Exploration du niveau de supervision et du contexte temporel pour la détection des personnes en infrarouge

par

Thomas DUBAIL

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE EN GÉNIE DE LA PRODUCTION AUTOMATISÉE
M. Sc. A.

MONTRÉAL, LE 27 NOVEMBRE 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. Éric Granger, directeur de mémoire
Département de génie des systèmes à l'École de technologie supérieure

M. Marco Pedersoli, codirecteur
Département de génie des systèmes à l'École de technologie supérieure

M. Christian Desrosiers, président du jury
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. Rafael Menelau Oliveira e Cruz, examinateur externe
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 27 OCTOBRE 2023

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

AVANT-PROPOS

Cher lecteur,

C'est avec enthousiasme et un immense honneur que je vous présenter ce mémoire, fruit d'efforts intenses et de dévouement envers la recherche et l'innovation. En parcourant ces pages, vous vous aventurerez dans l'univers passionnant de problématiques réelles, où la rencontre entre la théorie et l'industrie prend forme. Au cœur de ce mémoire, réside une aspiration profonde : celle de transcender les frontières traditionnelles de la recherche académique pour bâtir un pont solide entre la théorie et la pratique. À l'heure où les défis qui façonnent notre monde ne cessent de croître en complexité, il devient crucial de forger des liens plus étroits entre le savoir et son application concrète. C'est dans cette optique que mon travail s'est orienté vers la collaboration étroite avec l'entreprises partenaire Distech Controls Inc., pour les guider dans leurs décisions stratégiques. Mon objectif est clair : contribuer à leur succès en offrant des solutions innovantes et éclairées pour leurs défis présents et futurs. En adoptant cette approche, mes expérimentations et réflexions se transforment en des contributions à la fois académiques et précieuses pour l'industrie.

REMERCIEMENTS

Je tiens à exprimer ma sincère gratitude à mon école (ENSAM) pour m'avoir permis de réaliser un double diplôme à l'international, dans un domaine nouveau, ainsi qu'à l'ETS de m'avoir accueilli. Cette opportunité m'a permis d'élargir mes horizons et d'acquérir des connaissances précieuses qui auront un impact significatif sur ma carrière future. Je souhaite également adresser mes remerciements les plus chaleureux à mes parents pour leur soutien moral et financier inestimable tout au long de ce parcours académique. Leur encouragement constant a été la force motrice qui m'a permis de surmonter les défis et de persévérer dans mes efforts. Enfin, je tiens à exprimer ma reconnaissance envers l'ensemble de notre groupe de recherche pour m'avoir introduit dans ce monde passionnant. Particulièrement, je tiens à remercier Heitor, Masih, Fidel, Marco et Eric pour leurs différentes collaborations, discussions enrichissantes et leur supervision bienveillante. Leurs idées novatrices et leurs connaissances approfondies ont grandement contribué à la réussite de ce mémoire. Je tiens également à exprimer ma gratitude envers François, dont l'énergie débordante, l'enthousiasme communicatif et la bonne humeur contagieuse ont illuminé nos nombreuses réunions tout au long de notre collaboration avec Distech Controls Inc.

Exploration du niveau de supervision et du contexte temporel pour la détection des personnes en infrarouge

Thomas DUBAIL

RÉSUMÉ

La gestion intelligente des bâtiments est un domaine essentiel et prometteur, au regard des avantages économiques et écologiques qu'elle offre. En modulant activement les paramètres d'éclairage, de ventilation et de chauffage, il est possible de réduire la consommation d'énergie du bâtiment sans compromettre le confort des occupants. Dans ce mémoire, nous avons étudié plusieurs solutions pour estimer le nombre de personnes dans une pièce, une information cruciale pour une gestion proactive. Tout d'abord, nous avons étudié la détection des personnes en utilisant des images de basse résolution, évaluant ainsi le niveau de supervision nécessaire et les effets d'une mise en œuvre dans des environnements inconnus. Puis, nous avons étudié la temporalité afin de réaliser la reconnaissance des actions des occupants. Enfin, nous avons exploré l'intégration de contexte temporel afin d'améliorer la détection de personnes à haute résolution. L'ensemble de ces recherches ont été conduites en utilisant des images infrarouges pour préserver l'anonymat des individus, tout en permettant l'application de ces méthodes à grande échelle.

Mots-clés: apprentissage profond, détection, infrarouge

Exploring the level of supervision and temporal context for infrared person detection

Thomas DUBAIL

ABSTRACT

Smart building management is an essential and promising field, considering the economic and ecological advantages it offers. By actively adjusting lighting, ventilation, and heating parameters, it is possible to reduce a building's energy consumption without compromising occupants' comfort. In this thesis, we explored several solutions to estimate the number of people in a room, a crucial piece of information for proactive management. To begin with, we investigated people detection using low-resolution images, evaluating the necessary level of supervision and the effects of implementation in unknown environments. Next, we studied the optimal temporality for recognizing occupants' actions. Lastly, we delved into integrating temporal context to enhance high-resolution people detection. All of these research efforts were conducted using infrared images to preserve individuals' anonymity while enabling the widespread application of these methods.

Keywords: deep learning, detection, infrared

TABLE DES MATIÈRES

Page

INTRODUCTION 1

CHAPITRE 1 DESCRIPTION DE L'ENVIRONNEMENT 3
1.1 Développement d'un système embarqué 3
1.2 Protocole expérimental 4
1.3 Mesure de la qualité des résultats 5

CHAPITRE 2 REVUE DE LITTÉRATURE 9
2.1 Apprentissage non supervisé 12
2.1.1 Auto-encodeurs 12
2.1.2 Distillation 15
2.2 Apprentissage sous-supervisé 19
2.2.1 Cartes d'activations de classes 19
2.2.2 Estimation de densité 21
2.3 Apprentissage supervisé 22
2.3.1 Détecteur à deux étages 23
2.3.2 Détecteur à un étage 25
2.4 L'utilisation des images infrarouges basse résolution 26
2.5 Reconnaissance d'actions 29
2.5.1 Caractéristiques réalisées à la main 30
2.5.2 Extraction par apprentissage profond 30
2.5.3 Mécanisme d'attention et apprentissage auto-supervisé 31
2.5.4 Subdivision des actions 32
2.6 Utilisation de contexte temporel pour la détection 33
2.6.1 Cas d'une vue statique 33
2.6.2 Cas d'une vue non statique 35
2.6.3 Augmentation de données temporelles 38

CHAPITRE 3 ÉTUDE DU NIVEAU DE SUPERVISION POUR LA DÉTECTION DE PERSONNES EN INFRAROUGE ET BASSE RÉSOLUTION. 43
3.1 Introduction 43
3.2 Description du dispositif et de la base de données 45
3.2.1 FIR-Image-Action 47
3.2.2 Distech-Low-IR 49
3.3 Résultats 51
3.4 Conclusion 56

CHAPITRE 4 ÉTUDE DE LA TEMPORALITÉ 61
4.1 Reconnaissance d'actions 61

4.1.1 Introduction à la reconnaissance d'actions 61
4.1.2 Méthode proposée et adaptation des méthodes aux images de basse résolution 62
4.1.3 FIR-Image-Action 65
4.1.4 Résultats 68
4.1.5 Conclusion 71
4.2 Mise en valeur du contexte temporel 72
4.2.1 Méthode proposée 74
4.2.2 Distech-High-IR 76
4.2.3 FLIR ADAS Dataset 77
4.2.4 Résultats 77
4.2.5 Conclusion 84

CONCLUSION ET RECOMMANDATIONS 87

BIBLIOGRAPHIE 89

LISTE DES TABLEAUX

Page

Tableau 3.1 Performance des méthodes de détection sur le jeu de données FIR-Image-Action. Toutes les métriques sont calculées avec un IoU de 50% sur 3 répétitions 52

Tableau 3.2 Performance des méthodes de détection sur le jeu de données Distech-Low-IR. Toutes les métriques sont calculées avec un IoU de 50% sur 3 répétitions 53

Tableau 3.3 Résultats de la méthode GradCAM sur la validation avec un domaine de coté 55

Tableau 3.4 Comparaison des résultats de la méthode GradCAM en et hors domaine 55

Tableau 4.1 Performance des méthodes de classification d'actions sur le jeu de données FIR-Image-Action sur 3 répétitions 69

Tableau 4.2 Performance de la méthode TYOLO en fonction de l'écart entre images. Le modèle YOLOv5n a été utilisé sur 3 répétitions. 78

Tableau 4.3 Performance des augmentations de données sur différents détecteurs 79

Tableau 4.4 Dilatation des convolutions dans le modèle Yolov5 nano sur la base de donnée Distech-High-IR 84

LISTE DES FIGURES

Page

Figure 2.1 Exemple de boites englobantes (BBOX), il est notable que plusieurs classes sont présentes et que les annotations peuvent être subjectives, comme par exemple pour la localisation des bâtiments 11

Figure 2.2 Architecture d'un auto-encodeur 12

Figure 2.3 Les différents AE. Tirée de Baur, Wiestler, Albarqouni & Navab (2019) 13

Figure 2.4 AnoVAEGAN : modèle tiré de Baur *et al.* (2019) 14

Figure 2.5 Processus d'apprentissage de la métrique par l'enseignant. p^+ est un patch positif et p^- est un patch négatif pour p. Tirée de Bergmann, Fauser, Sattlegger & Steger (2020) 18

Figure 2.6 Exemple de décalage entre la métrique de l'enseignant et la distribution des élèves. Tirée de Bergmann *et al.* (2020) 18

Figure 2.7 Carte d'activation de classe. Tirée de Zhou, Khosla, Lapedriza, Oliva & Torralba (2016) 20

Figure 2.8 Fonctionnement de RCNN tiré de Girshick, Donahue, Darrell & Malik (2014) 23

Figure 2.9 Fonctionnement de Faster-RCNN tiré de Ren, He, Girshick & Sun (2016) 24

Figure 2.10 Réseau de neurones convolutifs en 2d tiré de Tateno, Zhu & Meng (2019) 28

Figure 2.11 Architecture utilisant des 3D-convolutions tiré de Tao *et al.* (2019) et reprises par Tateno, Meng, Qian & Hachiya (2020a) 28

Figure 2.12 Méthode tirée de Ullah, Muhammad, Haq & Baik (2019) 32

Figure 2.13 En a) l'image en nuances de gris, b) l'estimation des pixels n'appartenant pas au fond de l'image. Tirée de Cao, Sun, Odoom, Luan & Song (2016) 34

Figure 2.14 En a) l'image en couleur, b) l'extraction des contours de l'image a), c) l'estimation des contours de fond, d) le résultat de b-c, e est la

détection finale sur l'image RGB. Tirée de Yu, Chen, Sun & Xie (2008) 35

Figure 2.15 Architecture TYolo tirée de Duran-Vega, Gonzalez-Mendoza, Chang & Suarez-Ramirez (2021) 36

Figure 2.16 Utilisation du contexte tirée de Corsel, van Lier, Kampmeijer, Boehrer & Bakker (2023) 37

Figure 2.17 Mosaïque temporelle tirée de Duran-Vega *et al.* (2021) 38

Figure 2.18 Mixage temporel tiré de Duran-Vega *et al.* (2021) 39

Figure 2.19 Flou aléatoire tiré de Duran-Vega *et al.* (2021) 39

Figure 2.20 Suppression de région tirée de Duran-Vega *et al.* (2021) 40

Figure 2.21 Bruit aléatoire tiré de Duran-Vega *et al.* (2021) 40

Figure 3.1 Installation du dispositif au plafond 45

Figure 3.2 Défaut d'alignement entre l'image RGB et l'image IR. Pour la visualisation seuls les pixels supérieur à 20°C sont montré 46

Figure 3.3 Fonctions d'alignement entre les deux caméras : IR et RGB 47

Figure 3.4 En a) l'image RGB, b) la zone commune dans l'image IR correspondante a a), c) la superposition des zones communes 47

Figure 3.5 Voici des exemples d'images infrarouges avec annotations pour les ensembles de données FIR-Image-Action (a) et Distech-Low-IR (b)-(d) 48

Figure 3.6 Exemples de résultats de détection de personnes en IR à basse résolution. Superposition des modalités RGB et IR avec leurs annotations correspondantes (a), et les prédictions des boîtes englobantes de *dVAE* (b), *gradCAM* (c) et *Yolo v5* (d) 57

Figure 3.7 Sensibilité de la valeur de seuil pour un déploiement hors domaine sur la base de donnée Distech-Low-IR 58

Figure 3.8 Représentation des images de Distech-Low-IR dans l'espace latent d'un t-SNE en fonction de leur provenance 59

Figure 4.1 Bloc avec connexions résiduelles tiré de He, Zhang, Ren & Sun (2016). Ce bloc étant utilisé dans le ResNet avec des convolutions 2D, nous l'utilisons avec des convolutions 3D 63

Figure 4.2 Architecture 3D-CNN-Résiduel inspirée de Song *et al.* (2018) que nous avons adaptée pour l'utilisation d'images infrarouge et basse résolution 63

Figure 4.3 Architecture proposée par Tao *et al.* (2019) et Tateno *et al.* (2020a). Nous l'avons utilisé comme encodeur dans notre auto-encodeur 64

Figure 4.4 Architecture du décodeur, symétrique de l'encodeur, que nous avons réalisé pour adaptée la méthode proposée par Ullah *et al.* (2019) à nos images 64

Figure 4.5 Exemple d'une personne tombée (a) et d'une personne allongée en (b). Sans information temporelle, il nous est très difficile de différencier ces deux cas 66

Figure 4.6 Répartition des actions dans la base de données FIR-Image-Action 67

Figure 4.7 *N* nombre d'images d'entrée de modèle. P le pas de temps entre les images 67

Figure 4.8 Évolution du F1w sur l'ensemble de test en fonction de *N* la taille de séquence d'entrée. *P* est fixé à l'optimal pour l'ensemble de validation soit 1 70

Figure 4.9 Évolution du F1w sur l'ensemble de test en fonction de *P* le pas de temps. *N* est fixé à l'optimal pour l'ensemble de validation soit 17 71

Figure 4.10 Relation de contexte géométrique 72

Figure 4.11 Proposition d'augmentation de données pour le pré-entraînement 75

Figure 4.12 Proposition d'augmentation de données pour l'entraînement 75

Figure 4.13 Exemple d'images annotées de la base de données Distech-High-IR. RGB à gauche et IR à droite 76

Figure 4.14 Exemple d'images annotées de la base de données FLIR ADAS Dataset tirées de Group *et al.* (2018) 77

Figure 4.15 Effets de l'hyperparamètre *s* sur les deux bases de données 81

Figure 4.16 Effets de la profondeur du modèle 82

Figure 4.17 Exemple de détection sur l'ensemble de test, en (a) l'image IR en (b) en utilisant le contexte temporel. Le point de fonctionnement a été placé au maximum de score F1 sur l'ensemble de validation dans les deux cas 85

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

IA	intelligence artificielle
DL	apprentissage profond (deep learning)
ML	apprentissage automatique (Machine learning)
COCO	grande base de données de détection (Common object in context)
BBOX	boites englobantes
AE	auto-encodeur
PCA	analyse par composante principale
NN	réseau de neurones
CNN	réseau de neurones convolutif
MLP	réseau de neurones dense à multi-couches
IR	infrarouge
RGB	couleur (rouge, vert, bleu)
CAM	cartes d'activations de classe
GAN	réseaux antagonistes génératifs
LSTM	réseau de mémoire à long terme

LISTE DES SYMBOLES ET UNITÉS DE MESURE

x	Vecteur d'entrée
y	Vecteur de sortie
g	Vecteur d'annotation
W ou θ	Poids d'un modèle
$\mathcal{L}$	Fonction de coût

INTRODUCTION

L'intelligence artificielle (IA) a connu une évolution remarquable au cours des dernières décennies et a profondément transformé notre société dans de nombreux domaines. De la reconnaissance vocale aux voitures autonomes en passant par les recommandations personnalisées et la détection de fraude, l'IA est devenue omniprésente et ses applications sont de plus en plus variées. Plus récemment les évolutions dans le traitement du langage de ChatGPT ainsi que la génération d'images par Stable-diffusion ont touché le grand public, ce qui a suscité beaucoup d'intérêt pour le domaine. Dans ce mémoire, nous nous sommes intéressés à l'utilisation de l'apprentissage profond pour détecter des personnes en infrarouge. Les recherches ont été conduites en collaboration avec Distech Controls, pour développer une solution industrielle. Le but est de réaliser un nouveau capteur intelligent capable de compter le nombre de personnes présentes dans une pièce. Avec ce nouveau savoir, il sera donc possible de moduler dynamiquement les conditions d'une pièce : la lumière, la ventilation, la température, etc... Les applications sont nombreuses, on peut imaginer pour l'avenir à réaliser de la détection d'action qui peut prévenir d'une chute en maison de retraite ou encore le début d'un mouvement de foule. Des applications sont aussi possibles dans le domaine de la bureautique, par exemple dans le cadre d'un espace ouvert pour optimiser un taux d'occupation des bureaux. La recherche ainsi motivée, nous avons aussi introduit des contraintes matérielles d'un point de vue puissance de calcul qui sont réduites sur le système embarqué. Ce mémoire étudie plusieurs aspects des images infrarouges pour la détections de personnes. Nous avons mené une étude du niveau de supervision pour la détection afin de réduire le coût des annotations en basse résolution, ainsi qu'une étude sur la détection des actions pour prévenir d'éventuelles chutes. Et pour finir nous avons étudié l'impact du contexte temporel dans le but d'améliorer les performances de la détection.

CHAPITRE 1

DESCRIPTION DE L'ENVIRONNEMENT

Dans le cadre d'une collaboration avec l'entreprise avec Distech Controls Inc., cette recherche est motivée industriellement par les méthodes de vision par ordinateur appliquée à des images en infrarouge, plus particulièrement sur de la détection de personnes pour en estimer le nombre. L'objectif de cette collaboration est la réalisation d'un nouveau capteur intelligent qui remplace les détecteurs de présence binaire par une caméra infrarouge. En utilisant plusieurs méthodes d'apprentissage automatique, nous pouvons acquérir plus d'informations sur la scène perçue par le capteur. Celles-ci sont ensuite valorisées par les autres équipements que propose Distech Controls Inc. par un contrôle dynamique des paramètres du bâtiment. Par exemple la température, la ventilation, la climatisation peuvent être contrôlés dans notre cas. Le pari au long terme est d'augmenter les informations que nous procure ce capteur afin de réduire la consommation globale du bâtiment. Les motivations sont donc économiques et écologiques. Dans le cadre de ce mémoire nous utilisons uniquement un capteur à la fois, cependant des améliorations futures pourront recouvrir une zone de travail complète et ainsi améliorer le gain d'informations par exemple en proposant une estimation des flux de déplacement ainsi qu'une optimisation de la place applicable dans des zones de travail partagées. D'autres améliorations comme la reconnaissance d'actions pour la sécurité des occupants est aussi envisagée pour ce produit. En réalisant plusieurs capteurs de ce type là et en augmentant l'information acquise par le bâtiment nous pouvons réalisé des avancés vers un bâtiment intelligent et plus durable sans altérer la qualité de vie des ces occupants.

1.1 Développement d'un système embarqué

Le capteur imaginé par Distech Controls Inc. prend la forme d'un boîtier placé au milieu de la pièce. Étant fixé au plafond, ce dispositif permet d'avoir une vue d'ensemble de la pièce et ainsi en maximiser son gain d'informations. Son point de vue permet aussi de réduire les occlusions entres personnes qui seraient problématiques sur une vue de coté. Nos objectifs sont de réduire la consommation du bâtiment cependant cela repose sur de nombreux facteurs incontrôlés à ce

stade du projet, par exemple notre capacité à moduler les paramètres du bâtiment. Nous avons donc réduit le problème en considérant uniquement le capteur comme un système isolé que nous devons optimiser. Nos objectifs sont donc de maximiser la qualité des informations produites par ce capteur tout en réduisant sa consommation. Souvent évoluant de manière opposée, ces deux axes peuvent présenter un optimal entre complexité algorithmique et performance. Dans ce mémoire nous avons mesuré ces deux objectifs. Le produit étant un système embarqué, des contraintes matérielles sont aussi présentes, comme des limites de mémoire (RAM ainsi que Flash) ou encore la vitesse d'exécution. Bien que la grande inertie des systèmes commandés nous permet une légère latence dans nos mesures, la continuité des flux d'informations nous entraîne vers une solution en temps réel. Le projet étant au stade du développement, plusieurs de ces contraintes ne sont pas encore fixées ou peuvent encore faire évoluer les composants constituants le produit. Ces travaux de recherche ne se limitent donc pas à ces contraintes. L'extraction de l'information sera basée sur des images infrarouges pour assurer un minimum de confidentialité des individus présents sous le capteur. De plus, l'utilisation de l'infrarouge permet d'utiliser le capteur dans l'obscurité ce qui lui permet par exemple d'allumer les lumières en présence d'occupants. Dans ce mémoire nous nous sommes intéressés à la détection des personnes pour leur comptage ainsi que leur localisation sur l'image. Nous nous sommes aussi intéressés à la reconnaissance des actions. Bien qu'il est possible de traiter ces deux tâches avec un seul modèle nous les avons séparées pour en étudier plus simplement la faisabilité ainsi qu'analyser des méthodes présentes dans la littérature.

1.2 Protocole expérimental

Pour assurer une robustesse de nos résultats nous avons répété au moins 3 fois chaque expérimentation avec des séparations de données différentes pour en estimer la plage d'erreur. De plus, nous avons répété la plupart de nos études sur plusieurs bases de données pour permettre une reproductibilité externe. Bien que nous travaillons avec les données de Distech Controls Inc., leurs données sont privées ce qui exclue toute possibilité de publication. Ainsi nous avons

sélectionné des bases de données présentant des caractéristiques similaires et publiques afin de publier nos résultats.

1.3 Mesure de la qualité des résultats

Dans ce mémoire plusieurs tâches sont étudiées comme la détection et la reconnaissance d'actions. En fonction de la tâche à évaluer et des méthodes employées, plusieurs métriques sont utilisées dans la littérature. Commençons par une classification binaire : $y_i \in \{0, 1\}$ les labels associés aux vecteurs d'entrées x_i. Un modèle f_θ produit une prédiction $f_\theta(x_i) = d_i \in \{0, 1\}$. Ainsi pour chaque exemple (x_i, y_i) la prédiction peut être fausse si $d_i \neq y_i$ ou juste si $d_i = y_i$. Pour analyser plus précisément où sont commisses les erreurs, les quatre cas sont dissociées :

- si $d_i = 1$ et $y_i = 1$: l'exemple est compté comme vrai positif.
- si $d_i = 1$ et $y_i = 0$: l'exemple est compté comme faux positif.
- si $d_i = 0$ et $y_i = 0$: l'exemple est compté comme vrai négatif.
- si $d_i = 0$ et $y_i = 1$: l'exemple est compté comme faux négatif.

