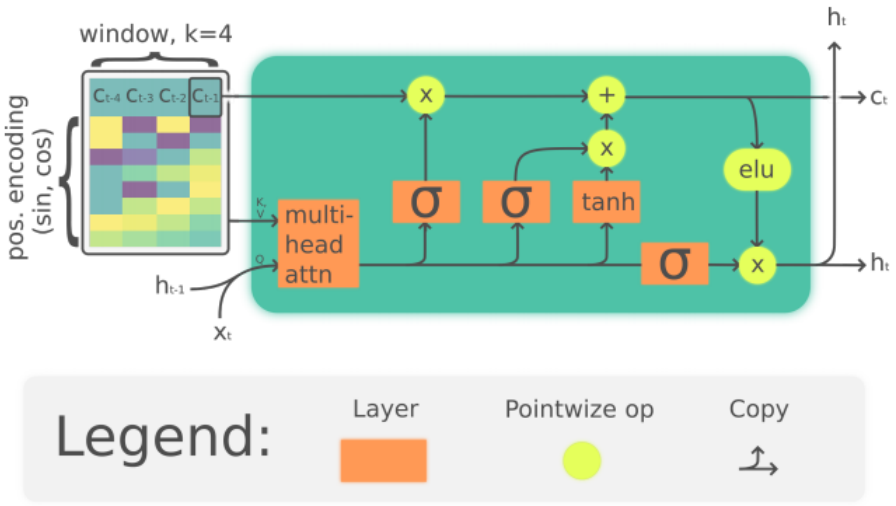

FIGURE 4.6 – Une implémentation concrète du LARNN [177] et son utilisation du mécanisme d'attention multi-têtes sur une file d'attente fenêtrée de ses états cellulaires passés. K, V et Q sont respectivement les clés d'entrée du mécanisme d'attention, les valeurs et une seule requête. Les états les plus récents $c_{[t-1\ldots t-k]}$ sont disposés dans une file d'attente "premier entré, premier sorti" de longueur k. Il y a aussi le fait non illustré que la requête est formulée à partir d'une couche de la concaténation de h_{t-1} et de x_t sur l'axe le plus interne (caractéristique).

Pour nos cellules LSTM, nous utilisons leur variante normalisée par lots (BN-LSTM) introduite dans [178] pour réduire le décalage interne des covariables [179]. Contrairement à la normalisation par lots utilisée dans les CNN, nous l'exploitons dans les connexions entrée-caché et caché-caché, au moyen de la fonction d'activation des unités linéaires exponentielles (ELUs) [180]. Étant donné une requête q et un ensemble de paires clé-valeur (K, V), l'attention peut être généralisée pour calculer une somme pondérée des valeurs dépendantes de la requête et des clés correspondantes. La requête détermine les valeurs sur lesquelles il faut se concentrer.

Plus particulièrement, le mécanisme d'attention est utilisé pour améliorer les caractéristiques des nœuds clés à chaque étape. Au lieu d'exécuter une seule fonction d'attention avec d clés, valeurs et requêtes à dimension du modèle, nous avons trouvé avantageux de projeter linéairement les requêtes, clés et valeurs h fois avec différentes projections linéaires apprises à

dk, dk et dv dimensions, respectivement. Sur chacune de ces versions projetées des requêtes, des clés et des valeurs, nous exécutons la fonction d'attention en parallèle, pour obtenir des valeurs de sortie à dv dimensions. Celles-ci sont concaténées et à nouveau projetées, pour obtenir les valeurs finales, comme le montre la figure 4.6 (équation 4.3) :

$$\begin{cases} i_t = \sigma(x_t U^i + h_{t-1} W^i) \\ f_t = \sigma(x_t U^f + h_{t-1} W^f) \\ o_t = \sigma(x_t U^o + h_{t-1} W^o) \\ \tilde{C}_t = tanh(x_t U^g + h_{t-1} W^g) \\ C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}) \\ h_t = tanh(C_t) * o_t \end{cases} \tag{4.3}$$

Où U et W sont les poids sous forme de matrice pour calculer l'attention multi-têtes de Q, K et V. v_t est présenté dans l'équation (4.4) :

$$v_t = [c_{t-1}, c_{t-2}, \cdots, c_{t-k}] \tag{4.4}$$

Où [.] signifie la concaténation le long de la "dimension des caractéristiques" [177]. La clé est donnée par l'équation (4.5) :

$$key = PE(v_t) \tag{4.5}$$

Où PE est *Positional Encoding*. La requête (the query) [177] est donnée par l'équation (4.6) :

$$query = W_{xh}([x_t, h_{t-1}]) \tag{4.6}$$

Le "Batch Normal" est présenté dans l'équation (4.7) :

$$BNELU_j(arg) = BatchNorm(elu(arg)) \tag{4.7}$$

a_t est présenté dans l'équation (4.8) :

$$a_t = S(\frac{query * BNELU_1(key)}{sqrt(d_k)}) * BNELU_2(values) \tag{4.8}$$

Où S est une fonction softmax à têtes multiples proposée dans le bloc "Transformer" introduit dans [147]. L'équation (4.9) présente c_t [177] :

$$c_t = c(h_t, a_t, z_t) \tag{4.9}$$

Où z_t est une couche cachée dans un codeur ou un décodeur typique de transduction de

séquence. Les cellules STAF utilisent l'attention multi-têtes du modèle Transformer pour exécuter plusieurs couches d'attention en parallèle sur ses valeurs d'état de cellule passées pendant un intervalle de temps limité. Par rapport à l'approche LSTM, STAF utilise la fonction d'attention multi-têtes à l'intérieur de chaque cellule LSTM afin qu'elle puisse interroger ses propres S valeurs passées non seulement avec attention mais aussi avec un fenêtrage restreint sur les S états cellulaires précédents les plus récents. Cela permet à la cellule d'effectuer des requêtes complexes pour exploiter ses mémoires internes précédentes, ce qui augmente l'effet à long terme du modèle STAF. Pour prouver l'efficacité de notre réseau STAF dans la compréhension à haut niveau de la route, nous avons réalisé plusieurs expériences sur les jeux de données JAAD [142, 143] et HDD [144]. Ces expériences avaient comme objectif la classification des comportements dits "de précondition" avant la traversée : se déplacer rapidement, se déplacer lentement ou rester debout.

Détails d'entrainement. Pour l'apprentissage de nos méthodes, nous avons utilisé la méthode d'optimisation Stochastic Gradient Descent (SGD) avec une mise à jour de l'apprentissage à pas constant, le paramètre γ étant fixé à 0.1. Le momentum, μ, et la décroissance du poids, ω, ont été fixés respectivement à 0.9 et 0.0005. La résolution d'apprentissage des images pour nos méthodes basées MobileNet et Inception est de 224×224 et 299×299 respectivement. L'Equation (4.10) présente les valeurs finales obtenues :

$$\begin{cases} i_t = \sigma(BN(x_t U^i + h_{t-1} W^i)) \\ f_t = \sigma(BN(x_t U^f + h_{t-1} W^f)) \\ o_t = \sigma(BN(x_t U^o + h_{t-1} W^o)) \\ \tilde{C}_t = tanh(BN(x_t U^g + h_{t-1} W^g)) \\ C_t = \sigma(BN(f_t * C_{t-1} + i^t * C^*)) \\ h_t = tanh(C_t * o_t) \end{cases} \tag{4.10}$$

4.4.5 Résultats expérimentaux

Jeux de données

Nous évaluons notre modèle STAF sur deux jeux de données public JAAD et HDD. Chacun d'entre eux est utilisé pour évaluer la performance du STAF sur la compréhension du comportement d'agents routiers. JAAD contient environ 300 séquences d'images d'une durée allant jusqu'à 15 secondes à 30 FPS acquises à l'aide de caméras positionnées à l'intérieur du véhicule et dans diverses conditions de circulation. Les paramètres de JAAD permettent de déterminer les comportements des piétons face aux informations fournies par les annotations spécifiques au véhicule (en particulier la marche, l'échouage, le regard vers le trafic

et le fait de ne pas regarder) ainsi que leur niveau d'attention correspondant, comme le fait de regarder, de se tenir sur le bord du trottoir et de vérifier le trafic. Ces derniers sont des indicateurs utiles de l'intention des agents de se déplacer, de traverser ou de se tenir debout, comme cela est présenté dans [171].

HDD comprend environ $10k$ séquences d'images, acquises dans différentes conditions météorologiques, et quatre couches d'annotation [144]. Dans nos expériences, nous utilisons deux couches appelées "Cause" et "Action orientée vers un but" (cause and goal-oriented action) proposant respectivement 5 et 11 classes pour les comportements des conducteurs (e.g. figure 4.7). L'ensemble de causes comprend un panneau, un embouteillage, un feu de circulation, une voiture garée et des classes de piétons qui sont les causes immédiates des classes de "Action orienté vers un but" : le franchissement d'une intersection, le virage à gauche, le virage à droite, le franchissement d'un passage piéton, le changement de voie à gauche, le changement de voie à droite, la fusion, le branchement sur la voie de droite, le franchissement d'un passage à niveau et le demi-tour (e.g. figure 4.7).

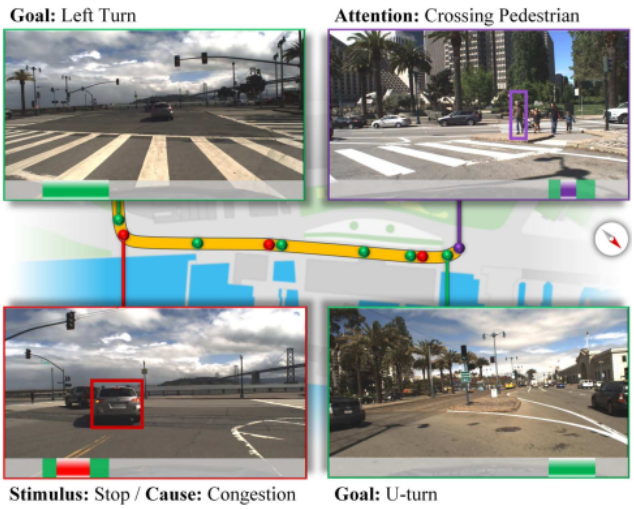

FIGURE 4.7 – Schéma d'annotation à 4 couches présentées dans [144] : Action orientée vers un but, Action guidée par un stimulus, Cause et Attention.

Analyse des résultats

Pour les trois ensembles de données dédiés aux tests (JAAD, HDD Goal-oriented et HDD Cause layer), et afin de présenter une comparaison équitable de notre framework STAF avec les approches de l'état de l'art [144, 142], nous proposons deux types d'extracteur de caractéristiques CNN : l'Inception-V3 [172] et MobileNets-V3 [181] pré-entraînés sur ImageNet [73].

Jeu de données HDD Cause Layer. Le Tableau 4.1 présente les résultats de la classifica-

tion des comportements de la couche "cause" qui sont les causes immédiates d'actions telles que la déviation ou l'arrêt. Nous avons observé que notre proposition, qui utilise les caractéristiques d'origine, surpasse les autres méthodes de l'état de l'art par une marge élevée de 38.96% en termes de mAP. Pour les classes "voiture garée" et "piéton", la performance de notre modèle STAF est plus importante. Pour la classe "piéton" par exemple, le conducteur réalise une manœuvre d'arrêt tout en effectuant des virages fréquents. Pour la classe "voiture garée", le conducteur dévie pour éviter une éventuelle collision. Visuellement, pour ces deux classes, l'action clé (key action) se produit sur une petite tranche de temps de la cause. Contrairement au modèle spatio-temporel, basé sur le réseau LSTM standard, qui ne parvient pas à modéliser le mouvement, notre modèle STAF, basé sur l'attention, peut se concentrer sur les images les plus importantes et ce, malgré leur durée par rapport à la séquence complète.

	sign (AP%)	congestion (AP%)	traffic light (AP%)	pedestrian (AP%)	parked car (AP%)	mAP%
Ramanishka [144]	46.83	39.72	45.31	2.15	7.24	28.25
STAF +MobileNet (ours)	71.97	74.21	**63.00**	47.54	45.81	60.50
STAF +Inception (ours)	**78.44**	**82.05**	61.84	**49.50**	**63.79**	**67.12**

TABLE 4.1 – Résultats pour la couche "Cause" de l'ensemble de données HDD. Les comportements de cette couche sont des causes immédiates pour les actions d'arrêt ou de déviation.

Jeu de données HDD Goal-oriented Layer. Les résultats de 11 actions de "Action orientée vers un but" sur le jeu de données HDD Gola-oriented layer sont présentés dans le tableau 4.2. La dernière colonne du tableau donne la valeur de précision mAP pour toutes les méthodes. Tout d'abord, nous décrivons les "baselines" utilisées dans cette expérience. La première baseline (*'Random'*) attribue simplement des étiquettes de comportement aléatoires à chaque image et sert de limite inférieure à la performance du modèle. La deuxième baseline ('CNN pool') encode chaque image en extrayant des caractéristiques convolutionnelles à l'aide d'un réseau InceptionResnet-V2 et en les regroupant spatialement dans un vecteur de longueur fixe. Ces représentations groupées des images sont introduites séquentiellement dans le réseau LSTM pour prédire l'étiquette du comportement. La troisième baseline ('Capteurs') utilise uniquement les données du capteur comme entrée pour le réseau LSTM. La méthode suivante ('CNN conv') est une variante de la deuxième méthode : au lieu de regrouper spatialement les encodages CNN des caractéristiques, nous avons utilisé un petit "convnet" pour réduire la dimensionnalité des encodages CNN des images avant de les passer dans le réseau LSTM. Enfin, la méthode "CNN+Sensors" ajoute des données de capteurs à la méthode "CNN conv".

STAF améliore les résultats de l'état de l'art avec une marge de précision (mAP) de 28.68%. Le modèle LSTM standard [144] n'a pas réussi à discerner les catégories *'merge'* et *'lane branch'* car elles sont assez similaires dans le domaine temporel et le taux de confusion entre elles devrait donc être plus élevé. Les améliorations apportées par notre modèle STAF prouvent sa capacité à repérer sélectivement les segments les plus pertinents de la séquence vidéo. Cette dernière classe "cause" permet une meilleure représentation pour toutes les autres classes évidentes. Il est intéressant de noter que le "modèle aveugle" (sans l'intervention de la caméra) peut deviner avec succès le *'passage d'intersection'* car la plupart de ces passages se produisent selon un schéma très spécifique : "décélération/attente/accélération". Un *'passage à niveau'* est étonnamment difficile à deviner pour le modèle CNN, car ce type de comportement comprend non seulement le passage des voies ferrées aux endroits désignés qui ont des caractéristiques visuelles discriminantes, mais aussi le passage à niveau du tramway. La confusion des classes de comportement avec une classe d'arrière-plan reste la source d'erreurs la plus fréquente pour toutes les couches.

	right turn	left turn	inter-section passing	rail-road passing	left lane branch	right lane change	left lane change	right lane branch	cross-walk passing	merge	u-turn	mAP%
Ramanishka[144]	77.47	76.16	76.79	3.36	25.47	23.08	41.97	1.06	11.87	4.94	17.61	32.71
STAF + MobileNet (ours)	82.41	80.52	75.27	29.98	**42.69**	50.77	69.03	42.63	39.05	56.32	**69.11**	57.98
STAF + Inception (ours)	**87.91**	**89.37**	**84.80**	**32.31**	41.05	**62.08**	**69.52**	**44.89**	**51.55**	**60.47**	68.58	**61.39**

TABLE 4.2 – Évaluation des performances d'un ensemble d'actions orientées vers un but, à partir d'un ensemble de données HDD. STAF+MobileNet (le nôtre) : 224×224 et STAF+Inception (le nôtre) : 299×299. Les valeurs sont exprimées en AP%.

Jeu de données JAAD. Nous avons mis en place aussi des expériences sur le jeu de données JAAD [142, 143]. Il s'agit du premier jeu de données consacré à la compréhension du comportement (conscience situationnelle) du véhicule autonome en ce qui concerne les piétons aux points de passage. JAAD comprend des informations comportementales et contextuelles ainsi que des informations de suivi de $337k$ piétons détectés à partir de 346 séquences vidéo recueillies en Europe et en Amérique du Nord. Grâce à l'outil BORIS (Behavioural Observation Research Interactive Software) [182], des annotations comportementales ont été fournies pour capturer les actions du conducteur et des piétons. L'étiquette d'action du conducteur comprend le déplacement rapide/lent comme premier état et le ralentissement/l'accélération comme réponse à l'environnement. Les actions des piétons sont classées en 3 ensembles : (i) Précondition : activités avant de traverser, comme se tenir debout, se

déplacer rapidement ou lentement (*ii*) Attention : indices gestuels et temporels qui servent à comprendre si la personne est consciente de l'approche du (ego-)véhicule et (*iii*) Réponse : pour vérifier si le piéton qui traverse réagit au comportement du véhicule en approche en identifiant l'une des actions/gestes élémentaires suivants : dégager la voie, accélérer, ralentir, geste de la main, hochement de tête et arrêt. Des annotations démographiques sont aussi fournies pour chaque personne détectée, telles que la classe d'âge et le sexe. En plus des étiquettes comportementales, des informations contextuelles sur la scène sont fournies : (*i*) configurations de la route, (*ii*) signaux de circulation, (*iii*) conditions météorologiques et (*iv*) heure de la journée afin de décrire les conditions d'éclairage.

Nous avons effectué une étude comparative des performances de notre modèle STAF avec un réseau LSTM générique sur JAAD. Nous nous sommes focalisés sur la classification des comportements dits de précondition avant la traversée à savoir : se déplacer rapidement, se déplacer lentement, ou rester debout. Les résultats obtenus prouvent encore une fois l'efficacité de notre modèle STAF qui surpasse les approches de pointe avec une marge mAP de 13% comme le montre le tableau 4.3.

	Without CI (mAP%)	**With CI** (mAP%)
Rasouli[142]	71.3	81.5
LSTM-baseline	51.4	56.0
STAF + MobileNet (ours)	82.9	84.1
STAF + Inception (ours)	**85.0**	**88.3**

Table 4.3 – Évaluation des performances sur JAAD et comparaison entre l'information visuelle et sa combinaison avec l'information contextuelle (Contextual Information (CI)).

Analyse & Synthèse. Nous avons présenté dans cette section une nouvelle approche innovante appelée STAF pour une compréhension de haut niveau des activités des agents routiers. Nous avons présenté une comparaison entre notre modèle STAF et les approches de l'état de l'art par la proposition de deux types de caractéristiques CNN : STAF +MobileNet et STAF +Inception. Les résultats de la classification des comportements de la couche "cause", qui sont les causes immédiates d'actions telles que dévier ou s'arrêter, montrent que notre approche STAF utilisant les caractéristiques d'inception surpasse les autres méthodes par une marge mAP importante de $38,96\%$, en particulier pour les classes "voiture garée" et "piéton". Le modèle LSTM standard proposé dans la littérature n'a pas réussi à discerner les catégories "fusion" et "embranchement" car elles sont assez similaires dans le domaine temporel. Les résultats obtenus par notre modèle STAF pour 11 actions orientées vers un but précis améliorent les résultats de l'état de l'art avec une marge mAP de $28,68\%$. Ces

travaux de recherche ont fait l'objet de plusieurs publications dans [16], [17] et [27].

4.5 Conclusion

Nous avons présenté dans ce chapitre une succession d'approches classiques et par deep learning de détection et tracking temps-réel d'objets et d'analyse et compréhension de scènes routières. Dans un premier temps, notre nouvelle approche Faster-DPM a été présentée et appliquée pour deux applications différentes à savoir la détection de piétons et l'asservissement visuel d'un drone. Par la suite, nous avons présenté une contribution basée sur trois approches de deep learning pour la détection, l'estimation de la distance et le suivi d'objets. Nous avons développé la détection d'objets par les algorithmes SSD et YOLOv3 afin de déterminer quel algorithme est le plus adapté à notre application. L'estimation de la distance basée sur l'algorithme monodepth a été développée. Nous avons fusionné les deux approches de détection et d'estimation de distance afin de partager les couches d'extraction de caractéristiques, ce qui avait améliorer leur efficacité. Nous avons validé notre approche sous différents jeux de données routiers et ferroviaires (Cityscape, KITTI et vidéos du tramway de Rouen) ainsi que dans des conditions réelles de circulation dans le centre-ville de Rouen.

Concernant l'analyse et compréhension de scènes, nous avons présenté notre nouvelle approche innovante STAF dédiée à l'analyse et compréhension de haut niveau des activités des agents routiers. STAF dépasse de loin en précision les approches de la littérature dont celui le plus connu, LSTM. Il apporte une amélioration de la précision avec une marge mAP de $28,68\%$ toutes situations confondues. Enfin, pour prouver l'efficacité de notre réseau STAF dans la compréhension de haut niveau des scènes routières, nous avons validé notre approche sur un jeu de données JAAD, plus adapté à la compréhension de la conscience situationnelle des véhicules autonomes. Là aussi notre modèle STAF a surpassé les approches de l'état de l'art avec une marge mAP de 13%.

Nous préparons aussi nos futurs travaux orientés vers l'analyse et la compréhension de scènes routières complexes par l'apprentissage profond multitâche (Multi-Task Learning (M-TL)). L'approche multitâche à "grain fin" en cours de développement (thèse CIFRE d'Alexandre Evain sur le train autonome monorail) vise à développer un modèle innovant d'apprentissage profond convolutifs et/ou récurrent. Les tâches sont à définir de telle sorte qu'elles répondent à des problématiques distinctes mais complémentaires. Bien entendu, le modèle STAF fera partie de ces tâches pour le développement de notre M-TL.

Cependant, dans le cadre du projet AntiHPert [3] ayant comme objectif l'amélioration des activités humaines dans l'industrie 5.0, nous allons développer une nouvelle approche dédiée à l'analyse du comportement d'agents dans les chaines de production. L'usage de notre méthode STAF permettra de définir des modèles de comportement d'agents. Cela représentera l'objectif des travaux du postdoctorant qui commencera en janvier 2023.

3. Opérateur 4.0 et anticipation dynamique de ses perturbations dans les ateliers de production.

Chapitre 5

Détection, Localisation et Suivi d'Objets pour un Fauteuil Roulant Intelligent

Sommaire

5.1 Introduction

Parmi les tâches de vision utiles à la perception d'environnement, nous pouvons citer la détection et localisation d'objets, l'estimation de la profondeur, le tracking et la segmentation sémantique. Toutes ces tâches sont importantes et nécessaires à une bonne compréhension des scènes complexes. Ce chapitre est dédié à la smart mobilité indoor et outdoor d'un véhicule autonome de type fauteuil roulant intelligent. Ces travaux ont été réalisé dans le cadre du projet ADAPT [1] (Assistive Devices for empowering disAbled People through robotic Technologies) dédié au développement d'un fauteuil roulant intelligent améliorant l'autonomie de personnes à mobilité réduite. Nous présentons une succession d'approches clés de la perception d'environnement à savoir : détection, localisation, tracking d'objets et segmentation sémantique. Ce chapitre comprend donc 3 parties : (*i*) détection d'objets par vision omnidirectionnelle appliquée à la détection et reconnaissance de visage et à la navigation autonome d'un fauteuil roulant, (*ii*) détection, localisation et tracking d'objets par deep learning pour fauteuil roulant intelligent et (*iii*) segmentations sémantiques pour fauteuil roulant intelligent naviguant sur les trottoirs.

Dans un premier temps, une nouvelle approche dédiée à la détection d'objets par vision omnidirectionnelle sera présentée. Elle sera déclinée en deux axes complémentaires : la détection et reconnaissance de visages pour des applications de contrôle d'accès biométriques d'un robot mobile et la navigation autonome d'un fauteuil roulant intelligent. Par la suite, nous présenterons notre démarche de détection, localisation et tracking d'objets par deep learning pour un fauteuil roulant intelligent. Afin d'assurer une navigation autonome, une localisation temps-réel indoor et outdoor du fauteuil roulant par SLAM visuel sera aussi présentée. La troisième et dernière partie affinera davantage notre démarche et sera dédiée à la segmentation sémantique d'un fauteuil roulant naviguant sur les trottoirs. Deux approches seront présentées : une temporelle et l'autre rapide ainsi que le développement d'un nouveau dataset dédié à la smart mobilité sur les trottoirs. L'ensemble des approches développées seront testées et validées sur notre plateforme de fauteuil roulant intelligent.

1. `http://adapt-project.com`

5.2 Détection et reconnaissance de visages par vision omnidirectionnelle

5.2.1 Contexte & objectifs

Ces travaux ont été réalisés dans le cadre du projet NOBA (NOmad Biometric Authentication) qui visait le contrôle d'accès biométrique pour un robot mobile en se basant sur la reconnaissance de la marche, du visage et de l'iris. Ces travaux ont été réalisés par le postdoctorant Jin-xin Liu ainsi que plusieurs stagiaires Master 2. Nous avons développé une plateforme innovante d'authentification biométrique basée sur la reconnaissance du visage. Le système de vision était innovant car il combine un capteur catadioptrique et une caméra Pan Tilt Zoom (PTZ) pour l'authentification biométrique. Le capteur catadioptrique est calibré et utilisé pour détecter et suivre les régions d'intérêt (RoI) dans son champ de vision de 360°, en particulier les régions du visage FOV (Field Of View). En utilisant une stratégie de calibration conjointe, les paramètres de la caméra PTZ sont automatiquement ajustés par le système afin de détecter et de suivre la RoI du visage dans une résolution plus élevée.

5.2.2 Reconnaissance de visage

Dataset biométrique NOBA. Nous avons développé notre propre jeu de données dédié à la détection et reconnaissance de visage appelé NOBA comprenant 3 sous-ensembles de données : la marche, le visage et l'iris. Les données ont été collectées via deux caméras PTZ et infrarouge. Le dataset NOBA a été développé en prenons en compte plusieurs paramètres : variations des poses de la tête et de la direction des yeux, diverses expressions faciales et diverses conditions d'éclairage. Cependant, comme beaucoup d'autres datasets de visages, nous avons eu le problème de Small Sample Size (SSS) - lorsque la taille de l'échantillon est inférieure à sa dimensionnalité, la matrice de dispersion intra-classe est singulière -. Notre objectif était donc de développer de nouvelles approches palliant à cette problématique.

Détection de visage par CS et LBP. Le Compressed Sensing (CS) dans Candès, Tao et Romberg [183, 184, 185] et Donoho [186], appelé aussi détection comprimée, signifie que le signal peut être récupéré à partir de beaucoup moins de mesures que ce qui est habituellement considéré comme nécessaire. Un framework CS est constitué de deux parties : le processus d'échantillonnage, qui n'est d'autre que le signal original $x \in R^N$ et la récupération CS (CS recovery process). Nous avons donc utilisé deux approches de reconnaissance de visage : Local Binary Patterns [187] (LBP) et CS. Dans notre démarche, l'approche LBP peut être considérée comme une méthode d'extraction de caractéristiques tandis que l'approche CS est considérée comme une stratégie de classification spéciale. Il existe plusieurs approches ty-

piques de reconnaissance faciale avec CS dont Sparsity Preserving Projections (SPP) [188]. Contrairement à de nombreuses techniques existantes, telles que la projection préservant la proximité (LPP : Local Preserving Projection) et l'incorporation (NPE : Neighborhood Preserving Embedding), la SPP vise à préserver la relation reconstructive éparse des données.

Dataset biométrique hétérogène. Nous avons développé un nouveau dataset hétérogène comprenant 3 jeux de données : NOBA, ORL [187] et CSU [187]. Il comprend 31 objets distincts et dix images différentes par objet. En général, il y a au moins deux étapes pivots dans tout processus de reconnaissance de formes, l'une est l'extraction de caractéristiques et l'autre est la classification. Les LBP et le Sub-Sampling (SS) sont utilisées comme deux méthodes d'extraction de caractéristiques. Nearest Neighbor (NN) et CS quant à elles sont utilisées pour la classification. La figure 5.1 illustre un exemple de résultats obtenus sur les deux datasets NOBA et "hétérogène" avec les trois combinaisons : LBP-NN, SS-CS et LBP-CS.

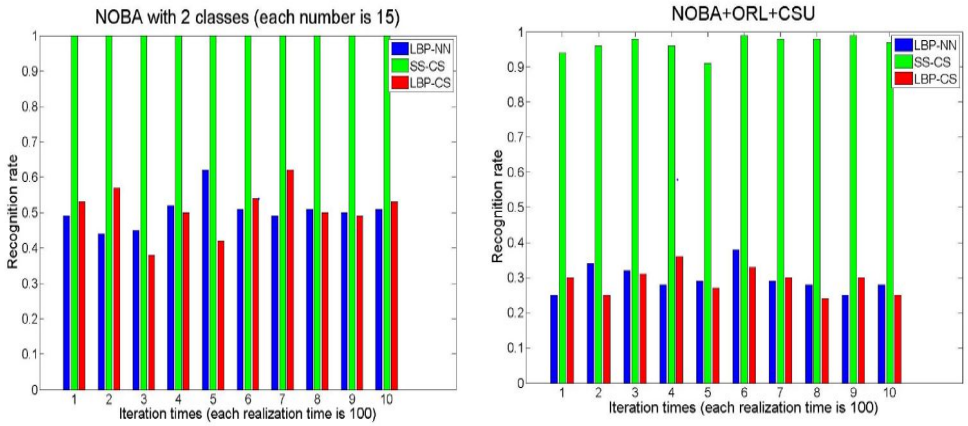

FIGURE 5.1 – Reconnaissance du visage. A gauche : dataset NOBA avec 2 classes (15 pour chacune). A droite : Dataset hétérogène NOBA+ORL+CSU.

5.2.3 Détection et tracking de visage par fusion multicapteurs

Détection et reconnaissance de visage. Les déformations optiques causées par le système catadioptrique ne nous permettaient pas d'utiliser les algorithmes de détection de visages sur les images brutes. Nous avons donc effectué des transformations géométriques pour obtenir des images panoramiques proches de celles perspectives. Nous avons utilisé l'approche de détection de visages basée Viola-Jones [189]. Ce dernier utilise des caractéristiques pseudo-Haar [190, 191] et un classifieur en cascade pendant l'étape d'apprentissage puis compare les images soumises et effectue la phase de détection. Nous avons aussi fait appel à plusieurs approches de détection de visage : Eigenface [192], LBP, le classifieur CS, etc. Nous avons combiné les deux méthodes d'extraction de caractéristiques, LBP et SS, et nous avons com-

paré la combinaison avec les algorithmes LBP et NN. Le système se concentre d'abord sur l'image omnidirectionnelle qui est dépliée (unwrapping) [191, 35, 26]. L'algorithme effectue une différenciation de trames (à t et $t-1$), si la différence entre deux valeurs successives dépasse un seuil, le pixel appartient à une zone d'intérêt. S'ensuivent l'érosion et la dilatation des groupes de pixels ainsi déterminés pour filtrer les pixels dispersés. Cette méthode permet de définir des groupes de pixels correspondants à des objets en mouvement dans le flux vidéo [191, 35, 26].

L'obtention du masque binaire (ou blob) nous a permis de focaliser la caméra PTZ sur l'objet détecté. Nous appliquons par la suite l'algorithme de Violas Jones pour la détection de visage. Pour la reconnaissance de visage, nous avons utilisé l'algorithme Eigenface [192]. Notre approche de détection et reconnaissance de visage comprenait 5 étapes : 1. détecter un nouveau visage à identifier et le transformer en ses composantes propres. 2. déterminer quelle classe de visage fournit la meilleure description de l'image d'entrée. 3. choisir un seuil qui définit la distance maximale permise par rapport à toute classe de visage ainsi qu'un seuil qui définit la distance maximale autorisée de l'espace des visages. 4. si le nouveau visage est classé comme un individu connu, cette image peut être ajoutée à l'ensemble original d'images de l'ensemble "familier", et les visages propres peuvent être recalculés. 5. donner la possibilité de modifier l'espace des visages au fur et à mesure que le système rencontre des visages connus. Pour tester les performances de notre approche, nous avons utilisé un dataset de visages public (Olivetti Research Laboratories (ORL)) comprenant 40 objets et 10 images échantillons par objet. Nous avons fusionné ORL avec notre dataset NOBA (dans les mêmes conditions) pour obtenir un dataset hétérogène. Trois classes ont été créées : 1. expression faciale, 2. position/orientation de la tête et des yeux et 3. conditions d'éclairage.

Résultats expérimentaux. Nous avons proposé un système de vision unique et efficace pour détecter et suivre automatiquement les RoIs de visages à un niveau de zoom plus élevé. Les résultats expérimentaux démontrent une amélioration significative de la précision en fournissant un regard plus proche de la cible à des fins de reconnaissance. L'intégration de processus d'authentification tels que la détection, le suivi et la reconnaissance du visage rend le système autonome pour l'authentification biométrique. La méthode de reconnaissance des visages CS (SS-CS) est généralement meilleure que celle traditionnelle (LBP-NN) qui est basée sur la distance euclidienne. La combinaison des approches LBP et CS n'est pas satisfaisante sur le dataset "hétérogène" car la redondance est un facteur important dans la méthode CS, mais pas dans l'approche LBP. SS-CS aura une bonne performance dans la reconnaissance de visages pour un dataset hétérogène. Ces travaux de recherche ont fait l'objet de plusieurs publications dans [26], [35] et [36].

5.3 Navigation autonome par vision omnidirectionnelle d'un fauteuil roulant

5.3.1 Contexte & objectifs

Dans le cadre des travaux de recherche du projet COALAS (Cognitive Assisted Living Ambiant System) sur le développement d'un fauteuil roulant intelligent et particulièrement ceux de Yassine Nasri et Emmanuel René, Ingénieurs de recherche, nous avons développé un nouveau ADAS de navigation autonome par odométrie visuelle à partir d'un capteur de vision omnidirectionnelle. L'objectif consistait à utiliser l'information de vision pour estimer le déplacement du fauteuil roulant intelligent. L'estimation du déplacement entre deux images est réalisée à partir d'appariements de points d'intérêts et d'estimation de la géométrie épipolaire. Les campagnes de mesures ont été réalisées dans notre laboratoire IRSEEM, équipé d'un système de capture du mouvement Vicon, pour deux types de trajectoires : en ligne droite et en courbe. Les résultats expérimentaux étaient très encourageants et confirmaient la pertinence de l'approche au regard des problématiques d'estimation de mouvement basée vision.

5.3.2 Navigation autonome par vision omnidirectionnelle

L'algorithme déployé sur le fauteuil roulant est divisé en 3 étapes clés : 1. détection et appariement des points d'intérêt entre deux images consécutives, 2. estimation de la matrice essentielle à partir des points reprojetés sur la sphère et 3. estimation de pose (levée de l'ambigüité sur la translation par triangulation).

Détection et appariement des points d'intérêt. Afin d'estimer le mouvement du fauteuil roulant, il fallait d'abord détecter et mettre en correspondance des points d'intérêts dans les images prises itérativement. La mise en correspondance est une étape fondamentale dont dépend la précision de l'algorithme. La contrainte majeure est de pouvoir détecter des points d'intérêts par un détecteur/descripteur adapté aux images omnidirectionnelles. Ces points d'intérêts détectés sont ensuite appariés directement sur la sphère de représentation. Nous avons évalué plusieurs méthodes : ORB, SIFT et SURF [193]. La méthode SURF (Speeded Up Robust Features) était la plus adaptée aux images omnidirectionnelles [194, 195] et a permis d'obtenir le meilleur compromis entre rapidité d'extraction et performance d'appariements.

Estimation de mouvement. La matrice essentielle contient la transformation géométrique entre les deux sphères de représentation. En adaptant l'hypothèse d'un mouvement plan, le problème se résume donc à estimer un vecteur de translation tW défini au facteur d'échelle prêt et la matrice de rotation $R_W = R_W(z)$. La résolution de ce problème est bien connue en vision par ordinateur via les algorithmes des "$8-points$" et RANSAC [196]. La méthode

consiste à sélectionner aléatoirement huit points parmi les couples de points appariés sur les sphères pour une première estimation de la matrice essentielle. Les autres points sont alors testés pour vérifier s'ils respectent la contrainte de la géométrie épipolaire à partir de la matrice essentielle précédemment estimée. Ainsi les points respectant la contrainte épipolaire sont sauvegardés.

Ce processus est répété plusieurs fois, afin d'obtenir la meilleure estimation de la matrice essentielle possible. L'ensemble de ces points cohérents est alors réutilisé pour le calcul final de la matrice essentielle. Nous avons ensuite appliqué une décomposition en valeur singulière pour extraire le mouvement en translation et en rotation donnant deux rotations $R1_W(z)$ & $R2_W(z)$ et deux translations $t1_W$ & $t2_W = -t1_W$. L'ambiguïté est levée en triangulant les points d'intérêt en 3D et en déterminant la transformation $T = \{R_W(z), t_W\}$ minimisant l'erreur de reprojection (e.g. figure 5.2)). Cet algorithme fonctionne de manière itérative et estime pour chaque nouvelle image le mouvement par rapport à celui calculé précédemment. La trajectoire peut donc être reproduite à partir des calculs de poses successives.

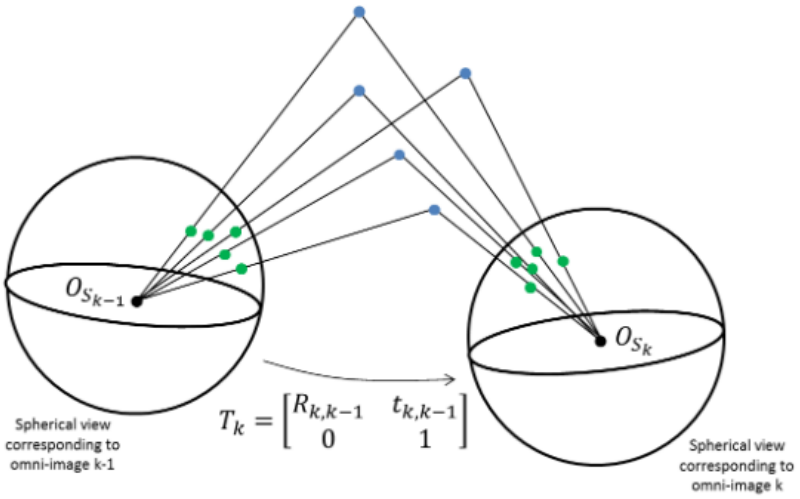

FIGURE 5.2 – Processus d'appariement des points d'intérêt à partir de leur représentation sphérique. Soit deux images prises à deux instants consécutifs k et $k+1$, l'appariement (en vert) est effectué pour estimer la transformation géométrique Tk entre les deux vues à partir de la matrice essentielle.

En collaboration avec notre partenaire sur le projet COALAS, l'Université de Kent (UK), nous avons par la suite développé un algorithme d'évitement d'obstacles basé fusion de capteurs infrarouges et ultrasons. La campagne de tests s'est déroulée dans notre laboratoire équipé d'un système de capture du mouvement Vicon, utilisé entre autre pour estimer le facteur d'échelle. Afin d'évaluer les performances de l'algorithme, deux scénarios ont été définis : ligne droite et courbe en forme de Z. Les trajectoires de référence sont reproduites

en noir, celles estimées par l'algorithme sont en rouge. Comme métrique, nous utilisons la distance euclidienne entre la position estimée par l'algorithme et le point le plus proche de la trajectoire de référence. L'évolution de l'erreur en fonction du nombre d'images est représentée sur la figure 5.3. Les résultats pour les deux trajectoires (en ligne droite et courbe) sont prometteurs. Nous remarquons néanmoins un écart entre la trajectoire de référence et celle reproduite par vision omnidirectionnelle. Ceci est en cohérence avec notre approche qui estime un déplacement relatif. Les erreurs se cumulent donc entre deux estimations successives et ce, sur la totalité de la trajectoire. Ces travaux de recherche ont fait l'objet de publications dans [24] et [197].

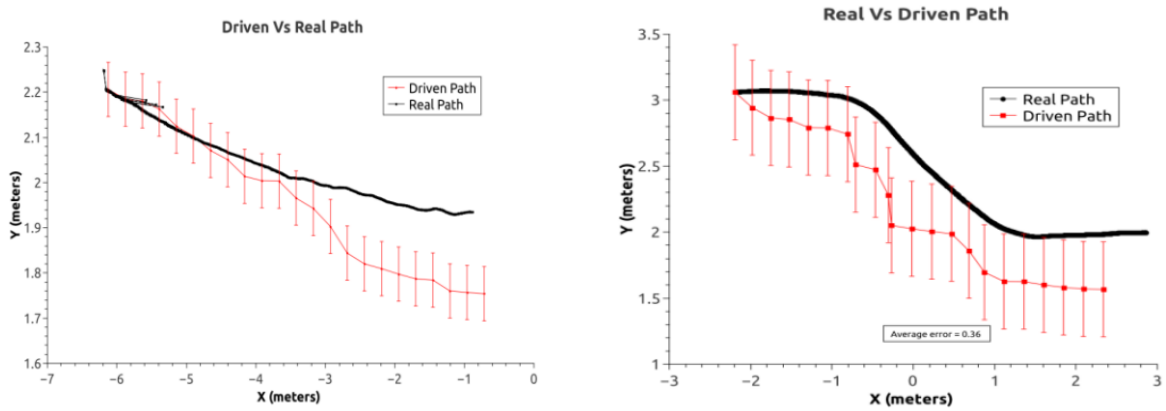

FIGURE 5.3 – Comparaison entre la trajectoire reproduite par vision omnidirectionnelle (rouge) et celle de référence (en noir) ainsi que l'intervalle d'erreurs. A gauche : trajectoire en ligne droite, A droite : trajectoire en courbe en forme de Z.

5.4 Détection, localisation et suivi d'objets pour un fauteuil roulant intelligent

5.4.1 Contexte & objectifs

Dans le cadre du projet de recherche ADAPT [2] (Assistive Devices for empowering disAbled People through robotic Technologie) sur le développement d'un fauteuil roulant intelligent, nous avons mené plusieurs travaux via quatre ingénieurs de recherche (Louis Lecrosnier, Aristide Laignel, Edgard Petit et Vishnu Pradeep) et huit stagiaires Master 2 qui portaient sur le développement d'un système ADAS pour un fauteuil roulant électrique intelligent afin d'améliorer l'autonomie des personnes à mobilité réduite. Dans un premier temps, nous nous sommes focalisés sur la détection, localisation et tracking d'objets dans l'environnement indoor du fauteuil roulant, à savoir : les portes et poignées de portes. L'objectif de ces travaux

2. http ://adapt-project.com

était de fournir une couche de perception au fauteuil roulant, permettant ainsi la détection de ces points clés dans son environnement immédiat (ellipsoïde de présence) ainsi que la construction d'une carte sémantique éphémère à courte durée de vie. Cela s'est traduit par une détection d'objets basée YOLOv3, une estimation de la profondeur ainsi qu'un suivi d'objets 3D basé sur SORT [57]. Par la suite, nous avons développé une approche de localisation temps-réel du fauteuil roulant basée sur un SLAM visuel (vSLAM) où nous avons comparé deux approches de la littérature. Afin de valider tous les développements, nous avons réalisé différentes expériences dans un environnement intérieur contrôlé en utilisant notre propre jeu de données IRSEEM.

5.4.2 Détection d'objets

Nous utilisons une plateforme de fauteuil roulant homologué à laquelle nous avons remplacé toute l'électronique propriétaire par notre système. Le fauteuil est ainsi instrumenté par un ensemble de capteurs dont deux caméras RealSense D435 (images RGB de profondeur) et T265 (odométrie visuelle temps-réel). La détection d'objets est basée YOLOv3. Les points clés que nous détectons sont les portes et les poignées de porte. Comme ces classes sont sous-représentées dans les jeux de données classiques dédiés à la détection d'objets par YOLOv3 (e.g. ImageNet [73]), nous n'avons pas pu trouver un dataset public avec une représentation suffisante des poignées de porte. Nous avons donc développé notre propre jeu de données personnalisé. Dans un premier temps, nous avons extrait 755 images de portes du dataset $MCIndoor20000$ [198], composé d'images étiquetées contenant divers objets d'intérieurs. Par la suite, nous avons collecté 1885 images, que nous avons combiné avec les images de portes du jeu de données $MCIndoor20000$. Nous avons supervisé l'étiquetage de 2640 images provenant des deux jeux de données combinés en utilisant un outil d'étiquetage semi-automatique que nous avons développé. Pour ce processus d'apprentissage par transfert, nous avons entrainé uniquement les couches de classification du réseau. Des résultats quantitatifs sur la détection des portes et poignées de portes sont présentés dans le tableau 5.1.

Class	Mean IOU	Median IOU	Std. dev. IOU	Precision (%)	Recall (%)
door	0.89	0.89	0.05	0.90	0.80
handle	NA	NA	NA	0.85	0.29

TABLE 5.1 – Evaluation de la détection d'objets : portes et poignées de portes.

5.4.3 Estimation de la profondeur

Pour l'évaluation de la distance entre le fauteuil roulant et les différents objets, nous avons utilisé des modèles CNN dédiés à l'estimation de la profondeur déjà validés dans les chapitres précédents comme Monodepth [55], Monodepth2 [6][129] et MADNET [30][56]. Nous avons effectué des mesures de distance en utilisant directement la caméra RealSense D435 sans aucun modèle d'apprentissage profond et ce, en raison des contraintes embarquées liées au fauteuil roulant. Nous comparons par la suite la distance obtenue via la caméra avec les données de vérité terrain issues du système de capture de mouvement Vicon (précision millimétrique). La figure 5.4 montre un exemple de résultats obtenus sur les distances mesurées, ainsi que l'erreur absolue de mesure de la distance, en comparaison avec la vérité terrain.

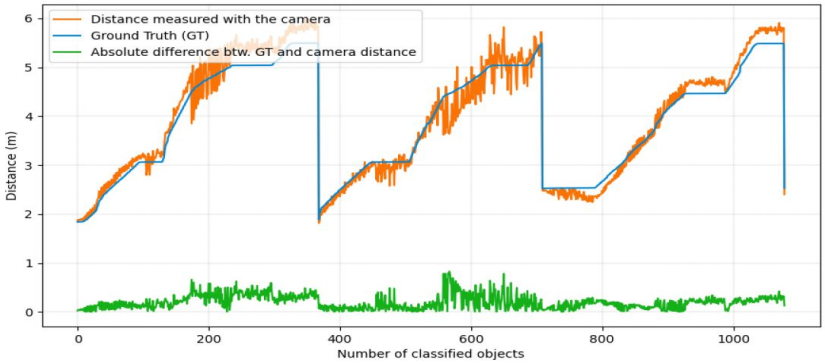

FIGURE 5.4 – Distance mesurée entre les objets correctement classés et la caméra RealSense $D435$ (orange), distance de vérité terrain (bleu), valeur absolue de la différence entre la distance vérité terrain et celle estimée par la caméra (vert). Les objets détectés sont triés par classe, puis par ordre de détection.

5.4.4 Tracking d'objets

La détection de portes et de poignées de portes se fait sur la base d'images couleurs alors que l'estimation de la distance est effectuée à l'aide des images de profondeur. A partir d'un objet classé et d'une profondeur associée, la position de l'objet 3D est ajoutée à la carte sémantique en utilisant les données d'odométrie fournies par la deuxième caméra $T265$. Nous utilisons la combinaison de la position et de la profondeur associées à chaque élément extrait par l'algorithme de détection, pour effectuer un tracking d'objets. Ce dernier a pour objectif d'asservir le déplacement du fauteuil vers les portes. Pour le tracking, nous faisons appel à l'algorithme SORT [57] déjà présenté dans les chapitres précédents. SORT inspecte les objets détectés et détermine si un objet donné est nouvellement vu, ou si le mouvement de l'objet est une conséquence des mouvements du fauteuil roulant. Enfin, nous utilisons les données odométriques de la caméra $T265$ pour estimer le déplacement du fauteuil roulant.