En définissant TP (respectivement : FP,TN,FN) le nombre total de vrais positifs (respectivement : faux positifs, vrais négatifs, faux négatifs), il est possible d'évaluer plusieurs métriques. Le taux de réussite : $(TP + TN)/N$ avec $N = TP + FP + TN + FN$ le nombre d'exemples, ainsi que la précision : $TP/(TP+FP)$ et le rappel : $TP/(TP+FN)$, sont accessibles à partir de ces éléments. Une métrique souvent utilisée est le score F1 qui est obtenu en faisant la moyenne harmonique entre précision et rappel. L'ensemble de ces mesures s'applique à des modèles de classification dur (sans score de confiance) mais aussi à des modèles de classification mous qui eux prédisent une probabilité $p_i = f_\theta(x_i)$ qui estime $P(y_i = 1|x_i)$. Dans ce cas on peut se ramener à un détecteur dur en sélectionnant un point de fonctionnement $\tau \in [0, 1]$ ainsi $d_i = H(p_i - \tau)$ où H est la fonction de Heaviside. Afin d'évaluer le modèle sur toute sa plage de fonctionnement, et de résoudre la sélection du paramètre τ, il est possible de définir AP (respectivement AR) la précision (respectivement rappel) moyenne pour $\tau \in [0, 1]$ (voir équation 1.1).

$$AP(y,p) = \int_{\tau=0}^{\tau=1} \text{Précision}(y, H(p-\tau)) \cdot d\tau$$
$$AR(y,p) = \int_{\tau=0}^{\tau=1} \text{Rappel}(y, H(p-\tau)) \cdot d\tau \tag{1.1}$$

Considérons maintenant une classification multi-classes à C classes : nous avons alors $y_i, d_i \in [\![1, C]\!]$ correspondant au numéro de la classe annotée ou prédite par le modèle. Dans ce cas il est uniquement possible de définir pour une classe donnée $c \in [\![1, C]\!]$ les vrais positifs ($d_i = y_i = c$), faux négatifs ($d_i \neq c$ et $y_i = c$) ainsi que les faux positifs ($d_i = c$ et $y_i \neq c$). Le cas où $d_i \neq c$ et $y_i \neq c$ ne concerne pas la classe c. Ainsi pour chaque classe les nombres TP_c,FP_c et FN_c sont accessibles et on définit la précision et le rappel pour chaque classe c de la même manière que précédemment. Afin de mesurer la performance globale d'un classifieur multi-classes, il est possible de calculer le taux d'exactitude : $(\sum TP_c)/N$, la moyenne des précisions, rappels et scores F1 selon les classes. Cependant ces mesures sont sensibles aux déséquilibres entre classes. Afin de remédier à cela, il est possible de faire une moyenne pondérée par le nombre d'exemples dans chaque classe pour avoir une mesure représentative de la réalité. Dans le cas de la reconnaissance d'actions, nous avons utilisé le score F1 pondéré selon les classes noté F1w (voir équation 1.2). La fonction indicatrice est représentée par $\mathbb{1}$.

$$\text{F1w} = \sum_c w_c \cdot \text{F1}_c \qquad \text{où} \quad w_c = \frac{\sum_i \mathbb{1}_c(y_i)}{N} \tag{1.2}$$

Dans le cadre d'un classifieur multi-classes mou, p_i n'est plus une probabilité mais un vecteur de probabilité tel que sa somme vaut $1 = \sum_c p_{i,c}$. Ainsi pour passer de p_i à d_i nous sélectionnons la classes c atteignant la plus grande probabilité $d_i = argmax_c(p_{i,c})$. Dans ce cas une mesure grandement utilisée est alors la moyenne des AP_c sur les classes portant alors le nom mAP.

Passons maintenant à la détection : non seulement chaque boite englobante a une classe, mais aussi une localisation $l_i \in \mathbb{R}^4$ dans l'image x_i. L'évaluation de chaque couple prédiction / annotation nécessite de prendre en compte la localisation de l'un par rapport à l'autre. Pour cela, la mesure la plus utilisée est intersection sur l'union (IoU : voir équation 1.3). Plus l'IoU entre deux boites est grande plus leur localisation est partagée.

$$IoU(a, b) = \frac{a \cap b}{a \cup b} \tag{1.3}$$

Ainsi pour qu'une prédiction soit vraie positive, les conditions sur les classes demeurent inchangées et la mesure de localisation doit est au dessus d'une valeur seuil τ_{loc}. La valeur la plus utilisée pour qu'une localisation soit considérée comme bonne, est IoU > 0.5, soit plus de 50% de recouvrement. Pour les détecteurs durs, nous avons utilisé l'oLRP qui est la borne inférieure de l'LRP (voir équation 1.4) pour $\tau \in [0, 1]$, métrique proposée par Oksuz, Cam, Kalkan & Akbas (2021). Cette nouvelle métrique permet d'évaluer les détections qui sont mal classées ainsi que d'évaluer la localisation des vrais positifs.

$$LRP_\tau(p, y) = \frac{1}{N} \times \left(FP + FN + \sum_{i=1}^{TP} \frac{1 - IoU(y_i, p_i)}{1 - \tau} \right) \tag{1.4}$$

Avec des détecteurs mous il est possible d'évaluer l'mAP avec un τ_{loc} fixé. Cette métrique est nommée $mAP@\tau_{loc}$. Plus τ_{loc} augmente plus l'évaluation de la localisation est stricte et ainsi la valeur $mAP@\tau_{loc}$ décroît. Afin d'avoir une évaluation indépendante du seuil de localisation, un moyen de résoudre ce problème est de faire la moyenne sur un ensemble de seuils. Bien qu'il serait possible de le faire avec $\tau_{loc} \in [0, 1]$, le coût algorithmique est trop élevé en pratique. Ainsi $mAP@0.5 : 0.95$ est souvent utilisé, la moyenne de $mAP@\tau_{loc}$ pour $\tau_{loc} \in \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$.

Dans ce mémoire nous avons utilisé la métrique de l'oLRP pour les comparaisons entre détecteurs comportant des détecteurs durs. En ce qui concerne l'évaluation des détecteurs mous nous

avons utilisé l'$mAP@0.5:0.95$ comme elle est la principale employée pour ces tâches dans la littérature. Pour les tâches de classifications multi-classes nous avons utilisé le score F1w. Enfin, afin d'évaluer la complexité des méthodes utilisées nous avons testé le temps d'inférence.

CHAPITRE 2

REVUE DE LITTÉRATURE

Dans ce mémoire, plusieurs niveaux de supervision pour la détection des personnes en infrarouge est étudié. Nous avons organisé la revue de la littérature en considérant les différents niveaux de supervision.

Dans le cadre d'un apprentissage automatique (ML) une base de données $\mathcal{D}$ est utilisée pour optimiser les paramètres θ d'un modèle $\mathcal{M}_\theta$. Généralement cette base de données est séparée en trois partitions $\mathcal{D}_e$ les données d'entraînement, $\mathcal{D}_v$ les données de validation et $\mathcal{D}_t$ l'ensemble des données de test. L'optimisation des paramètres θ est le plus souvent atteinte par la minimisation d'une fonction de coût $\mathcal{L}$ sur l'ensemble d'entraînement $\mathcal{D}_e$. Des itérations d'entraînement et de validation sur l'ensemble de validation $\mathcal{D}_v$ permettent de réaliser une sélection des hyperparamètres (paramètres d'apprentissage, architecture de modèle, traitement avant/ après modèle,...). Les deux ensembles $\mathcal{D}_e$ et $\mathcal{D}_v$ sont utilisés pour guider notre amélioration sur l'élaboration du modèle. Ainsi le modèle est biaisé sur ces deux ensembles, c'est pourquoi $\mathcal{D}_t$ n'est pas utilisé durant l'apprentissage pour avoir une évaluation réaliste de notre modèle sur des données qu'il n'a jamais vu. L'ensemble de test $\mathcal{D}_t$ évalue la performance du modèle sur la tâche qu'il doit résoudre (exemple : classification, régression, localisation, détection, segmentation,...) cependant les ensembles $\mathcal{D}_e$ et $\mathcal{D}_v$ ne sont pas nécessairement liés directement à la tâche finale. Ainsi les niveaux de supervisions définissent le niveau d'informations disponibles durant l'entraînement du modèle, comparé au niveau d'informations que le modèle va nous délivrer durant le test. Un entraînent faiblement supervisé utilise des annotations plus faibles que la tâche finale, inversement un entraînement sur-supervisé utilise plus d'annotations durant l'entraînement que pendant le test.

Dans le cadre de la détection, les annotations sont des boites englobantes (voir figure 2.1) qui entourent les objets d'intérêts ainsi que la classe correspondante pour chaque objet. De cette façon nous pouvons définir les niveaux de supervision suivant :

1. Non supervisé : Le modèle utilise uniquement les images d'entrée sans aucune annotation pour être optimisé. Il apprend à extraire des caractéristiques et à découvrir des motifs intrinsèques dans les données pour détecter la présence de personnes.
2. Auto-supervisé : L'approche auto-supervisée exploite des méta-données associées aux images pour guider le processus d'apprentissage. Dans le cas de la détection de personnes, les méta-données peuvent inclure des informations telles que les différentes chambres ou zones dans lesquelles les images ont été capturées. Il est aussi possible d'utiliser d'autres informations qui sont recueillies durant la capture des données comme la date et l'heure.
3. Sous-supervisé - binaire : Le modèle est entraîné avec des images accompagnées d'une annotation binaire indiquant la présence ou l'absence de personnes dans chaque image. Il apprend à classifier les images en fonction de la présence ou de l'absence de personnes sans fournir de localisation précise.
4. Sous-supervisé - nombre : Le modèle est entraîné avec des images accompagnées du nombre de personnes présentes dans chaque image. Cela permet au modèle d'apprendre à estimer le nombre de personnes sans fournir d'annotations précises sur leur emplacement.
5. Sous-supervisé - point : Le modèle est entraîné avec des images accompagnées d'une annotation de localisation représentée par un point. Ce point indique approximativement l'emplacement (ou le centre) de la personne dans l'image. Le modèle apprend à estimer la position approximative des personnes en se basant sur ces annotations de localisation ponctuelles.
6. Supervisé : Le modèle est entraîné avec des images accompagnées d'annotations de boîtes englobantes précises qui encadrent chaque personne dans l'image. Le modèle apprend à détecter et à localiser les personnes en se basant sur ces annotations comme exemple.
7. Sur-supervisé : Le modèle est entraîné avec des images accompagnées d'annotations plus détaillées, telles que des mesh 3D qui capturent la structure tridimensionnelle des personnes présentes dans l'image. Cette supervision supplémentaire permet au modèle de comprendre les informations spatiales et de forme plus complexes pour la détection des personnes.

De manière générale, un modèle qui a accès à un niveau de supervision plus élevé sera plus performant pour le même nombre de données. Cependant le coût des annotations augmente aussi avec le niveau de supervision, ce qui n'est pas pris en compte dans la recherche bien que limitant pour les petites entreprises. C'est pourquoi, les méthodes avec des niveaux de supervision basse, dites faiblement annotées, restent une grande motivation de recherche. De plus, pour des grands modèles tels que Stable-diffusion ou ChatGPT, la quantité de données est si grande qu'il serait trop coûteux de les annoter.

Dans notre cas, l'utilisation (avec Distech-Controls) d'un détecteur pour compter le nombre de personnes dans une pièce, peut être vue comme un apprentissage sur-supervisé. La localisation supplémentaire qui est apportée par le modèle, peut ajouter de la facilité à comprendre son fonctionnement, ainsi que de l'explicabilité. Dans notre cas, le projet n'est pas dans sa version finale, ce qui encourage l'utilisation de ces approches pour des versions futures. Des améliorations feront usage de la localisation fournie par les détections, par exemple en proposant une ré-identification des individus.

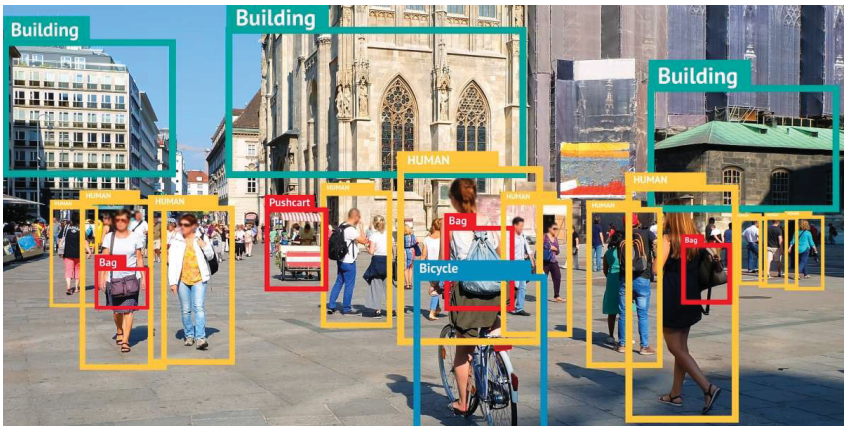


Figure 2.1 Exemple de boites englobantes (BBOX), il est notable que plusieurs classes sont présentes et que les annotations peuvent être subjectives, comme par exemple pour la localisation des bâtiments

2.1 Apprentissage non supervisé

Une façon de voir la détection d'objets avec un apprentissage non supervisé est la détection d'anomalies. Nous avons une distribution d'image $x \in \mathbb{R}^{H \times W \times C}$ (H hauteur, W largeur et C le nombre de canaux d'une image) et l'objectif est de détecter les objets qui sont hors de cette distribution.

2.1.1 Auto-encodeurs

Le principe de l'auto-encodeur (AE) est calqué sur l'analyse par composante principale (PCA). Cependant les transformations ne sont pas linéaires entre vecteur d'entrée et vecteur propre, mais elles sont complexes et automatiquement apprises par un réseau de neurones entre vecteur d'entrée et espace latent. L'auto-encodeur est séparé en deux parties. La première partie est constituée de l'encodeur $\mathcal{E}$ qui permet de passer d'une entrée x à une variable latente $z = \mathcal{E}(x)$. La seconde partie est le décodeur $\mathcal{D}$ qui permet d'inverser l'opération et donc de passer de z à $x' = \mathcal{D}(z)$. En forçant $Dim(z) < Dim(x)$ nous pouvons contraindre le modèle à réduire la dimension du vecteur x vers une représentation plus concise z en optimisant $x' = \mathcal{D}(\mathcal{E}(x))$ a être proche de x. Pour ce faire, nous pouvons utiliser la distance $\mathcal{L}_1 = |x - x'|$ ou encore la distance $\mathcal{L}_2 = ||x - x'||$. Cela explique l'architecture bien spécifique des AE (voir figure 2.2).

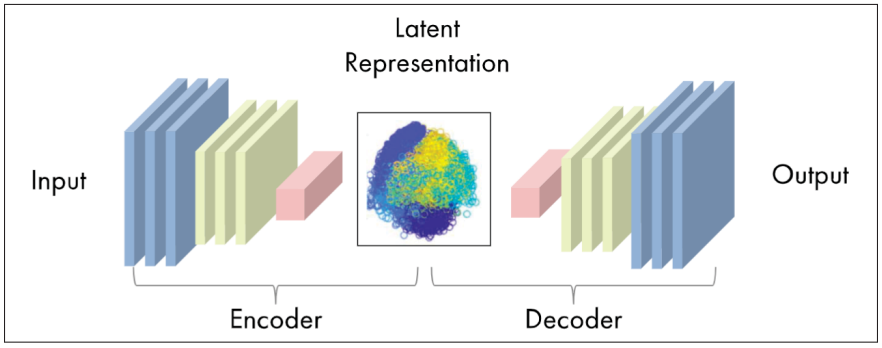


Figure 2.2 Architecture d'un auto-encodeur

Dans le cadre de la détection d'anomalies, on entraîne l'auto-encodeur sur des données ne comportant pas d'anomalie. Le modèle sera donc capable de reproduire ces images. Lorsque le

modèle sera utilisé en mode inférence et qu'une anomalie lui sera présentée il ne sera pas capable de la reproduire car elle ne fait pas part de son ensemble d'entraînement. De cette manière en évaluant la valeur de $|\mathcal{D}(\mathcal{E}(x)) - x|$ pour chaque pixel, il est donc possible d'estimer la localisation d'une anomalie. Cette idée est une des bases fondamentales des méthodes présentées par Baur *et al.* (2019). Il propose la comparaison de 4 AE différents pour la détection des tumeurs au cerveau. Il propose donc 4 dénominations. Le dAE pour auto-encodeur dense : un AE qui a un espace latent plat. Le sAE qui est un AE avec un espace latent respectant la géométrie de l'image d'entrée : pour ce faire, l'espace latent est formé uniquement de couches de convolutions. Une autre variante est le VAE pour auto-encodeur variationnel introduit par Kingma & Welling (2013) : l'espace latent a la particularité d'être probabiliste. On a alors $\mathcal{D}$ qui n'estime donc plus z mais l'écart-type σ et la moyenne μ tel que $z \hookrightarrow \mathcal{N}(\mu, \sigma^2)$. Ainsi il est possible d'échantillonner z en utilisant $\epsilon \hookrightarrow \mathcal{N}(0, 1)$ en employant la reparamétrisation $z = \mu + \epsilon \cdot \sigma$. De plus cette reparamétrisation est nécessaire à la rétropropagation des gradients. En effet, il est nécessaire d'avoir des opérations déterministes. Pour contraindre l'espace latent à respecter la distribution normale, la distribution de z est comparée à celle-ci avec la divergence de Kullback–Leibler $\mathcal{D}_{KL}$ qui est utilisée comme fonction de coût. En combinant les dAE / sAE aux VAE / AE il est donc possible d'avoir les 4 combinaisons : 2.3

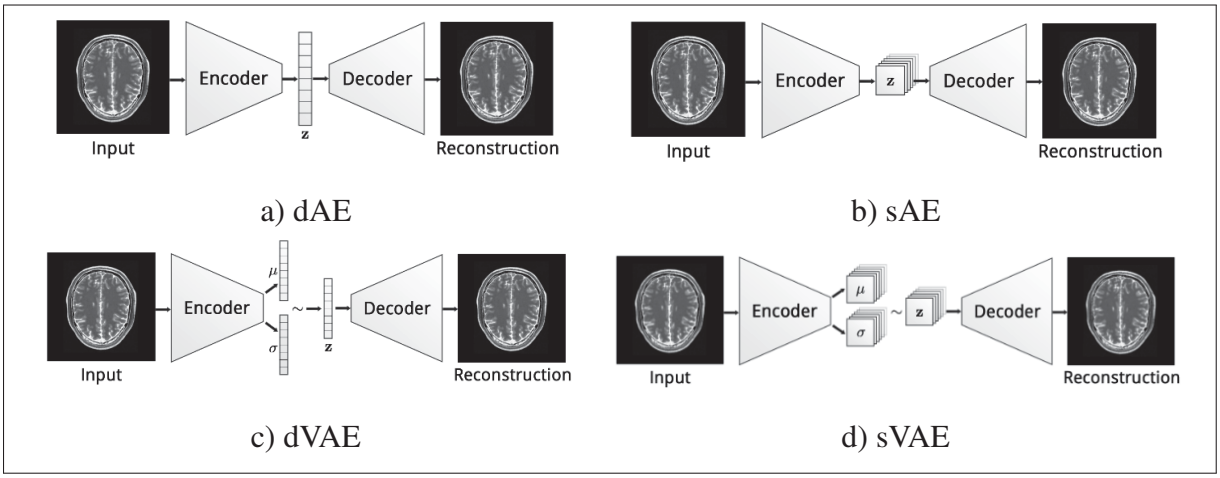


Figure 2.3 Les différents AE. Tirée de Baur *et al.* (2019)

Une amélioration proposée par Baur *et al.* (2019) est d'utiliser une fonction de coût supplémentaire en utilisant un réseau de neurones discriminant qui a pour but de différencier les images réelles des images générées par l'AE. Les deux modèles sont donc en concurrence : ce qui est appelé GAN (Generative adversarial network) dans la littérature. Ils ont été introduits par Goodfellow *et al.* (2014) . La méthode finalement proposée par Baur *et al.* (2019) est un sVAE avec une fonction de coût GAN 2.4.

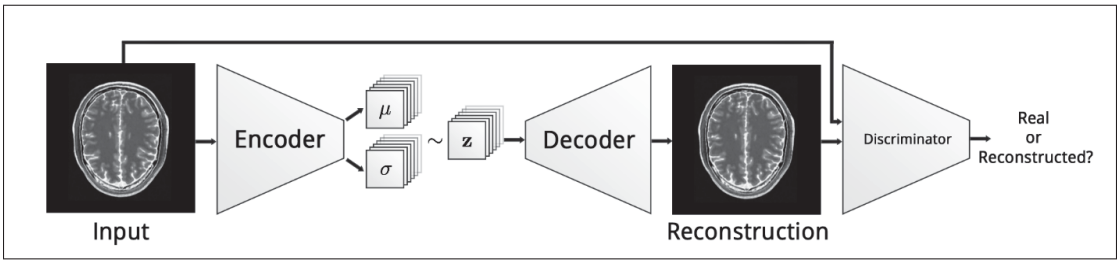


Figure 2.4 AnoVAEGAN : modèle tiré de Baur *et al.* (2019)

La fonction de coût de l'AE est présentée en équation 2.1, elle est composée du coût de reconstruction, complété par la fonction de coût induite par la critique faite par le discriminant sur les images générées.

$$\mathcal{L}_{VAEGAN} = \lambda_1 \times \underbrace{\|x - \mathcal{D}(\mathcal{E}(x))\|}_{\text{erreur de reconstruction}} - \lambda_2 \times \underbrace{log(Dis(\mathcal{D}(\mathcal{E}(x))))}_{\text{critique de la génération}} + \lambda_3 \times \underbrace{\mathcal{D}_{KL}(\mathcal{E}(x)||\mathcal{N}(0,1))}_{\text{contrainte probabiliste}} \tag{2.1}$$

Les hyperparamètres λ sont responsables de l'équilibre entre les différents objectifs. Le discriminant quant à lui, est entraîné pour différencier les images vraies, des images fausses en optimisant l'entropie croisée équation 2.2.

$$\mathcal{L}_{Discriminent} = -log(Dis(x)) - log(1 - Dis(\mathcal{D}(\mathcal{E}(x)))) \tag{2.2}$$

En mode inférence, uniquement le VAE sera utilisé pour détecter les anomalies.

Le problème de cette méthode est directement lié à l'utilisation du GAN. Ces modèles sont difficiles à entraîner, car instables et très sensibles aux hyperparamètres : il peut être difficile de trouver un équilibre entre le générateur et le discriminant. Ceci peut entraîner des problèmes de convergence et des performances peu fiables. De plus, il peut être difficile de déterminer à quel moment arrêter l'entraînement d'un modèle GAN. Contrairement à l'entraînement supervisé classique qui utilise des critères de performance explicites, comme la précision ou l'erreur de reconstruction, il n'existe pas de métrique claire pour évaluer la qualité des échantillons générés par le générateur. L'optimisation est souvent cyclique et incertaine. Pour remédier à ces problèmes, la littérature a introduit des termes supplémentaires dans la fonction de coût du discriminant comme une régularisation $||W||_2$. Cette régularisation pénalise les poids d'amplitude trop grands pour le discriminant, réduisant ainsi sa puissance. Les poids d'amplitude élevés sont connus pour être révélateurs du sur-apprentissage. En les contraignant à rester faibles, nous augmentons la capacité de généralisation du modèle, comme par exemple dans l'optimiseur AdamW proposé par Loshchilov & Hutter (2019). Un autre point négatif de cette méthode est l'architecture de l'AE qui est pratiquement 2 fois plus coûteuse qu'un réseau de neurones classique que ce soit en complexité de calcul ou de mémoire.

Une amélioration sur ce domaine est proposée par Gutoski, Aquino, Ribeiro, Lazzaretti & Lopes (2017) en coupant l'AE en deux, après son entraînement. Par cette astuce, il est possible de diminuer par un facteur 2 les calculs, cependant l'espace de sortie du modèle n'est plus une image mais l'espace latent z. Cette méthode propose donc l'utilisation d'un classifieur à une classe de type machine à vecteurs de support (one-SVM) proposé par Chen, Zhou & Huang (2001). Le one-SVM est déjà une méthode de détection d'anomalies. Néanmoins, l'utiliser directement sur des images ne donne pas de résultats convenables. La grande dimensionnalité de celles-ci nécessiteraient trop d'exemples pour fournir des résultats convenables.

2.1.2 Distillation

L'apprentissage profond a connu des avancées significatives ces dernières années, permettant aux modèles de devenir de plus en plus performants dans de nombreux domaines, tels que la vision par

ordinateur, le traitement du langage naturel ou la reconnaissance vocale. Cependant, ces modèles sont souvent très complexes et nécessitent des ressources considérables en termes de puissance de calcul et de taille de mémoire. La distillation du savoir vise à transférer les connaissances d'un modèle complexe, souvent appelé enseignant, vers un modèle plus simple, appelé élève. Cette approche permet de réduire la taille et les exigences de calculs du modèle enseignant, tout en maintenant ses performances au niveau de l'enseignant, voire en les améliorant dans certains cas. Le modèle élève étant plus simple, les concepts appris par l'enseignant sont souvent mieux généralisés par l'élève. Les motivations derrière l'utilisation de la distillation du savoir sont multiples. Tout d'abord, la réduction de la taille du modèle peut être essentielle dans des environnements où les ressources de calcul et de stockage sont limitées, tels que les appareils mobiles ou les systèmes embarqués comme dans notre cas. En comprimant les connaissances de l'enseignant dans un modèle plus petit, la distillation permet de déployer des modèles puissants sur des plateformes plus légères. En outre, la distillation du savoir peut également améliorer la généralisation du modèle élève en lui permettant d'apprendre des informations subtiles et complexes qui sont présentes dans les prédictions de l'enseignant. Ces informations peuvent être difficiles à apprendre directement à partir des données d'entraînement et peuvent fournir des connaissances supplémentaires qui améliorent les performances du modèle élève sur de nouvelles tâches ou de nouveaux ensembles de données. Enfin, la distillation du savoir peut être utilisée comme une méthode d'optimisation régularisante, aidant à éviter le sur-apprentissage en transférant des informations régulières et lisses de l'enseignant vers l'élève. Cela permet de réduire la sensibilité aux bruits et aux variations des données d'entraînement, renforçant ainsi la capacité du modèle élève à généraliser.