Nous combinons ces données avec la position de l'objet (classe et profondeur associée) afin de visualiser une carte sémantique éphémère 3D de l'environnement contenant les objets détectés et suivis. Une carte de coûts locaux (local cost map) construite à partir des images de profondeur permet d'éviter dynamiquement les obstacles à mesure qu'ils se présentent sur la trajectoire du fauteuil, y compris s'ils ne sont pas encore cartographiés. La figure 5.5 illustre un exemple de carte 2D sémantique construire dans un environnement indoor. Ces travaux de recherche ont fait l'objet des publications dans [18], [19] et [40].

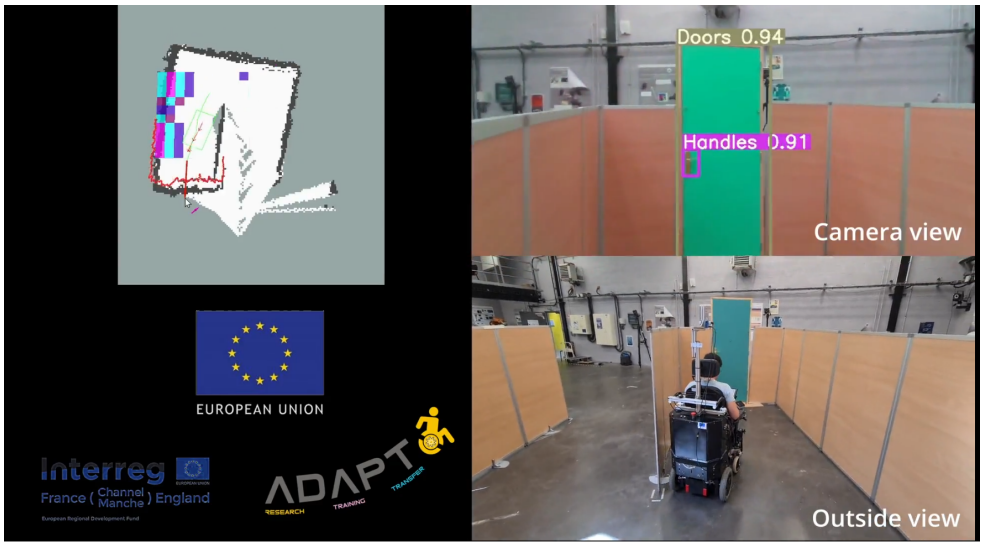

FIGURE 5.5 – Exemple d'essai de validation en indoor. Détection et tracking temps-réel d'objets, carte sémantique 2D, SLAM et planification de trajectoire.

5.4.5 Localisation du fauteuil roulant par SLAM Visuel

Contexte & objectifs

Dans le cadre du projet ADAPT, nous avons par la suite développé un benchmarking de localisation du fauteuil roulant par SLAM visuel (vSLAM) en comparant deux approches de la littérature : ORB-SLAM2 [199] et RTAB-Map [200]. L'ORB-SLAM2 a été implémenté en prenant en compte des caméras monoculaires, stéréoscopiques et RGB-D tandis que le RTAB-Map SLAM avec uniquement des caméras monoculaires et RGB-D. Plusieurs expériences ont été réalisées dans un environnement intérieur contrôlé équipé d'un système de capture de mouvement Vicon. Afin d'évaluer et valider les algorithmes vSLAM, différents scénarios ont été privilégiés dans diverses configurations à savoir : ligne droite, ligne droite et demi-tour, trajectoire circulaire avec fermeture de boucle, etc.

Comparaison entre ORB-SLAM et RTAB-MAP SLAM

ORB SLAM. ORB-SLAM2 est une librairie SLAM temps-réel pour les caméras monoculaires, stéréo et RGB-D qui calcule la trajectoire de la caméra et une reconstruction 3D éparse. Le système comprend un mode de localisation légère qui exploite les pistes d'odométrie visuelle pour les régions non cartographiées et les correspondances avec les points de la carte qui permettent une localisation sans dérive. Les principales fonctionnalités du ORB-SLAM sont le suivi des caractéristiques, la cartographie, la fermeture de boucle et la localisation. Nous avons implémenté l'ORB-SLAM avec ses 3 déclinaisons : monoculaire, stéréo et RGB-Depth. Pour l'ORB-SLAM monoculaire, la première étape consiste à détecter les caractéristiques et à initialiser la carte et sa position. Comme la profondeur ne peut pas être récupérée à partir d'une seule image, les points peuvent être initialisés, avec une grande incertitude sur la profondeur, en utilisant une paramétrisation inverse de la profondeur qui, in fine, convergent plus tard vers leurs positions réelles.

La détection des caractéristiques dans le SLAM stéréo est meilleure que le SLAM monoculaire. Pour la paire d'images stéréo et pour chaque ORB de gauche, nous cherchons une correspondance dans l'image de droite. Ceci peut être fait en supposant que les images stéréo sont rectifiées de sorte que les lignes sont horizontales. Nous générons ensuite le point clé stéréo avec les coordonnées de l'ORB de gauche et les coordonnées horizontales de la correspondance droite, qui est affinée au sous-pixel par corrélation de patch (patch correlation). Enfin, l'ORB-SLAM RGB-Depth nécessite une image RGB avec sa profondeur. Pour les images RGB-D, nous extrayons les caractéristiques ORB sur l'image RGB. Pour chaque caractéristique ayant des coordonnées $(uL\,;\,vL)$ nous transformons sa valeur de profondeur d en une coordonnée virtuelle droite (e.g. équation 5.1) :

$$uR = uL - \frac{fx * b}{d} \tag{5.1}$$

Où fx est la distance focale horizontale et b est la baseline entre le projecteur de lumière structurée et la caméra infrarouge. A l'issue du ORB SLAM, nous récupérons la cartographie de l'environnement pour la comparer avec la vérité terrain de localisation temps-réel du fauteuil.

RTAB-MAP SLAM. Real-Time Appearance Based Mapping (RTAB-MAP) utilise l'image de profondeur avec les images couleurs pour construire des cartes. Le graphe est créé où chaque nœud contient des images RGB et de profondeur avec la pose odométrique correspondante. Les liens sont des transformations entre chaque nœud. Lorsque le graphe est mis à jour, RTAB-Map compare la nouvelle image avec toutes les images précédentes dans le graphe pour trouver une fermeture de boucle. Lorsque cette dernière est trouvée, l'optimi-

sation du graphe est ainsi effectuée pour corriger les poses. Pour chaque nœud du graphe, nous générons un nuage de points à partir des images RGB et de profondeur. La carte 3D est alors créée.

Résultats expérimentaux. Nous avons donc effectué une étude comparative entre ORB-SLAM et RTABMAP SLAM avec les mêmes scénarios de tests (e.g. tableau 5.2). Afin d'améliorer la qualité de la localisation ainsi que celle de la cartographie obtenue par le SLAM de base, nous avons adapté ORB-SLAM à notre application en prenant en compte certains paramètres. Nous avons testé ORB-SLAM en utilisant une caméra monoculaire, une stéréo et une RGB-D. L'ORB-SLAM est désormais capable de sauvegarder la carte et de la charger ultérieurement à des fins de localisation. L'ORB-SLAM a également été amélioré pour fournir des estimations précises de l'odométrie. Nos essais de validation ont montré que RTABMAP est meilleur pour l'estimation de la trajectoire que ORB-SLAM, limité par les images monoculaires, même si ce dernier reste relativement précis sur les données odométriques et meilleur en mesure de distance. Nous avons validé l'ensemble des développements sur notre plateforme de fauteuil roulant en comparant, dans deux environnements intérieurs et extérieurs, les données ORB-SLAM et RTAB-MAP avec celles de vérité terrain issues du système Vicon. Ces travaux de recherche ont fait l'objet de publications dans [31] et [42].

Scenario	Ground Truth	RTAB RGBD	ORBSLAM Stereo	ORBSLAM RGBD
Indoor 1	6.48	11.83	5.33	8.94
Indoor 2	24.5	37.61	30.64	27.78
Indoor 3	23.26	31.31	20.82	21.43

TABLE 5.2 – Comparaison entre ORB-SLAM et RTABMAP SLAM (les valeurs sont exprimées en mètre).

5.5 Segmentation sémantique temporelle pour un fauteuil roulant intelligent

5.5.1 Contexte & objectifs

Dans la smart mobilité, la détection d'objets, l'estimation de profondeur et la segmentation sémantique sont des tâches importantes pour une bonne compréhension de l'environnement. Ces dernières années, de nombreuses études ont été réalisées dans le domaine de la segmentation sémantique pour véhicule autonome. Certains jeux de données annotés sont désormais disponibles pour les tâches de segmentation sémantique permettant ainsi le déve-

loppement de solutions performantes. En revanche, pour d'autres types de smart mobilité, comme les fauteuils roulants intelligents, il n'existe malheureusement pas de jeux de données spécifiques dédiés par exemple à la navigation sur les trottoirs (où le point de prise de vue caméra est différent de celui sur la route). En plus, la plupart des modèles de segmentation sémantique utilisent des images uniques or nous pensons que l'information temporelle dans les séquences d'images est importante. Dans le cadre du projet ADAPT, et afin de renforcer la perception d'environnement outdoor du fauteuil roulant sur les trottoirs, nous avons développé notre propre jeu de données de courtes séquences d'images virtuelles extérieures de scènes de rue prises depuis des points de vue situés sur des trottoirs. Afin de valider notre dataset et d'assurer la segmentation sémantique de l'environnement du fauteuil, nous avons développé un nouveau réseau CNN adapté au traitement temporel incluant des techniques supplémentaires pour améliorer ses performances en précision.

5.5.2 Etat de l'art

Datasets. Les deux jeux de données de référence pour la segmentation sémantique sont Cityscapes [201] et Camvid [202]. D'autre jeux existent comme NYUDv2 [203] qui comprend des séquences vidéos enregistrées avec une caméra RGB-D. Cityscapes est disponible en 2 versions : images isolées et séquences. La version image comprend $5k$ d'images annotées et 19 classes. Les images ont été prises dans un environnement urbain, dans des conditions météorologiques variables et dans 50 villes différentes. Les extraits vidéos sont composés de 30 images, prises à 17 FPS, dont la 20ième est annotée pour la segmentation sémantique. Camvid comprend 701 images annotées et 5 séquences vidéos. Ces jeux de données sont destinés aux véhicules autonomes avec un point de vue situé sur les routes. Ils ne sont donc pas adaptés à notre application, avec un point de vue situé sur le trottoir. C'est d'ailleurs la raison principale qui nous a conduit à développer notre propre jeu de données. Une autre façon d'obtenir des séquences d'images annotées est d'utiliser des images virtuelles 3D générées par un logiciel : GTAV [204] (24966 images et 19 classes), SYNTHIA [104] (9400 images et 16 classes), Virtual KITTI, etc. Le principal avantage des jeux de données virtuels est qu'ils peuvent être générés facilement, dans des conditions variées et sans contraintes techniques. En revanche, elles souffrent du manque du réalisme ce qui nécessite souvent le recours aux techniques d'adaptation du domaine. Dans notre travail, nous proposons un jeu de données virtuel généré via CARLA, un simulateur public dédié aux véhicules autonomes.

Segmentation sémantique par CNN. Les travaux pionniers sur la segmentation sémantique par CNN comprennent diverses approches comme les réseaux déconvolutionnels [205], les caractéristiques hiérarchiques [206], les CNN récurrents [207], les caractéristiques zoom-out [208], etc. Un grand pas est fait avec le réseau d'apprentissage par déconvolution (LDN) [209],

UNet [210], FCN (Fully Convolutional Network) [211] et SegNet [212]. Ces modèles ont une structure d'encodeur-décodeur. Parmi les solutions alternatives à la partie décodeur, PSP-Net [213] effectue un regroupement de pyramides spatiales pour capturer les informations multi-échelles. DeepLab [214] utilise la convolution "à trous" pour capturer les informations des images à différentes échelles spatiales (Atrous Spatial Pyramid Pooling - ASPP). Dans DeepLabV3+ [215], un module de décodage est ajouté après l'étape ASPP pour affiner les résultats de la segmentation. Dans DenseASPP [216], le principe de l'ASPP est étendu mais avec une connexion plus dense, conduisant à une meilleure précision. Dans [217], la fusion complète par portes (Gated Fully Fusion - GFF) fusionne sélectivement les caractéristiques de plusieurs niveaux en utilisant des portes de manière entièrement connectée. Dual Attention Network (DANet) [218] capture les dépendances des caractéristiques dans les dimensions spatiales et de canal.

5.5.3 Génération d'un dataset virtuel

Pour générer notre jeu de données, nous avons fait appel au simulateur CARLA [219]. Il comprend 8 cartes d'environnement différents dont 6 ont été utilisées dans le développement de notre dataset. Nous avons utilisé le nombre maximum de véhicules disponibles par carte pour simuler tous les scénarios possibles (pas de trafic, faible trafic, forte charge de trafic, embouteillage). Les images de segmentation sémantique sont composées de 13 classes : aucune, bâtiment, clôture, autre, piéton, poteau, marquage de voie, route, trottoir, végétation, véhicule, mur et panneau de signalisation. La catégorie "aucune" correspond aux textures qui ne font pas partie d'un objet (e.g. pelouses qui ne font pas partie de "végétation"). Dans la catégorie "autre", on trouve des objets qui ne sont pas inclus dans les autres classes (e.g. plantes). Pour notre application, les catégories les plus importantes sont les classes "trottoirs" et "routes" pour trouver le chemin à suivre, ainsi que "bâtiments" et "poteaux" pour éviter les obstacles.

Les séquences sont constituées de 4 images prises avec un écart de $0,05$s entre chacune d'elles. Le jeu de données est composé de 46436 images (i.e. 11609 séquences) partitionnées en 41024 images pour l'entrainement, 2696 images pour la validation et 2716 images pour le test (e.g. figure 5.6). De plus, nous avons généré un autre jeu de données plus petit (7692 images et 1923 séquences) avec des images prises depuis deux points de vue différents : l'un situé sur la route et l'autre sur le trottoir. Il a pour objectif de montrer l'importance du point de vue dans la segmentation sémantique. Ceci peut être fait par validation croisée : apprentissage sur des images prises depuis un point de vue route et test sur des images avec un point de vue trottoir, et vice-versa.

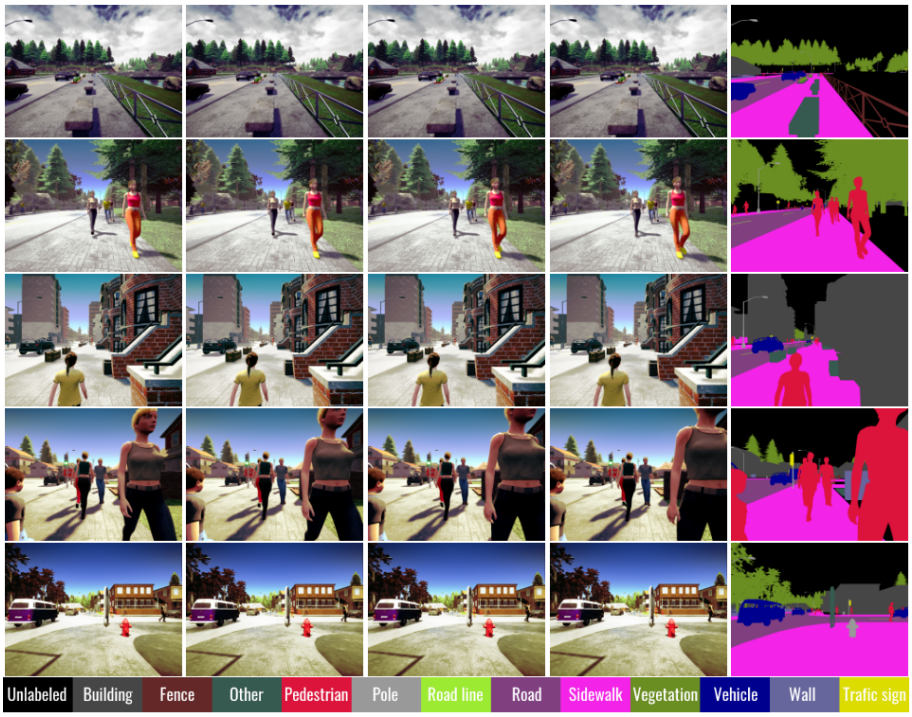

FIGURE 5.6 – Exemples de séquences d'images de notre jeu de données. La dernière colonne représente la vérité terrain correspondante à la dernière image RGB de chaque séquence.

5.5.4 Segmentation sémantique temporelle

Afin de réaliser des expériences avec notre jeu de données, nous avons conçu un nouveau réseau CNN basé sur l'architecture DeepLabV3. Nous avons introduit de nouvelles techniques pour adapter le modèle au traitement de séquences d'images et pour améliorer ses performances en terme de précision.

Couches temporelles. Afin d'intégrer les informations temporelles dans les entrées du modèle, nous utilisons des blocs de réseaux temporels (Temporal Network - TN), comme proposé dans le réseau de convolution temporel (TCN) [220]. Dans un bloc, TN a deux couches de convolution dilatées avec la fonction ReLU. Pour des raisons d'efficacité, nous n'utilisons qu'une seule couche de convolution. Nous utilisons également des connexions résiduelles. La sortie finale d'un bloc TN est la sortie des convolutions ajoutées à l'entrée du bloc.

Architecture du réseau. Pour la partie encodeur du réseau, nous utilisons des blocs ResNest [221], qui est une variante des blocs ResNet [222] où des modules à attention partagée sont utilisés à l'intérieur de chaque bloc ResNet. Pour intégrer l'aspect temporel des données, nous utilisons des couches temporelles de type TCN [220]. Comme le meilleur endroit des couches temporelles dans le réseau n'est pas connu *a priori*, nous avons effectué différents tests entre trois modes de fonctionnement : avec une couche temporelle entre l'encodeur et

le décodeur, avec des couches temporelles supplémentaires entre chaque bloc de l'encodeur et avec les deux.

Fonction de perte. La fonction de perte est un autre moyen efficace pour améliorer la précision du modèle. Une étude des fonctions de perte dédiées à la segmentation sémantique a été réalisée dans [223]. Nous avons testé plusieurs fonctions : la perte focale [224], initialement proposée pour la détection d'objets, et la fonction de perte d'entropie croisée catégorielle définie par l'équation (5.2) :

$$Loss = -\sum_{i=1}^{n} y_i * log\hat{y}_i \tag{5.2}$$

Où n est le nombre de classes, $\mathrm{y_i}$ la ième valeur cible et $\hat{\mathrm{y}}_\mathrm{i}$ la ième valeur de sortie (i.e. valeur estimée). La perte focale est basée sur le principe de la réduction de la pondération des exemples faciles et de la concentration sur les exemples difficiles. Un poids est attribué à chaque classe de la segmentation en fonction de l'importance de son taux de prédiction correcte. Il est défini dans l'équation (5.3) :

$$FL = -\sum_{i=1}^{n} \alpha_i (1 - \hat{P}_i)^\gamma \log \hat{P}_i \tag{5.3}$$

Où n est le nombre de classes, α_i le coefficient de pondération statique de la ième classe, $\hat{P}_i$ la distribution de la prédiction de cette même classe et γ un hyper-paramètre à régler ($\gamma > 0$). Dans nos expériences, nous avons pris γ=2.

Equilibrage des classes. L'un des principaux problèmes de la formation d'un réseau de segmentation sémantique est le déséquilibre des classes. Ainsi le réseau sera biaisé vers les classes les plus largement représentées et peu performant sur les classes les plus rares. Nous utilisons la pondération de classe pour atténuer ce déséquilibre en rendant les poids pour les classes rares plus grands qu'avec l'équilibrage de fréquence médiane, comme dans [225] (équation (5.4)).

$$w_i = 1/\log(c + N_i/N) \tag{5.4}$$

Avec w_i le poids pour la ième classe, N_i le nombre de pixels de la ième classe, N le nombre total de pixels et c (pris à $1,02$) un hyper-paramètre permettant de limiter les poids.

Résultats expérimentaux

Nous avons présenté un nouveau modèle CNN de traitement temporel pour la segmentation sémantique d'images extérieures prises à partir du point de vue trottoir. Nous avons également développé un nouveau dataset de séquences d'images virtuelles annotées pour la

segmentation sémantique. Afin de valider le nouveau réseau de segmentation sémantique, nous avons testé trois variantes de ce dernier : (*i*) sans couche temporelle, (*ii*) avec une couche temporelle entre l'encodeur et le décodeur et (*iii*) avec les couches temporelles entre chaque couche spatiale de la partie encodeur et entre l'encodeur et le décodeur. Nous avons également testé deux encodeurs différents : ResNet et ResNest. Le meilleur résultat a été obtenu via la troisième variante et l'encodeur ResNest, avec une précision $mIoU$ de 84.06%. La figure 5.7 montre quelques exemples d'inférence de notre modèle sur des images extraites de l'ensemble de validation. Ces travaux de recherche ont fait l'objet de publication dans [11].

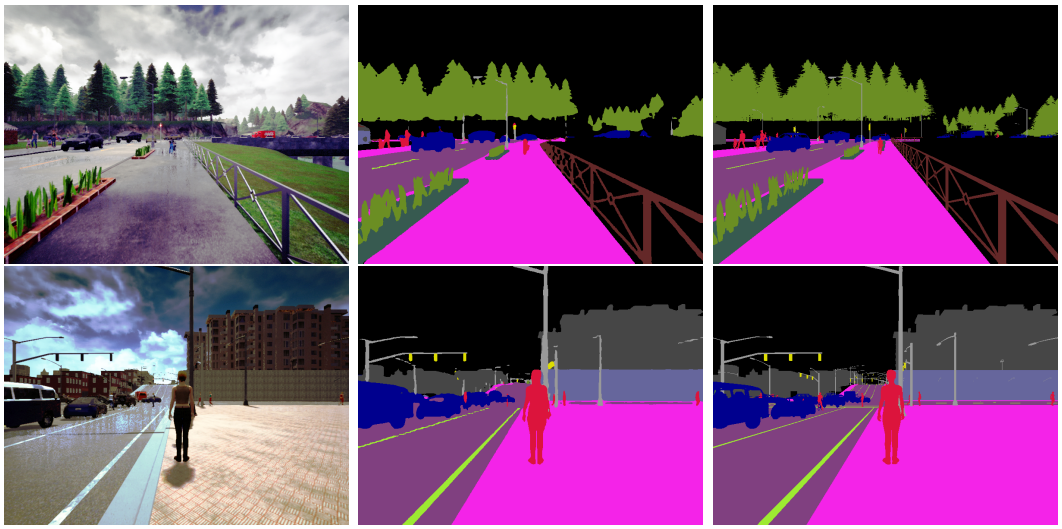

FIGURE 5.7 – Deux exemples d'inférence du modèle. De gauche à droite : image brute, prédiction du modèle et vérité terrain.

5.6 Segmentation sémantique rapide pour un fauteuil roulant intelligent

5.6.1 Contexte & objectifs

Dans la continuité des travaux ADAPT sur la segmentation sémantique du fauteuil roulant naviguant sur un trottoir, réalisé par l'ingénieur de recherche Vishnu Pradeep sur la segmentation sémantique rapide de vidéos, nous présentons une approche robuste temps-réel offrant une perception visuelle avancée dans des scènes complexes. La clé de cette proposition est d'avoir une méthode avec des estimations de flux légères et des extractions de caractéristiques fiables. Nous abordons cette question en choisissant une approche basée sur les tendances récentes en matière de segmentation vidéo. Bien que ces dernières démontrent des performances efficaces, elles nécessitent des procédures supplémentaires pour mettre en

pratique leurs caractéristiques remarquables.