Dans notre cas, la distillation du savoir n'est pas utilisée dans ce but car la méthode proposée par Bergmann *et al.* (2020) utilise la même architecture de modèle et donc n'a pas l'avantage d'être réduite.

Le principe repose sur la distillation du savoir d'un modèle enseignant sur les modèles élèves et uniquement sur les données sans anomalie. Ainsi les modèles vont être en accord sur les données

sans anomalie, et en présence d'anomalies : le changement de distribution des prédictions des modèles sera détectable.

L'entraînement du modèle enseignant $\hat{T}$ est réalisé par patchs d'images p, le but étant de créer une représentation dans un espace latent z de plus petite dimension que l'espace d'entrée $x \in \mathbb{R}^{H \times W \times C}$ (H hauteur, W largeur et C le nombre de canaux d'une image). Il faut noter que cet apprentissage n'est réalisé que sur des données de pré-entraînement. Ainsi les capacités de description sont très puissantes. L'espace latent doit être représentatif des caractéristiques présentes sur l'image et par la même occasion se rapprocher d'une métrique. Pour se faire l'enseignant est entraîné avec des exemples positifs p^+ qui doivent être rapprochés et des exemples négatifs p^- qui doivent être éloignés. L'optimisation de l'enseignant repose donc sur les patchs p, p^+ et p^- avec la fonction de coût 2.3. Le paramètre $\delta > 0$ est un hyperparamètre qui a l'effet d'une marge entre exemples positifs et négatifs.

$$\mathcal{L}_{\hat{T}} = \lambda_1 \times \underbrace{max(0, \delta + \delta^+ + \delta^-)}_{\text{erreur de métrique}} + \lambda_2 \times \underbrace{\sum_{i \neq j} c_{ij}}_{\text{erreur de redondance}} \tag{2.3}$$

$$\delta^+ = ||\hat{T}(p) - \hat{T}(p^+)||^2 \tag{2.4}$$

$$\delta^- = min(||\hat{T}(p) - \hat{T}(p^-)||^2, ||\hat{T}(p) - \hat{T}||^2) \tag{2.5}$$

L'erreur de redondance fait intervenir la corrélation c_{ij} entre les descripteurs i et j sur une optimisation d'un paquet de données. Celle-ci permet de guider le modèle à apprendre des descriptions décorrélées et ainsi plus concises.

La méthode proposée par Bergmann *et al.* (2020) utilise plusieurs élèves S_i qui vont distiller le savoir de $\hat{T}$. L'objectif est d'avoir la moyenne des descriptions des étudiants en accord avec l'enseignant : $1/n \sum^n S_i(x) \approx \hat{T}(x)$. Cet apprentissage est réalisé uniquement sur les images

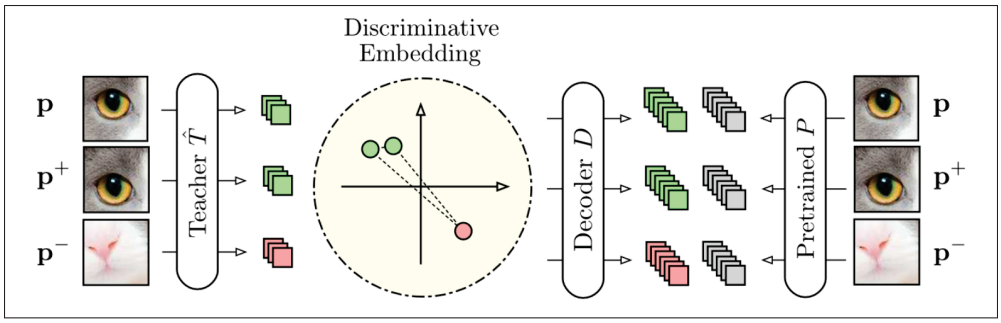


Figure 2.5 Processus d'apprentissage de la métrique par l'enseignant. p^+ est un patch positif et p^- est un patch négatif pour p. Tirée de Bergmann *et al.* (2020)

qui ne comportent pas d'anomalie. La différence de performance élève / enseignant peut être expliquée par l'utilisation d'une autre base de données pour l'enseignant. Pour finir, il est possible de détecter une anomalie dans le patch d'entrée en mesurant un décalage entre la moyenne des élèves et la valeur de l'enseignant figure 2.6.

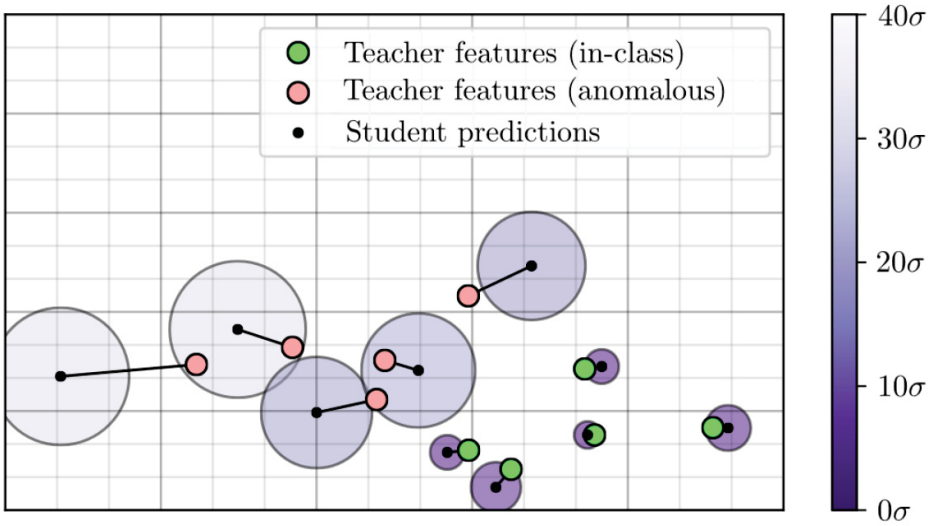


Figure 2.6 Exemple de décalage entre la métrique de l'enseignant et la distribution des élèves. Tirée de Bergmann *et al.* (2020)

Les limitations de cette méthode réside dans l'utilisation de plusieurs réseaux de neurones ; par conséquent le temps d'inférence est très coûteux. Bergmann *et al.* (2020) utilise 3 étudiants pour un total de 4 réseaux de neurones durant le déploiement de la méthode ce qui la rend

inutilisable sur des systèmes avec peu de ressources. Bien que les résultats sont prometteurs, ils dépendent grandement du changement de distribution entre l'ensemble de pré-entraînement et les données d'entraînement pour l'enseignant. Ainsi, lors de la distillation, si ces distributions sont trop proches, les étudiants seront encore en accord avec l'enseignant même en présence d'une anomalie.

2.2 Apprentissage sous-supervisé

Dans le cadre d'un apprentissage faiblement supervisé, les classes binaires informent la présence ou non d'un objet d'intérêt durant l'entraînement. De cette manière l'ensemble d'entraînement $\mathcal{D}_t$ est composé des images x_i ainsi que des annotations c_i, la classe de l'image. L'objectif de l'apprentissage est alors de déduire la corrélation entre la présence des objets d'intérêts dans l'image et les annotations. En mode inférence, le modèle sera donc capable de localiser l'objet sur l'image en se basant sur ces déductions.

2.2.1 Cartes d'activations de classes

Les cartes d'activations de classes (CAM) on été développées dans un premier temps pour expliquer la décision prise par un réseau de neurones, par exemple dans le cas d'une classification. Ces méthodes permettent d'estimer une carte de scores aux dimensions de l'image d'entrée pour chaque classe. Cette carte peut être utilisée pour expliquer et localiser les zones les plus discriminantes de l'image. Soit $\mathcal{M}$ un modèle permettant la classification d'images, il est courant de le séparer en deux parties : les couches de convolution (CNN) permettent d'extraire les caractéristiques pertinentes des images $\mathcal{M}_1$, puis un réseau de neurones dense (MLP) pour prendre la décision finale $\mathcal{M}_2$. Nous avons donc $\mathcal{M} = \mathcal{M}_1 \circ \mathcal{M}_2$. La méthode proposée par Zhou *et al.* (2016) est d'utiliser les activations des neurones entre $\mathcal{M}_1$ et $\mathcal{M}_2$ car ils sont sous la forme d'une image de taille $H/2^n \times W/2^n$ ou n est le nombre d'opérations de mise en commun (pooling ou strides) dans $\mathcal{M}_1$ et H est la hauteur de l'image d'entrée et W sa largeur. Il est donc possible d'utiliser une interpolation pour augmenter la dimension de cette carte afin de revenir aux dimensions de l'image d'origine. Plusieurs cartes sont disponibles en sortie de $\mathcal{M}_1$ pour

chaque filtre de convolution. Zhou *et al.* (2016) propose donc de faire une somme des cartes pondérées par le poids w associé à chaque carte pour la prise de décision finale dans $\mathcal{M}_2$. Il est donc possible de réaliser une carte de localisation pour chacune des classes i avec l'équation 2.6.

$$CAM(x)_i = \frac{1}{n} \sum_{j=0}^{n} w_{ij} \times \mathcal{M}_{1,ij} \tag{2.6}$$

Le modèle est donc en capacité de localiser l'objet avec la carte de localisation générée par la CAM. De plus le modèle produit un score de classification entre présence et l'absence de l'objet dans l'image.

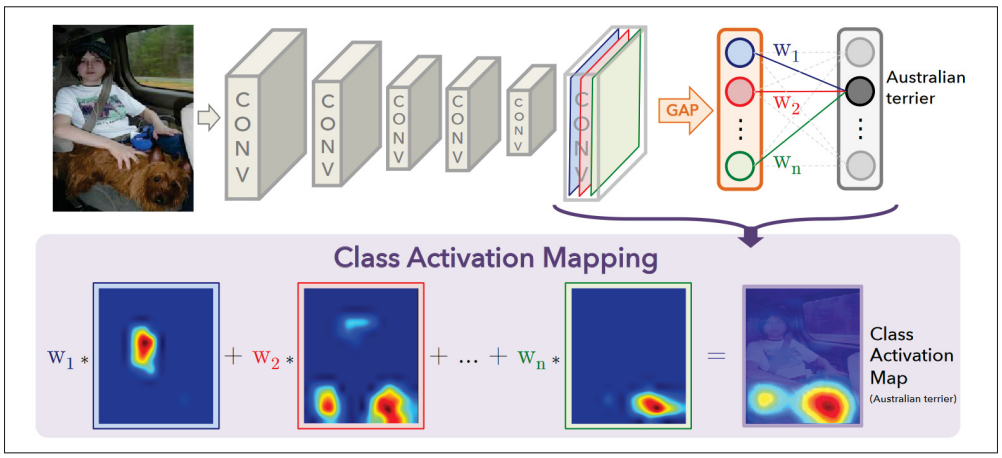


Figure 2.7 Carte d'activation de classe. Tirée de Zhou *et al.* (2016)

Une amélioration de l'original CAM est proposée par Selvaraju *et al.* (2017) : gradCAM. A la place de baser la création des cartes sur les activations, l'auteur propose d'utiliser la rétropropagation du gradient lors d'une passe à contre sens pour estimer une carte de localisation. Cette méthode est plus permissive en terme d'architecture et propose par la même occasion de fusionner des cartes de caractéristiques pertinentes à différents niveaux. Cela permet d'affiner la localisation sur les objets d'intérêts en utilisant des cartes en plus haute résolution au début du modèle tout en gardant le niveau d'abstraction des couches finales du modèle.

De récentes améliorations de ces méthodes proposent de modifier les flux d'informations que ce soit pour la rétropropagation dans le cas de gradCAM++ proposée par Chattopadhay, Sarkar, Howlader & Balasubramanian (2018) ou encore dans les deux sens pour la méthode layerCAM proposée par Jiang, Zhang, Hou, Cheng & Wei (2021). Ces méthodes prennent uniquement en compte les flux d'informations positifs en appliquant l'opération $relu(x) = max(0, x)$ sur le flux d'informations. La scoreCAM proposée par Wang *et al.* (2020) propose des évaluations successives du modèle afin de remplacer la rétropropagation.

Le principal problème de ces méthodes provient de la construction de ces cartes d'activations : elles mettent en valeur les zones les plus discriminantes de l'image. Ce qui a des implications directes : si un objet est très facilement détectable sur une image il va atténuer la localisation de tous les autres objets possiblement présent dans la même image. Cela est un problème majeur pour la détection de plusieurs objets.

2.2.2 Estimation de densité

L'estimation de densité de personnes intéresse toute une branche de la littérature dédiée. Mais bien que le problème diffère du nôtre, des techniques similaires peuvent être employées. L'objectif final étant souvent d'estimer le nombre de personnes présentes dans une image, les méthodes proposées utilisent des annotations sous forme de boites et de points, renseignant de la tête d'un individu. Les images traitées sont souvent plus denses lorsqu'il s'agit de foules qui sont particulièrement difficiles à analyser car présentant un grand nombre d'occlusions. De plus de grands facteurs d'échelles peuvent être présents par des effets de perspective ce qui rend ce domaine bien spécifique.

Une méthode de ces méthodes d'estimation de densité est proposée par Song *et al.* (2021) et se base sur des annotations sous forme de points pour chaque tête. Le modèle est composé d'un extracteur de caractéristiques de l'image VGG16. Ces cartes de caractéristiques sont ensuite traitées pour deux tâches indépendantes : la classification de la région qui renseigne de la présence ou non d'une tête, et la régression pour estimer la localisation de la tête. La fonction de

coût proposée par Song *et al.* (2021) se compose de deux parties pour guider les deux têtes de prédictions : $\mathcal{L} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{reg}$. La partie responsable de la classification $\mathcal{L}_{cls}$ est l'entropie croisée entre les prédictions et la présence d'annotations dans les régions respectives. La partie responsable de la régression $\mathcal{L}_{reg}$ est la distance euclidienne $\mathcal{L}_2$. Celle-ci est utilisée entre la localisation de la tête la plus proche de l'annotation, et l'annotation elle même. Le paramètre λ permet d'ajuster le poids de chaque terme en fonction de la base de données utilisée.

Bien que ces méthodes ne soient pas directement liées à notre problème, la tâche finale peut être confondue. Cependant les défis présentés dans ce domaine de la littérature, semblent être trop distants de nos cas d'applications. Ces méthodes sont conçues pour des images comportant un grand nombre de personnes, par exemple en présence de foules.

2.3 Apprentissage supervisé

Les détecteurs d'objets jouent un rôle crucial dans le dénombrement de personnes, notamment lorsqu'il s'agit de la détection en temps réel de personnes à partir de caméras de surveillance. Les avancées récentes dans le domaine de la vision par ordinateur ont permis le développement de détecteurs performants tels que YOLO (You Only Look Once), SSD (Single Shot MultiBox Detector) ou encore Faster R-CNN (Region-based Convolutional Neural Networks). Ces détecteurs utilisent des réseaux de neurones convolutifs (CNN) pour localiser et classifier rapidement les objets d'intérêt, en l'occurrence dans notre cas, les personnes. Un modèle $\mathcal{M}_\theta$ prend alors une image en entrée x_i et prédit une série de boites englobantes b_i associées à une série de classes c_i pour chaque objet potentiellement détecté dans l'image. L'objectif est donc d'apprendre les paramètres du détecteur θ. Pour se faire $\mathcal{M}_\theta : x \mapsto (b|c)$ est appris en optimisant 2 termes : le coût de classification $\mathcal{L}_{classe}$ associé à chaque type d'objet, ainsi qu'un coût de localisation $\mathcal{L}_{localisation}$ associé aux emplacements des boites englobantes. De cette manière un détecteur peut être vu comme un classifieur (indiquant le type d'objet) ainsi qu'un régresseur (donnant les coins de la boite englobante) en même temps. Deux grandes familles de détecteurs sont disponibles dans la littérature : les détecteurs à un seul étage et ceux à deux étages. Les détecteurs à deux étages séparent la tâche de régression puis la tâche de classification.

2.3.1 Détecteur à deux étages

Le modèle R-CNN (Region-based Convolutional Neural Network) a été l'une des premières avancées majeures dans le domaine de la détection d'objets basée sur les réseaux de neurones convolutifs. Girshick *et al.* (2014) a introduit une approche en plusieurs étapes pour détecter et classifier les objets dans une image. Le fonctionnement de R-CNN est divisé en deux étapes principales. Tout d'abord, une étape de génération de propositions est utilisée pour générer un ensemble de régions candidates susceptibles de contenir des objets. Ces régions sont souvent décrites comme régions d'intérêt. Ensuite, chaque région candidate est extraite de l'image et est redimensionnée à une taille fixe pour être utilisée en entrée d'un réseau de neurones convolutifs. La région extraite est ensuite propagée à travers ce réseau pour extraire des caractéristiques. Enfin, les caractéristiques extraites sont utilisées pour classifier la région et prédire la classe de l'objet qu'elle contient. Cette classification peut être réalisée à l'aide d'un classificateur linéaire ou d'un réseau de neurones supplémentaire. Ainsi on a $\mathcal{M}_1$ qui propose des régions d'intérêt $\mathcal{M}_1 : x \mapsto b$. Une interpolation de ces régions sur une dimension fixe est réalisée pour ensuite avoir $\mathcal{M}_2$ qui classifie ces régions $\mathcal{M}_2 : x \mapsto c$. Le premier réseau $\mathcal{M}_1$ est donc entraîné pour la régression en minimisant $\mathcal{L}_{localisation} = ||b - \hat{b}||$. La classification est donc entraînée de manière classique en optimisant l'entropie croisée $\mathcal{L}_{classe} = \sum \hat{c_j} \times log(p_j)$ où $\hat{c}$ est le vecteur encodant la classe réelle et p_j est la probabilité prédite par le modèle.

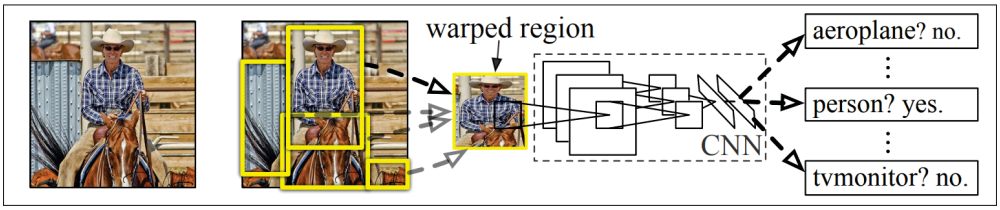


Figure 2.8 Fonctionnement de RCNN tiré de Girshick *et al.* (2014)

Bien que R-CNN ait été une avancée majeure, il présente certaines limitations importantes, ce qui a conduit au développement de modèles plus efficaces, comme par exemple la méthode fast-RCNN Girshick (2015) ou encore plus tard faster-RCNN Ren *et al.* (2016). Ces méthodes sont des améliorations de la méthode R-CNN et sont encore activement utilisées aujourd'hui.

Faster-RCNN de Ren *et al.* (2016) proposent de fusionner l'extraction des caractéristiques de l'image utilisée à la régression et l'extraction des caractéristiques de l'image utilisée à la classification. Les caractéristiques pouvant être redondantes ou encore mieux : complémentaires. Ainsi dans faster-RCNN nous avons une étape d'extraction des caractéristiques puis une tête de régression pour obtenir les régions d'intérêt. Puis la tête de classification est utilisée directement sur les caractéristiques (précédemment calculées) de la région d'intérêt. La tête de proposition des régions d'intérêt est optimisée avec la fonction de coût : $\mathcal{L} = \mathcal{L}_{region} + \mathcal{L}_{objet}$ avec $\mathcal{L}_{region}$ caractérisant la régression pour la tête de proposition des régions. Et $\mathcal{L}_{objet}$ est la classification binaire entre région d'intérêt et de non intérêt. Le premier terme du coût de la tête de classification est composé de $\mathcal{L}_{classe}$ qui est le coût de classification du type d'objet : comme les problèmes de classification classique, l'entropie croisée est utilisée. Le second est $\mathcal{L}_{regression}$ qui est responsable de l'ajustement de région d'intérêt par le réseau de classification. Ainsi $\mathcal{L} = \mathcal{L}_{classe} + \mathcal{L}_{regression}$ pour le réseau classifieur.

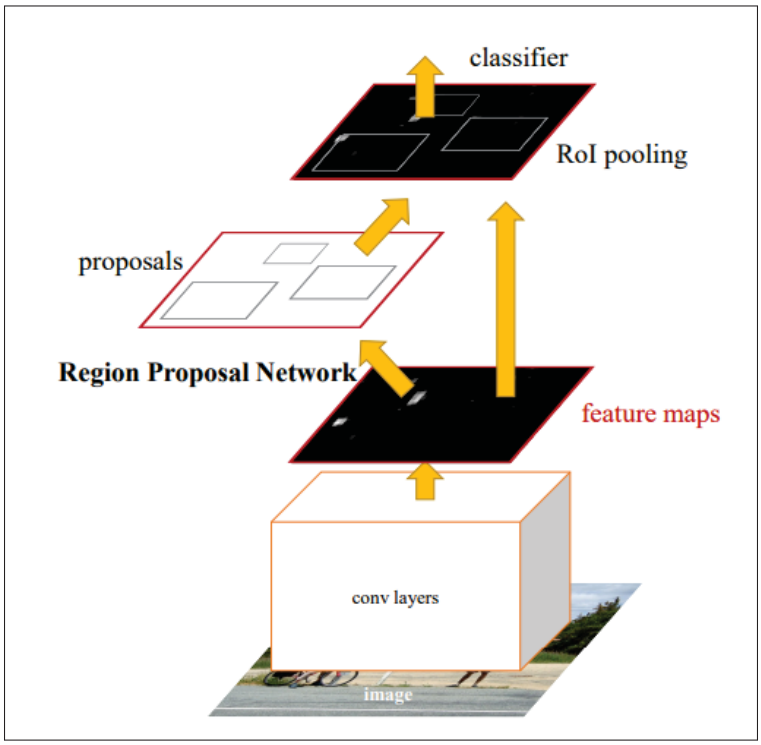


Figure 2.9 Fonctionnement de Faster-RCNN
tiré de Ren *et al.* (2016)

La séparation de la tâche de détection en deux (proposition de régions puis classification) permet de traiter des problèmes isolés et plus simples. Cependant l'augmentation du nombre de calculs que cela implique, rend ce genre de détecteurs inutilisable sur des appareils avec de faibles performances.

2.3.2 Détecteur à un étage

Les détecteurs en un seul étage, tels que YOLO (You Only Look Once) et SSD (Single Shot MultiBox Detector), sont des modèles de détection d'objets qui se distinguent par leur approche efficace et rapide. Contrairement aux détecteurs en deux étapes comme R-CNN, ils effectuent la détection en un seul passage à travers le réseau de neurones convolutifs, ce qui les rend plus rapides et moins complexes. Pour se faire le premier réseau qui extrait les régions d'intérêt est remplacé par un découpage uniforme de l'image pour ainsi former une grille de détection.

YOLO proposé par Redmon & Farhadi (2018) divise l'image en une grille de cellules et prédit directement les boîtes englobantes et les classes d'objets pour chaque cellule. Grâce à sa méthode de détection en un seul passage, YOLO est extrêmement rapide, ce qui lui permet d'effectuer des détections en temps réel. De plus, YOLO excelle dans la détection de petits objets en capturant les contextes globaux des objets dans l'image. Son approche de bout en bout simplifie le processus de détection en intégrant la détection et la classification dans un modèle unique.

De son côté, SSD adopte une approche basée sur des boîtes d'ancrage. Il prédit plusieurs boîtes d'ancrage pour chaque cellule de la grille, à différentes échelles et ratios, puis ajuste ces boîtes pour obtenir des prédictions plus précises. SSD offre une bonne précision et polyvalence, grâce à son équilibre entre la vitesse et la précision. De plus, il s'adapte bien aux différentes formes d'objets grâce à ses boîtes d'ancrage multiples et peut détecter des objets de différentes échelles.

Les avancées sur les modèles de la famille YOLO ont commencé à fusionner les idées avec SSD bien qu'il n'y a pas de publication après la version 3. Ainsi Yolov5 a incorporé l'ancrage des boites utilisées dans SSD. Actuellement les différences entres YOLO est SSD sont faibles en terme d'architecture de modèle. Les principales différences sont dans l'implémentation de SSD

qui est plus rapide et optimisé par l'équipe de Tensorflow. De l'autre coté, YOLO est développé de manière open-source sur Pytorch ce qui a conduit à l'élaboration d'un artifice d'outils pour en améliorer les performances. Ainsi YOLO est plus performant mais souvent négligé pour la recherche car trop complexe pour en étudier et comprendre tous les aspects. Bien que les résultats de ces deux familles se rapprochent, il est commun de dire que les détecteurs à un étage sont plus rapides mais moins précis que les détecteurs à deux étages, leur utilisation dépendra donc de notre application.

2.4 L'utilisation des images infrarouges basse résolution

L'anonymat des personnes est un enjeu majeur dans notre société moderne, où la protection de la vie privée est devenue une préoccupation croissante. Dans de nombreux contextes où l'on veut développer une technologie qui doit être utilisée par un grand nombre de personnes, il est essentiel de préserver l'anonymat des individus tout en garantissant la sécurité de leurs données.

L'une des approches les plus avancées pour garantir l'anonymat consiste à se concentrer sur la phase d'acquisition des images plutôt que sur le traitement ultérieur de celles-ci. Par exemple He *et al.* (2021) propose de flouter les visages avant l'utilisation de l'image dans le modèle. Le problème est partiellement résolut car image d'origine reste disponible sur l'appareil. Cela signifie que la capture des images elle-même doit être conçue de manière à protéger l'identité des individus. Dans cette optique, l'utilisation d'images à basse résolution et de capteurs infrarouges jouent un rôle clé. Les images à basse résolution offrent un niveau de détail réduit, ce qui rend plus difficile l'identification précise des individus. Associées à des capteurs infrarouges, ces images permettent de capturer les mouvements et les actions des personnes sans révéler leurs caractéristiques faciales spécifiques. Les capteurs infrarouges détectent les émissions de chaleur du corps humain, ce qui permet de visualiser les personnes même dans l'obscurité totale ou lorsque la visibilité est limitée ce qui est notre cas d'utilisation par exemple en utilisant nos méthodes pour détecter la présence et ainsi allumer la lumière. Cependant, malgré leur potentiel pour préserver l'anonymat, la détection des personnes dans des images à basse résolution et infrarouges reste un domaine de recherche relativement peu exploré. La plupart des études

existantes se concentrent principalement sur la détection des actions plutôt que sur la détection et localisation des personnes.

La détection des chutes, par exemple, est essentielle dans les environnements médicaux tels que les hôpitaux, les maisons de retraite et les urgences. Les chutes peuvent avoir des conséquences graves pour les patients et les personnes âgées, et une détection rapide peut permettre une intervention immédiate des secours. En utilisant des techniques de détection d'actions, il est possible de surveiller en temps réel les mouvements des individus et de détecter les chutes dès qu'elles se produisent, déclenchant ainsi une alerte pour les soignants ou les équipes médicales.

Ainsi Tateno, Meng, Qian & Li (2020b) propose d'utiliser le seuil de Ostu (publié dans Otsu (1979)) afin de segmenter l'image et de localiser la personne sans pour autant l'évaluer. Ce travail réalisé par Tateno *et al.* (2020b) se concentre d'avantage sur l'utilisation d'un réseau récurrent type LSTM pour fusionner les informations d'un CNN suivant l'axe de temps afin d'améliorer les résultats proposés par lui même dans Tateno *et al.* (2019). Cependant son précédent travail avait pour but la classification de gestes d'une main avec le même type d'images en 32x24 pixels en infrarouge. Ainsi Tateno parvient à améliorer le taux d'exactitude de 72,2% à 81,8% en utilisant un modèle temporel.

Karayaneva, Baker, Tan & Jing (2018) propose une étude initiale sur la détection des actions en infrarouge basse résolution en proposant 3 modèles qui sont naïvement appliqués directement aux données afin d'en évaluer la difficulté : une machine à vecteurs de support (SVM), une forêt d'arbres décisionnels (RF) et la méthode des k plus proches voisins (K-NN) sont ainsi introduites. Tao *et al.* (2018) propose une amélioration au travail précédent en essayant n'atténuer le bruit du capteur par un filtrage avec un noyau Gaussien ainsi que la suppression de fond de l'image. Un traitement pré-classification est aussi introduit par Tao *et al.* (2018) par l'utilisation de la transformée en cosinus discrète afin d'extraire les informations temporelles et spatiales. Un SVM est ensuite utilisé pour assurer la classification. De cette manière Tao *et al.* (2018) améliore les résultats proposés par Karayaneva *et al.* (2018) en passant de 94.2% taux d'exactitude à 97.9%.

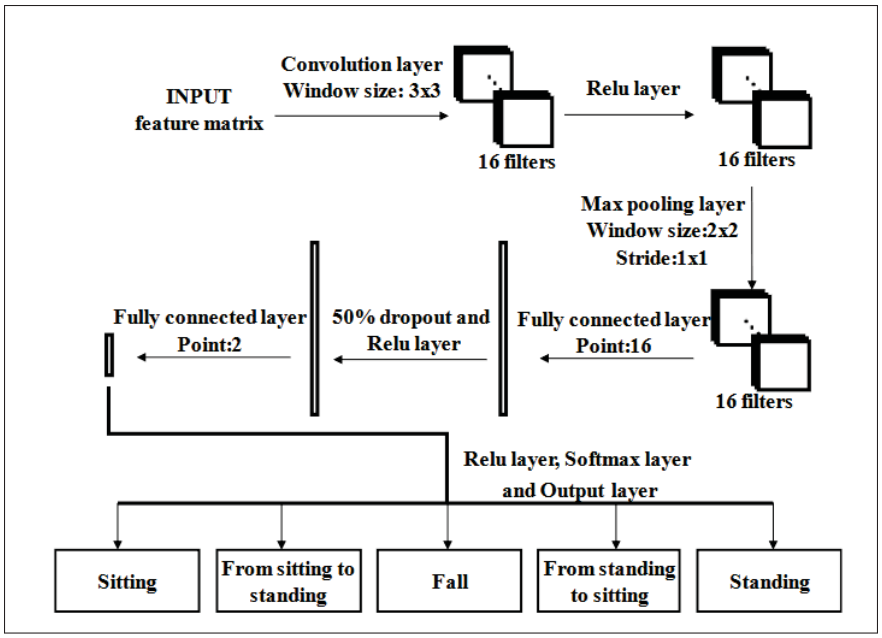


Figure 2.10 Réseau de neurones convolutifs en 2d tiré de Tateno *et al.* (2019)

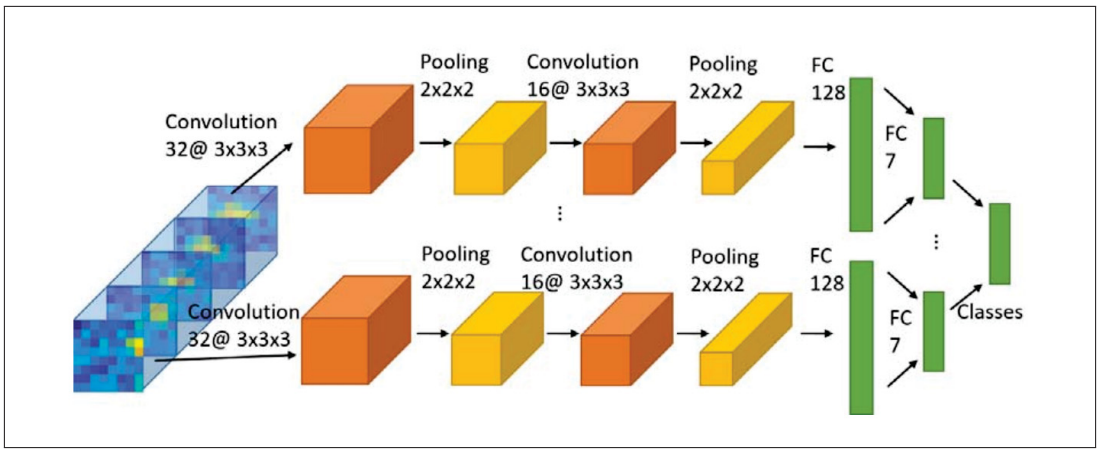


Figure 2.11 Architecture utilisant des 3D-convolutions tiré de Tao *et al.* (2019) et reprises par Tateno *et al.* (2020a)

La dernière amélioration que l'on observe dans la littérature au sujet de la classification d'actions a été proposée par deux groupes de recherche dans une même période de temps : Tateno *et al.* (2020a) et Tao *et al.* (2019). Ils proposent d'utiliser des 3D-convolutions (voir figure 2.11) afin d'extraire l'information temporelle en même temps que l'information spatiale. De cette

manière nous n'avons pas les difficultés à optimiser un réseau de neurones récurrents comme précédemment avec le LSTM. De plus, la fusion de l'information dès le début du réseau permet d'extraire plus de caractéristiques dépendantes du temps. Généralement, plus de calculs sont réalisés après leurs agrégations, comparé à une approche utilisant un LSTM à la fin. Ces deux publications sont en accord et proposent une architecture très proche qui surpasse l'état de l'art.

Le défaut récurrent dans ces méthodes proposées par Tateno *et al.* (2020a) et Tao *et al.* (2019) est un manque d'évaluation de la localisation des personnes bien qu'une partie y est réservée dans leurs travaux. Nous pensons que c'est grandement lié au manque de bases de données publiques dans ce domaine. Nous avons donc essayé de combler ce vide.

2.5 Reconnaissance d'actions

Comme vu précédemment la littérature utilisant des images infrarouges de basse résolution s'intéresse principalement à la reconnaissance d'actions. Nous en avons étudié une partie, cependant nous n'avons pas traité tous les domaines qui la compose. Celle-ci est parfois incompatible avec notre application, par exemple en utilisant les coordonnées d'un squelette comme entrée pour classifier l'action. La taxonomie proposée par Morshed, Sultana, Alam & Lee (2023) sépare les méthodes en quatre groupes pour faire de la reconnaissance d'actions. Les trois premiers se basent sur de l'extraction de caractéristiques avant de classer l'image. Il y a les caractéristiques réalisées à la main, les caractéristiques issues d'un apprentissage profond, et les caractéristiques d'attention souvent issues d'un apprentissage auto-supervisé. Le dernier groupe ne se base pas sur de l'extraction de caractéristiques mais plutôt sur une subdivision plus fine des actions. Ce dernier groupe se base donc sur les interactions inter-individus et inter-subdivision pour en prédire les actions présentes. Plusieurs approches que nous avons présentées précédemment peuvent être incluses dans ces groupes.

2.5.1 Caractéristiques réalisées à la main

Dans les caractéristiques réalisées à la main, Morshed *et al.* (2023) place des méthodes d'extractions de caractéristiques comme proposée par Vishwakarma, Kapoor & Dhiman (2016). Sa méthode repose sur une projection des caractéristiques de l'image dans un espace plus simple pour une classification des actions. Vishwakarma *et al.* (2016) utilise la transformée de Radon (proposée par Radon (1917)) suivie d'une réduction de dimension par la méthode d'incorporation linéaire locale (LLE proposé par Saul & Roweis (2000)). La transfomée de Radon, souvent utilisée en tomographie médicale, permet d'extraire de l'information sur la topologie de la scène. De cette manière, l'information temporelle et géométrique contenue dans la séquence est transformée et caractérisée par sa topologie. La réduction par LLE permet une meilleure généralisation de l'étape de classification. Une autre méthode proposée par Tao *et al.* (2018), présentée précédemment, utilisant une transformée en cosinus discrète se place ainsi aussi dans ce groupe. La transformée en cosinus discrète, est très proche de la transformée de Fourier bien connue. Elle permet de transformer l'information spatiale et temporelle contenue dans l'image dans le domaine fréquentiel, de cette manière, la vitesse d'exécution des actions ainsi que leurs répétitions sont utilisées pour mieux prédire le type d'action réalisée.

Bien que ces méthodes soit simples d'utilisation et rapides, leur point faible est le niveau des performances qu'elles obtiennent. En effet les caractéristiques réalisées à la main permettent d'améliorer un classifieur naïf directement sur les données. Leurs manque de flexibilité ne leurs permettent pas de profiter du grand nombre d'exemples présent dans les données d'entraînements.

2.5.2 Extraction par apprentissage profond

Les méthodes que présente Tateno *et al.* (2020b) et Tao *et al.* (2019) se base sur de l'extraction de caractéristiques utilisant un réseau de neurones. Ainsi ces méthodes se placent dans ce même groupe. Plusieurs manières de fusionner l'information existent. La fusion avant l'extraction de caractéristiques permet cette agrégation, comme le propose Ji, Xu, Yang & Yu (2012) en utilisant des opérations de convolutions en trois dimensions pour extraire les informations

géométriques ainsi que temporelles. Une autre méthode est la fusion après l'extraction des caractéristiques. Shi, Tian, Wang & Huang (2017) proposent d'utiliser un LSTM pour fusionner l'information temporelle afin d'obtenir les trajectoires des caractéristiques profondes du réseau CNN. De cette manière Shi *et al.* (2017) est capable de reconnaître les actions. Une amélioration dans l'architecture du modèle est proposée par Song *et al.* (2018) en ajoutant des connections résiduelles aux 3d-convolutions ainsi qu'une tête de décision utilisant un réseau récurrent pour la décision finale. Ce modèle utilise donc une fusion avant et après l'extraction des caractéristiques, améliorant ainsi les performances des méthodes précédentes.

Les désavantages de ces méthodes sont la complexité algorithmique élevée ainsi qu'un apprentissage conséquent. Cependant les résultats que proposent ces méthodes, permettent à ce groupe d'obtenir les meilleurs résultats.

2.5.3 Mécanisme d'attention et apprentissage auto-supervisé

Ce groupe repose sur un apprentissage en deux étapes. La première est l'entraînement d'un extracteur de caractéristiques souvent entraîné de manière auto-supervisée. La seconde étape utilise les caractéristiques précédemment apprises ainsi que les annotations afin d'affiner le modèle et de prédire les différentes classes. La première étape n'a pas besoin d'annotations, il est donc souvent possible d'utiliser plus de données. De plus, cet apprentissage permet d'apprendre des caractéristiques intrinsèques aux données ce qui peut rendre le modèle plus robuste. Une méthode proposée par Ullah *et al.* (2019) utilise un extracteur de caractéristiques pré-entraîné ainsi qu'un auto-encodeur pour en réduire l'information. L'auto-encodeur est entraîné à projeter les caractéristiques issues du réseau pré-entraîné dans un espace de plus petite dimension. Pour se faire, l'auto-encodeur a pour but de réduire l'erreur moyenne des carrées pour la tâche de reconstruction des caractéristiques. Le réseau extracteur, adjoint à l'encodeur ainsi entraîné, permet de projeter les entrées dans un espace de faible dimension et permet une meilleure classification. Un SVM est ensuite utilisé par Ullah *et al.* (2019) pour finaliser la prédiction. La prédiction directe par un SVM serait de moins bonne qualité considérant la dimensionnalité de l'espace d'entrée.

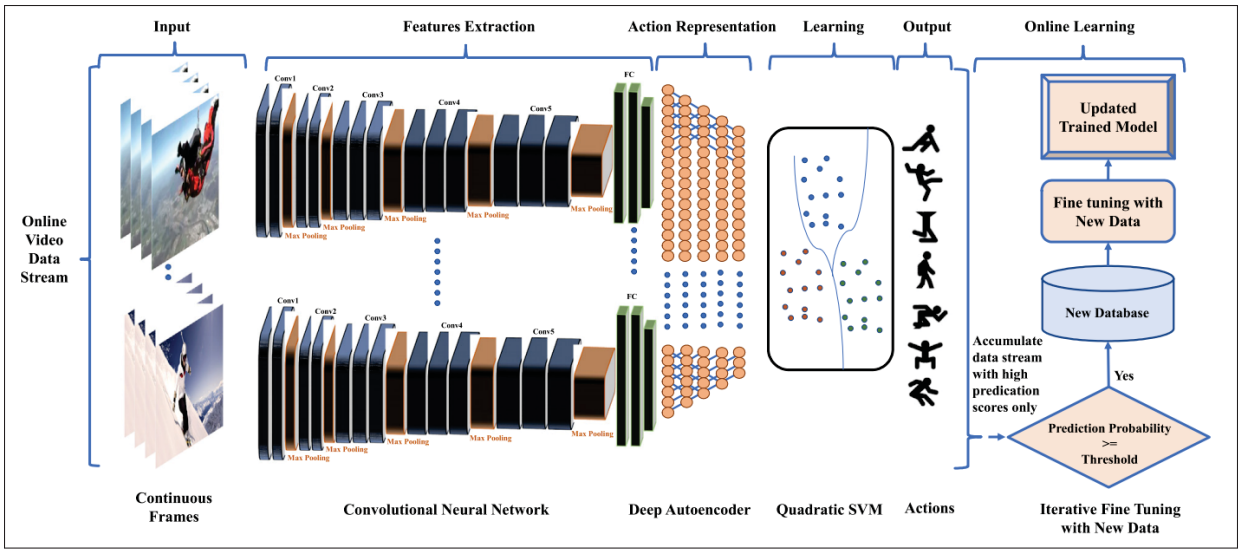


Figure 2.12 Méthode tirée de Ullah *et al.* (2019)

Wang *et al.* (2021) et E. Santos & Pedrini (2019) proposent d'utiliser dans leurs méthodes les récentes avancées sur les mécanismes d'attentions utilisés principalement dans les transformers proposés par Vaswani *et al.* (2017). Cette nouvelle architecture qui est maintenant généralisée par Dosovitskiy *et al.* (2020) pour la vision par ordinateur, permet de changer la façon dont les modèles de vision fonctionnent : les caractéristiques sont maintenant traitées indépendamment et mises en relation par des systèmes d'attentions. Wang *et al.* (2021) et E. Santos & Pedrini (2019) les utilisent dans un entraînement auto-supervisé pour résumer la scène puis la reconstruire en minimisant l'erreur moyenne carrée. Wang *et al.* (2021) propose de se limiter à une utilisation en ligne ce qui ne lui permet pas d'utiliser des données futures pour la prédiction. E. Santos & Pedrini (2019) propose de séparer l'information actuelle de l'information temporelle (déduite d'une estimation du flux optique) dans le but de fusionner l'information en fin de réseau.

2.5.4 Subdivision des actions

Les méthodes employées dans ce groupe reposent sur la subdivision des actions à plusieurs échelles que Morshed *et al.* (2023) catégorise en micro-actions qui sont qualifiée d'atomiques, le regroupement comportemental, les interactions entre individus, ainsi que les actions de groupe.

Les méthodes qui appartiennent à ces catégories, permettent donc de prédire les actions à plusieurs échelles. Par exemple, on peut prédire qu'un joueur joue au tennis avec des micro-actions (mouvement de raquette, service, courir) ainsi que par l'interaction avec son adversaire (deux personnes qui s'envoient une balle).

Marszalek, Laptev & Schmid (2009) proposent une méthode pionnière dans le domaine, en jouant sur le contexte physique de la scène. Pour se faire, ils utilisent un ensemble de caractéristiques extraites par différents CNN ainsi que les sous-titres présents dans les films pour apprendre le contexte dans lequel les actions sont réalisées. Ainsi, cette méthode réalise un apprentissage sous privilège du texte. Les informations contextuelle et physique permettent à Marszalek *et al.* (2009) une meilleure classification des actions réalisées dans les films.

2.6 Utilisation de contexte temporel pour la détection

Les réseaux de neurones convolutifs utilisés dans le cadre de la vision par ordinateur reposent sur l'application successive de filtres de convolution. Ces applications permettent d'extraire des caractéristiques en se reposant sur des motifs. Ainsi ce n'est pas tant la valeur d'un pixel qui nous donne de l'information mais la relation entre les pixels de l'image. Il est possible de le voir simplement en évaluant la corrélation entre les pixels dans une même zone (voir figure 4.10). Cependant dans le cadre de la surveillance et de la détection de personnes nous sommes souvent amenés à traiter des images provenant de vidéos. De cette manière, la corrélation temporelle est aussi très forte. Les méthodes se basant sur l'utilisation du temps permettent d'utiliser cette information pour améliorer les résultats de détection.

2.6.1 Cas d'une vue statique

Dans le cas d'une caméra statique il est possible d'extraire le fond de l'image pour pouvoir le soustraire de l'image actuelle, et ainsi en déduire les objets qui sont nouveaux comme dans le cadre de l'auto-encodeur.

Lee, Hull & Erol (2003) propose d'utiliser un modèle probabiliste de mélange gaussien afin d'estimer le fond de l'image. Ainsi pour chaque pixel $p_{i,j}$ ($i \in W$ et $j \in H$) de l'image nous avons une densité de probabilité caractérisant son appartenance au fond de l'image ou non. En appliquant une valeur seuil nous pouvons segmenter l'image en deux. Ce travail est reprit par Zivkovic (2004) en proposant une amélioration du temps d'inférence de 30%.

Cette méthode est employée par Cao *et al.* (2016) pour compter les personnes passant sous une caméra fixée au plafond.

a) b)

Figure 2.13 En a) l'image en nuances de gris, b) l'estimation des pixels n'appartenant pas au fond de l'image. Tirée de Cao *et al.* (2016)

Le principal problème de cette technique est la grande quantité de calculs nécessaire pour l'inférence. Pour chaque pixel, il y a un modèle de mélange gaussien à évaluer.

Une autre méthode proposée par Yu *et al.* (2008) permet de résoudre ce problème. A la place d'estimer un modèle pour chaque pixel, Yu *et al.* (2008) propose de s'intéresser uniquement aux contours des personnes. Comme la localisation ne fait intervenir que les contours d'une personne, sa méthode commence donc par une extraction des bords de l'image. Ainsi Yu *et al.* (2008) définit une image des bords appartenant au fond de l'image en considérant les pixels extraits dans un contour à plus de 80% du temps. Enfin, en mode inférence, il met à jour la carte des bords appartenant au fond de l'image et en faisant la soustraction de celle-ci avec l'image initiale, seuls les contours des personnes en mouvement restent.