Inspirés donc par les récentes approches de segmentation sémantique vidéo (VSS), nous proposons une nouvelle méthode qui comprend : 1. un modèle de segmentation sémantique de base qui adopte l'épine dorsale ResNet18, 2. GSVNET [226], un réseau de propagation qui effectue une convolution spatiale variable guidée sur des cartes de caractéristiques brutes à échelle réduite provenant du réseau de segmentation de base, 3. une convolution dilatée multi-échelle pour améliorer encore le champ réceptif de la carte de caractéristiques et 4. une méthode de mise à l'échelle bilinéaire à faible rang pour obtenir les cartes de segmentation dans la résolution d'origine. Afin de valider notre approche, nous développons notre propre dataset synthétique (amélioration du jeu de données CARLA présenté dans la section précédente) avec une distribution de cibles de mélange.

5.6.2 Etat de l'art

Nous nous concentrons dans cette section sur les approches de segmentation vidéo [227, 228, 229] qui ajustent la perte entre vitesse et précision. La plupart des méthodes appliquent le même modèle CNN à chaque image et agrègent temporellement les caractéristiques avec des couches supplémentaires [230, 231, 232]. Bien que ces méthodes obtiennent de bonnes valeurs de précision par rapport aux approches à une seule image, elles entraînent un surcout de calcul considérable. Ainsi, d'autres approches visant à tirer parti des coûts de calcul élevés ont été proposées pour maintenir la continuité temporelle [233, 234, 235, 236]. Néanmoins, le mouvement continu des objets et leur occlusion dans les vidéos sont des obstacles difficiles à surmonter pour propager de manière robuste les prédictions du pixel dans le temps.

Pour résoudre ce problème, [234, 237] réutilise directement les caractéristiques de haut niveau relativement stables extraites des images réduites dans les couches profondes. Une autre approche consiste à utiliser le flux optique pour envelopper les caractéristiques de haut niveau des images clés dans les images non clés [233] et à mettre à jour les cartes de caractéristiques enveloppées dans le flux avec celles peu profondes extraites des images actuelles [236]. Cependant, l'utilisation du flux optique entraîne des coûts de calcul importants et peut échouer avec de grandes variations et des régions non texturées. Pour y remédier, Li *et al.* [235] utilisent une convolution variante dans l'espace (SVC) ainsi qu'un extracteur de caractéristiques léger pour les images non clés. Bien que ces méthodes offrent une réduction globale des coûts de calcul par rapport à leurs bases de segmentation d'image, leur précision a évidemment aussi diminuée [233, 235, 236]. De plus, en raison d'extractions moins fiables au niveau des images-clés, ces méthodes sont sujettes à des vitesses incohérentes avec une latence équivalente à celle des modèles simple image. Par conséquent, pour effectuer une

segmentation sémantique légère et rapide sur une vidéo, il est crucial d'utiliser les caractéristiques extraites au maximum de leur potentiel et de rendre cette extraction sur les images non clés aussi simple que possible.

GSVNET est un simple réseau de propagation temporelle qui effectue des convolutions spatiales sur les cartes de segmentation d'un réseau de base [238]. Il effectue un enveloppement temporel par flux optique sur des images réduites pour estimer les cartes de caractéristiques des images actuelles et des convolutions spatiales pour atténuer les erreurs dues à des flux optiques imparfaits. Il concatène les estimations brutes des canaux d'un réseau de segmentation de base pour former une segmentation réduite de chaque image. Cependant, l'extraction de caractéristiques souffre d'une grande perte d'informations contextuelles due aux estimations de flux sur les images à échelle réduite.

5.6.3 Segmentation sémantique rapide de la vidéo

Segmentation sémantique basée GSVNET. Pour la segmentation des images clés, nous utilisons les modèles SN-R18 et BN-R18, basés sur les réseaux SwiftNet-R18 [239], BiSeNet [240], et GSVNET comme cadre de propagation. GSVNET commence par une estimation spatio-temporelle sur une segmentation à échelle réduite obtenue à partir des modèles susmentionnés. S_{t-1} est la segmentation de l'image précédente I_{t-1}. Pour obtenir une estimation initiale S_t pour l'image courante I_t, GSVNET effectue un enveloppement temporel par flux optique sur S_{t-1}, comme exprimé dans l'équation (5.5) :

$$S_t(c, x, y) = S_{t-1}(c, x + m_{tx}, y + m_{ty}), \forall c \epsilon C \tag{5.5}$$

Où x et y désignent les emplacements des pixels dans la segmentation réduite, m_{tx} et m_{ty} sont des attributs issus de la fusion hiérarchique des caractéristiques du réseau du flux optique. Afin de corriger les erreurs lors de l'estimation du flux optique, l'enveloppement de la segmentation est affiné en appliquant un SVC guidé. Tout d'abord, on effectue une convolution séparable sur S_t pour la décaler, par canal de pixels, dans une certaine direction, puis on l'additionne avec les estimations brutes de l'image actuelle I_t. Un SVC est appliqué sur les canaux de chaque bloc C, puis additionné pour obtenir l'estimation à échelle réduite de la segmentation de l'image courante $S_t(c, x, y)$.

Convolution dilatée. La convolution dilatée permet de comprendre la relation positionnelle et sémantique entre les objets [241]. Afin d'affiner la carte des caractéristiques de notre modèle basé GSVNET, nous appliquons la convolution multi-échelle à la carte de segmentation, à échelle réduite, sur $S_t(c, x, y)$. Cette méthode permet d'extraire efficacement

les informations contextuelles globales et d'élargir le champ réceptif sans perdre en résolution [242, 243, 244]. Les cartes de caractéristiques raffinées de différents facteurs générées par la convolution dilatée sont concaténées avec l'image d'entrée. Grâce à cette concaténation, les informations brutes des caractéristiques et les informations de la structure hiérarchique peuvent être combinées. Ensuite, la carte de caractéristiques fusionnée est utilisée pour le processus de sur-échantillonnage. Une convolution dilatée $2D$ peut être exprimée par l'équation (5.6) :

$$Y_t = \sum_{i=0}^{x} \sum_{j=0}^{y} S_t(c, x + r \times i, y + r \times j) \tag{5.6}$$

Où r est le taux de dilatation. Afin d'obtenir l'image de sortie à la même résolution, un noyau de taille $k \times k$ est agrandi à $k + (k-1)(r-1)$ avec un pas de dilatation r. Cela nous permet donc d'améliorer le contexte capturé à différentes échelles spatiales.

5.6.4 Résultats expérimentaux

Construction du jeu de données. Le modèle GSVNET natif utilise pour son entrainement les jeux de données Cityscapes [245] et CamVid [246]. Ces deux datasets sont dédiés aux véhicules autonomes sur route et non pas sur le trottoir comme c'est le cas pour le fauteuil roulant. Nous avons donc amélioré notre dataset CARLA avec des extraits de vidéo de distribution fixe dans la prévisualisation du trottoir. Notre dataset a été divisé en trois sous ensemble de données dédiés à l'entrainement (11904 images), validation (8020 images) et test (10076 images). Il comprend des séquences d'images (i.e. résolution 2048×1024) avec 30 images par séquence, la 19ème image étant finement annotée pour 22 classes. Ces classes reprennent celles du jeu de données Cityscapes auxquelles nous avons ajouté des classes supplémentaires cruciales à notre contexte telles que les lignes de route, les glissières de sécurité et les objets statiques (e.g. bouches d'incendie, bancs fixes, fontaines et arrêts de bus). Les scènes sont capturées dans différents scénarios de visibilité, de météo et de complexité.

Implémentation. Nous utilisons les poids pré-entraînés SN-R18 et BN-R18 [238] de GSVNET pour la segmentation des images clés. Le processus commence par la segmentation de la première image clé pour obtenir une segmentation sémantique grossière à partir de BiseNet/SwfitNet comme réseau de segmentation de base. Notre framework de propagation, qui est une combinaison de GSVNET avec le processus de convolution dilatée, propage par la suite temporellement les cartes de prédiction de segmentation des images précédentes pour aider à prédire les cartes de segmentation des images non clés actuelles. À chaque étape, les cartes de segmentation du GSVNET des images précédentes sont soumises à une convolution dilatée pour affiner et améliorer les cartes de segmentation. La figure 5.8 illustre la méthode proposée.

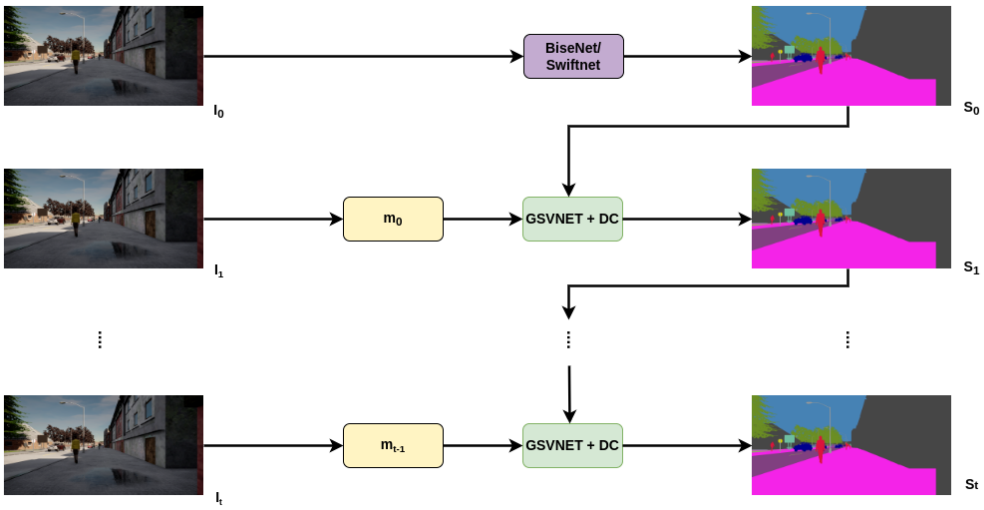

FIGURE 5.8 – Architecture du modèle proposé. BiseNet/SwiftNet est le réseau de segmentation de base, GSVNET avec convolution dilatée (DC) est le framework de propagation. Les cartes de prédiction de segmentation S des images I sont propagées temporellement jusqu'à l'image clé suivante.

Résultats et analyse. Le Tableau 5.3 présente une comparaison des compromis "précision/débit" des modèles GSVNET et de nos modèles. Le SN-R18 de notre méthode surpasse celui de GSVNET de 21% en mIoU et de 19, 4% en mPA. De même, notre BN-R18, dans une configuration similaire, surpasse celui de GSVNET de 27, 1% en mIoU et de 8, 3% en mPA. Bien qu'il y ait une perte de vitesse de 19, 5% dans SN-R18 et de 10, 9% dans BN-R18, notre approche est plus précise en mIoU avec une vitesse, en FPS, acceptable pour notre application. En plus, notre méthode n'introduit aucune surcharge et conserve la caractéristique clé de légèreté de GSVNET tout en améliorant la précision.

Method	Model	Scale Factor	Avg. mIoU	mPA	FPS
GSVNET	SN-R18	0.75	38.4	65.8	65
	BN-R18	0.75	34.2	66.7	59.5
Ours	**SN-R18**	0.75	**46.5**	**78.6**	52.3
	BN-R18	0.75	**43.5**	**72.3**	49.5

TABLE 5.3 – La comparaison de la précision et du débit obtenus par GSVNET et notre approche sur notre dataset CARLA.

La figure 5.9 illustre un exemple d'évaluation qualitative de notre modèle sur notre dataset CARLA. La vue agrandie illustre la comparaison de la perte de détails par rapport à l'annotation de la vérité terrain. La carte de segmentation de GSVNET souffre d'une perte considérable d'informations en unités de pixels. En outre, les limites de l'objet sont déformées

avec la perte d'informations vitales telles que les informations sur les feux de circulation, les piétons, le contexte des bâtiments et les poteaux d'éclairage public. Grâce à notre technique de convolution dilatée et de récupération des pixels, nos modèles donnent de meilleures performances avec des cartes de prédiction de segmentation plus détaillées. Comme le montre la vue agrandie de la carte obtenue par notre méthode, les informations perdues dans GSVNET sont désormais reproduites avec une meilleure ségrégation des limites des objets. Ces travaux de recherche ont fait l'objet de publication dans [15].

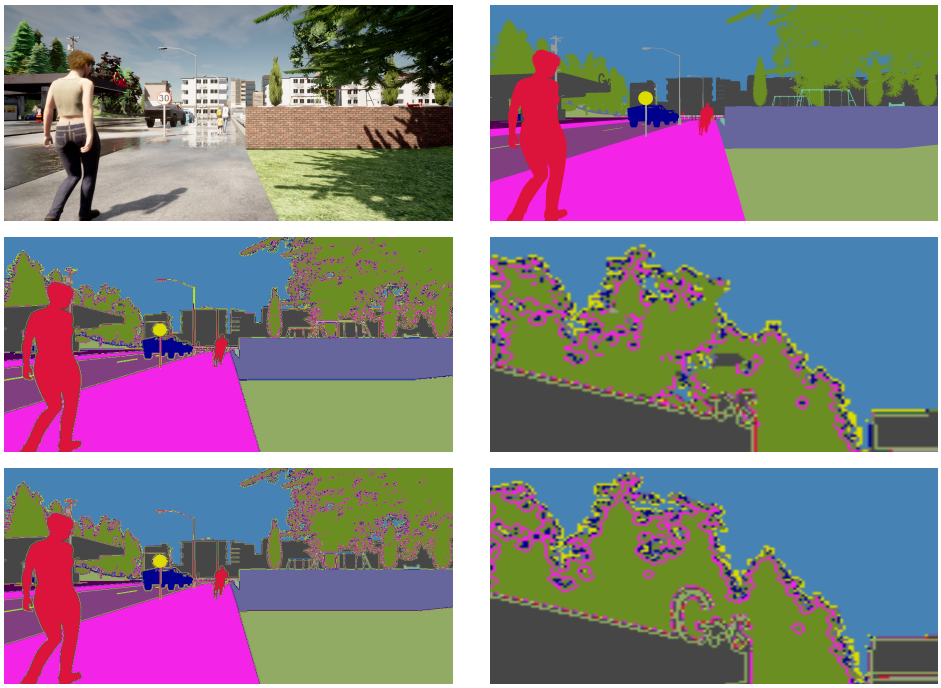

FIGURE 5.9 – Évaluation qualitative. Ligne du haut : un échantillon d'image du jeu de données CARLA avec sa vérité terrain (image à droite), ligne du milieu : l'image de sortie GSVNET SN-R18 et sa vue zoomée à droite, ligne du bas : l'image de sortie de notre GSVNET modifié et sa vue zoomée à droite. En comparant les images zoomées des lignes du milieu et du bas, les contours du bâtiment appelés "GAS" sont complètement déformées dans la ligne du milieu, tandis qu'ils sont bien préservés dans la ligne du bas. Ainsi, notre modèle conserve plus d'informations contextuelles dans les images de segmentation de sortie.

5.7 Conclusion

Nous avons présenté dans ce chapitre une chaine complète de perception d'environnement comprenant la détection et localisation d'objets, estimation de profondeur, tracking et segmentation sémantique pour la smart mobilité de fauteuils roulants intelligents. Tout d'abord, nous avons présenté notre approche de détection et reconnaissance de visage par vision omnidirectionnelle pour des applications de contrôle d'accès biométriques de robots mobiles. Nous avons étendu l'utilisation de la vision omnidirectionnelle à la navigation auto-

nome d'un fauteuil roulant intelligent. Nous avons présenté notre approche de détection de portes et poignées de portes par deep learning. Nous avons montré qu'en combinant la position du fauteuil et celle des objets détectés sur chaque image, il était possible de regrouper les instances des objets par un algorithme de clustering, et de les intégrer à une carte 2D locale (sémantique) construite par l'algorithme de SLAM RTABMAP. Une carte de coûts locaux construite à partir des images de profondeur a permis d'éviter dynamiquement les obstacles dans l'environnement proche du fauteuil. Par la suite, nous avons présenté notre étude comparative de deux algorithmes vSLAM dédiés à la localisation du fauteuil roulant. Nous avons amélioré les deux approches pour une meilleure localisation temps-réel. Nous avons validé l'ensemble des développements sur une plateforme de fauteuil roulant en comparant les données de nos approches RTABMAP et ORB-SLAM avec celles de vérité terrain.

Par la suite, nous avons présenté un nouveau modèle CNN avec un module de traitement temporel pour la segmentation sémantique d'images extérieures prises à partir du point de vue trottoir. Nous avons également développé un nouveau jeu de données de séquences d'images virtuelles annotées pour la segmentation sémantique. Enfin, nous avons présenté une autre méthode de segmentation vidéo légère et efficace pour la perception visuelle dans les environnements de trottoirs urbains pour les fauteuils roulants intelligents. Nous avons proposé une approche auto-supervisée qui facilite les applications inter-domaines des modèles CNN populaires dans des cas d'utilisation distincts. Nous utilisons un simulateur public CARLA pour développer un ensemble de données de séquences vidéo synthétiques dans divers environnements de trottoirs. Par la suite, nous avons prouvé la faisabilité de la réutilisation de modèles deep learning existants, GSVNET et modifier les caractéristiques intrinsèques de la précision des résultats en fonction des besoins et ce, sans coût supplémentaire de ressources informatiques. Afin de préserver la légèreté de nos modèles, nous avons amélioré l'approche GSVNET par l'introduction de deux processus supplémentaires effectuant une convolution dilatée pour affiner la carte de caractéristiques.

Chapitre 6

Système de Navigation Intelligent d'Optimisation de la Consommation d'Energie

Sommaire

6.1 Introduction

Souvent les systèmes ADAS visent l'amélioration de la sécurité des véhicules autonomes. Or depuis l'émergence du Véhicule Electrique (VE), les systèmes ADAS doivent non seulement assurer la sécurité à bord, mais aussi optimiser la consommation afin de prolonger l'autonomie du VE. Nous allons présenter dans ce chapitre le développement de systèmes de perception par fusion multicapteurs dédiés à l'optimisation de la consommation d'énergie de VEs autonomes. Par la suite, nous présenterons deux autres systèmes dont un est dédié au diagnostic et détection de défauts à distance des parcs éoliens et l'autre au tracking temps-réel de véhicules. Ce chapitre comprend donc 4 parties : 1. système fusion multicapteurs pour l'optimisation de la consommation d'énergie, 2. amélioration de l'optimisation de la consommation par machine learning, 3. diagnostic et détection de défauts à distance et 4. tracking temps-réel de véhicules.

Dans un premier temps, une nouvelle approche basée fusion multicapteurs sera présentée pour l'optimisation de la consommation d'énergie d'un VE. Le système prend en compte la dynamique du véhicule ainsi que la topologie de la route pour estimer une vitesse censée optimiser la consommation à bord. Par la suite, nous allons améliorer l'approche développée afin d'offrir une meilleure estimation de la consommation d'énergie. Un modèle machine learning par apprentissage paresseux sera présenté. Par la suite, nous présenterons notre modèle V2G modélisant l'interaction entre les VEs et le réseau électrique. En troisième partie, nous présenterons notre plateforme de diagnostic et de détection de défauts à distance. Enfin, nous présenterons une solution de tracking temps-réel de véhicules dans la chaine logistique portuaire.

6.2 Système fusion multicapteurs d'optimisation de la consommation d'énergie

6.2.1 Contexte & objectifs

Ces travaux de recherche ont été développés dans le cadre du projet VIRTUOSE (VéhIcule électRique intelligenT à prolongateUr d'AutOnomie et Sources d'Energie multiples) qui visait à développer un démonstrateur de VE urbain intelligent équipé d'un prolongateur d'autonomie et sources d'énergie multiples. Nous avons développé de nouvelles génération d'ADAS basés perception d'environnement par fusion multicapteurs afin d'améliorer non seulement la sécurité mais aussi et surtout d'optimiser la consommation d'énergie à bord pour une meilleure autonomie. Notre approche innovante prend en compte plusieurs para-

mètres liés à la dynamique du véhicule, les conditions météorologiques et l'infrastructure via la topologie de la route. Tout d'abord, le système se limite à conseiller le conducteur par une vitesse adaptative. Par la suite, et afin de répondre aux contraintes temps-réel et de consommation d'énergie de certains scénarios, le système pourra reprendre le contrôle du VE si nécessaire. La plateforme a été validée non seulement dans des phases de simulation mais aussi dans des environnements intérieurs (banc d'essai) et extérieurs (trafic réel).

6.2.2 Architecture du système

L'approche permet de générer une vitesse consigne optimisant la consommation d'énergie à bord. Le système de perception basée fusion multicapteurs permet d'enrichir davantage les données ce qui améliore la décision et permet donc par la suite, de déclencher des actions avec une grande précision. Notre système de perception comprend les éléments suivants :

— Un calculateur embarqué
— Quatre caméras fisheye dédiées à la vue panoramique (birdview) autour du VE
— Un RADAR à 24 GHz dédié à la détection fine/précise des obstacles et véhicules
— Module GPS dédié à la géolocalisation fine pour le contrôle intelligent de la vitesse
— Une caméra Mobileye

En général, les systèmes ADAS sont basés sur la combinaison de capteurs (RADAR, LiDAR, caméras, IMU) et d'algorithmes qui assurent la sécurité du véhicule, du conducteur, des passagers et des piétons en fonction de différents paramètres tels que le trafic, la météo, etc. [247]. Dans le cadre de nos travaux, l'ADAS vise non seulement à améliorer la sécurité, mais aussi à optimiser l'énergie à bord. Notre contribution vise donc le développement d'un système fusion multicapteurs dédié à la gestion optimisée de l'énergie embarquée. Il prend en compte les propriétés topologiques de l'itinéraire (profil de la route), le comportement du conducteur (réduit à la vitesse appliquée au véhicule) et la géolocalisation. Cette dernière basée GPS couplé à des cartes routières permet d'obtenir une description détaillée de la trajectoire à effectuer par le VE à savoir : route plane, route en pente, croisement, etc. Cela permet d'ajuster la vitesse en fonction du profil de la route, ce qui permet au conducteur d'adapter son comportement de conduite en fonction de l'itinéraire : informé d'une consigne de vitesse en montée (ou en descente), le conducteur adopte un nouveau comportement lui permettant de réduire sa consommation d'énergie. Informé de la signalisation à l'approche d'un carrefour à 500 mètres devant, lui permettant d'adopter une conduite souple (ne pas accélérer) et ainsi prolonger l'autonomie du VE.

6.2.3 ADAS basé fusion multicapteurs pour l'optimisation d'énergie

Le système comprend donc trois modules ADAS destinés à optimiser la consommation à bord : (*i*) contrôleur de vitesse adaptative, (*ii*) vue panoramique autour du véhicule et (*iii*) réalité augmentée afin d'enrichir la perception avec des données liées à la route.

ADAS de régulation adaptative de vitesse

Sur un itinéraire donné, et dans un premier temps, les vitesses locales sont définies en minimisant la quantité d'énergie nécessaire pour effectuer le trajet. Ceci est nécessaire pour estimer si le trajet est compatible avec l'autonomie du VE et donc les contraintes de temps fixées par l'utilisateur. Dans la deuxième étape, l'approche est basée sur un calcul dynamique permettant de prendre en compte les données réelles afin d'affiner l'estimation de la vitesse locale. Dans la troisième étape, nous utilisons une approche de modulation de la vitesse locale d'instruction afin de prendre en compte la topologie de la route. Notre algorithme comprend donc les étapes suivantes : 1. calcul des différentes forces appliquées au véhicule (résistance au roulement, à l'air, à la pente, au vent et à l'accélération). 2. calcul de l'énergie totale par segment, 3. calcul de la vitesse consigne correspondante à la minimisation de l'énergie pour l'ensemble du trajet et 4. calcul itératif des vitesses correspondantes aux minimum de l'énergie globale du trajet.

Le système diminue donc la vitesse dans les pentes ascendantes (segments à dérivation haute) afin de consommer moins d'énergie, et augmente la vitesse dans les pentes descendantes (segments à dérivation faible) pour récupérer plus d'énergie. L'algorithme est itératif et nécessite le recalcule de l'énergie globale du trajet d'un point A à un point B. Pour déterminer la consommation d'énergie, on calcule d'abord la somme des forces appliquées au véhicule sur un trajet donné (équation (6.1)) :

$$\begin{aligned}\sum F(n) = m.g.sin(\alpha_n) + (0.01 + 10^{-5}(\frac{V(n)+V(n-1)}{2})^2)m.g \\ + \frac{1}{2}.C_x.S.\rho.\,[(\,\frac{V(n)+V(n-1)}{2}) - V_{air}(N)]^2 \\ + m.\frac{V^2(N)-V^2(N-1)}{2.\Delta L}\end{aligned} \tag{6.1}$$

Où C_x est le coefficient de traînée aérodynamique dans l'air, S est la projection sur la surface avant du véhicule (m^2), ρ est la densité de l'air ambiant ($1,227kg.m^{-3}$) et V_{air} est la vitesse de l'air par rapport au sol (ms^{-1}). $V(N)$ est la vitesse initiale sur le segment numéro N. Après plusieurs transformations en prenant en compte la vitesse initiale, le temps du trajet, l'accélération, l'énergie mécanique, la dérivation de l'énergie sur le trajet, nous pouvons décrire notre modèle d'estimation de vitesse optimale via l'équation (6.2) :

$$V_{optimized} = \frac{V(N).T^{'}(N)}{T(N)} \tag{6.2}$$

Où $V_{optimized}$ est la vitesse optimisée qui sera donnée comme vitesse d'instruction au conducteur. $T'(N) = 1 - k.(DE(N) - DE_{moy})$ et $T(N) = DE_{moy}$ est la valeur moyenne de l'énergie dérivée $DE(N)$. k est le coefficient de convergence à fixer en fonction de la vitesse/sensibilité souhaitée (0.05 dans notre cas).

Afin de valider notre approche, plusieurs scénarios ont été expérimentés en banc de test et en conditions de trafic réel à savoir : (i) optimisation d'énergie, (ii) optimisation de la distance (sans contrainte de temps) et (iii) optimisation de l'énergie et du temps de parcours (scénario hybride qui optimise, sous contraintes, l'énergie en prenant en compte le temps de parcours). Un des scénarios comprenait un trajet d'essai de $3km$ de long, avec 31 mètres de variation d'altitude et une pente de 4% maximum. Les résultats de simulation montraient un gain de consommation d'énergie de 18% (12.5% en banc à rouleau). La figure 6.1 illustre un exemple de résultats obtenus.

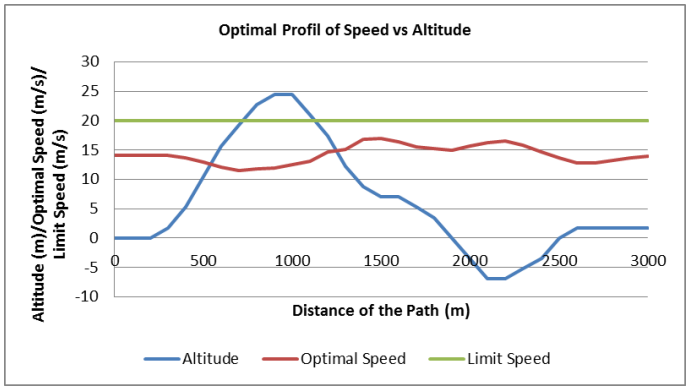

FIGURE 6.1 – Exemple de profile de la vitesse optimale estimée par notre modèle.

ADAS basé vue panoramique autour du véhicule

Cette fonctionnalité est assurée par le système de fusion multicapteurs basé sur 4 caméras fish-eye afin de fournir une vue panoramique autour du véhicule (i.e. vue d'oiseau ou bird-view). Ce système ADAS ne contribuait pas directement à l'optimisation de la consommation d'énergie, mais aidait le conducteur à adopter une écoconduite censée réduire indirectement sa consommation comme par exemple l'aide au stationnement, le changement de voie, détection d'obstacles, piétons et angles morts.
Les caméras sont placées de chaque côté du véhicule et orientées vers le sol à environ 45° ce qui permet de couvrir non seulement la surface du sol, mais aussi l'environnement de

5 à 8 mètres à la verticale. Notre algorithme birdview comprenait 3 étapes : 1. correction de la distorsion causée par l'objectif fish-eye (une matrice des coefficients de distorsion est calculée), 2. transformation en rotation des images pour les exprimer dans le même repère recalage et 3. recalage d'images, où les différents points d'entrées sont mappés dans l'image de sortie finale dans les coordonnées respectives pour former la vue d'ensemble recalée (birdview). Nous avons aussi effectué un birdview avec une deuxième approche basée Speeded Up Robust Features (SURF) [248] donnant de meilleurs résultats en précision mais au détriment de la vitesse ($0.5s$ par image), non adaptée aux contraintes temps-réel de l'embarqué. La figure 6.2 illustre un exemple de résultat de birdview obtenu.

FIGURE 6.2 – Exemple de birdview obtenu à partir de 4 images caméras fish-eye.

ADAS basé réalité augmentée

Le système de réalité augmentée repose sur la superposition de l'environnement réel du conducteur, matérialisé par le flux vidéo (champ de vision du conducteur), et des données virtuelles (guidage GPS, vitesse optimisée, énergie consommée, autonomie, etc.). Le système développé permet ainsi de superposer l'itinéraire de la trajectoire à suivre par le véhicule (i.e. Google Maps) sur le flux vidéo temps-réel, obtenu par la caméra embarquée. Cela permet au conducteur de visualiser le flux vidéo enrichi de données virtuelles utiles pour la navigation comme : instruction de vitesse, énergie estimée, autonomie, etc. Ce module ADAS (comme le système birdview) contribue indirectement dans l'optimisation énergétique permettant ainsi

au conducteur d'adopter un comportement d'écoconduite. La figure 6.3 illustre un exemple de l'interface de réalité augmentée développée. Ces travaux de recherche ont fait l'objet de publications dans [23] et [32].

FIGURE 6.3 – Exemple d'interface de réalité augmentée développée.

6.3 Amélioration de l'estimation de la consommation d'énergie par machine learning

6.3.1 Contexte & objectifs

Trois classes principales sont considérées comme des facteurs significatifs lors de la prédiction de la consommation d'énergie d'un VE : le véhicule, le comportement du conducteur et l'environnement. Ces classes tiennent compte des paramètres constants ou variables qui influencent la consommation d'énergie à bord. Dans le cadre des travaux de recherche du projet VIRTUOSE, nous avons développé un nouveau modèle prenant en compte les trois classes ainsi que leur interaction afin d'améliorer l'estimation de la consommation d'énergie des VEs. Le modèle dépend désormais d'une nouvelle approche basée machine learning et notamment l'algorithme $k-NN$. En suivant un paradigme d'apprentissage paresseux (Lazy Learning Paradigm - LLP), notre modèle permet une meilleure performance d'estimation. L'avantage de notre modèle, par rapport aux approches déterministes développées dans la section précédente, est sa prise en compte de la situation réelle de l'écosystème et ce, sur la base de données historiques. En effet, le comportement du conducteur (style de conduite, utilisation du chauffage, l'air conditionné, l'état de la batterie) a un impact direct sur la consommation d'énergie et donc l'autonomie du véhicule. Les résultats obtenus montrent que nous pouvons atteindre une précision de $96,5\%$ sur l'estimation de la consommation

d'énergie. La méthode proposée est utilisée afin de trouver le chemin optimal entre deux points (départ-destination) en termes de consommation d'énergie [20].

6.3.2 Algorithme d'estimation de la consommation d'énergie

Les modèles existants étaient basés à la fois sur des données partagées en temps-réel, comme le trafic, et sur un modèle mathématique pour la consommation d'énergie afin de calculer le chemin optimal. Dans notre approche, nous proposons de considérer les données historiques afin de prédire la consommation d'énergie sur un trajet. Pour cela, nous avons fait appel à l'algorithme k-NN [249] d'apprentissage paresseux. Le modèle prend en entrée plusieurs paramètres : ID du conducteur, point de départ, point d'arrivée, date, heure, type du véhicule, trafic, météo, ainsi que l'énergie consommée et fournit en sortie une prédiction de la consommation relative à un trajet donné par le conducteur. Le modèle propose le chemin optimal en termes de consommation d'énergie. Notre solution innovante combine deux approches complémentaires déterministe et stochastique pour calculer la moyenne des valeurs des k voisins les plus proches. Si par contre le système se trouve face à une situation ne comprenant pas assez de données (données extraites inférieurs à k), il bascule vers l'approche déterministe basée sur nos modèles [32, 23], dans le cas contraire, il reprend l'approche stochastique par machine learning pour prédire la consommation d'énergie du trajet qui est égale à (i.e. équation (6.3)) :

$$\widehat{E_c}(A, B) = \sum_{i=1}^{k} \alpha_i E_c(A_i, B_i) \tag{6.3}$$

Où $E_c(A_i, B_i)$ est la fonction qui estime la consommation d'énergie du i-ème trajet du point A_i au point B_i, k est le nombre de points voisins sélectionnés. α_i représente le poids du i-ème trajet où $\sum_{i=1}^{k} \alpha_i = 1$. Nous attribuons le plus grand poids aux voisins les plus proches du trajet $A - B$ qu'aux voisins les plus éloignés.

6.3.3 Algorithme du plus court chemin en consommation d'énergie

Afin de proposer le chemin optimal en termes de consommation d'énergie, nous nous sommes inspirés de notre algorithme. La fonction de coût est présentée dans l'équation (6.4) :

$$f(n) = t(n) + h(n) \tag{6.4}$$

Où $t(n)$ représente la valeur réelle du coût pour atteindre le nœud n. $h(n)$ représente l'estimation de la consommation d'énergie calculée par notre algorithme. Notre algorithme du plus court chemin (Shortest Path Algorithm (SPA)) maintiendra également deux listes,

ouverte (openlist) et fermée (closelist). Les nœuds de la liste ouverte sont ceux qui vont être visités tandis que les nœuds de la liste fermée sont ceux qui ont déjà été visités.

La première étape consiste à générer les données d'entraînement. Tout d'abord, dans la partie de la simulation du flux de trafic, nous utilisons le script RandomTrips fourni par SUMO [250] ainsi que le comportement du VE généré conforme à la distribution de Poisson. Nous ajoutons le VE au trafic, les positions de départ et de destination sont distribuées aléatoirement. Nous supposons ici que le nombre de VEs nouvellement ajoutés est relativement faible par rapport au flux de trafic initial dans la zone d'étude. Pendant la simulation, nous devons enregistrer les trois paramètres à savoir : la distance parcourue, la durée du trajet et la consommation d'énergie du VE jusqu'à atteinte sa destination. Le modèle garde toujours en entrée ses 6 paramètres significatifs (coordonnées du point de départ et d'arrivé, l'heure de départ, état du trafic, type de véhicule ainsi que son identifiant). La deuxième étape consiste à appliquer notre approche basée $k-NN$ pour estimer la consommation d'énergie. Quant à la troisième et dernière étape, elle consiste à appliquer l'algorithme proposé d'estimation de la consommation d'énergie à l'algorithme du plus court chemin. Il suffit ainsi de modifier la partie de la fonction heuristique par le calcul de la distance euclidienne.

6.3.4 Résultats expérimentaux

Dans la figure 6.4, nous présentons la variance expliquée des données de test en choisissant différentes valeurs de $k-values$ (de 1 à 10) afin d'obtenir le meilleur choix (ici 5). Plus la valeur est proche de 1, plus le modèle est précis. De même, nous pouvons voir l'erreur quadratique moyenne en fonction des valeurs de $k-values$.

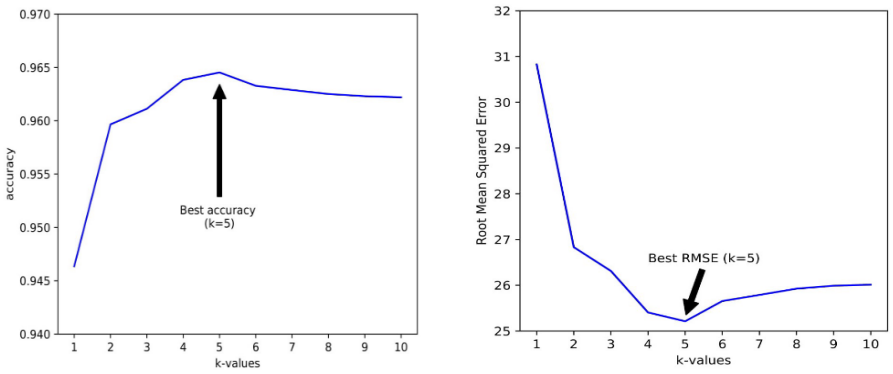

FIGURE 6.4 – Exemple de résultats de précision et d'erreurs quadratique moyenne. A gauche : score de régression de la variance en fonction des valeurs de k. A droite : erreur quadratique moyenne (RMSE) en fonction des valeurs de k.

Dans la figure 6.5, un exemple de distribution des données obtenues par notre modèle est illustré où chaque point de la courbe représente un trajet. Les valeurs $x-value$ et $y-value$

représentent simultanément la consommation d'énergie réelle (obtenue par simulation) et estimée (obtenue par notre modèle). Nous avons fourni une solution basée sur une approche d'apprentissage paresseux afin d'obtenir la meilleure estimation de la consommation d'énergie et de proposer un chemin optimal, en terme d'énergie, pour atteindre une destination. Les résultats obtenus étaient très prometteurs. Ces travaux ont fait l'objet de plusieurs publications dans [20], [23] et [32].

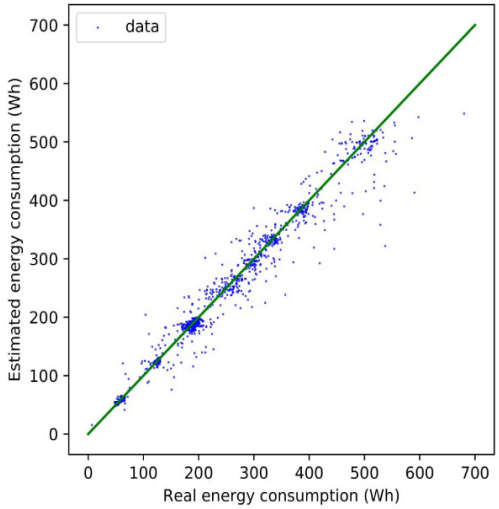

FIGURE 6.5 – Distribution de la consommation d'énergie estimée pour $k = 5$.

6.3.5 ADAS V2G pour l'interaction véhicule autonome et Smart Grid

Ces travaux de recherche ont été développé dans le cadre du projet SAVEMORE (The Smart Autonomous Vehicle for Urban Mobility using Renewable energy) visant à développer et à démontrer la viabilité et l'efficacité des systèmes de transport et de logistique urbaine. Ces systèmes se basent sur des VEs robotisés autonomes fonctionnant dans le cadre d'un réseau intelligent de distribution d'énergie électrique. Nous nous sommes donc intéressés au couplage des VEs autonomes avec le Smart Grid. À l'échelle d'une ville, les VEs peuvent être considérés comme un moyen de stockage intermittent de l'énergie électrique qui peut être distribuée au réseau lorsqu'il est nécessaire. J'ai donc développé un nouveau ADAS exploitant un modèle Vehicle to Grid (V2G) [251] qui met en œuvre l'interaction entre le VE et le Smart Grid. Le modèle prend en compte 14 paramètres liés à la batterie, à la station de recharge, la taille de la flotte, coefficient d'expansion du réseau électrique, etc., et génère en sortie l'estimation de la consommation d'énergie ainsi que celle qui pourrait être envoyée au Smart Grid depuis le VE.

Nous avons pu alléger notre modèle V2G en utilisant les plans d'expériences (Design of

Experiments - DOE [252]) afin de passer de 14 paramètres initialement identifiés, à seulement 4 paramètres dits "significatifs". Plusieurs simulations ont été réalisées pour valider le modèle proposé. Dans un premier temps, nous avons étudié le comportement du modèle sur une journée type d'un trajet de VE domicile-travail. Par la suite, et afin d'étudier le comportement du modèle, nous l'avons testé sur plusieurs saisons. Les résultats montrent l'efficacité du modèle développé (e.g. figure 6.6).

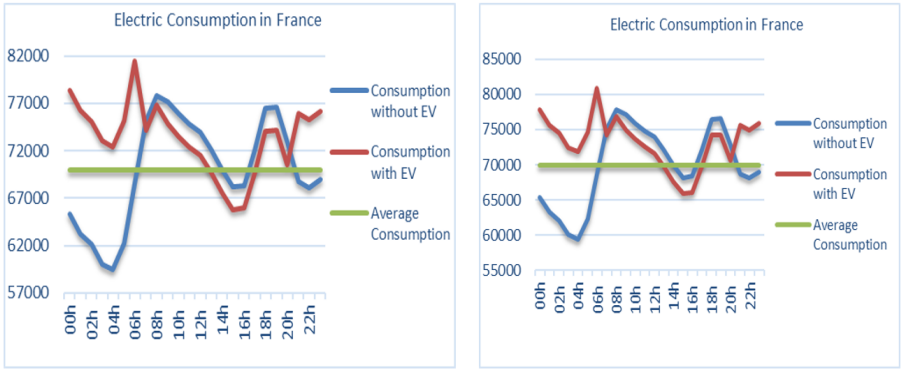

FIGURE 6.6 – Consommation d'énergie : (à gauche) : par coefficient d'expansion, (à droite) : par temps charge/décharge batterie (axe des x en heures et axe des y en MWh).

Nous avons constaté, au cours de nos différentes simulations, que malgré les 2 millions de VEs ajoutés au parc automobile à l'horizon de 2020, et malgré la consommation supplémentaire d'énergie causée par le grand nombre de VEs, le comportement général du Smart Grid n'a pas changé d'une manière très significative. L'énergie des VEs renvoyée dans le réseau est importante et pourrait, à terme, soulager le Smart Grid. Cependant, le retour régénératif dans le réseau n'affecte pas la durée de vie des batteries lithium-ion. Notre modèle V2G montre que pour une flotte de 3 millions de VEs, nous pouvons injecter dans le réseau environ $31,6$GWh par jour, ce qui représente par exemple la production d'énergie de deux réacteurs nucléaires ou d'un parc éolien de 645 éoliennes. Cela illustre l'efficacité de notre modèle basé V2G et les avantages de l'utilisation des VEs dans un environnement Smart Grid. Ces travaux ont fait l'objet de publication dans [33].

6.4 Diagnostic et détection de défauts à distance

6.4.1 Contexte & objectifs

Ces travaux ont été effectués dans le cadre de la thèse en cotutelle, entre l'ESIGELEC et l'Université Politehnica de Bucarest [1], de Lavinius Ioan GLIGA (2016-2019) [1]. Les éoliennes

1. `https://upb.ro/`

à entraînement direct comprennent plusieurs défauts lors de leur fonctionnement. L'analyse de la signature du courant est souvent utilisée pour rechercher des problèmes du générateur. La Transformée de Fourier Rapide (TFR) est utilisée pour calculer le spectre des courants mais présentait des limites. Le Filtre de Kalman Étendu (FKE) est lui aussi proposé comme solution pour une prédiction robuste du courant de l'éolienne. Lors de l'utilisation du FKE, un défi consistait à estimer la matrice de covariance pour le bruit du processus. Nous avons donc développé une nouvelle approche pour le calcul de la matrice de covariance du bruit. Cependant, les parcs éoliens (Wind Farm - WF) se situent dans des zones isolées où l'infrastructure de communication nécessaire au système de contrôle et de surveillance demeure coûteuse. Les WF comprennent souvent des turbines éoliennes (Wind Turbines - WT) qui sont réparties géographiquement. Les Réseaux de Capteurs Sans Fil (RCSF) et l'Internet des Objets (Internet of Thing - IoT) ont été présentés comme des solutions à ces problèmes. Nous avons développé une architecture de RCSF basée sur LoRa et validée sur une plateforme à échelle réduite. C'est donc sur cette partie de RCSF que cette section sera focalisée.

6.4.2 Prédiction des paramètres

Le recours au FKE est indispensable afin d'estimer l'ensemble des paramètres du WF et ce, particulièrement lorsque le système subit des problèmes liés à une coupure de communication, défaillance capteur ou cyberattaque. La matrice de covariance R du FKE peut être calculée comme suit (équation (6.5)) [253] :

$$\begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & ... & \sigma_n^2 \end{bmatrix} \tag{6.5}$$

Où σ_i, $i = \{1, 2, ..., n_y\}$ est l'écart-type pour chaque canal de mesure. Le modèle d'espace d'état non linéaire du processus réel est présenté dans l'équation (6.6)) [253] :

$$x_k = f(x_{k-1}, u_k) + w_k y_k = h(x_k) + v_k \tag{6.6}$$

Où $w_k \in R^{n_x}$, est le bruit ou l'incertitude qui affecte le processus et $v_k \in R^{n_y}$, est la perturbation des mesures. L'estimation de l'erreur est présentée dans l'équation (6.7)) [253] :

$$\epsilon_k = y_k - \widehat{y}_k = h(x_k) + v_k - h(\widehat{x}_k) \tag{6.7}$$

Où $h : R^{n_x} \longrightarrow R^{n_y}$ est une fonction linéaire définie comme $h(x) = Ax+b$, où $A \in R^{n_y * n_x}$ est une matrice inversible et $b \in R^{n_y}$ est un vecteur.

6.4.3 Architecture du système de diagnostic à distance par RCSF

Vue d'ensemble

La chaine de surveillance, de contrôle et de supervision d'un WF nécessite non seulement de collecter les données pour assurer le diagnostic du système (état de fonctionnement, production temps réelle de l'éolienne, défaillance, etc.), mais aussi leur traitement (identification, décision) ainsi que le déclenchement d'actions de haute précision. Même avec les contraintes imposées par la localisation des WTs, il est nécessaire d'assurer une communication fiable et temps-réel entre le WF et son opérateur. Pour faciliter la gestion des WF, les IoT basés RCSF pourraient être de bons candidats en tant qu'outils de communication robustes et peu coûteux. Nous avons non seulement effectué plusieurs simulations pour étudier le comportement de la plateforme, mais aussi développé une plateforme, à échelle réduite, afin d'étudier les différents scénarios de diagnostic et de surveillance d'éoliennes : production énergétique temps-réel, état de santé du parc éolien, contrôle à distance d'éolienne et télémaintenance. Elle comprend deux grandes parties : (*i*) système de contrôle et de diagnostic à distance (poste de commande) et (*ii*) système de mesure et de collecte de données (WF). Les deux parties communiquent via LoRa. La figure 6.7 illustre la plateforme développée.

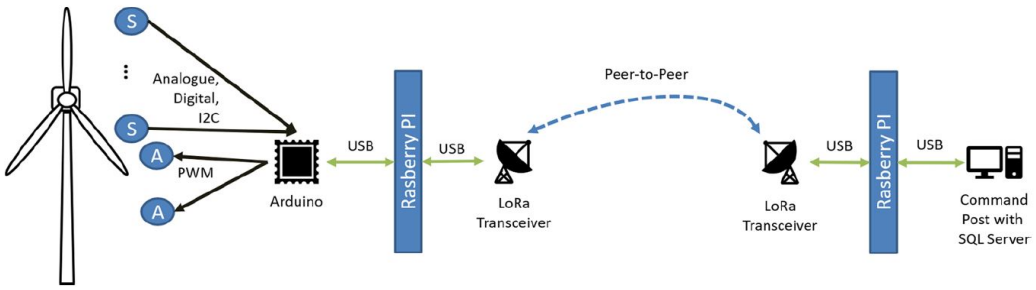

FIGURE 6.7 – Plateforme de diagnostic et de surveillance à distance basée IoT et LoRa.