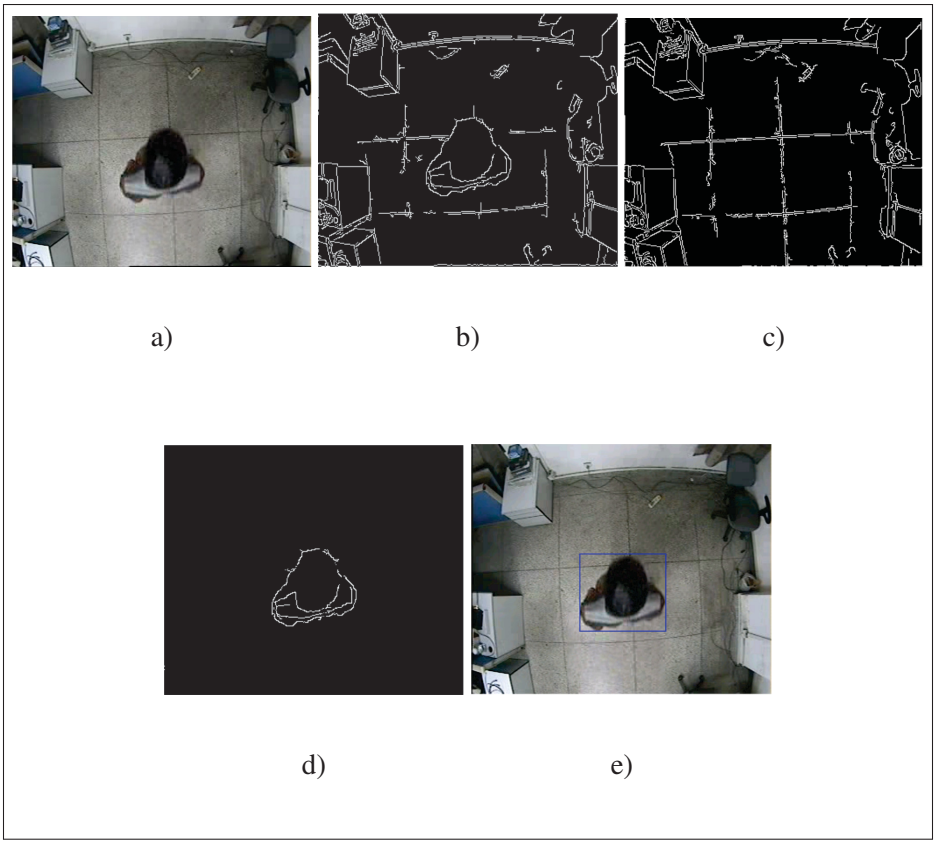


Figure 2.14 En a) l'image en couleur, b) l'extraction des contours de l'image a), c) l'estimation des contours de fond, d) le résultat de b-c, e est la détection finale sur l'image RGB. Tirée de Yu *et al.* (2008)

Bien que ces méthodes soient simples d'utilisation et de mise en oeuvre, leurs détections reposent uniquement sur la cohérence temporelle des images. Ainsi ces méthodes sont uniquement en capacité de détecter des personnes en mouvement et confondraient tout objet en mouvement avec une personne.

2.6.2 Cas d'une vue non statique

Dans le cas d'une vue non statique, il est bien plus compliqué d'utiliser l'information temporelle que dans le cas statique précédemment. Par exemple Duran-Vega *et al.* (2021) propose d'utiliser une architecture de détecteur Yolov5 puis une architecture de réseau de neurones récurents (RNN) pour fusionner l'information temporelle. Duran-Vega *et al.* (2021) propose d'étudier

plusieurs type de RNN : ConvLSTM (proposé par Shi *et al.* (2015)), LSTM (proposé par Hochreiter & Schmidhuber (1997), RNN et QRNN. Beaucoup de travaux antécédents ont été proposés, calqués sur les différents type de détecteurs. Par exemple le travail de Kang *et al.* (2017) repose sur l'architecture de R-CNN en fusionnant l'information temporelle après le modèle. Liu & Zhu (2018) propose une architecture semblable sur la base du détecteur SSD.

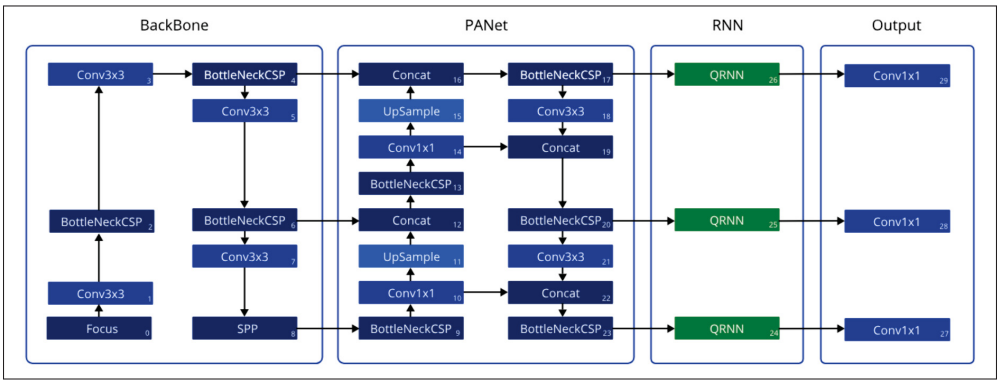


Figure 2.15 Architecture TYolo tirée de Duran-Vega *et al.* (2021)

L'ensemble de ces techniques repose sur le même principe : l'utilisation d'un réseau de neurones CNN pour extraire des caractéristiques qui sont ensuite fusionnées dans la direction du temps pour en produire la détection sur l'image actuelle.

Un des problèmes dont ces techniques souffrent, est le sur-apprentissage : étant donné que les entrées du modèle ne sont plus des images mais des courtes vidéos, le nombre d'exemples réduit drastiquement. De plus, l'augmentation du nombre de paramètres présents dans le modèle, amplifie ce problème.

Corsel *et al.* (2023) propose une solution, bien que ce ne soit pas sa première motivation. Il propose d'utiliser le contexte temporel pour améliorer la détection des petits objets. L'idée est d'utiliser les 3 canaux de couleur : rouge, vert , bleu pour incorporer l'information temporelle. Ainsi pour une image couleur, Corsel *et al.* (2023) propose d'utiliser une transformation en nuances de gris. De cette manière il condense l'information en 1 canal de couleur, puis il utilise l'image provenant de l'acquisition antérieure et postérieure à l'instant actuel.

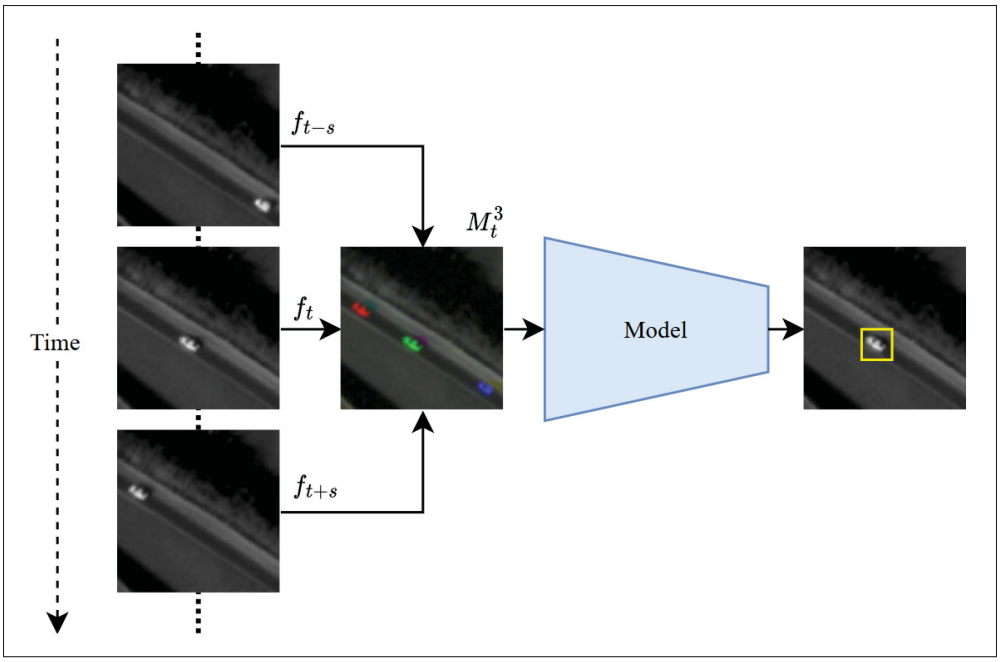


Figure 2.16 Utilisation du contexte tirée de Corsel *et al.* (2023)

De cette manière nous pouvons voir figure 2.16 que l'objet en mouvement apparaît en rouge vert et bleu. Bien que cette image peut être plus confuse pour nous, le modèle a la capacité d'apprendre à reconnaître l'objet sur l'image du milieu. La technique proposée par Corsel *et al.* (2023) exploite la négligence du contexte chromatique pour utiliser le contexte temporel à la place. Il montre dans sa publication qu'il arrive à augmenter la qualité de détection pour des objets en mouvement et petits, sans détériorer la détection pour les objets statiques.

Cette méthode est particulièrement intéressante dans notre cas, car les modèles de détection utilisent 3 canaux pour les trois couleurs. Dans le cas de la détection en infrarouge, nous n'avons qu'un seul canal de température qui nous permet d'utiliser cette méthode sans perdre de l'information. A la place de tripler l'image en IR pour la faire correspondre au modèle de détection, nous pouvons directement utiliser 3 images consécutives afin d'améliorer la détection. Ainsi, le modèle de détection est le même, le coût supplémentaire de cette technique est donc nul.

2.6.3 Augmentation de données temporelles

L'ajout d'information peut aider le modèle à prendre des décisions mais dans certains cas une utilisation d'un espace d'entrée trop grand peut conduire à une mauvaise généralisation. Ainsi une méthode employée par celles susmentionnées, utilise de l'augmentation de données pour rendre l'entraînement du modèle plus robuste.

Duran-Vega *et al.* (2021) propose 5 augmentations de données : la mosaïque temporelle, le mixage temporel, le flou, la suppression de région et l'ajout de bruit.

Figure 2.17 Mosaïque temporelle tirée de Duran-Vega *et al.* (2021)

Cette augmentation de mosaïque temporelle (voir figure 2.17 est une augmentation de données qui vient coller plusieurs clips vidéos les uns à coté des autres. Cette augmentation de données est connue pour être très puissance dans le cadre de la détection et a été adaptée par Duran-Vega *et al.* (2021) pour être utilisée avec des clips vidéos.

L'augmentation de mixage temporel (voir figure 2.18) est aussi inspirée de l'augmentation de mixage proposée par Zhang, Cisse, Dauphin & Lopez-Paz (2017). Elle améliore la généralisation globale d'un modèle en permettant une plus grande diversité d'exemples. Cet exemple augmenté s'obtient en faisant une moyenne pondérée des valeurs d'entrée et de sortie : $\tilde{x} = \lambda \cdot x_i + (1-\lambda) \cdot x_j$ avec $\lambda \in [0, 1]$.

Figure 2.18 Mixage temporel tiré de Duran-Vega *et al.* (2021)

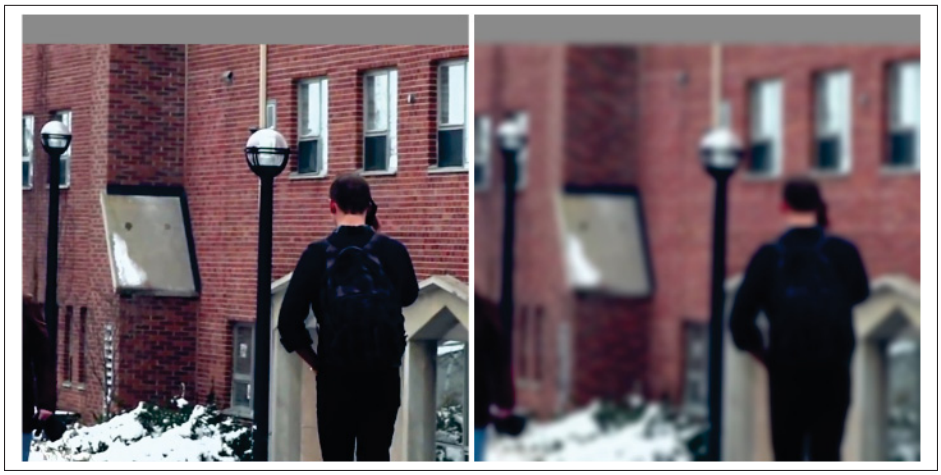


Figure 2.19 Flou aléatoire tiré de Duran-Vega *et al.* (2021)

L'augmentation de données de flou proposée par Duran-Vega *et al.* (2021) est motivée par un phénomène physique que l'on peut rencontrer lors de l'utilisation des caméras : la mise au point automatique. Cette mise au point est appliquée de manière cyclique par la caméra. Ainsi cette augmentation de données (voir figure 2.19) vise à simuler cet effet.

Cette augmentation de données permet de simuler l'occlusion d'un objet à détecter dans la vidéo.

Figure 2.20 Suppression de région tirée de Duran-Vega *et al.* (2021)

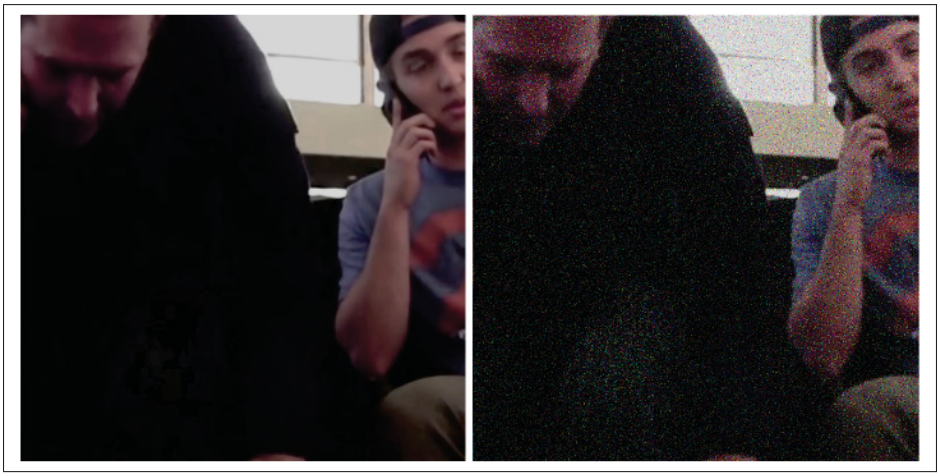


Figure 2.21 Bruit aléatoire tiré de Duran-Vega *et al.* (2021)

L'augmentation d'ajout de bruit aléatoire (voir figure 2.21) permet de rendre le détecteur plus robuste au bruit qui est plus important dans le cadre de vidéos par comparé aux images. Cela est expliqué par un temps d'exposition plus contraint dans le cadre des vidéos.

Bien que ces augmentations de données soient souvent motivées par des phénomènes physiques (mise au point automatique, bruit de capteur,...), l'évaluation de leurs caractéristiques physiques avec les données d'entraînement pourrait être bénéfique pour mieux en simuler la distribution.

De plus la suppression de régions dépend grandement de la base de données utilisée car les occlusions ne sont que rarement annotées dans le cadre d'une détection. Cette augmentation de données fera donc gonfler le nombre de faux positifs dans ces cas là.

CHAPITRE 3

ÉTUDE DU NIVEAU DE SUPERVISION POUR LA DÉTECTION DE PERSONNES EN INFRAROUGE ET BASSE RÉSOLUTION.

Cette étude a été réalisée dans le cadre d'une collaboration avec la compagnie Distech controls. Ces résultats ont aussi fait l'objet d'une publication lors d'un atelier à la conférence européenne sur la vision par ordinateur ECCV : Dubail *et al.* (2022). Ce travail a également été présenté à la conférence AI Symposium 2022 au laboratoire de recherche Mila ainsi qu'au Colloque REPARTI 2023 à l'université Laval où il a été récompensé par un prix visant à encourager la participation.

3.1 Introduction

Les solutions intelligentes de gestion des bâtiments visent à maximiser le confort des occupants tout en minimisant la consommation d'énergie. Ces solutions sont essentielles pour réduire l'utilisation des combustibles fossiles, contribuant ainsi directement à la réduction de notre impact direct sur l'environnement. Cette économie d'énergie est généralement réalisée en contrôlant de manière adaptative l'éclairage, le chauffage, la ventilation et la climatisation en fonction de l'occupation du bâtiment, en modulant ces flux en fonction du nombre de personnes présentes dans une pièce donnée. Pour cela, des méthodes peu coûteuses énergiquement sont nécessaires pour évaluer le niveau d'occupation d'une pièce et contrôler efficacement les différents systèmes du bâtiment.

Parmi les différents niveaux d'informations sur l'occupation pouvant être extraites dans un bâtiment intelligent Sun, Zhao & Zou (2020). définissent la localisation des occupants comme la caractéristique la plus importante pour le contrôle intelligent des espaces. Grâce aux récents progrès de l'apprentissage automatique et de la vision par ordinateur, la plupart des solutions s'appuient généralement sur des réseaux neurones convolutionnels profonds (CNN) pour détecter les personnes Gao, Li, Zhang, Liu & Wang (2016); Chen, Wang & Liu (2018). Malgré le haut niveau de précision pouvant être atteint avec les CNN pour la détection d'objets basée sur des images RGB (couleurs), leur implémentation pour des applications réelles de vidéosurveillance

entraîne une complexité numérique élevée, des problèmes de confidentialité et des biais de genre et de race (Buolamwini & Gebru (2018); Schwemmer *et al.* (2020)). Enfin, les solutions de gestion de l'occupation des bâtiments sont généralement mises en œuvre sur des dispositifs embarqués compacts, installés de manière rigide au plafond ou aux entrées des pièces, et intégrant des caméras peu coûteuses capables de capturer des images IR (infrarouge) de faible résolution.

Pour respecter ces contraintes de liberté individuelle, He *et al.* (2021) ont proposé un détecteur d'objets préservant la confidentialité qui floute les visages des personnes avant d'effectuer la détection. Pour réduire les biais du détecteur reposant sur le genre/ethnie, les mêmes auteurs ont proposé une variation de remplacement de visage qui préserve également la confidentialité au prix d'une complexité informatique accrue. Malgré de bonnes performances, leur approche ne garantit pas la confidentialité au niveau de l'acquisition, car elle repose sur des capteurs RGB pour construire la solution. De plus, leur détecteur a été conçu pour des paramètres entièrement annotés en utilisant COCO Lin *et al.* (2014) comme base de données. Cela rend difficile sa généralisation à la détection de personnes dans des conditions de capture différentes (comme lorsque les caméras sont situées au plafond, sur des systèmes embarqués compacts) et à des changements extrêmes dans l'environnement. De plus, il est difficile de collecter et d'annoter des données d'images pour entraîner ou affiner des détecteurs d'objets basés sur des CNN pour une application donnée, c'est pourquoi l'apprentissage faiblement supervisé ou non supervisé est une approche prometteuse dans ce cas d'application.

Notre travail aborde le problème de localisation des occupants en détectant les personnes dans des images infrarouges (IR) à basse résolution (24x32), ce qui évite la plupart des problèmes de confidentialité mentionnés ci-dessus. La basse résolution permet non seulement de réduire la complexité numérique, mais aussi d'améliorer la confidentialité, c'est-à-dire qu'une détection sur des images infrarouges haute résolution ne serait pas satisfaisant car il serait possible de ré-identifier les personnes comme le montre Zheng *et al.* (2022). Plus spécifiquement, nous analysons la détection des personnes avec différents niveaux de supervision. Dans ce travail, nous comparons des solutions non supervisées, faiblement supervisées et entièrement supervisées. Il

s'agit d'un aspect essentiel de la chaîne de détection, car la production d'annotations de boîtes englobantes est très coûteuse et il existe que peu de bases de données de détection d'objets en open source pour les scénarios infrarouges à basse résolution. En fait, réduire le niveau de supervision peut permettre un meilleur cas d'usage pour les applications réelles et une réduction de la complexité informatique, ce qui est important compte tenu de l'utilisation des algorithmes proposés sur des dispositifs embarqués dans notre cas.

3.2 Description du dispositif et de la base de données

L'entreprise Distech controls propose donc un système embarqué qui va être installé sur le plafond et muni d'une caméra infrarouge basse résolution (24x32). Cette solution permet au dispositif d'avoir une vue d'ensemble de la pièce pour en détecter le nombre de personnes (voir figure 3.1).

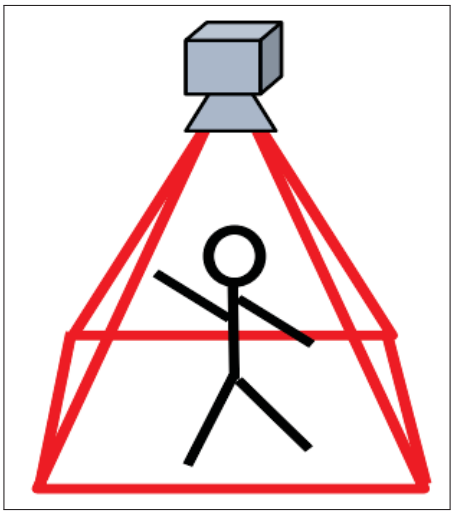


Figure 3.1 Installation du dispositif au plafond

Le défi étant de concevoir un produit qui va consommer moins d'énergie que le gain qu'il va induire par la gestion dynamique des paramètres du bâtiment. Comme le contrôle des paramètres de température, flux d'air et de lumière ne sont pas de notre ressort, nous nous sommes concentrés sur l'optimisation de la détection, tout en réduisant le coût de calcul.

Afin de pouvoir faire l'acquisition d'une base de données d'exemples et de pouvoir être capable d'annoter les images nous avons fait le choix d'utiliser une seconde caméra RGB et haute résolution. De cette manière, un humain peut facilement annoter les images couleurs. Dans un second temps nous pouvons transférer la boite englobante de l'image haute résolution RGB à l'image basse résolution IR.

Figure 3.2 Défaut d'alignement entre l'image RGB et l'image IR. Pour la visualisation seuls les pixels supérieur à 20°C sont montré

La distorsion étant proche pour les deux caméras, la fonction de transfert des coordonnées RGB aux coordonnées IR peut être exprimée par une fonction linéaire $x_{IR} = a \cdot x_{RGB} + b$. Afin de déterminer ces paramètres nous avons annoté des images représentant une scène simple et sans confusion possible sur les deux modalités. Il est donc possible de déduire les fonctions d'alignement.

Comme nous pouvons le voir sur les fonctions d'alignements figure 3.3, l'ordonnée à l'origine n'est pas nulle. Ce qui témoigne d'un champ de vision plus grand pour la caméra RGB que la camera IR. Afin de retirer toute confusion du modèle sur la zone de détection, nous avons décidé de zoomer et de couper les bords des images pour se limiter uniquement à la zone commune.

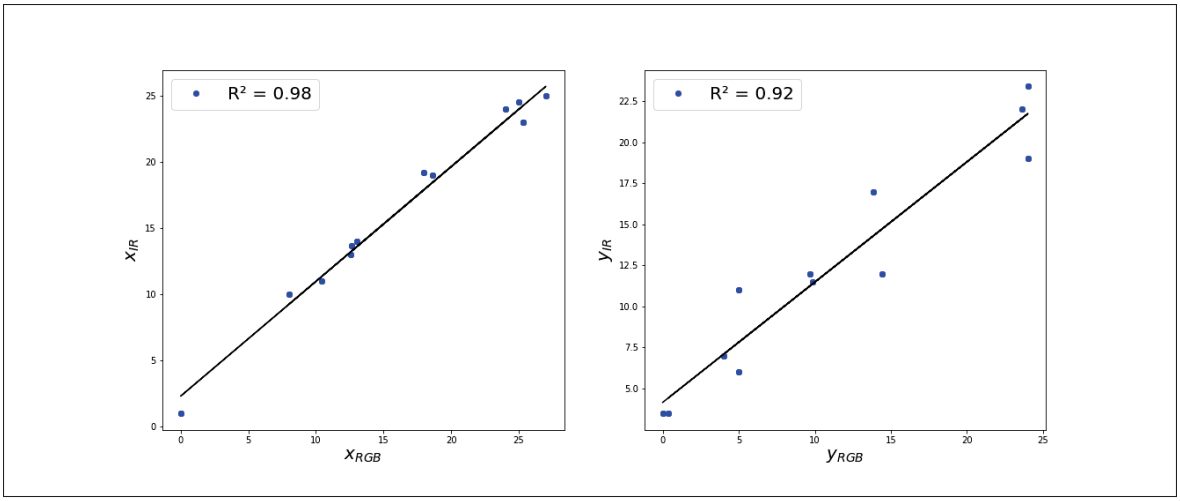


Figure 3.3 Fonctions d'alignement entre les deux caméras : IR et RGB

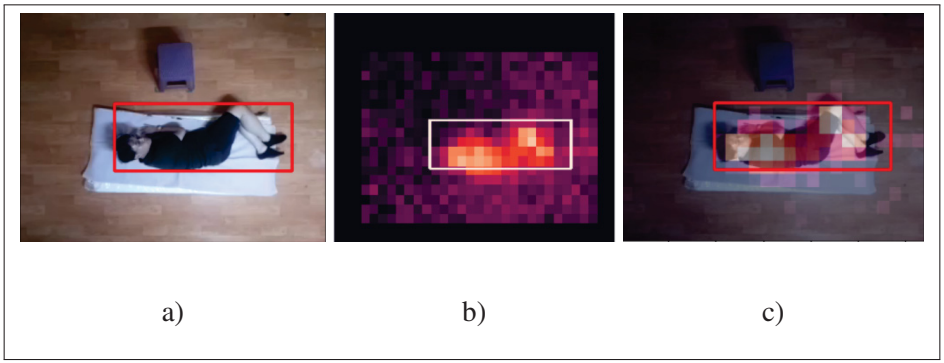


Figure 3.4 En a) l'image RGB, b) la zone commune dans l'image IR correspondante a a), c) la superposition des zones communes

Pour des soucis de répétabilité, nous avons réalisé les expérimentations sur deux bases de données : la base de donnée privées Distech-Low-IR ainsi qu'une base de données publique FIR-Image-Action-Dataset qu'on a annotée afin de publier ce travail.

3.2.1 FIR-Image-Action

Le jeu de données FIR-Image-Action Zhang (2020) comprend 110 vidéos annotées. Nous avons sélectionné au hasard 36 vidéos de cet ensemble pour l'ensemble de test $\mathcal{D}_t$ et les 74 autres pour

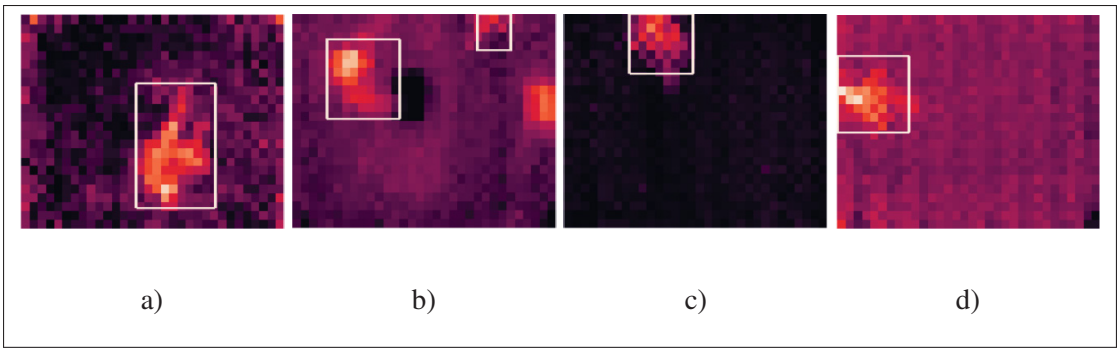


Figure 3.5 Voici des exemples d'images infrarouges avec annotations pour les ensembles de données FIR-Image-Action (a) et Distech-Low-IR (b)-(d)

l'entraînement $\mathcal{D}_e$ et la validation $\mathcal{D}_v$. De plus, les ensembles d'entraînement et de validation ont été séparés en utilisant une sélection aléatoire des images (70% et 30%, respectivement). Toutes les approches ont été entraînées en utilisant la même partition des données pour assurer la comparabilité des résultats.

À notre connaissance, il n'existe pas de jeu de données infrarouge à basse résolution avec des annotations de boîtes englobantes pour la détection de personnes. Par conséquent, nous avons annoté ce jeu de données au niveau des boîtes englobantes. Le jeu de données a été créé par Haoyu Zhang de Visiongo Inc. pour la reconnaissance d'actions basée sur la vidéo. Ce jeu de données propose 126 vidéos d'une durée totale d'environ 7 heures. Étant donné que cette étude vise à évaluer les performances de différentes techniques de localisation de personnes basées sur l'infrarouge, nous n'avons utilisé que les images IR fournies par les auteurs pour nos expériences. Néanmoins, il convient de mentionner que deux modalités sont disponibles dans le jeu de données : RGB avec une résolution spatiale de 320×240 acquise à 24 images par seconde, et IR avec une résolution de 32×24 échantillonnée à 8 images par seconde. Bien que le RGB ne soit pas pertinent pour ce travail, nous l'avons utilisé pour obtenir les annotations des boîtes englobantes, comme décrit précédemment. Dans le cadre des contributions de ce travail, les annotations de localisation pour 110 vidéos sur les 126, pour les modalités RGB et IR, ont été crées et rendues publiques. Étant donné qu'il y a redondance entre les images voisines et que

notre application ne nécessite pas de données temporelles, nous avons échantillonné davantage le jeu de données IR pour obtenir un équivalent à 2 images par seconde.

Nous avons utilisé une approche semi-automatique pour obtenir les annotations des boîtes englobantes pour les images IR à basse résolution du jeu de données FIR-Image-Action. Tout d'abord, il a fallu créer manuellement des annotations de boîtes englobantes pour un sous-ensemble d'images RGB sélectionnées au hasard. Nous avons soigneusement annoté ces boîtes englobantes afin de réduire l'impact des erreurs de placement lors de la diminution de la résolution pour la modalité IR. Ensuite, un détecteur *SSD* (proposé par Liu *et al.* (2016)), noté h_ϑ, a été entraîné sur le jeu de données RGB annoté, puis utilisé pour obtenir des pseudo-annotations sur la partition non annotée restante du jeu de données RGB. Un nouveau sous-ensemble sélectionné au hasard est ensuite annoté et l'entraînement de h_ϑ est répété en utilisant une plus grande partition des données. Ce processus a été répété trois fois, résultant en une version entièrement annotée du jeu de données RGB.

Enfin, les annotations des boîtes englobantes pour le jeu de données IR sont obtenues en associant des images des deux modalités, suivies d'une procédure d'alignement des coordonnées décrite précédemment. Étant donné que les vidéos IR et RGB étaient désynchronisées, le décalage temporel initial a été déterminé manuellement en utilisant une superposition visuelle des deux modalités (voir figure 3.4 c). Une telle synchronisation a été effectuée individuellement pour chaque vidéo. Les annotations de localisation finales pour l'IR ont été obtenues en effectuant une interpolation linéaire des coordonnées des boîtes englobantes à partir du jeu de données RGB annoté. Un exemple d'annotation des boîtes englobantes obtenues pour les modalités RGB et IR peut être observé dans la figure 3.4 a et b.

3.2.2 Distech-Low-IR

Le deuxième jeu de données, appelé Distech-Low-IR, a suivi les mêmes proportions de séparation avec 1500 images pour l'entraînement, 500 pour la validation et 800 pour les tests. Ce jeu de données, similaire à FIR-Image-Action, contient deux modalités d'images (RGB et IR) avec

leurs annotations de boîtes englobantes correspondantes fournies par Distech Controls Inc. Le jeu de données reflète l'intérêt croissant de l'industrie pour les solutions de préservation de la confidentialité vis-à-vis de la localisation des personnes. Le jeu de données Distech-Low-IR a également montré qu'il reflétait mieux les scénarios réels, car il est composé de 7 pièces présentant différents niveaux de difficulté, tels que des appareils rayonnant de la chaleur, des fenêtres exposées au soleil et des cas avec plus d'une personne par pièce. Pour simuler le déploiement, nous avons utilisé des pièces qui n'ont pas été vues pendant l'entraînement pour l'ensemble de test. La figure 3.5 montre quelques exemples d'images provenant des deux jeux de données. Par exemple sur l'image figure 3.5 c, un four chaud est présent dans l'image ce qui représente un réel défi pour le différencier d'un humain.

Afin de garantir la reproductibilité de nos résultats, nous les avons répétés sur deux bases de données différentes. De plus, pour chaque expérimentation, trois répétitions ont été effectuées afin de mesurer la précision des mesures et d'en améliorer la robustesse. Pour chacune de ces répétitions, nous avons effectué une séparation des données en ensembles d'entraînement $\mathcal{D}_e$, de validation $\mathcal{D}_v$ et de test $\mathcal{D}_t$ différents. Ces séparations sont restées constantes pour toutes les méthodes évaluées. Nous avons rendu la base de donnée FIR-Image-Action publique ce qui permettra d'assurer une reproductibilité externe.

Dans la littérature plusieurs métriques sont présentes pour évaluer la détection. La plus utilisée mAP@0.5 :0.95 est basée uniquement sur des détecteurs mous qui proposent des boites associées à des scores de confiance. Cette métrique n'est pas utilisable notre cas, car plusieurs de nos méthodes sont de type durs. Dans cette étude, nous avons donc utilisé l'oLRP (voir équation 1.4, proposée par Oksuz *et al.* (2021)) afin de caractériser la capacité de chaque méthode à détecter la présence de personnes et à les localiser. Cette métrique nous permet d'évaluer les méthodes de mesure qui fournissent des détections de boîtes englobantes sans score associé (comme les AEs et les CAMs) et des méthodes avec un score de détection (comme les détecteurs SSD et Yolo v5). De plus, comme l'indique Oksuz *et al.* (2021), cette métrique reflète plus précisément la qualité de localisation que les autres mesures, en fournissant des mesures distinctes pour les différentes erreurs qu'une méthode de détection peut commettre. La métrique prend des valeurs entre 0 et

100, les valeurs plus basses étant les meilleures. Dans l'évaluation de la métrique, nous calculons également les composantes de localisation (oLRP$_{loc}$), de faux positifs (oLRP$_{FP}$) et de faux négatifs (oLRP$_{FN}$), qui fournissent des informations supplémentaires sur le comportement de la méthode évaluée. On pourra donc interroger ces différents termes afin d'accroître explicabilité de cette métrique.

$$LRP_{loc}(G,Y) = \frac{1}{N_{TP}} \times \left(\sum_{i=1}^{N_{TP}} 1 - lq(g_i, y_i) \right) \tag{3.1}$$

$$LRP_{FP}(G,Y) = 1 - \text{Précision} = \frac{N_{FP}}{|\mathcal{Y}|} \tag{3.2}$$

$$LRP_{FN}(G,Y) = 1 - \text{Rappel} = \frac{N_{FN}}{|\mathcal{G}|} \tag{3.3}$$

Enfin, le temps d'exécution sur le même matériel a été calculé pour obtenir une approximation de la complexité temporelle. Dans l'idéal, nous voulons diminuer l'erreur oLRP et réduire le temps de calcul. Comme la décision de la pondération de ces deux objectifs dépendra de l'application, nous nous sommes contentés de les reporter pour chacune des méthodes.

3.3 Résultats

Les méthodes que nous avons retenues pour cette étude se situent à différents niveaux de supervision. Le modèle seuil utilise le moins de supervision. Les modèles de détection d'anomalies de type auto-encodeur utilisent uniquement des images vides sans annotation. Au niveau supérieur, nous avons les CAM qui utilisent l'information binaire : si une personne est présente ou non dans l'image. Le niveau maximum est le détecteur Yolov5 et SSD qui utilise toutes les images et annotations d'entraînement.

Dans le tableau 3.1, la métrique oLRP nous est donnée avec que ses différents termes oLRP$_{loc}$, oLRP$_{FP}$ et oLRP$_{FN}$ ainsi que le temps d'inférence. Nous pouvons voir que les méthodes utilisant le plus de supervision demandent plus de ressources, cela peut s'expliquer par la complexité des méthodes. Des annotations plus complexes demandent au modèle une architecture capable

Tableau 3.1 Performance des méthodes de détection sur le jeu de données FIR-Image-Action. Toutes les métriques sont calculées avec un IoU de 50% sur 3 répétitions

Méthode	oLRP↓	oLRP$_{loc}$ ↓	oLRP$_{FP}$ ↓	oLRP$_{FN}$ ↓	Temps (ms)↓
Seuil	86.5 ± 0.1	32.3	45.3	44.2	0.4
Seuil d'Otsu	83.5 ± 0.1	31.6	45.7	27.7	0.7
dVAE	74.7 ± 1.3	31.2	26.6	24.5	13.0
dAE	77.4 ± 1.3	30.4	31.2	29.1	12.3
CAM	85.1 ± 1.1	34.3	41.2	29.0	11.4
GradCAM	85.5 ± 3.2	34.5	43.0	32.1	24.3
LayerCAM	84.8 ± 2.2	34.9	37.2	33.1	25.6
SSD	63.8 ± 2.7	**25.3**	12.6	18.6	46.6
Yolo v5	**56.9** ± 1.8	25.5	**6.3**	**6.2**	45.9

de les imiter et donc une architecture plus complexe. Il est donc normal qu'une méthode CAM qui classifie de manière binaire, ait une architecture plus simple qu'un détecteur. Cependant l'auto-encodeur utilise deux fois plus de couches de convolution dû à son architecture ; il reste quand même plus rapide que GradCAM et LayerCAM qui ont besoin de rétropropager le gradient de la décision. Ainsi, il est notable que le temps de détection dépendra du contenu des images. Cela est aussi vrai pour les méthodes avec plus de supervision Yolov5 et SSD qui utilisent l'algorithme NMS (Suppression non maximale) pour fusionner les boites.

Le tableau 3.1 nous montre que les méthodes utilisant plus de supervision, sont plus performantes. Cependant les méthodes auto-encodeurs performent mieux que les méthodes CAM alors que leurs niveaux de supervision sont supérieurs.

Dans le cas de la base de données Distech-Low-IR les résultats sont globalement moins bons : cela est dû à deux complexités qui n'étaient pas présentes dans les données FIR-Image-Action. La première est la complexité naturelle des données, la présence de four chaud, micro-onde, ainsi qu'une plus grande densité de personnes dans les images : jusqu'à 8 pour Distech-Low-IR contre 1 personne maximum dans le FIR-Image-Action. La seconde complexité réside dans la méthode de validation des résultats.Dans le cadre de Distech, nous entraînons le modèle sur

Tableau 3.2 Performance des méthodes de détection sur le jeu de données Distech-Low-IR. Toutes les métriques sont calculées avec un IoU de 50% sur 3 répétitions

Méthode	oLRP↓	oLRP$_{loc}$ ↓	oLRP$_{FP}$ ↓	oLRP$_{FN}$ ↓	Temps (ms)↓
Seuil	93.6 ± 2.4	37.1	72.7	54.1	0.4
Seuil d'Otsu	95.5 ± 1.2	34.2	83.3	50.0	0.7
dVAE	83.3 ± 8.9	33.2	32.7	40.6	13.0
dAE	82.7 ± 9.0	32.4	33.7	40.0	12.3
CAM	93.1 ± 1.9	37.6	59.7	52.3	11.4
GradCAM	91.6 ± 2.3	37.5	50.3	48.8	24.3
LayerCAM	91.1 ± 2.5	37.7	45.5	50.3	25.6
SSD	82.0 ± 7.2	31.1	**26.3**	44.7	46.6
Yolo v5	**80.2** ± 7.7	**30.2**	31.4	**37.4**	45.9

4 des 6 pièces puis on évalue le modèle dans les pièces qui n'ont pas été vues. Un problème de généralisation de domaine se cache derrière cette base de données. Cela explique aussi la plus grande incertitude des résultats : des pièces sont plus "simples" que d'autres ainsi en fonctions du découpage entre entraînements/validations et test l'ensemble des méthodes performent mieux ou moins bien. Il est possible de voir les deux bases de données comme la performance en domaine : développement du modèle dans le domaine connu et contrôlé, alors que dans la seconde situation le modèle est développé dans un environnent non contrôlé et plus complexe. Ce contraste intéressant entre ces deux bases de données nous montre par la même occasion qu'une méthode supervisée qui fonctionne convenablement dans un environnent, va plus souffrir en changeant d'environnement que des méthodes moins supervisées dans le cadre des auto-encodeurs. Numériquement l'écart absolu en domaine est de 31,4% entre les méthodes non supervisées et les méthodes entièrement supervisées, contre 3,8% hors domaine de contrôle. Ces résultats montrent la difficulté rencontrée durant l'élaboration de modèles dans le cadre de la collaboration avec l'entreprise Distech : les modèles performent très bien dans des espaces connus du modèle et les performances sont grandement altérées lors de sont développement hors distribution. La figure 3.7 nous montre les performances de plusieurs méthodes en fonction du seuil de déploiement sélectionné. Nous avons choisi de représenter la métrique de score F1

avec un IoU de 0.5 afin d'en faciliter la compréhension. Ce graphique nous renseigne sur la stabilité des méthodes vis-à-vis de la sélection du seuil pour leur déploiement hors distribution. Bien que la performance hors domaine chute, cette altération des performances n'est pas dû à un potentiel décalage du seuil optimal, ces courbes présentent des plateaux de performance autour des performances optimales. A noter que le seuil pour les détecteurs mous, est le seuil de confiance associé à chaque détection qui varie donc de 0 à 1. Pour les détecteurs durs, le seuil correspond à la valeur utilisée pour passer d'une carte de localisation à une segmentation binaire, nécessaire à l'élaboration de boites englobantes. Ainsi nous avons normalisé ces valeurs pour les représenter sur un même axe (voir figure 3.7) de manière à avoir un score F1@0.5 nul en 0 et 1. Il est cependant impossible de comparer ces courbes entre elles rigoureusement, car l'axe des abscisses représente par exemple une probabilité pour Yolo v5 et une température pour le dVAE. Ces deux mesures ne sont pas équivalentes.

La visualisation par un t-SNE (proposé par Van der Maaten & Hinton (2008)) en fonction des pièces permet de visualiser ce changement de distribution (voir figure 3.8). Cet algorithme permet de visualiser les données de grandes de dimensions dans un espace latent à la manière d'une analyse par composante principale. Cependant les transformations ne sont plus linéaires. Il est clairement identifiable dans cet espace latent, l'écart entre les différents domaines que représente les pièces. Par exemple la pièce n°4 est grandement différente de la pièce n°3 contrairement à la pièces n°6. Les distributions 3 et 4 sont complètement disjointes contrairement à 6 qui est incluse dans l'ensemble 4. Ainsi un modèle entraîné sur la pièce n°4 performera bien sur 6 contrairement à 3.

Cette visualisation permet de mettre en lumière le changement de distribution entre les différents environnements. Cependant des changements dans la distribution peuvent être observés au niveau des images d'entrée ainsi que dans la distribution du nombre de personnes. Bien que ces deux vecteurs sont grandement corrélés, le t-SNE repose uniquement sur le vecteur d'entrée : l'image. Pour évaluer plus précisément la différence entres les différents domaines, un autre schéma de validation est possible : laisser un domaine de coté. Pour chacune des 7 différentes

pièces, nous entraînons un modèle sur les 6 autres pièces et évaluons la performance sur la pièce qui n'a pas été vue durant l'entraînement.

Tableau 3.3 Résultats de la méthode GradCAM sur la validation avec un domaine de coté

F1@0.5 ↑	N° pièce hors entraînement						
N° pièce de test	1	2	3	4	5	6	7
1	0,286	0,246	0,469	0,684	0,265	0,218	0,327
2	0,229	0,205	0,242	0,338	0,226	0,223	0,234
3	0,607	0,545	0,464	0,688	0,546	0,491	0,645
4	0,403	0,320	0,591	0,456	0,405	0,358	0,321
5	0,427	0,328	0,453	0,650	0,363	0,407	0,385
6	0,349	0,332	0,470	0,650	0,356	0,313	0,359
7	0,276	0,341	0,386	0,584	0,400	0,334	0,419

Dans ce tableau 3.3 la méthode GradCAM a été utilisée pour montrer la différence entre les pièces. Pour chaque numéro de pièce hors entraînement, nous avons reporté le score F1 pour toutes les pièces. Ainsi l'ensemble de ces scores peut montrer des caractéristiques similaires ou opposées entre les images provenant de pièces différentes. Une manière d'analyser ce tableau est d'étudier les résultats hors domaine et en domaine pour chacune des pièces. La performance hors domaine est donnée par la diagonale (i=j dans le tableau 3.3) et la performance en domaine est la moyenne (pour i où j $\neq$ i dans le tableau 3.3).

Tableau 3.4 Comparaison des résultats de la méthode GradCAM en et hors domaine

N° pièce	résultats en domaine (ID)	résultats hors domaine (OD)	ID-OD
1	0,368	0,286	0,082
2	0,249	0,205	0,043
3	0,587	0,464	0,123
4	0,400	0,456	-0,057
5	0,442	0,363	0,078
6	0,419	0,313	0,107
7	0,387	0,419	-0,032

La différence de performance en domaine, versus hors domaine, dans le tableau 3.4, nous témoigne de l'écart de distribution entre la pièce en question et les autres pièces. Il est donc notable que la distribution la plus éloignée dans ce cas est la pièce n°3, ce qui correspond à la distribution la plus isolée dans l'espace latent du t-SNE figure 3.8. De plus les valeurs sont négatives pour les pièces 4 et 7 ; le gain de performance impliqué par la généralisation sur d'autres domaines surpasse le faible changement de distribution. A noter que les distributions 4 et 7 sont centrales dans la figure 3.8 et partagent beaucoup de caractéristiques avec les autres pièces. A contrario, la pièce n°6 est placée au centre dans la figure 3.8 mais présente un grand écart entre performance en domaine et hors domaine. Cela peut s'expliquer par un changement de distribution, non pas dans l'image, mais dans les annotations. Ceci qui est uniquement mis en lumière dans cette approche plus quantitative. En complément, cette approche nous permet d'affirmer que la pièce n°2 est la plus compliquée des pièces, mais la complexité n'est pas due au changement de domaine.

3.4 Conclusion

Dans ce chapitre nous avons étudié le niveau de supervision dans le cadre de la détection de personnes en basse résolution IR. Dans le cas d'un environnement connu, les méthodes supervisées sont les plus performances bien qu'elles utilisent le plus de ressources de calcul en mode inférence. Cependant dans le cas d'un environnement inconnu, les résultats sont largement dégradés. Les méthodes utilisant des AE semblent moins souffrir de cette dégradation. Selon nous, cela est dû à l'apprentissage des caractéristiques plus intrinsèques aux images par rapport à l'apprentissage supervisé qui est guidé pas les annotations. Ainsi les performances des AE sont très proches des performances des détecteurs. Sachant qu'ils ont été uniquement entraînés sur les images vides, soit 30% des données : cela peut remettre en question le niveau d'annotation considérant leur coût élevé.