Système de mesure et de collecte de données

L'éolienne est instrumentée avec plusieurs capteurs permettant la collecte de données : vitesse et direction du vent, température, humidité, pluie, luminosité, tension, courant et puissance produite. En plus, chaque éolienne se voit octroyée un identifiant assuré par un capteur dédié afin de l'identifier dans le parc éolien. Si nécessaire, le système peut aussi contrôler la direction de l'éolienne. L'ensemble de ces capteurs est connecté à un calculateur, qui récupère et traite les données issues du WF, afin de les transmettre, via le réseau LoRa,

vers le système de contrôle et de diagnostic à distance. Les RCSF sont rapides à installer, faciles à entretenir et s'adaptent facilement. Les exigences d'un RCSF potentiel pour un champ éolien WF ont été étudiées. Différentes technologies de communication sans fil ont été comparées de manière approfondie. Les protocoles de faible puissance à longue portée et les protocoles à grande vitesse à courte portée sont considérés.

Système de contrôle et de diagnostic à distance

Pour des raisons de fiabilité, la communication entre le système de contrôle/diagnostic à distance et le système de collecte des données est assurée par deux protocoles de communication différents : WIFI et LoRa. La communication via WIFI permet de tester rapidement différents scénarios, afin de vérifier la faisabilité de l'architecture proposée (courte portée). La communication via LoRa, quant à elle, vise à reproduire les conditions d'un fonctionnement réel à grande échelle avec une communication à large portée (dizaines de kilomètres). Le réseau mis en œuvre a une portée de 30km et une fréquence de 868Mhz, avec un mode RF point à point. Une application web a été développée intégrant une IHM pour montrer l'état du WT simulé et afficher l'ensemble des données. Elle permet aussi le contrôle de l'éolienne : mise en marche, arrêt et changement de direction. Les données reçues via le réseau LoRa sont transférées vers une base de données MySQL.

Résultats expérimentaux

Dans l'environnement urbain, la communication temps-réel avec un taux d'échantillonnage d'une seconde ainsi que la distance de $7km$ demeurent suffisants pour la surveillance à distance d'un parc éolien. L'envoie des données à intervalles de $10mn$ est appropriée avec une perte négligeable ($< 1\%$). LoRa peut être utilisé pour les milieux sans couverture $4G$. Autre innovation de nos travaux est celle de la faculté du système à continuer à estimer l'ensemble des paramètres du parc éolien même en mode dégradé (coupure de communication, défaillance d'un capteur, etc.) et ce, grâce au filtre FKE développé. Ces travaux ont fait l'objet de plusieurs publications dans [1], [2] et [29].

6.5 Tracking temps-réel de véhicules par RFID

6.5.1 Contexte & objectifs

Les travaux de tracking temps-réel par RFID ont été développés dans le cadre du projet collaboratif Roll-On Roll-Off (RORO-MAX) qui avait pour objectif le développement de l'activité logistique du Grand Port Maritime du Havre (GPMH). Ce développement passait

par une augmentation significative du trafic RORO et donc une gestion plus intelligente de ce flux. Dans ce contexte, nous avons développé une solution de traçabilité de véhicules basée sur la RFID pour mieux gérer le grand nombre de véhicules à l'import et à l'export, et permettre ainsi aux différents acteurs du terminal RORO d'utiliser un outil efficace et robuste. Afin de valider et qualifier la solution développée, des mesures et tests en environnement intérieur et extérieur ont été effectués.

6.5.2 Tracking temps-réel par RFID

Dans le cadre du développement de la solution technique de traçabilité, une campagne de mesures a été réalisée, d'une part pour la caractérisation et d'autre part pour valider leur utilisation dans le contexte d'un port RORO. Dans un premier temps, les mesures ont été réalisées en environnement contrôlé (intérieur), en chambre semi-anéchoïque (IRSEEM). Ces mesures ont été réalisées pour caractériser le diagramme de rayonnement et l'immunité aux sources d'interférence RF (immunité rayonnée). Dans un deuxième temps, des mesures ont été effectuées en outdoor, dans des conditions proches de celles du terminal portuaire.

Nous avons fait appel à deux technologies RFID : la première basée sur le lecteur RFU63 [254] qui est adapté aux applications industrielles [255, 256]. La deuxième solution se composait de deux lecteurs compacts de type MRU RFID utilisés avec des tags UHF passifs ou semi-passifs. Les lecteurs MRU sont spécialement conçus pour l'identification automatique des véhicules (Automated Vehicle Identification - AVI). Ils sont particulièrement adaptés aux applications où la technologie passive et la longue portée sont requises. Les essais ont été conçus pour vérifier la détection de véhicules à différentes vitesses (de 5 à $35 km/h$) de passage devant le lecteur RFID ainsi que différentes orientations de l'antenne (10 à 70°). Plusieurs tests ont été réalisés en indoor (i.e. diagramme de rayonnement et immunité rayonnée) et outdoor (i.e. déploiement de l'étiquette RFID, distance lecteur/étiquette et détection du passage du véhicule) en tenant compte de trois configurations principales : antenne, étiquette et lecteur (e.g. figure 6.8). Ces travaux ont fait l'objet de plusieurs publications dans [25], [34] et [45].

FIGURE 6.8 – Protocole de tests en indoor (ligne du haut) et outdoor (ligne du bas).

6.6 Conclusion

Nous avons présenté dans ce chapitre un système prédictif innovant pour l'optimisation de la consommation d'énergie d'un VE. Le modèle prend en compte les différentes forces appliquées au VE mais aussi la topologie de la route pour proposer une vitesse adaptative réduisant la consommation d'énergie embarquée. Les résultats obtenus montrent un gain de consommation d'énergie de $12,5\%$ sur un trajet de $3km$. D'autres modules associés au système ont permis d'optimiser davantage la consommation d'énergie, comme le système ADAS vue panoramique autour du VE ainsi que le système ADAS réalité augmentée. Afin d'améliorer notre approche et d'offrir une meilleure estimation de la consommation d'énergie, nous avons développé une deuxième approche stochastique par apprentissage paresseux. Les résultats obtenus en simulation sont très prometteurs. Par la suite, nous avons présenté notre modèle V2G incluant plusieurs paramètres significatifs qui a fait l'objet de plusieurs simulations pour démontrer sa faisabilité. Nous avons allégé, par plans d'expérience, notre modèle V2G en passant de 14 à 4 paramètres significatifs. Les résultats obtenus montrent l'intérêt de l'interaction entre les VE et le Smart Grid.

Par la suite, nous avons présenté notre plateforme de diagnostic et de détection de défauts à distance pour un champ éolien. Notre architecture basée LoRa et IoT a montré son efficacité et ce, sur plusieurs aspects : une communication temps-réel avec un taux d'échantillonnage d'une seconde sur des distances d'environ $7km$, une perte des données négligeable ($< 1\%$) ainsi que la possibilité d'envoyer les données à intervalles de $10mn$. Notre approche est innovante car même en cas de coupure de communication ou de défaillance des capteurs, le système continue à estimer l'ensemble des paramètres du parc éolien et ce, grâce à notre

filtre FKE développé. Enfin, nous avons présenté notre étude de faisabilité du déploiement d'une solution de tracking temps-réel de véhicules dans la chaine logistique portuaire. Notre solution de suivi par RFID a été testée et caractérisée dans un environnement contrôlé et à l'extérieur.

Afin d'aller plus loin dans nos travaux, nous sommes en train de développer de nouvelles approches dans la cadre de deux nouveaux projets de recherche : CETRIA [2] (Carnot ESP) et ArtIsmo [3] (ANR). CETRIA a pour objectif de développer une carte d'énergie temps-réel basée IA pour l'écomobilité. Cette dernière permet d'afficher en temps réel le coût énergétique des différents tronçons. Elle sera capable non seulement de visualiser la consommation temps-réel de l'énergie du trafic, mais aussi d'optimiser la consommation en favorisant des itinéraires écologiques à faible consommation. ArtIsmo quant à lui vise à estimer l'ensemble des paramètres d'un véhicule autonome par une approche hybride innovante mêlant des modèles déterministes et stochastiques.

2. Carte d'Energie Temps-Réel basée-IA pour l'Ecomobilité.
3. Intelligent Estimation Algorithms for Smart Mobility.

Chapitre 7

Bilan & Perspectives

Sommaire

7.1 Bilan

J'ai présenté dans ce mémoire un aperçu de mes activités pédagogiques, administratives et de mes travaux de recherche pour la période 2009-2022.
La détection, localisation et tracking d'objets sont des tâches indispensables pour la perception d'environnement. Depuis 2012, le deep learning est devenu un outil très puissant en raison de sa capacité à traiter de grandes quantités de données. L'apparition de nombreuses méthodes basées sur l'apprentissage profond a conduit à des progrès significatifs. Malgré cet engouement à l'IA, peu de méthodes se concentrent sur l'aspect temps-réel, essentiel pour les applications réelles et ce, en raison des coûts de calcul élevés. En plus, ces algorithmes présentent des lacunes évidentes, dans les scènes complexes en partie, à cause du manque de données vérité terrain comme c'est le cas pour la smart mobilité ferroviaire et la santé.
Mes travaux de recherche touchent à la détection, localisation et tracking temps-réel d'objets 2D et/ou 3D. Ils se concentrent donc sur la perception d'environnement pour deux domaines de la smart mobilité : routier/ferroviaire et robotique mobile/santé. L'objectif est d'atteindre un niveau d'analyse et de compréhension de scènes complexes permettant d'assurer une smart mobilité de très haut niveau de sécurité, de confort et d'énergie optimale.

J'ai tout d'abord commencé par les approches classiques de vision par ordinateur avant d'adopter les algorithmes d'apprentissage profond dédiés à la smart mobilité. Cela repose sur deux briques essentielles et complémentaires : (*i*) système fusion multicapteurs permettant d'enrichir davantage la perception avec des données hétérogènes, (*ii*) perception d'environnement basée IA permettant l'exploitation des données collectées pour une meilleure prédiction de l'ensemble des situations. J'ai toujours voulu développer une chaine complète de collecte et traitement de données permettant une perception d'environnement robuste. J'ai commencé par la détection d'objets 2D, l'estimation de distance, le tracking temps-réel et la détection d'objets 3D. Par la suite, j'ai développé des approches de segmentation sémantique et de compréhension de comportement d'agents routiers afin de mieux analyser et comprendre une scène. Cela s'est traduit par le développement de plusieurs plateformes génériques ouvertes pour expérimenter et valider des concepts technologiques et scientifiques du monde académique et industriel.

Ces travaux ont fait l'objet de publications dans douze revues internationales avec comité de lecture (rang A), dans une vingtaine de conférences et ont permis le co-encadrement de cinq thèses, l'encadrement de cinq ingénieurs de recherche, trois postdoctorants ainsi qu'une trentaine de stages Master 2 et Ingénieurs. Ces travaux ont été menés via une dizaine de projets de recherche régionaux, nationaux et européens. J'ai aussi pu mener mes travaux vis des partenariats et contrats industriels particulièrement dans la smart mobilité multimodale.

J'ai pu développer et publier quatre datasets dont une représentant une première mondiale. J'ai aussi développé plusieurs plateformes à forte valeurs ajoutées (e.g. fauteuil roulant intelligent (TRL8), véhicules autonomes (TRL6), etc.).
Dans les sections suivantes, je présenterai un bilan de mes travaux de recherche, quelques perspectives concernant mes futurs travaux avant de terminer avec les moyens mis en œuvre pour mener à bien mes activités de recherche.

7.2 Projet de recherche - Perspectives

Mon projet de recherche s'articule autour de trois axes : 1. Détection et tracking temps-réel d'objets 3D, 2. Segmentation sémantique de scènes multimodales et 3. Analyse et compréhension de scènes complexes multimodales. L'objectif est donc d'améliorer l'analyse et la compréhension de scènes complexes par perception en faisant appel, non seulement à un système de fusion multicapteurs, mais aussi et surtout au deep learning.

7.2.1 Détection et tracking temps-réel d'objets 3D

Pour la smart mobilité, il est nécessaire d'avoir une perception précise de l'environnement pour garantir une prise de décision fiable, et de pouvoir étendre les résultats obtenus à d'autres domaines comme la santé, l'industrie 5.0, etc. À cette fin, mes travaux actuels ont porté sur la détection, localisation et tracking temps-réel d'objets 2D/3D pour une meilleure perception de l'environnement (travaux de thèse d'Antoine Mauri, Samar Chebbi, Hicham BESSAFA et Alexandre Evain, travaux de postdoc de Rim Trabelsi, travaux d'ingénieurs de recherches : Yassine Nasri, Louis Lecrosnier, Aristide Laignel, Edgar Petit et Vishnu Pradeep). J'ai pu donc développer de nouveaux réseaux CNN monoculaires et stéréoscopiques de détection d'objets 2D et 3D temps-réel optimisés aux contraintes embarquées (légèreté, vitesse et précision). Je souhaite donc continuer dans cette voie afin d'optimiser les modèles CNN à des applications embarquées temps-réel. Cela passera aussi par le développement de nos propres datasets afin de prendre en compte l'aspect temps-réel lors de l'entrainement de nos modèles à l'image du dataset hybride (virtuel et réel) et multimodale (routier et ferroviaire) ESRORAD développé dans le cadre des travaux de Thèse d'Antoine Mauri.

7.2.2 Segmentation sémantique de scènes multimodales

Dans la smart mobilité, la segmentation sémantique est une tâche importante pour la compréhension de l'environnement. Une meilleure description sémantique de la scène permet d'améliorer la prise de décision, ce qui, à son tour, permet des actions de très haute précision.

C'est l'un des objectifs de mes travaux effectués dans deux domaines distincts de la smart mobilité :

1. Santé : via les travaux d'ingénieurs de recherche A. Laignel, E. Petit et V. Pradeep sur le développement de cartes sémantiques pour un fauteuil roulant en indoor et outdoor. Un nouveau jeu de données de courtes séquences d'images extérieures de scènes de rue prises depuis des points de vue situés sur des trottoirs dans un environnement virtuel 3D (simulateur CARLA) a été développé. Cela a permis de pallier à la défaillance des datasets pris du point de vue trottoir. Nous avons aussi développé une nouvelle architecture de réseau CNN avec un traitement temporel de segmentation sémantique améliorant les performances en précision.
2. Routier/ferroviaire : via les travaux de Rim Trabelsi dans son postdoc sur l'analyse de comportement d'agents pour la sécurité multimodale où nous avons développé un nouveau réseau spatio-temporel appelé STAF [16] basé sur les réseaux LSTM[141].

L'amélioration des performances en termes de précision conduit en général à une augmentation de la complexité des CNNs et donc à une diminution du taux d'images par seconde (FPS). De nombreux algorithmes privilégient la précision, quel qu'en soit le coût de temps de traitement, mais de plus en plus d'entre eux cherchent à optimiser le compromis entre précision et vitesse d'inférence, qui est directement lié au nombre de paramètres des modèles [4]. Mes travaux à court terme traite donc cette problématique d'optimisation des modèles deep learning pour qu'ils soient compatibles avec les contraintes de l'embarqué. Autre verrous à lever dans le domaine de la segmentation sémantique est celui de la disponibilité des datasets. Celles disponibles dans la littérature sont dédiés à des environnements particuliers et ne peuvent donc pas être transposées à d'autres. Les travaux de thèse d'Antoine Mauri ont permis de développer un nouveau dataset ESRORAD hybride (virtuel et réel) et multimodale pour la smart mobilité routière et ferroviaire. Mes travaux de recherche effectués dans le cadre du projet ADAPT ont aussi permis de développer deux datasets, un virtuel pour l'outdoor sur trottoirs (V. Pradeep et B. Decoux) et l'autre réel pour l'indoor (E. Petit, L. Lecrosnier et A. Laignel), dédiés au fauteuil roulant intelligent.

7.2.3 Analyse et compréhension de scènes complexes multimodales

Depuis les deux dernières décennies, l'analyse des données visuelles a connu un progrès très important ce qui a permis l'apparition d'un grand nombre de techniques et d'algorithmes innovants à caractère évolutif. Parmi ces algorithmes, le deep learning a révolutionné la vision par ordinateur [257][258][259]. Dans la perception d'une scène par exemple, le deep learning apporte une détection et une localisation d'objets avec une très grande précision au regard

des approches classiques de vision par ordinateur. Cela fait plusieurs années que je travaille sur les problématiques liées à la détection, localisation et tracking temps-réel d'objets dans des scènes indoor mais aussi et surtout outdoor multimodales routières et/ou ferroviaires. Ces travaux alimentent la perception d'environnement pour la smart mobilité qui elle, in fine, vise l'analyse et la compréhension de scènes complexes.

Cela s'inscrit dans les travaux de Thèse d'Antoine Mauri (détection et tracking temps-réel d'objets 3D), de Samar Chebbi (analyse et compréhension de scènes complexes par deep learning multitâches), d'Alexandre Evain (perception d'environnement par deep learning multitâches pour un train autonome monorail), de Hicham Bessafa (architectures d'estimation hybride basées sur l'apprentissage en ligne pour le suivi de véhicule intelligent), les travaux de la postdoctorante Rim Trabelsi (compréhension du comportement d'agents dans les scènes routières) ainsi que les travaux d'ingénieurs de recherche de Louis Lecrosnier, Edgar Petit, Aristide Laignel et Vishnu Pradeep (détection d'objets 2D et segmentation sémantique 3D d'un fauteuil roulant intelligent).
En revanche, l'exploration des données vidéo reste moins maitrisée et demeure donc difficile à traiter par ces approches et ce, à cause de leur complexité spatio-temporelle. Dans ce contexte, la problématique d'extraction de caractéristiques fines pour la description de vidéos (et donc de scènes) représente un défi à surmonter par la communauté scientifique.

Perspectives

7.2.4 Détection d'objets 3D par deep learning adaptée à l'embarqué

Mon objectif est de développer de plus en plus de modèles CNN légers à l'image des travaux de thèse d'Antoine Mauri sur le développement de CNN léger pour la détection d'objets 3D temps-réel dans des environnements routiers et ferroviaires. L'objectif est de prendre en compte l'aspect spatio-temporel dans nos modèles CNN afin d'explorer davantage les données vidéo et ce, sans baisser la vitesse des traitements qui doit avoisiner les 30 FPS pour les applications embarquées. L'optimisation des approches d'apprentissage profond doit être effectuée afin de rendre nos modèles compatibles avec les applications embarquées en terme de temps de calcul et d'espace mémoire.

7.2.5 Tracking temps-réel et description sémantique

D'un point de vue plus général, la plupart des modèles de segmentation sémantique utilisent des images uniques [260]. Nous pensons que l'information temporelle dans les séquences d'images est utile dans les algorithmes [261], car la variation d'une image à l'autre

est généralement importante. Je continuerai donc dans cette voie pour améliorer nos modèles de segmentation sémantique pour qu'il soient plus rapide et donc temps-réel pour des séquences vidéos (travaux de l'ingénieur de recherche Vishnu Pradeep par exemple), GSV-NET [226][262]. L'objectif est d'économiser les calculs liés à l'extraction de caractéristiques par apprentissage par transfert.

Pour les travaux futurs, nous nous focaliserons sur l'allègement des modèles de tracking pour une meilleure adaptation à nos applications de l'embarquée : smart mobilité routière, ferroviaire, robotique d'assistance, etc. Cela passera tout d'abord par une optimisation du temps de calcul lié à la détection d'objets 3D sans baisser la qualité de détection et ce, contrairement à ce qui a été proposé dans la littérature [263][264]. L'objectif est de renforcer le compromis ente la vitesse et la performance de tracking avec un allègement de nos modèles. L'augmentation de la précision des détecteurs d'objets se fait au détriment de la vitesse, ce qui implique la mise en place de critères de choix de détecteurs d'objets prenant en compte ces contraintes.

L'idée est d'assurer une vitesse élevée tout en maintenant une précision de qualité assurant un bon tracking sans autant baisser le nombre d'images par frame (FPS). Mon objectif est de développer une plateforme générique de tracking et segmentation sémantique qui prendra en compte plusieurs détecteurs d'objets et un modèle CNN de tracking, SORT [263], C++SORT [265] couplé à un autre modèle CNN de segmentation sémantique. Cela reviendra à choisir le meilleur détecteur selon la situation (Filtre de Kalman, CNN, etc.). Pour cela, la thèse CIFRE d'Alexandre Evain a comme objectif de renforcer cette piste. Trois post-doctorants à recruter, en septembre/octobre 2022, sur trois projets de recherche (ArtIsmo ANR, CETRIA Carnot ESP et AntiHPert RIN Recherche), traiteront en grande partie cette problématique liée au tracking et description sémantique.

7.2.6 Analyse de scènes complexes multimodales par apprentissage profond multitâches

Mon objectif est d'améliorer mes approches deep learning spatio-temporelles pour aller représenter à fine-grained (caractéristiques élémentaires) et de manière sophistiquée les séquences spatio-temporelles et employer correctement ces représentations pour détecter, classifier et extraire de plus amples analyses pour la compréhension de scènes. Ces applications visent donc à renforcer la compréhension sémantique de haut niveau d'une « scène complexe ».
La contribution majeure de ces travaux de recherche est de proposer un modèle basé deep learning convolutif et/ou récurrent additionnant/combinant plusieurs types de problèmes à

résoudre par le même modèle et ce, afin d'améliorer les performances du réseau. On parle donc d'apprentissage multitâches (Multi-Task Learning (M-TL))[266]. Les « tâches » sont à définir de façon à ce qu'elles répondent à des problèmes distincts mais complémentaires les uns des autres. En guise d'exemple dans l'analyse de scènes routière : détection de l'ensemble des objets, leur localisation ainsi que leur tracking afin de prédire leur comportement dans la scène. Les tâches du modèle M-TL peuvent être : un véhicule qui va s'arrêter devant le carrefour, un piéton qui va traverser la route, un véhicule qui va changer de file, etc. En guise d'exemple, la possibilité d'effectuer une description textuelle fine de la scène analysée sur la base d'une carte sémantique 2D/3D.

Cela passera certainement par les étapes suivantes : 1. argumentation des choix d'orientation par rapport aux différentes tâches à cohabiter dans le même réseau CNN, 2. collecte des données adéquates et (re-)annotation des différentes séquences pour pouvoir répondre aux problématiques prédéfinies, 3. modélisation multitâches non-linéaire des primitives visuelles soit avec un seul type de réseau soit en adaptant un modèle hybride incorporant deux types différents de réseaux, 4. adaptation des modèles déjà entrainés disponibles dans la littérature, pour résoudre notre problématique et réciproquement, essayer notre modèle, déjà entrainé, pour résoudre d'autres problématiques (apprentissage par transfert), 5. évaluation et qualification des approches proposées sur des datasets issus de la littérature et comparaison avec les approches de l'état de l'art, 6. essais et validation selon les scénarios prédéfinis touchants à la description d'une scène complexe que ce soit en indoor et/ou outdoor (multimodale routière et ferroviaire, santé, robotique d'assistance, industrie 5.0, etc.).

Le système devra être générique pour être utilisé dans tout type d'environnement. Il devra donc être indépendant de la plateforme utilisée. L'objectif est donc de montrer la faisabilité de concepts scientifiques et technologiques qui touchent non seulement à la perception d'environnement, mais aussi à sa compréhension pour une analyse fine et précise.

7.3 Développement du projet : moyens mis en œuvre

Afin de mettre en œuvre mon projet de recherche décrit ci-dessus, plusieurs moyens humains et matériels sont nécessaires.

7.3.1 Mutualisation entre les projets existants

La mutualisation des ressources humaines et des moyens matériels était une démarche fortement adoptée lors de la réalisation de mes travaux de recherche. Cela m'a permis de capitaliser sur les compétences et de mutualiser les équipements de certains projets avec

d'autres dont les financements n'étaient pas suffisants. Plusieurs briques logicielles et matérielles ont été développées dans des projets et réutilisées (avec adaptation) dans d'autres. Un certain nombre de mes travaux de recherche actuels comme par exemple la segmentation sémantique pour un fauteuil roulant intelligent ou encore le tracking temps-réel pour smart mobilité pourraient faire appel aux ressources des projets en cours (ANR ArtIsmo, RIN AntiHPehrt et Carnot ESP CETRIA). Outre les moyens matériels, le projet ANR ArtIsmo prévoit un postdoctorant de 18 mois (qui travaillera sur la détection d'objets 3D par apprentissage profond multitâches) et deux stages Master 2, le projet RIN AntiHPert prévoit un postdoctorant de 18 mois (qui travaillera sur l'analyse du comportement des opérateurs en indoor par tracking temps-réel) et le projet Carnot ESP CETRIA prévoit un postdoctorant de 12 mois qui travaillera sur les ADAS à vocation d'optimisation de la consommation de l'énergie dans le trafic routier et le développement de cartes d'écomobilité.

7.3.2 Renforcement des collaborations

Comme cela a déjà été évoqué dans le chapitre 1 (Section 1.4), plusieurs de mes travaux de recherche sont le fruit de collaborations effectuées via les projets de recherche ou sous forme de valorisations avec nos partenaires industriels (SEGULA, FAAR, CERTAM, CEREMA, etc.). Tout d'abord via les projets de recherche, j'ai pu développer plusieurs collaborations régionales (LITIS, CESI, CEREMA, CERTAM, GREYC, CHU Rouen, ISEL, Unilasalle, Heatself, SITIA, etc.), nationales (INSA de Rennes, CRAN Université de Lorraine, IBISC Université d'Evry, MIS UPJV, Pôle ST Hélier Rennes, NEOMA, etc.) et internationales (University of Kent, UCL, University of Leads, University of Bournemouth, University of Latvia, Essex University, University of Leeds, University of Minnesota, Polytechnique Montréal, SUP'COM, etc.).

Cependant, concernant les collaborations industrielles, certaines sont inscrites dans une logique de partenariat à long terme comme c'est le cas par exemple pour SEGULA Technologies où je continue à développer mes travaux sur la perception d'environnement pour le train autonome. L'entreprise vient de prolonger ce partenariat pour développer de nouvelles solutions innovantes pour le train autonome monorail. Ce partenariat est décliné en 2 thèses et une dizaine de stages Master 2. Un partenariat avec l'entreprise FAAR [1] qui m'a permis de renforcer mes travaux de détection et localisation d'objets 2D dans des scènes routières. Je contribue dans le développement d'un nouveau partenariat entre l'IRSEEM et Faurecia [2] sur le développement de nouveaux ADAS dédiés à l'aide au stationnement (park assist). Je souhaite donc renforcer davantage ces collaborations en développant de nouveaux partenariats

1. https://www.faar-industry.com/
2. https://www.faurecia.com/

via les activités contractuelles ou les projets de recherche. Cela me permettra non seulement de disposer de plus de moyens pour mener à bien mes travaux de recherche, mais aussi de monter en compétences sur des domaines stratégiques : véhicule autonome (voiture, train), robotique collaborative et smart mobilité.

7.3.3 Renforcement de contrats industriels

Comme cela a été mentionné dans le chapitre 1 (Section 1.4), j'ai mené mes travaux via les projets de recherche ainsi les partenariats et contrats industriels. D'ailleurs certains de ces contrats sont encore en cours de développement comme par exemple le partenariat avec SEGULA Technologies qui date depuis 2016 et qui a permis le développement de plusieurs plateformes de smart mobilité routière et ferroviaire. Ce contrat est stratégique car il permet encore d'aller plus loin dans mon axe de recherche sur la smart mobilité routière et ferroviaire en travaillant sur le train autonome monorail. Un nouveau partenariat stratégique est en cours de mise en place avec Faurecia sous forme de contrats R&D et thèses CIFRE. Parmi les projets à développer pour le véhicule autonome : Park Assist Cloud base et Automotive Camera Sensor Fusion. Je souhaite renforcer mes activités contractuelles actuelles en développant de nouveaux contrats et partenariats. Cela me permettra de renforcer mes activités de recherche, en ayant de nouveaux moyens matériels, mais aussi de pouvoir attirer de nouvelles compétences.

7.3.4 Montage de nouveaux projets

L'IRSEEM est un laboratoire de recherche dont le modèle repose sur deux piliers : les projets de recherche et les partenariats avec les industriels. J'ai donc pu effectuer l'essentiel de mes travaux de recherche via les projets de recherche dans lesquels j'ai participé. Selon que l'IRSEEM était porteur du projet ou partenaire, les projets de recherche avaient des financements régionaux (RIN, FEDER, Carnot ESP), nationaux (ANR) ou européens (INTERREG). Je prépare actuellement plusieurs montages de projets : 1. INTERREG North Sea sur la robotique collaborative avec le CESI [3] Rouen, 2. INTERREG North Sea sur la maintenance de la chaine logistique ferroviaire avec NEOMA [4] Paris, 3. Horizons Europe sur le train autonome et 4. RIN Recherche émergent sur le développement d'une carte d'énergie pour véhicules électriques. Les montages de projets me permettent de renforcer les collaborations existantes, de développer de nouveaux partenariats, de monter en compétences sur de nouvelles thématiques de recherche et d'avoir les ressources humaines et matérielles nécessaires à l'avancement de mes travaux de recherche.

3. https://rouen.cesi.fr/

4. https://neoma-bs.fr/

Bibliographie

[1] L. ioan Gliga *et al.*, "Diagnostic d'une turbine eolienne à distance à l'aide du réseau de capteurs sans fil," Ph.D. dissertation, Normandie, 2019.

[2] L. I. Gliga, R. Khemmar, H. Chafouk, and D. Popescu, "A survey of wireless communication technologies for an iot-connected wind farm," *Wireless Personal Communications*, vol. 122, no. 3, pp. 2253–2272, 2022.

[3] A. Mauri, "Détection d'objets 3d et tracking temps-réel par deep learning. application à la smart mobilité routière et ferroviaire," Ph.D. dissertation, Normandie, 2022.

[4] A. Mauri, R. Khemmar, B. Decoux, M. Haddad, and R. Boutteau, "Lightweight convolutional neural network for real-time 3D object detection in road and railway environments," *Journal of Real-Time Image Processing*, Feb. 2022. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03592337

[5] ——, "Real-time 3d multi-object detection and localization based on deep learning for road and railway smart mobility," *Journal of imaging*, vol. 7, no. 8, p. 145, 2021.

[6] A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Boutteau, J.-Y. Ertaud, and X. Savatier, "Deep learning for real-time 3d multi-object detection, localisation, and tracking : Application to smart mobility," *Sensors*, vol. 20, no. 2, p. 532, 2020.

[7] R. Khemmar, A. Mauri, B. Decoux, M. Haddad, R. Boutteau, N. Ragot, L. Lecrosnier, Y. Duchemin, and R. Rossi, "Environment Perception-based Deep Learning. Application to Road and Railway Smart Mobility." in *V-ASET2021. 5th Edition of Applied Science, Engineering and Technology Virtual.*, Teams, United Kingdom, Dec. 2021. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03525808

[8] A. Mauri, R. Khemmar, B. Decoux, T. Benmoumen, M. Haddad, and R. Boutteau, "A comparative study of deep learning-based depth estimation approaches : Application to smart mobility," in *2021 8th International Conference on Smart Computing and Communications (ICSCC)*. IEEE, 2021, pp. 80–84.

[9] A. Mauri, R. Khemmar, R. Boutteau, B. Decoux, J.-Y. Ertaud, and M. Haddad, "A new evaluation approach for deep learning-based monocular depth estimation methods,"

in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.

[10] C. D. J. G. V. V. M. H. Antoine Mauri, Redouane Khemmar and R. Boutteau, "Road and railway smart mobility : A high-definition ground truth hybrid dataset," *Sensors*, vol. 22, no. 10, p. 3922, 2022.

[11] B. Decoux, R. Khemmar, N. Ragot, A. Venon, M. Grassi-Pampuch, A. Mauri, L. Lecrosnier, and V. Pradeep, "A dataset for temporal semantic segmentation dedicated to smart mobility of wheelchairs on sidewalks," *Journal of Imaging*, vol. 8, no. 8, p. 216, 2022.

[12] S. Chebbi, "Analyse de scènes complexes par deep learning multitâches. application à la smart mobilité," Ph.D. dissertation, Normandie, 2020-2021.

[13] H. BESSAFA, "Architectures d'estimation hybride basées sur l'apprentissage en ligne pour le suivi de véhicule intelligent," Ph.D. dissertation, Normandie, 2021-2024.

[14] A. Evain, "Contrôle/commande robuste et perception d'environnement par intelligence artificielle pour véhicules autonomes," Ph.D. dissertation, Normandie, 2022-2025.

[15] V. Pradeep, R. Khemmar, L. Lecrosnier, Y. Duchemin, R. Rossi, and B. Decoux, "Self-supervised sidewalk perception using fast video semantic segmentation for robotic wheelchairs in smart mobility," *Sensors*, vol. 22, no. 14, p. 5241, 2022.

[16] R. Trabelsi, R. Khemmar, B. Decoux, J.-Y. Ertaud, and R. Boutteau, "STAF : Spatio-Temporal Attention Framework for Understanding Road Agents Behaviors," *ieee access*, no. 10, pp. 55 794–55 804, May 2022. [Online]. Available : https://hal.archives-ouvertes.fr/hal-xxxx

[17] R. Trabelsi, R. Khemmar, B. Decoux, J.-Y. Ertaud, and R. Butteau, "Recent advances in vision-based on-road behaviors understanding : A critical survey," *Sensors*, vol. 22, no. 7, p. 2654, 2022.

[18] L. Lecrosnier, R. Khemmar, N. Ragot, R. Rossi, J. Y. Ertaud, B. Decoux, and Y. Dupuis, "Object detection, localisation, and tracking-based deep learning for smart wheelchair," *AMSE, Journal of the Association for the Advancement of Modelling and Simulation Techniques in Enterprises*, 2021.

[19] L. Lecrosnier, R. Khemmar, N. Ragot, B. Decoux, R. Rossi, N. Kefi, and J.-Y. Ertaud, "Deep learning-based object detection, localisation and tracking for smart wheelchair healthcare mobility," *International journal of environmental research and public health*, vol. 18, no. 1, p. 91, 2021.

[20] A. Cabani, P. Zhang, R. Khemmar, and J. Xu, "Enhancement of energy consumption estimation for electric vehicles by using machine learning," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, p. 215, 2021.

[21] L. Ménard, A. Petit, É. Leblong, M. Stein, E. Hatzidimitriadou, R. Khemmar, S. Manship, R. Morris, N. Ragot, and P. Gallien, "Novel Robotic Assistive Technologies : Choosing Appropriate Training for Healthcare Professionals," *Modelling, measurement and control C*, vol. 81, no. 1-4, pp. 43–48, Dec. 2020. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03138290

[22] R. Khemmar, L. Delong, and B. Decoux, "Real Time Pedestrian Detection-based Faster HOG/DPM and Deep Learning Approaches," *International Journal of Computer Applications*, vol. 176, no. 42, pp. 34–38, Jul. 2020. [Online]. Available : https://hal.archives-ouvertes.fr/hal-02910408

[23] A. Cabani, R. Khemmar, J.-Y. Ertaud, R. Rossi, and X. Savatier, "Adas multi-sensor fusion system-based security and energy optimisation for an electric vehicle," *International Journal of Vehicle Autonomous Systems*, vol. 14, no. 4, pp. 345–366, 2019.

[24] Y. Nasri, V. Vauchey, R. Khemmar, N. Ragot, K. Sirlantzis, and J.-Y. Ertaud, "Ros-based autonomous navigation wheelchair using omnidirectional sensor," *International Journal of Computer Applications*, vol. 133, no. 6, pp. 12–17, 2016.

[25] R. Khemmar, F. Bouzbouz, N. Ragot, and X. Savatier, "The application of rfid technology in a port," *International Journal of Computer Applications*, vol. 86, no. 7, 2014.

[26] R. Khemmar, J. Ertaud, and X. Savatier, "Face detection & recognition based on fusion of omnidirectional & ptz vision sensors and heteregenous database," *International Journal of Computer Applications*, vol. 61, 2013.

[27] R. Trabelsi, R. Khemmar, R. Boutteau, B. Decoux, and J. Y. Ertaud, "Toward Comprehensive Road Agents BehaviorUnderstanding," in *CSTI 2020 : 1er Colloque Francophone des Systèmes de Transports Intelligents*, tunis, Tunisia, Nov. 2020. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03016779

[28] R. Khemmar, M. Gouveia, B. Decoux, and J.-Y. Ertaud, "Real time pedestrian and object detection and tracking-based deep learning. application to drone visual tracking," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 01 2019.

[29] R. Khemmar and H. Chafouk, "Diagnostic et détection de défauts à distance des eoliennes à l'aide de l'internet des objets (iot)," in *COFMER'03 (Energie solaire, Energie éolienne, Biomasse & Bioénergie, Efficacité énergétique & Stockage d'énergie)*, 2019.

[30] Z. Chen, R. Khemmar, B. Decoux, A. Atahouet, and J.-Y. Ertaud, "Real time object detection, tracking, and distance and motion estimation based on deep learning : Application to smart mobility," in *2019 Eighth International Conference on Emerging Security Technologies (EST)*. IEEE, 2019, pp. 1–6.

[31] N. Ragot, R. Khemmar, A. Pokala, R. Rossi, and J.-Y. Ertaud, "Benchmark of visual slam algorithms : Orb-slam2 vs rtab-map," in *2019 Eighth International Conference on Emerging Security Technologies (EST)*. IEEE, 2019, pp. 1–6.

[32] A. Cabani, R. Khemmar, J.-Y. Ertaud, and J. Mouzna, "Intelligent navigation system-based optimization of the energy consumption," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 785–789.

[33] R. Khemmar, J.-Y. Ertaud, K. Sirlantzis, and X. Savatier, "V2g-based smart autonomous vehicle for urban mobility using renewable energy," June, 21-26, 2015.

[34] N. R. X. S. Redouane Khemmar, Fadoua Bouzbouz, "Vehicle tracking based rfid technology and its application on automotive supply chain," *10th ITS European Congress, 16–19 June 2014, Helsinki, Finland*, 2014.

[35] A. Raj, R. Khemmar, J. Y. Eratud, and X. Savatier, "Face detection and recognition under heterogeneous database based on fusion of catadioptric and ptz vision sensors," in *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Springer, 2013, pp. 171–185.

[36] J.-X. Liu, R. Khemmar, J.-Y. Ertaud, and X. Savatier, "Compressed sensing face recognition method in heterogeneous database with small sample size problem," in *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*. IEEE, 2012, pp. 80–84.

[37] E. Hirsch, A. Lallement, and R. Khemmar, "Steps Towards an Intelligent Self-Reasoning System for the Automated Vision-Based Evaluation of Manufactured Parts," in *Workshop on Applications of Computer Vision, in conjunction with ECCV, May 12, Graz, Austria, 2006*, Graz, Austria, May 2006. [Online]. Available : https://hal.archives-ouvertes.fr/hal-02948239

[38] A. Mauri, R. Khemmar, B. Decoux, J.-Y. Ertaud, M. Haddad, and R. Boutteau, "Une nouvelle approche pour l'évaluation des méthodes monoculaires d'estimation de la profondeur basées sur l'apprentissage profond," in *ORASIS 2021*, 2021.

[39] J. Y. Ertaud, N. Ragot, R. Da Silva Moura, P. Alias, M. Babel, S. Guégan, L. Devigne, F. Pasteau, R. Khemmar, and R. Rossi, "Modélisation énergétique d'un fauteuil roulant électrique pour prédire la faisabilité d'effectuer un trajet," in *handicap 2020*, paris, France, Nov. 2020. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03050656

[40] R. Khemmar, L. Lecrosnier, R. Rossi, J.-Y. Ertaud, F. Rouvray, B. Decoux, and D. Yohan, "Detection, localisation et tracking d'objets basé deep learning pour un fauteuil roulant intelligent." in *Handicap 2020 Technologies pour l'autonomie et l'inclusion*, 2020.

[41] N. Ragot, R. Khemmar, A. Pokala, R. Rossi, B. Decoux, and J.-Y. Ertaud, "TRAJECTOGRAPHY ESTIMATION FOR A SMART POWERED WHEELCHAIR ORB-SLAM2 VS RTAB-MAP A PREPRINT," in *M2SC (modélisation systémique de systèmes cyber-physiques)*, Poitiers, France, Jun. 2019. [Online]. Available : https://hal.archives-ouvertes.fr/hal-02343517

[42] A. Pokala, R. Khemmar, N. Ragot, R. Rossi, J. Y. Ertaud, and B. Decoux, "Estimation de trajectoire pour un fauteuil roulant électrique," in *Modélisation systémique de systèmes cyber-physiques*, 2019.

[43] R. Khemmar, F. Bonardi, J.-Y. Ertaud, and X. Savatier, "Biometrie authentication platform-based multisensor fusion," in *2017 Seventh International Conference on Emerging Security Technologies (EST)*. Canterbury, France : IEEE, Sep. 2017. [Online]. Available : https://hal.archives-ouvertes.fr/hal-01871013

[44] R. E. Datondji, N. Ragot, Y. Nasri, R. Khemmar, and R. Boutteau, "Odométrie visuelle par vision omnidirectionnelle pour la navigation autonome d'une chaise roulante motorisée," in *Journées francophones des jeunes chercheurs en vision par ordinateur*, Amiens, France, Jun. 2015. [Online]. Available : https://hal.archives-ouvertes.fr/hal-01161916

[45] N. R. X. S. Redouane Khemmar, Fadoua Bouzbouz, "La rfid au service du terminal roulier du gpmh," *ATEC ITS France, Jan 2014, paris, France*, 2014.

[46] R. Khemmar, N. Ragot, F. Bouzbouz, J.-Y. Ertaud, A.-M. Kokosy, O. Labbani-Igbida, P. Sajous, E. Niyonsaba, D. REGUER, H. Hu, K. McDonald-Maier, K. Sirlantzis, G. Howells, M. Pepper, and M. Sakel, "Enhancing the Autonomy of Disabled Persons : Assistive Technologies Directed by User Feedback," in *2013 Fourth International Conference on Emerging Security Technologies (EST)*. Cambridge, France : IEEE, Sep. 2013, pp. 71–74. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03369528

[47] R. Khemmar, R. Trabelsi, B. Decoux, and R. Boutteau, "Development in vision-based on-road behavior understanding," *E scholarly community encyclopedia*, 2022.

[48] R. Khemmar and H. Chafouk, "Diagnostic et détection de défauts à distance des eoliennes à l'aide de l'iot." in *COFMER'03 (Energie solaire, Energie éolienne, Biomasse & Bioénergie, Efficacité énergétique & Stockage d'énergie)*, 2021.

[49] R. Khemmar, A. Lallement, and E. Hirsch, "Design of an intelligent self-reasoning system using a situation graph tree for the automated vision-based 3d reconstruction of manufactured parts," *Design of an Intelligent Self-Reasoning System Using a Situation Graph Tree for the Automated Vision-Based 3D Reconstruction of Manufactured Parts*, 2007.

[50] R. Khemmar, "Extraction contrôlée d'indices images et automatisation de la reconstruction 3D : Application à la mesure dimentionnelle par vision par ordinateur ." Theses, université de strasbourg, Dec. 2005. [Online]. Available : https://hal.archives-ouvertes.fr/tel-03084952

[51] ——, "Compression d'images fixes par critères d'informations." Laboratoire Signal, Image et Commiunication (SIC), Université de Poitiers. ; Mémoire de Master 2 Recherche (DEA Recherche), Research Report, 2002. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03648535

[52] ——, "Développement d'une carte 2D d'environnement statique pour un robot mobile autonome." Mémoire du projet de fin d'études Ingénieur, Université de Batna (Algérie), 2000, Research Report, 2000. [Online]. Available : https://hal.archives-ouvertes.fr/hal-03648538

[53] J. Redmon and A. Farhadi, "Yolov3 : An incremental improvement," *arXiv preprint arXiv :1804.02767*, 2018.

[54] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv :1806.01260*, 2018.

[55] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[56] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano, "Real-time self-adaptive deep stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 195–204.

[57] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[58] S. Yang and M. Baum, "Extended kalman filter for extended object tracking," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4386–4390.

[59] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics : The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[60] A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Boutteau, J.-Y. Ertaud, and X. Savatier, "Deep Learning for Real-Time 3D Multi-Object Detection, Localisation, and Tracking : Application to Smart Mobility," *Sensors*, vol. 20, no. 2, p. 532, Jan. 2020. [Online]. Available : https://hal-normandie-univ.archives-ouvertes.fr/hal-02632529

[61] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd : Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[62] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once : Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[63] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[64] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[65] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn : Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[66] J. Palacín, T. Pallejà, M. Tresanchez, R. Sanz, J. Llorens, M. Ribes-Dasi, J. Masip, J. Arno, A. Escola, and J. R. Rosell, "Real-time tree-foliage surface estimation using a ground laser scanner," *IEEE transactions on instrumentation and measurement*, vol. 56, no. 4, pp. 1377–1383, 2007.

[67] B. Kang, S.-J. Kim, S. Lee, K. Lee, J. D. Kim, and C.-Y. Kim, "Harmonic distortion free distance estimation in tof camera," in *Three-Dimensional Imaging, Interaction, and Measurement*, vol. 7864. International Society for Optics and Photonics, 2011, p. 786403.

[68] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking : A survey," 2019.

[69] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance : The clear mot metrics," 2008.

[70] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," in *CVPR Workshop on the Future of Datasets in Vision*, vol. 2, 2015.

[71] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge : A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[72] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco : Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[73] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[74] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig *et al.*, "The open images dataset v4 : Unified image classification, object detection, and visual relationship detection at scale," *arXiv preprint arXiv :1811.00982*, 2018.

[75] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[76] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9799–9809.

[77] D. Min and K. Sohn, "Cost aggregation and occlusion handling with wls in stereo matching," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1431–1442, 2008.

[78] Kalman, "A new approach to linear filtering and prediction problems," in *Transactions of the ASME Journal of Basic Engineering*, 1960, pp. 35–45.

[79] A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Boutteau, J.-Y. Ertaud, and X. Savatier, "Deep Learning for Real-Time 3D Multi-Object Detection, Localisation, and Tracking : Application to Smart Mobility," *Sensors*, vol. 20, no. 2, p. 532, Jan. 2020. [Online]. Available : https://hal.archives-ouvertes.fr/hal-02446258

[80] R. Khemmar, M. Gouveia, B. Decoux, and J.-Y. Ertaud, "Real time pedestrian and object detection and tracking-based deep learning. application to drone visual tracking," in *WSCG'2019 - 27. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision'2019, May 2019, Plzen, Czech Republic.* Václav Skala-UNION Agency, 2019.

[81] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes : A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[82] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small : Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv :1907.10326*, 2019.

[83] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[84] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[85] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[86] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017.

[87] T. Koch, L. Liebel, F. Fraundorfer, and M. Körner, "Evaluation of cnn-based single-image depth estimation methods," *CoRR*, vol. abs/1805.01328, 2018. [Online]. Available : http://arxiv.org/abs/1805.01328

[88] A. Mauri, R. Khemmar, R. Boutteau, B. Decoux, J.-Y. Ertaud, and M. Haddad, "A new evaluation approach for deep learning-based monocular depth estimation methods," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.