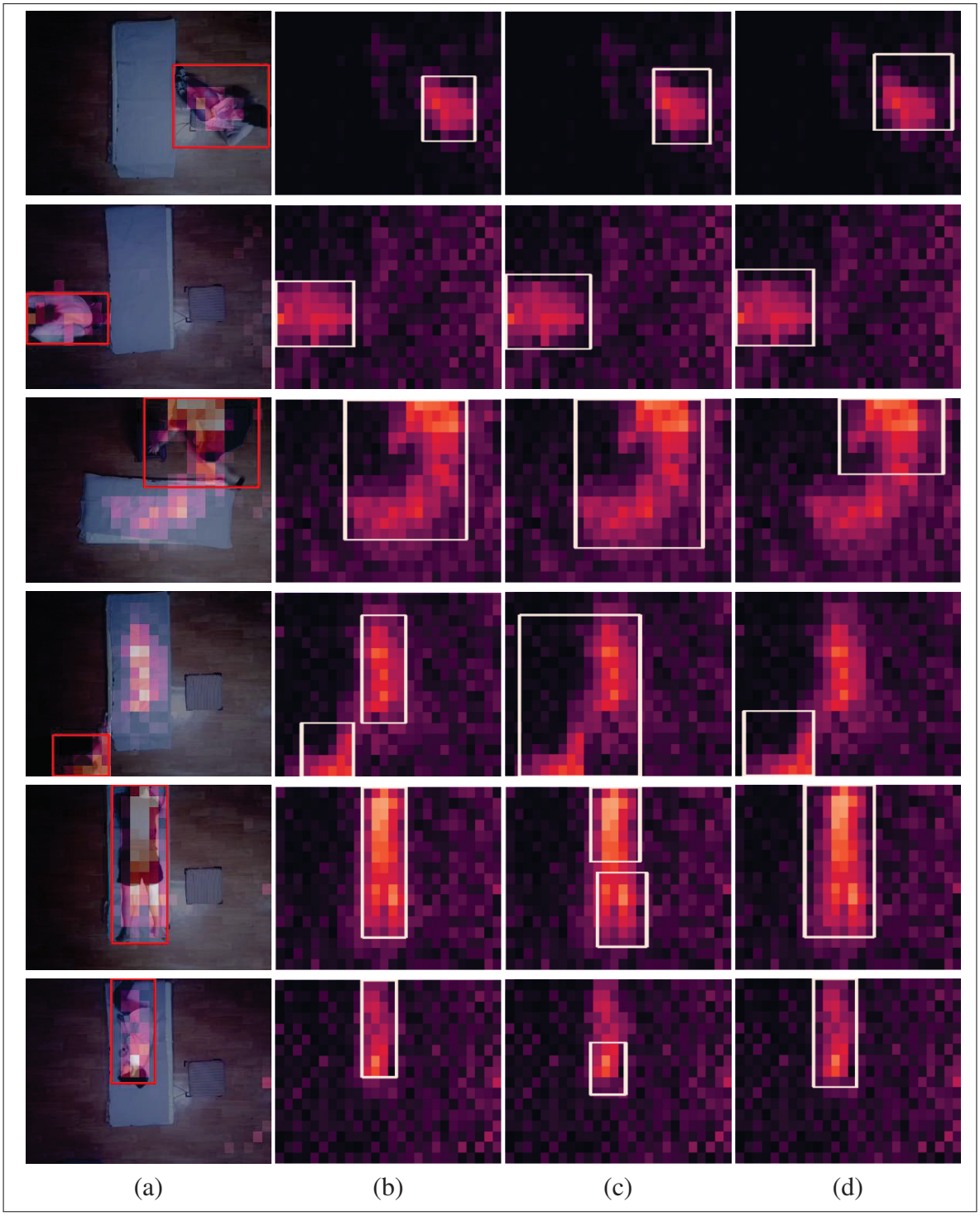


Figure 3.6 Exemples de résultats de détection de personnes en IR à basse résolution. Superposition des modalités RGB et IR avec leurs annotations correspondantes (a), et les prédictions des boîtes englobantes de *dVAE* (b), *gradCAM* (c) et *Yolo v5* (d)

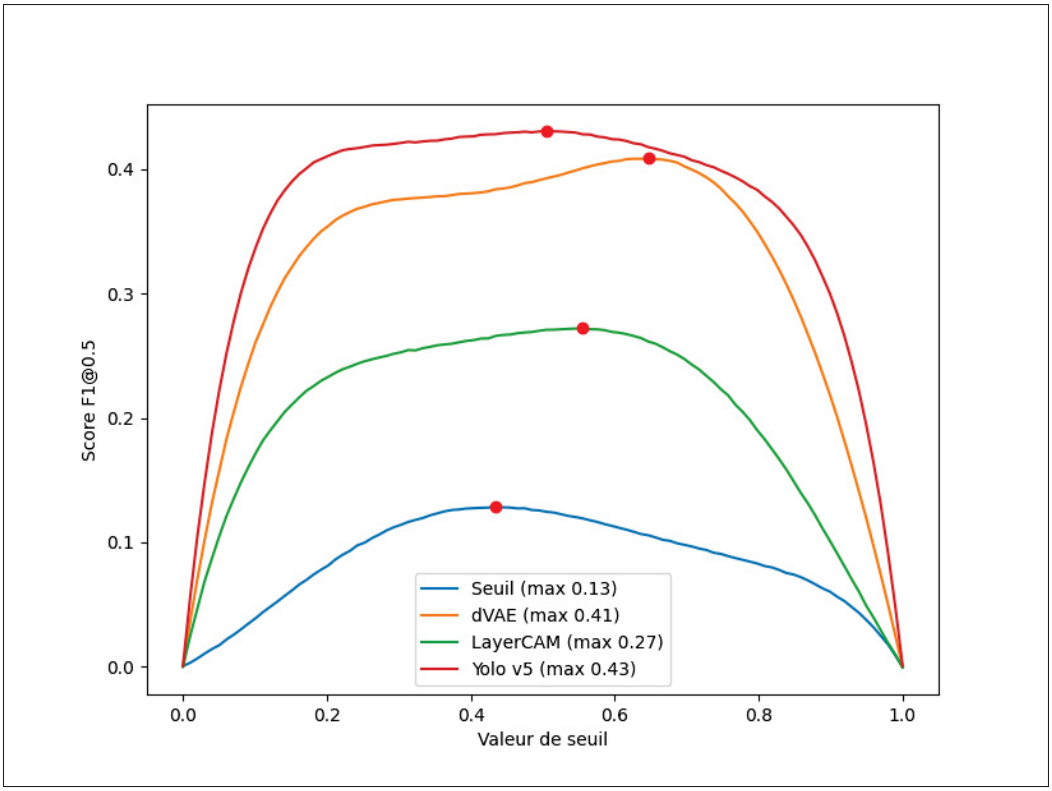


Figure 3.7 Sensibilité de la valeur de seuil pour un déploiement hors domaine sur la base de donnée Distech-Low-IR

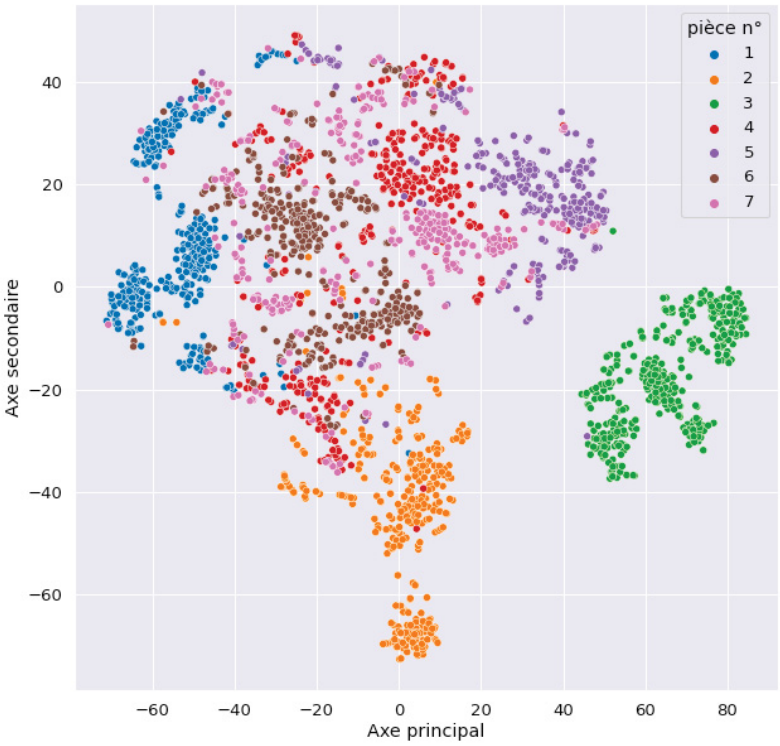


Figure 3.8 Représentation des images de Distech-Low-IR dans l'espace latent d'un t-SNE en fonction de leur provenance

CHAPITRE 4

ÉTUDE DE LA TEMPORALITÉ

Dans le chapitre précédent nous avons étudié le niveau de supervision pour la détection. Les modèles AE utilisant un principe de suppression du fond, généralisent mieux que les détecteurs dans un environnement nouveau. Cependant en ajoutant du contexte temporel, les méthodes qui estiment le fond de l'image peuvent être remplacées par une image enregistrée en mémoire à la place d'un modèle AE. Cela rend ces approches peu utilisées en pratique. Un modèle probabiliste de mélange gaussien proposé par Lee *et al.* (2003) utilise les données de test pour estimer le fond, ce qui permet une plus grande généralisation. Ainsi dans cette partie nous avons étudié les cas où nous avons accès à un clip vidéo réalisé à partir de quelques images successives. En pratique ces images peuvent être enregistrées sur le système embarqué dans une file mémoire. Dans un premier, temps nous avons étudié la reconnaissance d'actions étant donné qu'elle intègre la temporalité sur des modèles plus simples que ceux pour la détection. De plus, la littérature proche de notre cas d'étude, est uniquement focalisée sur la détection d'actions. Bien que notre projet avec Distech Controls n'a pas pour but de reconnaître les actions, cette nouvelle capacité peut être une motivation long terme pour de potentiels améliorations du produit. Ce travail permet aussi de compléter notre contribution sur la base de données FIR-Image-Action en évaluant tous ses aspects. Dans un second temps, nous avons utilisé le contexte temporel dans le modèle de détection pour en améliorer la performance.

4.1 Reconnaissance d'actions

4.1.1 Introduction à la reconnaissance d'actions

La reconnaissance d'actions, également connue sous le nom de détection d'actions, est un domaine de recherche en vision par ordinateur qui vise à identifier et à comprendre les actions ou les activités humaines dans des séquences d'images ou de vidéos. Cette technologie trouve de nombreuses applications pratiques, notamment dans les domaines de la sécurité des personnes, de la surveillance, de l'analyse du comportement et de l'interaction homme-machine.

La détection d'actions joue un rôle crucial dans la sécurité des personnes, en particulier dans les environnements sensibles tels que les zones publiques, les bâtiments, les installations industrielles et les espaces urbains. En détectant et en reconnaissant les actions des individus, il devient possible de prévenir et de réagir rapidement aux situations dangereuses ou suspectes, contribuant ainsi à la prévention des incidents et à la protection des personnes. Les avantages de la détection d'actions dans le domaine de la sécurité sont nombreux. Tout d'abord, elle permet une surveillance proactive en identifiant les comportements anormaux ou les actions potentiellement menaçantes, tels que l'intrusion, le vol, la violence, ou les comportements suspects. Cela permet aux systèmes de sécurité d'alerter rapidement les autorités compétentes et de prendre les mesures appropriées pour garantir la sécurité des personnes.

Bien que la détection d'actions présente beaucoup d'enjeux, elle présente aussi plusieurs défis. Les actions humaines peuvent être très variées et complexes, ce qui rend la tâche de détection et de reconnaissance difficile. De plus, les performances des systèmes de détection d'actions peuvent être affectées par des facteurs tels que les variations d'éclairage, les angles de vue, les occlusions et le bruit des données. Il est donc essentiel de développer des méthodes robustes et efficaces pour relever ces défis et améliorer les performances des systèmes de détection d'actions.

Dans cette étude nous avons analysé la détection d'actions comme étude de transition entre le niveau de supervision et l'ajout de la temporalité pour améliorer la détection. En effet, le problème de classification est plus simple en théorie que la détection. De plus les travaux antérieurs proposés en basse résolution infrarouge sont également concentrés sur ce sujet.

4.1.2 Méthode proposée et adaptation des méthodes aux images de basse résolution

En se basant sur l'étude précédente, nous avons utilisé la même architecture que proposée dans notre étude sur la supervision. L'architecture est légère compte tenu de la basse résolution 24x32 pixels. Ainsi nous avions utilisé des blocs composés de 2 convolutions (de noyau 3x3x3) avec une connexion résiduelle (amélioration proposée par He *et al.* (2016) voir figure 4.1) suivie d'une couche de regroupement maximum (de noyau 2x2x2). Ce schéma a été répété 2 fois

avant d'aplatir toutes les caractéristiques de la séquence, suivi d'une couche de classification dense. A noter que les entrées sont des séquences de N images. Ainsi les convolutions et les regroupements sont des opérations en 3 dimensions, contrairement à l'étude précédente en 2 dimensions. L'activation de type $ReLu(x) = max(0, x)$ a été utilisée comme non linéarité au sein du réseau de neurones ainsi que la fonction $Softmax$ en fin de classification pour transformer les activations en probabilités. L'architecture que nous avons proposée peut être apparentée à celle proposée par Song *et al.* (2018) adaptée pour une résolution plus basse ainsi qu'une utilisation avec un seul canal de couleur pour l'infrarouge. Nous avons appelé cette architecture 3D-CNN-Résiduel (voir figure 4.2).

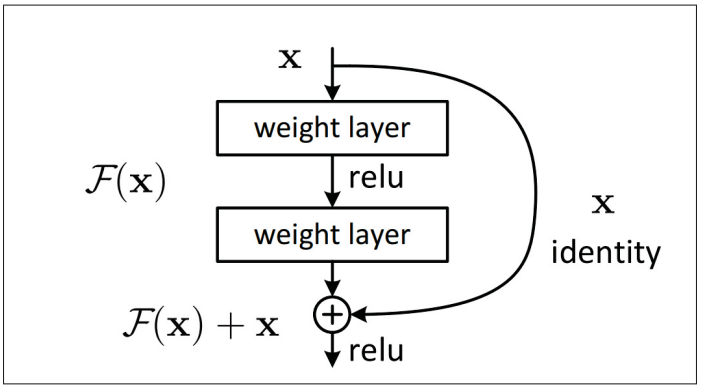


Figure 4.1 Bloc avec connexions résiduelles tiré de He *et al.* (2016). Ce bloc étant utilisé dans le ResNet avec des convolutions 2D, nous l'utilisons avec des convolutions 3D

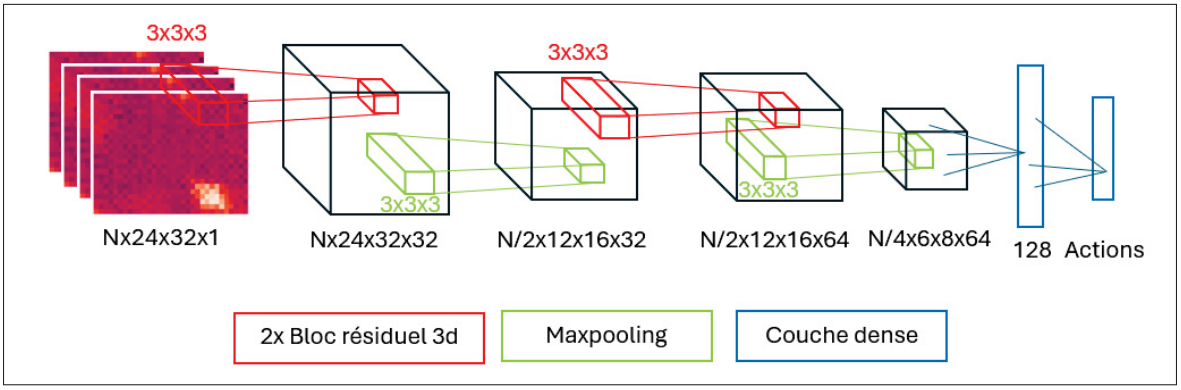


Figure 4.2 Architecture 3D-CNN-Résiduel inspirée de Song *et al.* (2018) que nous avons adaptée pour l'utilisation d'images infrarouge et basse résolution

Plusieurs autres techniques ont aussi été adaptées comme la méthode proposée par Ullah *et al.* (2019) qui utilise une architecture d'auto-encodeur suivie d'un SVM. En effet nous avons adapté l'architecture proposée pour permettre son utilisation avec des images infrarouges en basse résolution. Comme la méthode 3D-CNN proposée par Tao *et al.* (2019) et Tateno *et al.* (2020a) permet déjà une utilisations de ces images, nous donc réalisé un auto-encodeur en utilisant leur 3D-CNN comme encodeur (voir figure 4.3). Nous l'avons symétrisé pour réaliser un décodeur (voir figure 4.4) en inversant les opérations de Maxpooling par des opérations de duplication des caractéristiques dans toutes les directions.

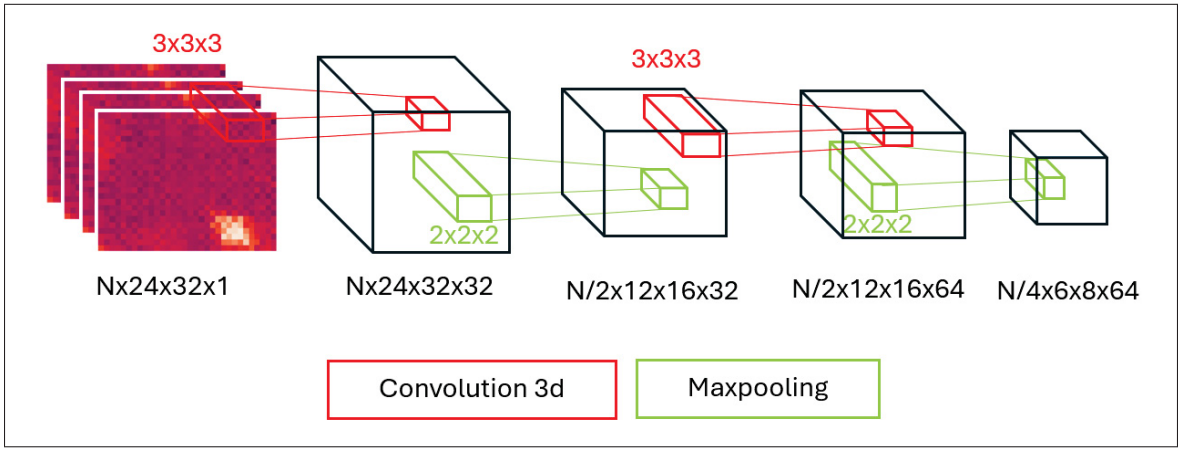


Figure 4.3 Architecture proposée par Tao *et al.* (2019) et Tateno *et al.* (2020a). Nous l'avons utilisé comme encodeur dans notre auto-encodeur

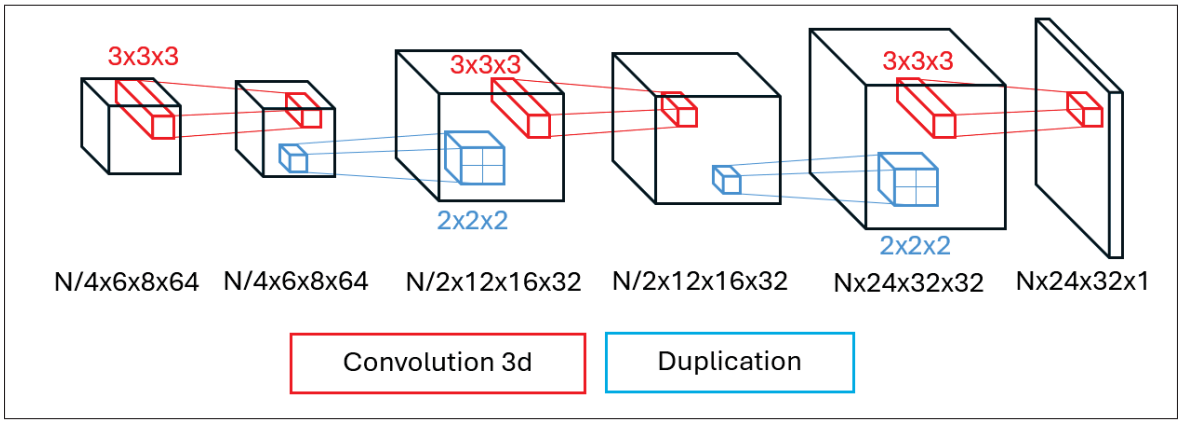


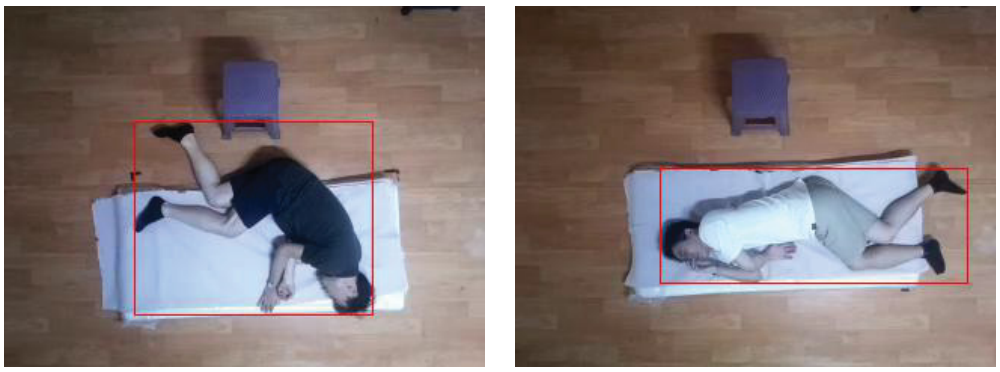
Figure 4.4 Architecture du décodeur, symétrique de l'encodeur, que nous avons réalisé pour adaptée la méthode proposée par Ullah *et al.* (2019) à nos images

Nous avons aussi adapté la méthode proposée par Corsel *et al.* (2023). Celle-ci propose d'utiliser le contexte temporel afin d'améliorer la détection. En l'appliquant dans cette expérience, nous avons voulu évaluer si son utilisation peut être généralisable sur d'autres tâches comme ici dans le cadre de la reconnaissance d'actions. Bien que Corsel *et al.* (2023) ait prouvé qu'on peut utiliser les différents canaux de couleurs pour incorporer le contexte temporel à défaut de retirer l'information chromatique de l'image. Dans notre cas, en infrarouge, aucune information n'est retirée. Cependant le modèle propose une détection basée sur 3 images. Ainsi pour des séquences où $N > 3$, nous avons utilisé la première, la dernière ainsi que l'image centrale pour imiter au mieux sa méthode. Par conséquent, l'information n'est pas dense comme dans les autres méthodes mais plutôt lacunaire. A noter que Corsel et al. a nommé sa méthode TYolo. Cependant nous avons dû adapter l'architecture pour qu'elle soit compatible avec nos données. Ainsi nous avons appelé cette méthode T-CNN. Le T-CNN utilise donc la même architecture pour l'extraction des caractéristiques que le 2D-CNN+LSTM proposé par Tateno *et al.* (2020b). La tête a, quant à elle, été remplacée par une couche dense.

4.1.3 FIR-Image-Action

La base de données FIR-Image-Action Zhang (2020) est une base de données pour la reconnaissance d'actions. Nous l'avons annotée d'un point de vue localisation pour réaliser notre première étude. A présent nous utiliserons les annotations d'origine proposées par Haoyu Zhang de Visiongo Inc. Pour chaque image une classe décrit l'action que la personne réalise. L'objectif est à partir de plusieurs images, de classifier l'action réalisée par l'individu. Les actions étant : assis, debout, couché, en marche, en chute, tombé, endormi, se lève et s'assoit (voir répartition figure 4.6). Ces différentes actions sont un vrai défi car la temporalité de l'information est nécessaire par exemple pour discriminer une personne qui s'est allongée d'une personne tombée allongée au sol. L'accès à cette information peut être déterminante par exemple dans le cadre d'une alerte pour les secours.

Le jeu de données FIR-Image-Action Zhang (2020) comprend 110 vidéos annotées. Nous avons sélectionné au hasard 36 vidéos de cet ensemble pour l'ensemble de test $\mathcal{D}_t$ et les 74 autres pour



(a) (b)

Figure 4.5 Exemple d'une personne tombée (a) et d'une personne allongée en (b). Sans information temporelle, il nous est très difficile de différencier ces deux cas

l'entraînement $\mathcal{D}_e$ et la validation $\mathcal{D}_v$. De plus, les ensembles d'entraînement et de validation ont été séparés en utilisant une sélection aléatoire des images (70% et 30%, respectivement). Toutes les approches ont été entraînées en utilisant la même partition des données pour assurer la comparabilité des résultats.

À notre connaissance, les données proposée par Haoyu Zhang de Visiongo Inc. pour la reconnaissance d'actions basée sur la vidéo est une des seules bases publiques dans le domaine de l'infrarouge de très basse résolution.

Pour les images utilisées sous forme de paquets d'images, deux paramètres sont le plus souvent employés : N le nombre d'images et P le pas entre chaque exemple (voir figure 4.7). Pour chaque séquence de N images d'entrée nous voulons prédire l'annotation correspondante à la dernière image. Ainsi le paramètre N est un hyperparamètre qui peut être optimisé en faisant une boucle entre entraînement et validation.