[89] O. Zendel, M. Murschitz, M. Zeilinger, D. Steininger, S. Abbasi, and C. Beleznai, "Railsem19 : A dataset for semantic rail scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[90] G. Jocher, A. Stoken, and J. B. et al., "ultralytics/yolov5 : v4.0 - nn.SiLU() activations, Weights & Biases logging, PyTorch Hub integration," Jan. 2021. [Online]. Available : https://doi.org/10.5281/zenodo.4418161

[91] G. Singh, S. Akrigg, M. Di Maio, V. Fontana, R. J. Alitappeh, S. Saha, K. Jeddisaravi, F. Yousefi, J. Culley, T. Nicholson *et al.*, "Road : The road event awareness dataset for autonomous driving," *arXiv preprint arXiv :2102.11585*, 2021.

[92] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *European conference on computer vision*. Springer, 2008, pp. 44–57.

[93] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The apolloscape open dataset for autonomous driving and its application," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2702–2719, 2019.

[94] Z. Che, G. Li, T. Li, B. Jiang, X. Shi, X. Zhang, Y. Lu, G. Wu, Y. Liu, and J. Ye, "D^2-city : A large-scale dashcam video dataset of diverse traffic scenarios," *arXiv preprint arXiv :1904.01975*, 2019.

[95] L. Ding, J. Terwilliger, R. Sherony, B. Reimer, and L. Fridman, "Mit driveseg (manual) dataset for dynamic driving scene segmentation."

[96] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[97] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 794–801.

[98] L. Neumann, M. Karg, S. Zhang, C. Scharfenberger, E. Piegert, S. Mistr, O. Prokofyeva, R. Thiel, A. Vedaldi, A. Zisserman *et al.*, "Nightowls : A pedestrians at night dataset," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 691–705.

[99] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse : 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.

[100] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska *et al.*, "Lyft level 5 av dataset 2019," *url :https ://level-5.global/data/*, 2019.

[101] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving : Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.

[102] S. Malla, B. Dariush, and C. Choi, "Titan : Future forecast using action priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 186–11 196.

[103] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA : An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[104] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset : A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243.

[105] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data : Ground truth from computer games," in *European Conference on Computer Vision (ECCV)*, ser. LNCS, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer International Publishing, 2016, pp. 102–118.

[106] UM and F. C. for Autonomous Vehicles. (2021) umautobots/gtavisionexport. URL: https://github.com/umautobots/GTAVisionExport
usepackage,Online ; Accessed July, 2021.

[107] J. Harb, N. Rébéna, R. Chosidow, G. Roblin, R. Potarusov, and H. Hajri, "Frsign : A large-scale traffic light dataset for autonomous trains," *CoRR*, vol. abs/2002.05665, 2020. [Online]. Available : https://arxiv.org/abs/2002.05665

[108] H.-N. Hu, Q.-Z. Cai, D. Wang, J. Lin, M. Sun, P. Krahenbuhl, T. Darrell, and F. Yu, "Joint monocular 3d vehicle detection and tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5390–5399.

[109] M. Račinský. (2018) Gtavisionexport. URL:https://github.com/racinmat/GTAVisionExport/tree/master.

[110] Jotrius. (2015) Railroad engineer. URL:https://www.gta5-mods.com/scripts/railroad-engineer, Online ; Accessed July, 2021.

[111] R. Khemmar, A. Mauri, C. Dulompont, J. Gajula, V. Vauchey, M. Haddad, and R. Boutteau, "Road and railway smart mobility : a high-definition ground truth hybrid dataset," *Sensors*, vol. 22, no. 10, p. 3922, 2022.

[112] "Supervisely Homepage," https://supervise.ly/lidar-3d-cloud/, [Online ; accessed 06 April 2022].

[113] "Scale Homepage," https://scale.com/, [Online ; accessed 06 April 2022].

[114] W. Zimmer, A. Rangesh, and M. Trivedi, "3d bat : A semi-automatic, web-based 3d annotation toolbox for full-surround, multi-modal data streams," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1816–1821.

[115] A. Mauri, R. Khemmar, B. Decoux, M. Haddad, and R. Boutteau, "Lightweight convolutional neural network for real-time 3d object detection in road and railway environments," *Journal of Real-Time Image Processing*, pp. 1–18, 2022.

[116] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[117] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082.

[118] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," 2016.

[119] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teuliere, and T. Chateau, "Deep manta : A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from mono-

cular image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2040–2049.

[120] B. Xu and Z. Chen, "Multi-level fusion based 3d object detection from monocular images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2345–2353.

[121] Z. Liu, Z. Wu, and R. Tóth, "Smoke : Single-stage monocular 3d object detection via keypoint estimation," 2020.

[122] Z. Qin, J. Wang, and Y. Lu, "Monogrnet : A general framework for monocular 3d object detection," 2021.

[123] G. Brazil and X. Liu, "M3d-rpn : Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.

[124] Y. Liu, Y. Yixuan, and M. Liu, "Ground-aware monocular 3d object detection for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 919–926, 2021.

[125] P. Li, H. Zhao, P. Liu, and F. Cao, "Rtm3d : Real-time monocular 3d detection from object keypoints for autonomous driving," 2020.

[126] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[127] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," 2019.

[128] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[129] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv :1806.01260*, 2018.

[130] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *Advances in Neural Information Processing Systems*. Citeseer, 2015, pp. 424–432.

[131] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4 : Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available : https://arxiv.org/abs/2004.10934

[132] M. Tan and Q. Le, "Efficientnet : Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[133] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[134] H. Li, "A Brief Tutorial On Recursive Estimation With Examples From Intelligent Vehicle Applications (Part II) : System Models," Jul. 2014, working paper or preprint. [Online]. Available : https://hal.archives-ouvertes.fr/hal-01018124

[135] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze : A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *European Conference on Computer Vision*. Springer, 2020, pp. 365–381.

[136] M. Taiana, J. C. Nascimento, and A. Bernardino, "An improved labelling for the inria person data set for pedestrian detection," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2013, pp. 286–295.

[137] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3198–3205.

[138] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007.

[139] J. Marin, D. Vázquez, D. Gerónimo, and A. M. López, "Learning appearance in virtual scenarios for pedestrian detection," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 137–144.

[140] G. Mohamed, A. A. Sofiane, and L. Nicolas, "Adaptive super twisting extended state observer based sliding mode control for diesel engine air path subject to matched and unmatched disturbance," *Mathematics and Computers in Simulation*, vol. 151, pp. 111–130, 2018.

[141] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.

[142] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross ? a benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 206–213.

[143] ——, "It's not all about size : On the role of data properties in pedestrian detection," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[144] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding : A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[145] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[146] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[147] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[148] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[149] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8739–8748.

[150] Y. Bengio, "The consciousness prior," *arXiv preprint arXiv :1709.08568*, 2017.

[151] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "Dr(eye)ve : A dataset for attention-based tasks with applications to autonomous and assisted driving," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2016, pp. 54–60.

[152] A. Palazzi, D. Abati, F. Solera, R. Cucchiara *et al.*, "Predicting the driver's focus of attention : the dr (eye) ve project," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 41, no. 7, pp. 1720–1733, 2018.

[153] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *Asian conference on computer vision*. Springer, 2018, pp. 658–674.

[154] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 773–786, 2018.

[155] T. Yu, H. Gu, L. Wang, S. Xiang, and C. Pan, "Cascaded temporal spatial features for video action recognition," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1552–1556.

[156] W. Yang, J. Feng, G. Xie, J. Liu, Z. Guo, and S. Yan, "Video super-resolution based on spatial-temporal recurrent residual networks," *Computer Vision and Image Understanding*, vol. 168, pp. 79–92, 2018.

[157] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look : Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[158] Z. Yang, Y. Han, and Z. Wang, "Catching the temporal regions-of-interest for video captioning," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 146–153.

[159] Y. Tu, X. Zhang, B. Liu, and C. Yan, "Video description with spatial-temporal attention," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1014–1022.

[160] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4584–4593.

[161] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 487–12 496.

[162] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE transactions on circuits and systems for video technology*, vol. 27, no. 12, pp. 2527–2542, 2016.

[163] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2174–2182.

[164] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video : A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[165] I. Dua, A. U. Nambi, C. Jawahar, and V. Padmanabhan, "Autorate : How attentive is the driver ?" in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[166] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "Traphic : Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8483–8492.

[167] C. Choi, A. Patil, and S. Malla, "Drogon : A causal reasoning framework for future trajectory forecast," *arXiv e-prints*, pp. arXiv–1908, 2019.

[168] K. Nakamura, S. Yeung, A. Alahi, and L. Fei-Fei, "Jointly learning energy expenditures and activities using egocentric multimodal signals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1868–1877.

[169] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding : A dataset for learning driver behavior and causal reasoning," in *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7699–7707.

[170] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping : A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[171] N. Guéguen, S. Meineri, and C. Eyssartier, "A pedestrian's stare and drivers' stopping behavior : A field experiment at the pedestrian crossing," *Safety science*, vol. 75, pp. 87–89, 2015.

[172] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[173] C. Wei, R. Romano, N. Merat, Y. Wang, C. Hu, H. Taghavifar, F. Hajiseyedjavadi, and E. R. Boer, "Risk-based autonomous vehicle motion control with considering human driver's behaviour," *Transportation Research Part C : Emerging Technologies*, vol. 107, pp. 1–14, 2019.

[174] J. Xie and M. Zhu, "Maneuver-based driving behavior classification based on random forest," *IEEE Sensors Letters*, vol. 3, no. 11, pp. 1–4, 2019.

[175] C.-H. Hu, Y. Zhang, F. Wu, X.-B. Lu, P. Liu, and X.-Y. Jing, "Toward driver face recognition in the intelligent traffic monitoring systems," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[176] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2097–2103.

[177] G. Chevalier, "Larnn : linear attention recurrent neural network," *arXiv preprint arXiv :1808.05578*, 2018.

[178] T. Cooijmans, N. Ballas, C. Laurent, Ç. Gülçehre, and A. Courville, "Recurrent batch normalization," *arXiv preprint arXiv :1603.09025*, 2016.

[179] S. Ioffe and C. Szegedy, "Batch normalization : Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[180] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv :1511.07289*, 2015.

[181] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[182] O. Friard and M. Gamba, "Boris : a free, versatile open-source event-logging software for video/audio coding and live observations," *Methods in Ecology and Evolution*, vol. 7, no. 11, pp. 1325–1330, 2016.

[183] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[184] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections : Universal encoding strategies ?" *IEEE transactions on information theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[185] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine learning research*, vol. 8, no. Jul, pp. 1519–1555, 2007.

[186] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[187] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns : Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[188] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognition*, vol. 43, no. 1, pp. 331–341, 2010.

[189] M. J. Jones, P. Viola *et al.*, "Robust real-time object detection," in *Workshop on statistical and computational theories of vision*, vol. 266, 2001, p. 56.

[190] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[191] A. Iraqui, Y. Dupuis, R. Boutteau, J.-Y. Ertaud, and X. Savatier, "Fusion of omnidirectional and ptz cameras for face detection and tracking," in *2010 International Conference on Emerging Security Technologies*. IEEE, 2010, pp. 18–23.

[192] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[193] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. Ieee, 2004, pp. I–I.

[194] A. C. Murillo, J. J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*. IEEE, 2007, pp. 3901–3907.

[195] F. Fraundorfer and D. Scaramuzza, "Visual odometry : Part ii : Matching, robustness, optimization, and applications," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 78–90, 2012.

[196] M. A. Fischler and R. C. Bolles, "Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[197] R. E. Datondji, N. Ragot, Y. Nasri, R. Khemmar, and R. Boutteau, "Odométrie visuelle par vision omnidirectionnelle pour la navigation autonome d'une chaise roulante motorisée," in *Journées francophones des jeunes chercheurs en vision par ordinateur*, 2015.

[198] F. S. Bashiri, E. LaRose, P. Peissig, and A. P. Tafti, "Mcindoor20000 : A fully-labeled image dataset to advance indoor objects detection," *Data in brief*, vol. 17, pp. 71–75, 2018.

[199] R. Mur-Artal and J. D. Tardos, "Orb-slam2 : an open-source slam system for monocular," *Stereo and RGB-D Cameras. arXiv preprint arXiv*, vol. 1610, 2016.

[200] M. Labbé and F. Michaud, "Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[201] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available : http://dx.doi.org/10.1109/CVPR.2016.350

[202] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV*, 2008.

[203] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[204] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data : Ground truth from computer games," *Lecture Notes in Computer Science*, p. 102–118, 2016. [Online]. Available : http://dx.doi.org/10.1007/978-3-319-46475-6_7

[205] R. Mohan, "Deep deconvolutional networks for scene parsing," 2014.

[206] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[207] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," 2014. [Online]. Available : http://infoscience.epfl.ch/record/199822

[208] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," 2014.

[209] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," 2015.

[210] O. Ronneberger, P. Fischer, and T. Brox, "U-net : Convolutional networks for biomedical image segmentation," 2015.

[211] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1–1, 05 2016.

[212] 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*, vol. 39, no. 12, p. 2481–2495, Dec 2017. [Online]. Available : http://dx.doi.org/10.1109/TPAMI.2016.2644615

[213] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. [Online]. Available : http://dx.doi.org/10.1109/CVPR.2017.660

[214] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, p. 834–848, Apr 2018. [Online]. Available : http://dx.doi.org/10.1109/TPAMI.2017.2699184

[215] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Lecture Notes in Computer Science*, p. 833851, 2018. [Online]. Available : http://dx.doi.org/10.1007/978-3-030-01234-2_49

[216] M. Yang, Y. Kun, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," 07 2018.

[217] X. Li, H. Zhao, L. Han, Y. Tong, and K. Yang, "Gff : Gated fully fusion for semantic segmentation," 2020.

[218] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2019.

[219] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "Carla : An open urban driving simulator," *ArXiv*, vol. abs/1711.03938, 2017.

[220] R. Sibechi, O. Booij, N. Baka, and P. Bloem, "Exploiting temporality for semi-supervised video segmentation," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 933–941, 2019.

[221] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "Resnest : Split-attention networks," 2020.

[222] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available : http://dx.doi.org/10.1109/CVPR.2016.90

[223] S. Jadon, "A survey of loss functions for semantic segmentation," 2020.

[224] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 07 2018.

[225] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet : A deep neural network architecture for real-time semantic segmentation," *ArXiv*, vol. abs/1606.02147, 2016.

[226] S.-P. Lee, S.-C. Chen, and W.-H. Peng, "Gsvnet : Guided spatially-varying convolution for fast semantic segmentation on video," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[227] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz, "Std2p : Rgbd semantic segmentation using spatio-temporal data-driven pooling," 2016. [Online]. Available : https://arxiv.org/abs/1604.02388

[228] A. Kundu, V. Vineet, and V. Koltun, "Feature space optimization for semantic video segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3168–3175.

[229] S. Tripathi, S. Belongie, Y. Hwang, and T. Nguyen, "Semantic video segmentation : Exploring inference efficiency," in *2015 International SoC Design Conference (ISOCC)*, 2015, pp. 157–158.

[230] R. Gadde, V. Jampani, and P. V. Gehler, "Semantic video cnns through representation warping," 2017. [Online]. Available : https://arxiv.org/abs/1708.03088

[231] X. Jin, X. Li, H. Xiao, X. Shen, Z. Lin, J. Yang, Y. Chen, J. Dong, L. Liu, Z. Jie, J. Feng, and S. Yan, "Video scene parsing with predictive feature learning," 2016. [Online]. Available : https://arxiv.org/abs/1612.00119

[232] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation." arXiv, 2016. [Online]. Available : https://arxiv.org/abs/1409.1556

[233] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," 2016. [Online]. Available : https://arxiv.org/abs/1611.07715

[234] E. Shelhamer, K. Rakelly, J. Hoffman, and T. Darrell, "Clockwork convnets for video semantic segmentation," 2016. [Online]. Available : https://arxiv.org/abs/1608.03609

[235] Y. Li, J. Shi, and D. Lin, "Low-latency video semantic segmentation," 2018. [Online]. Available : https://arxiv.org/abs/1804.00389

[236] S. Jain, X. Wang, and J. Gonzalez, "Accel : A corrective fusion network for efficient semantic segmentation on video," 2018. [Online]. Available : https://arxiv.org/abs/1807.06667

[237] J. Carreira, V. Patraucean, L. Mazare, A. Zisserman, and S. Osindero, "Massively parallel video networks," 2018. [Online]. Available : https://arxiv.org/abs/1806.03863

[238] S.-P. Lee, S.-C. Chen, and W.-H. Peng, "Gsvnet : Guided spatially-varying convolution for fast semantic segmentation on video," 2021.

[239] H. Wang, X. Jiang, H. Ren, Y. Hu, and S. Bai, "Swiftnet : Real-time video object segmentation," 2021. [Online]. Available : https://arxiv.org/abs/2102.04604

[240] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet : Bilateral segmentation network for real-time semantic segmentation," 2018. [Online]. Available : https://arxiv.org/abs/1808.00897

[241] Y. Sakai, Y. Nakayama, H. Lu, Y. Li, and H. Kim, "Recognition of surrounding environment for electric wheelchair based on wideseg," in *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, 2019, pp. 816–820.

[242] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available : https://arxiv.org/abs/1511.07122

[243] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," pp. 834–848, 2018.

[244] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[245] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," Jun 2016. [Online]. Available : http://dx.doi.org/10.1109/CVPR.2016.350

[246] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV (1)*, 2008, pp. 44–57.

[247] M. Zhao, A. Mammeri, and A. Boukerche, "Distance measurement system for smart vehicles," in *2015 7th International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2015, pp. 1–5.

[248] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf : Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[249] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[250] T. Kurczveil, P. Á. López, and E. Schnieder, "Implementation of an energy model and a charging infrastructure in sumo," in *Simulation of Urban MObility User Conference*. Springer, 2013, pp. 33–43.

[251] S. E. Letendre and W. Kempton, "The v2g concept : A new model for power ?" *Public Utilities Fortnightly*, vol. 140, no. 4, pp. 16–27, 2001.

[252] U. C. Okonkwo, C. E. Chukwunyelu, B. U. Oweziem, A. Ekuase *et al.*, "Evaluation and optimization of tensile strength responses of coir fibres reinforced polyester matrix composites (cfrp) using taguchi robust design," *Journal of Minerals and Materials Characterization and Engineering*, vol. 3, no. 04, p. 225, 2015.

[253] L. Gliga, H. Chafouk, D. Popescu, and C. Lupu, "A method to estimate the process noise covariance for a certain class of nonlinear systems," *Mechanical Systems and Signal Processing*, vol. 131, pp. 381–393, 2019.

[254] K. S. Leong, M. L. Ng, and P. H. Cole, "Operational considerations in simulation and deployment of rfid systems," in *2006 17th International Zurich Symposium on Electromagnetic Compatibility*. IEEE, 2006, pp. 521–524.

[255] D. Mullen, "The application of rfid technology in a port," *Port Technology International*, vol. 1, no. 1, pp. 181–182, 2005.

[256] P. V. Nikitin and K. S. Rao, "Theory and measurement of backscattering from rfid tags," *IEEE Antennas and Propagation Magazine*, vol. 48, no. 6, pp. 212–218, 2006.

[257] T. Raiko, H. Valpola, and Y. LeCun, "Deep learning made easier by linear transformations in perceptrons," in *Artificial intelligence and statistics*. PMLR, 2012, pp. 924–932.

[258] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving beyond feature design : Deep architectures and automatic feature learning in music informatics." in *ISMIR*. Citeseer, 2012, pp. 403–408.

[259] Y. LeCun, "L'apprentissage profond, une révolution en intelligence artificielle," *La lettre du Collège de France*, no. 41, p. 13, 2016.

[260] M. Thoma, "A survey of semantic segmentation," *arXiv preprint arXiv :1602.06541*, 2016.

[261] Y.-S. Xu, T.-J. Fu, H.-K. Yang, and C.-Y. Lee, "Dynamic video segmentation network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6556–6565.

[262] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[263] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[264] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, "Poi : Multiple object tracking with high performance detection and appearance feature," in *European Conference on Computer Vision*. Springer, 2016, pp. 36–42.

[265] S. Murray, "Real-time multiple object tracking-a study on the importance of speed," *arXiv preprint arXiv :1709.03572*, 2017.

[266] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks : A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

Résumé. Les véhicules autonomes sont de plus en plus présents dans notre quotidien, ouvrant de nouvelles perspectives pour la smart mobilité. Un véhicule autonome doit comprendre 3 fonctions essentielles : perception, décision et actions. Plus le système est capable de percevoir son environnement, plus il prendra de meilleures décisions lui permettant, in fine, de déclencher des actions répondant aux exigences de sécurité, confort et d'énergie. La détection, localisation et tracking d'objets sont des tâches indispensables pour la perception. Depuis 2012, le deep learning est devenu un outil très puissant en raison de sa capacité à traiter de grandes quantités de données. L'apparition de nombreuses méthodes basées sur l'apprentissage profond a conduit à des progrès significatifs. Malgré cet engouement à l'IA, peu de méthodes se concentrent sur l'aspect temps-réel, essentiel pour les applications réelles et ce, en raison des coûts de calcul élevés. En plus, ces algorithmes présentent des lacunes évidentes dans les scènes complexes en partie à cause du manque de données vérité terrain comme pour la smart mobilité ferroviaire et la santé. En plus de la précision et la vitesse, les algorithmes de perception doivent prendre en compte la contrainte d'énergie liée aux systèmes embarqués. Mes travaux de recherche sont concernés par cette problématique et se concentrent donc sur la perception d'environnement pour deux domaines de la smart mobilité : routier/ferroviaire et robotique mobile/santé. L'objectif est d'atteindre un niveau d'analyse et de compréhension de scènes complexes permettant d'assurer une smart mobilité de très haut niveau de sécurité, de confort et d'énergie optimale. Cela repose sur deux briques essentielles et complémentaires : 1. Système fusion multicapteurs permettant d'enrichir davantage la perception avec des données hétérogènes, 2. Perception d'environnement basée IA permettant l'exploitation des données collectées pour une meilleure prédiction de l'ensemble des situations. Il s'agit donc du développement de plateformes génériques ouvertes pour expérimenter et valider des concepts technologiques et scientifiques du monde académique et industriel.

Mots-clés. Perception d'environnement, détection d'objets, estimation de distance, suivi d'objets, localisation, jeux de données, intelligence artificielle, apprentissage profond, véhicule autonome, robotique mobile, analyse et compréhension de scènes, smart mobilité.

Abstract. Autonomous vehicles are increasingly present in our daily lives, opening new perspectives for smart mobility. An autonomous vehicle must include three essential functions : perception, decision, and actions. The more the system can perceive its environment, the better decisions it will make, allowing it to trigger high-quality actions that meet safety, comfort, and energy requirements. Object detection, localization, and tracking are essential tasks for environment perception. Since 2012, deep learning has become a very powerful tool due to its ability to process large amounts of data. The emergence of methods based on deep learning has led to significant progress. Despite this AI craze, few methods focus on the real-time aspect that is essential for real applications due to the high computational costs. In addition, these algorithms have obvious shortcomings in complex scenes due in part to the lack of ground-truth data, as is the case for railway and healthcare smart mobility. In addition to accuracy and speed, perception algorithms must take into account the energy constraint related to embedded systems. My research work is concerned with this issue and therefore focuses on the environment perception for two smart mobility fields : road/railway and mobile robotics/healthcare. The objective is to reach a level of analysis and understanding of complex scenes allowing to ensure smart mobility of a very high level of safety, comfort, and optimal energy. This requires two essential and complementary axis : 1. a multi-sensor fusion system to further enrich perception with heterogeneous data, 2. AI-based environment perception to exploit the data collected for better prediction of all situations. It is therefore the development of open generic platforms to experiment and validate technological and scientific concepts for the academic and industrial world.

Key-words. Environment perception, object detection, distance estimation, tracking, localization, multimodal dataset, artificial intelligence, deep learning, autonomous vehicles, mobile robotics, scenes analysis and understanding, smart mobility.