Le paramètre P est fixé à 1 pour l'espace de test cependant pour l'espace d'entraînement il peut être vu comme un compromis : P proche de 1 va permettre plus d'exemples d'entraînement. Inversement, lorsque le recouvrement ($N - P$) entre exemples augmente, les chances de sur-

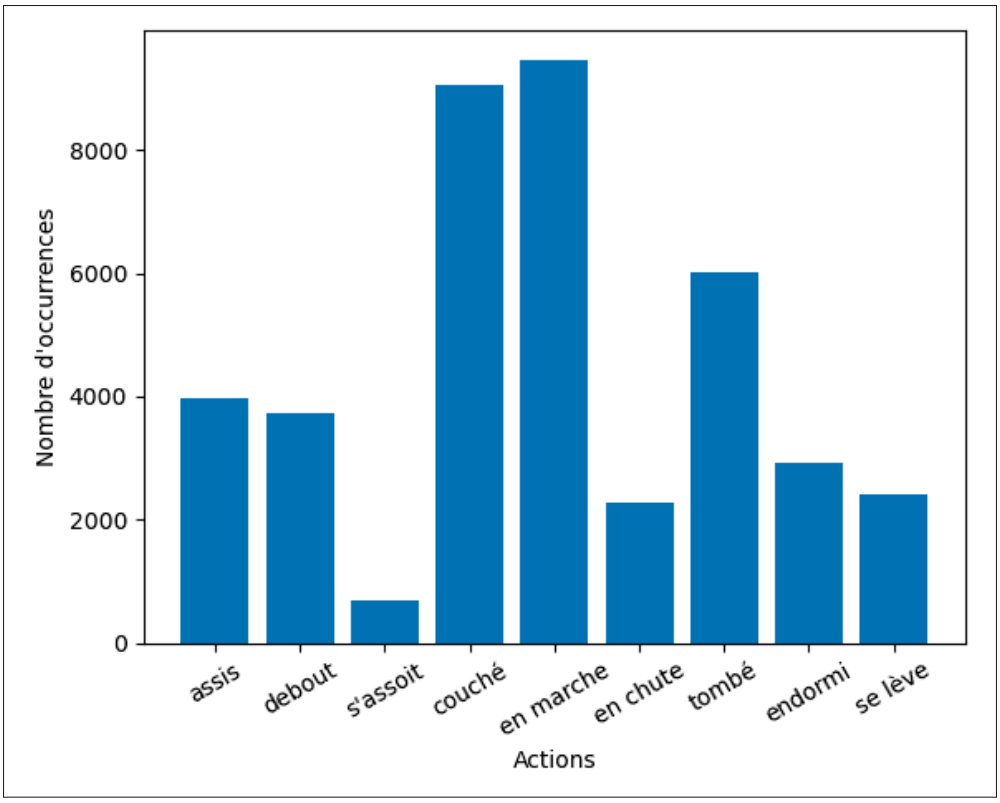


Figure 4.6 Répartition des actions dans la base de données FIR-Image-Action

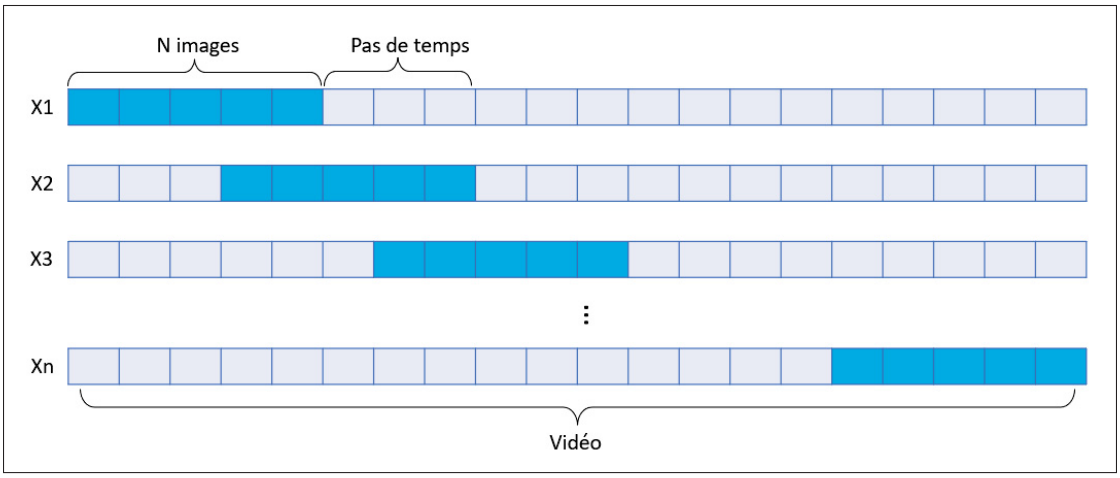


Figure 4.7 *N* nombre d'images d'entrée de modèle. P le pas de temps entre les images

apprentissage augmentent. Ainsi pour chaque méthode nous avons traité N et P comme des hyperparamètres à optimiser.

Pour chaque expérimentation, nous avons effectué trois répétitions afin de évaluer la précision des mesures et ainsi d'améliorer la robustesse. Pour chacune de ces répétitions, nous avons effectué une séparation des données en ensembles d'entraînement $\mathcal{D}_e$ (70%), de validation $\mathcal{D}_v$ (10%) et de test $\mathcal{D}_t$ (20%) différents, mais elles sont restées constantes pour toutes les méthodes évaluées. L'ensemble des résultats ont aussi été comparés aux résultats des approches compétitives. Nous avons uniquement analysé les tendances globales car les bases de données utilisées par les compétiteurs ne sont pas publiques.

La métrique utilisée dans cette étude de la classifications des actions est le score F1w (voir équation 1.2). Les méthodes concurrentes proposées par Tao *et al.* (2019) et Tateno *et al.* (2020a) on été évaluées en utilisant le taux d'exactitude et ne sont donc pas comparable avec notre étude. Leurs données non publiques semblent être balancées ce qui peut justifier le choix de cette métrique.

4.1.4 Résultats

Les expériences que nous avons réalisées nous mènent aux mêmes conclusions exprimées par Tao *et al.* (2019) et Tateno *et al.* (2020a) : l'architecture la plus performante pour classifier les actions avec des données sous forme de séquences, est une fusion de l'information en début de modèle avec des convolutions 3d comme dans le modèle 3D-CNN et 3D-Résidual. Contrairement au 2D-CNN + LSTM qui extrait les caractéristiques de l'image puis les fusionne sur la dimension temporelle avec des cellules LSTM (mémoire long et court terme). Nous avons aussi rajouté le modèle de classification linéaire (ainsi que SVM, RF et K-NN) afin de réaliser une échelle de performance, comme initialement réalisée par Karayaneva *et al.* (2018). Il est surprenant que le modèle 2D-CNN + LSTM performe moins bien que la régression linéaire. Tateno *et al.* (2020a) compare uniquement les modèles 2D-CNN + LSTM avec les modèles

3D-CNN ainsi les mauvaises performances proposées par l'utilisation du LSTM sont difficiles à relever.

Tableau 4.1 Performance des méthodes de classification d'actions sur le jeu de données FIR-Image-Action sur 3 répétitions

Méthode	F1w ↑	Paramètres (m)	Temps (ms)↓
2D-CNN (sans temporalité)	0.477 ± 0.022	0.315	1.39
2D-CNN + LSTM	0.479 ± 0.071	0.315	2.20
K-NN	0.503 ± 0.029	0	325
Régression linéaire	0.512 ± 0.011	0.037	0.17
T-CNN	0.520 ± 0.016	0.315	1.39
3D-AE+SVM	0.545 ± 0.010	0.702	12.7
RF	0.547 ± 0.033	2e-4	69.2
SVM	0.603 ± 0.023	0.032	107
3D-CNN	0.666 ± 0.022	0.770	152
3D-CNN-Résiduel	**0.684** ± 0.015	1.127	159

Dans notre étude, un maximum de F1w est atteint pour un $N = 17$ (voir figure 4.8) dans notre ensemble de validation. Ainsi celui-ci a été choisi pour évaluer le modèle sur l'ensemble de test $\mathcal{D}_t$. Le réel maximum sur l'ensemble de test est atteint en $N = 18$ ce qui est plutôt proche de nos résultats. L'étude menée par Tateno *et al.* (2020a) et Tao *et al.* (2019) affirme avoir un maximum atteint pour une valeur de $N = 15$, une valeur relativement proche de nos résultats compte tenu de la base de données qui est différente et d'une métrique différente également.

Avec cette figure 4.8 nous pouvons également visualiser la contribution de la temporalité en comparant le cas $N = 1$ avec le cas $N = 17$. Dans le cas de notre approche avec $N = 1$ nous avons un score F1w de 51,9 % contre 68,4 % à $N = 17$, ainsi l'ajout de contexte temporel nous donne un gain net de 16,5 % en absolu, amélioration significative en terme de F1w. La principale différence entre notre architecture et celle proposée par Tao *et al.* (2019) est l'utilisation de connexions résiduelles. Plusieurs hypothèses sont formulées dans la littérature pour expliquer ces résultats mais la plus commune est une meilleure circulation du gradient à travers l'optimisation par rétropropagration. Cette hypothèse est aussi confortée par les mauvais résultats obtenus par la fusion de l'information temporelle réalisée à la fin par le LSTM. Les LSTM souffrent

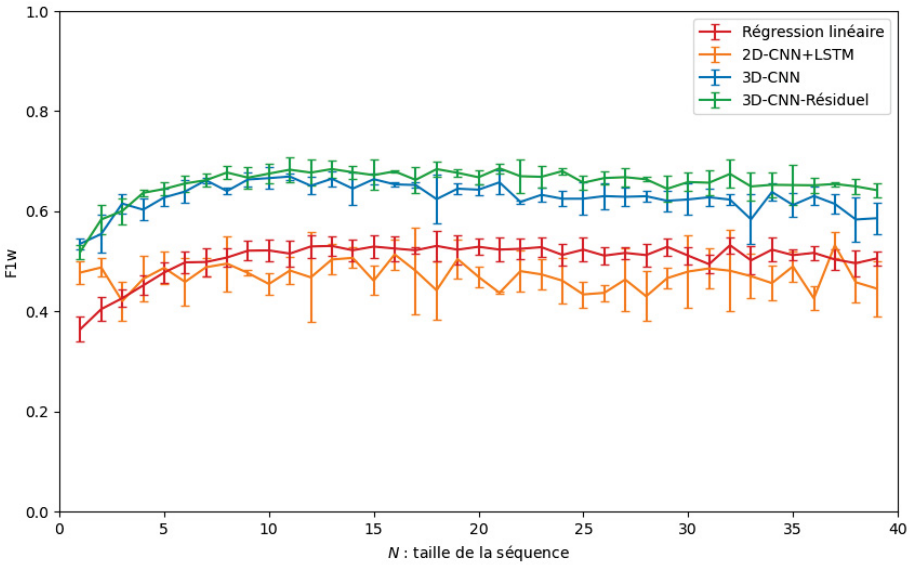


Figure 4.8 Évolution du F1w sur l'ensemble de test en fonction de N la taille de séquence d'entrée. P est fixé à l'optimal pour l'ensemble de validation soit 1

d'une mauvaise circulation des gradients ce qui a encouragé le développement des récents transformers.

Nous avions émis l'hypothèse que le paramètre P présentait un compromis entre quantité de données et sur-apprentissage. Cependant le graphique figure 4.9 montre pour $N = 17$ optimal l'évolution des résultats pour différents pas de temps. L'influence est très légère et F1w décroît avec N grandissant. Ainsi nous n'avons pas de problème de sur-apprentissage et nous avons toujours utilisé $P = 1$ pour maximiser le nombre de données disponibles.

Bien que les résultats de la méthode T-CNN ne sont pas prometteurs comparées aux méthodes utilisant une information dense, ces résultats présentent une amélioration du 2d-CNN qui n'utilise pas la temporalité. Étant donné du niveau de supervision égal, un nombre de paramètres égal ainsi qu'un temps d'inférence égal, cette méthode permet d'améliorer les résultats d'une approche classique sans pour autant la complexifier. Celle-ci n'est cependant pas la meilleure méthode que nous avons testée.

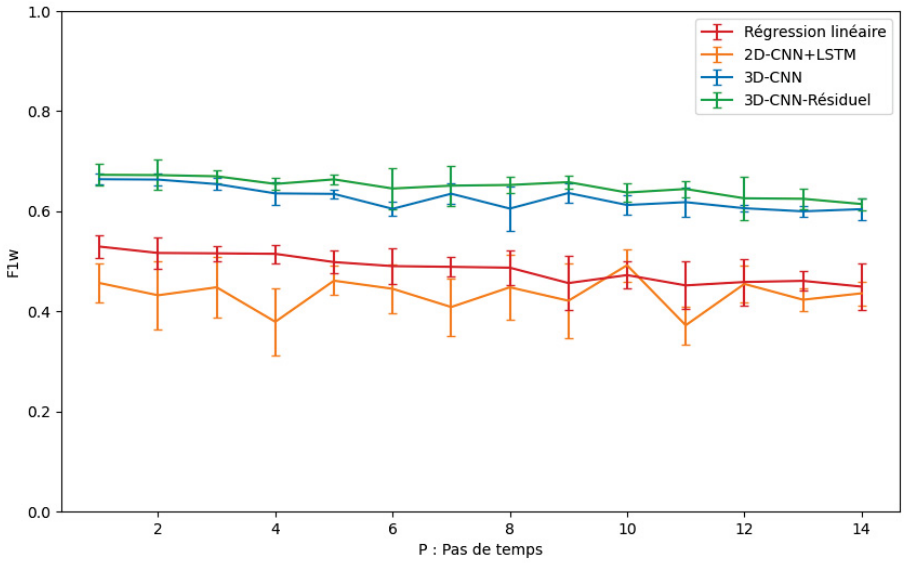


Figure 4.9 Évolution du F1w sur l'ensemble de test en fonction de P le pas de temps. N est fixé à l'optimal pour l'ensemble de validation soit 17

4.1.5 Conclusion

Dans cette partie nous avons étudié la classification d'actions sur une base de données en infrarouge basse résolution. Nous avons proposé une nouvelle architecture qui surpasse les modèles proposés dans la littérature en améliorant de 1.8 % le F1 score pondéré sur une base de donnée publique FIR-Image-Action par Zhang (2020). Cette amélioration propose une nouvelle solution plus robuste pour la détection d'actions en offrant ainsi plus d'applications disponibles avec un dispositif à bas coût. L'utilisation de la basse résolution permet par la même occasion de garder l'anonymat des personnes et ainsi facilite un déploiement de plus grande échelle de systèmes de sécurité comme par exemple la détection de chutes dans les hôpitaux ou maisons de retraites. Ce type d'images nous permet aussi de réduire la quantité de calculs nécessaires à la détection comparée à une image en haute résolution. Ainsi, cette méthode constitue une parfaite candidate pour réduire la puissance utilisée et gérer un bâtiment de manière intelligente.

4.2 Mise en valeur du contexte temporel

Lorsque l'on aborde la détection d'objets ou de personnes dans une image, les détecteurs sont généralement entraînés à extraire des caractéristiques géométriques spécifiques. Il est important de reconnaître que ces informations proviennent grandement de la corrélation entre les pixels de l'image, d'où l'utilisation courante des couches de convolutions pour extraire des caractéristiques. Ainsi, le contexte dans lequel se situent les pixels de l'image joue un rôle crucial dans la détection des objets.

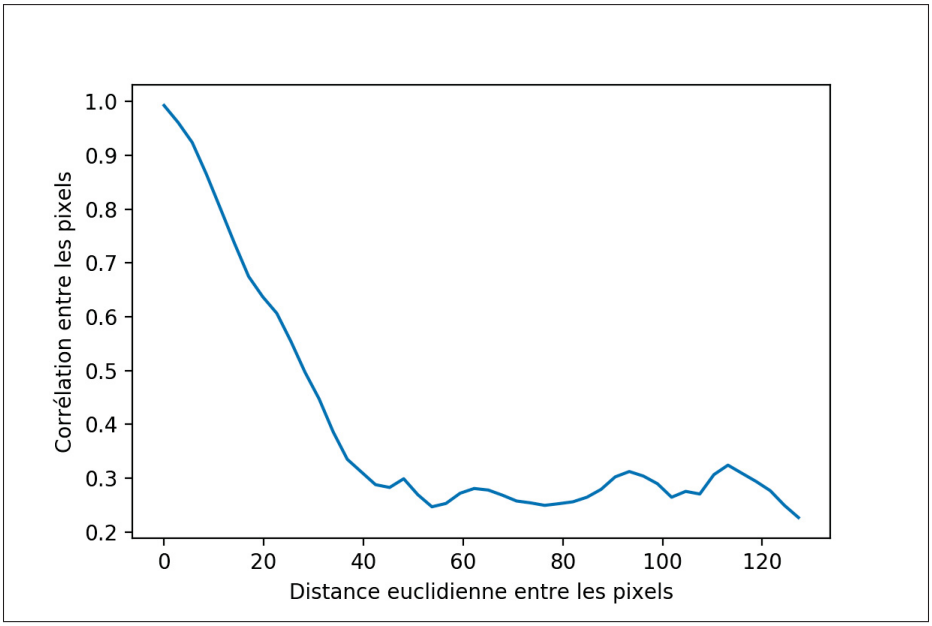


Figure 4.10 Relation de contexte géométrique

De manière similaire, nous pouvons considérer l'importance du contexte temporel dans l'analyse vidéo. Les caractéristiques visuelles d'une vidéo évoluent au fil du temps, il est donc naturel de constater une corrélation temporelle au niveau des pixels. En intégrant ce contexte temporel dans un modèle de détection, nous avons la possibilité d'améliorer les résultats sans avoir à modifier sa fonction ou à augmenter le nombre d'annotations requises.

Il convient de souligner que l'ajout du contexte temporel est particulièrement utile dans le cas d'objets en mouvement. Bien que cette tâche puisse être confondue avec le suivi d'objets,

notre intérêt se concentre exclusivement sur l'amélioration de la détection. Toute amélioration apportée dans le cadre du post-traitement des détections peut être utilisée de la même manière dans notre approche, renforçant ainsi l'efficacité et la précision de la détection des personnes.

En incorporant le contexte temporel dans notre modèle, nous ouvrons de nouvelles perspectives pour une détection plus robuste et précise, offrant ainsi des avantages significatifs dans des domaines tels que la sécurité, la surveillance, la détection d'urgences et bien d'autres. Comme nous l'avons observé dans notre première étude, les performances des différentes méthodes sont grandement impactées par le déploiement dans un environnent inconnu. Les méthodes utilisant la temporalité souffrent d'une mauvaise circulation des gradients, par exemple le 2D-CNN+LSTM dans notre seconde étude. Nous nous sommes donc intéressés à la valorisation du contexte temporel selon deux axes : d'une part par l'augmentation de données pour en améliorer la généralisation ; d'autre part par l'utilisation d'une architecture permettant une meilleure circulation du gradient. Le contexte temporel peut aider le modèle à généraliser à des environnements inconnus. De plus, l'augmentation de données est une méthode cruciale et largement utilisée pour augmenter la capacité de généralisation d'un modèle. A noter que d'autres méthodes de détection utilisant le contexte temporel existent comme la méthode proposée par Liu & Zhu (2018), celle-ci se base sur un grand nombre d'images d'entrées et utilise ainsi beaucoup de contexte temporel. La complexité algorithmique de ces méthodes est donc grandement augmentée ce qui les rend inutilisables sur un système embraqué. De plus, notre caractérisation de la temporalité optimale pour la détection nous oriente vers l'adoption d'un contexte plus restreint, déconseillant ainsi l'usage de méthodes telles que les 3D-CNN ou les 2D-CNN associés à des LSTM. C'est pourquoi nous nous somme concentrés sur la méthode proposée par Corsel *et al.* (2023).

Dans ce travail nous proposons plusieurs augmentations de données qui jouent sur l'aspect temporel des données. Nous avons déjà vu plusieurs méthodes concurrentes proposées par Duran-Vega *et al.* (2021) : la mosaïque temporelle, le mixage temporel, le flou, la suppression de régions, et l'ajout de bruit. Dans cette étude nous avons employé la méthode proposée par Corsel *et al.* (2023) car peu d'augmentations de données ont été utilisées dans son étude. Cependant

toute méthode utilisant du contexte pour la détection, peut bénéficier de nos contributions. Par ailleurs, la méthode proposée par Corsel *et al.* (2023) a été pensée pour des images RGB. Elle augmente le contexte temporel à défaut de l'information chromatique. Dans notre cas, en IR, nous avons un seul canal d'acquisition ce qui permet uniquement d'ajouter du contexte temporel.

4.2.1 Méthode proposée

Nous proposons deux augmentations de données jouant sur l'aspect temporel des données. L'une utile pour le pré-apprentissage sur une base de données d'images classiques ; l'autre pour la période d'ajustement des poids avec des clips vidéos.

Comme proposé par Corsel *et al.* (2023) pour le pré-entraînement, nous transformons l'image $x_{RGB} \in \mathbb{R}^{W \times H \times 3}$ en nuances de gris x_G. Cette image est ensuite déplacée de manière aléatoire pour simuler le mouvement (voir figure 4.11). Afin de réaliser cette opération nous avons utilisé les opérations de translations, rotations, parallaxe, et zoom, qui sont réalisables en parallèle à l'aide d'une matrice de transformation homographique en 2D. Soit p un pixel de coordonnées $(i, j) \in \mathbb{N}^2$. En géométrie projective, ces coordonnées peuvent être reformulées en coordonnées homogènes $[i\ j\ 1]^T$ une multiplication avec la matrice de transformation T (voir équation 4.1) permet d'obtenir les nouvelles coordonnées $[i'\ j'\ w']^T$. Celle-ci nous donne accès à la nouvelle localisation du pixel p en $[i'/w', j'/w']^T$.

$$T(t_x, t_y, s_z, s_x, s_y, \theta) = \begin{bmatrix} 1+s_z & 0 & t_x \\ 0 & 1+s_z & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \tan(s_x\pi) & 0 \\ \tan(s_y\pi) & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4.1)$$

L'ensemble des paramètres (t_x,t_y,s_z,s_x,s_y,θ) est issu d'une distribution uniforme $\mathcal{U}(-s, s)$ contrôlée par le paramètre s, seul hyperparamètre de notre augmentation de données. A noter que la sélection des poids du modèle ainsi entraîné, sera lui aussi réalisé avec des images qui ont subit cette transformation, sinon la performance du modèle décroîtra durant l'optimisation.

Figure 4.11 Proposition d'augmentation de données pour le pré-entraînement

Nous proposons une seconde augmentation de données qui est dédiée à l'apprentissage avec les données sous format temporel. Le principe est le même que la première augmentation sauf que dans ce cas, nous avons déjà 3 images différentes qui ont une cohérence temporelle. Notre augmentation de données consiste en un décalage relatif de l'image antérieure et postérieure (voir figure 4.12). Dans ce cas, la sélection du modèle est réalisée sur l'ensemble de validation sans transformation.

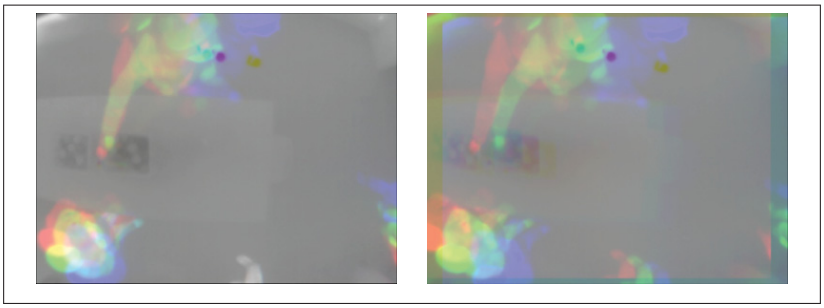


Figure 4.12 Proposition d'augmentation de données pour l'entraînement

En utilisant des clips vidéos, le modèle est amené à faire plus de sur-apprentissage, compte tenu de la complexité de l'entrée par rapport à une détection classique. Nos augmentations de données se basent alors sur la création de trajectoires aléatoires afin de rendre le modèle plus performant. Bien qu'il est souvent possible d'avoir les annotations pour l'image précédente et suivante, nous avons préféré utiliser uniquement l'annotation de l'image actuelle ce qui peut être utilisé dans tous les cas. A noter que les paramètres de mouvements relatifs des images peuvent être changés avec l'hyperparamètre s. Dans nos expérimentations avons utilisé $s = 5\%$.

4.2.2 Distech-High-IR

Le premier jeu de données, appelé Distech-High-IR, est pratiquement une reproduction de la base de données Distech-Low-IR présentée précédemment à la grande différence d'avoir une haute résolution : soit 256x192 pixels. L'acquisition des données est réalisée dans les deux modalités : IR et RGB. Les images RGB sont utilisées pour annoter les images IR en utilisant une fonction d'alignement des deux modalités. La base de données présente 1500 exemples de séquences vidéos enregistrées à une vitesse de 4 images par seconde. Pour chaque séquence, uniquement l'image centrale de la séquence est annotée. La séparation des données est réalisée de manière aléatoire en respectant les proportions suivantes : 70% pour l'entraînement, 10% pour la validation et 20% pour le test.

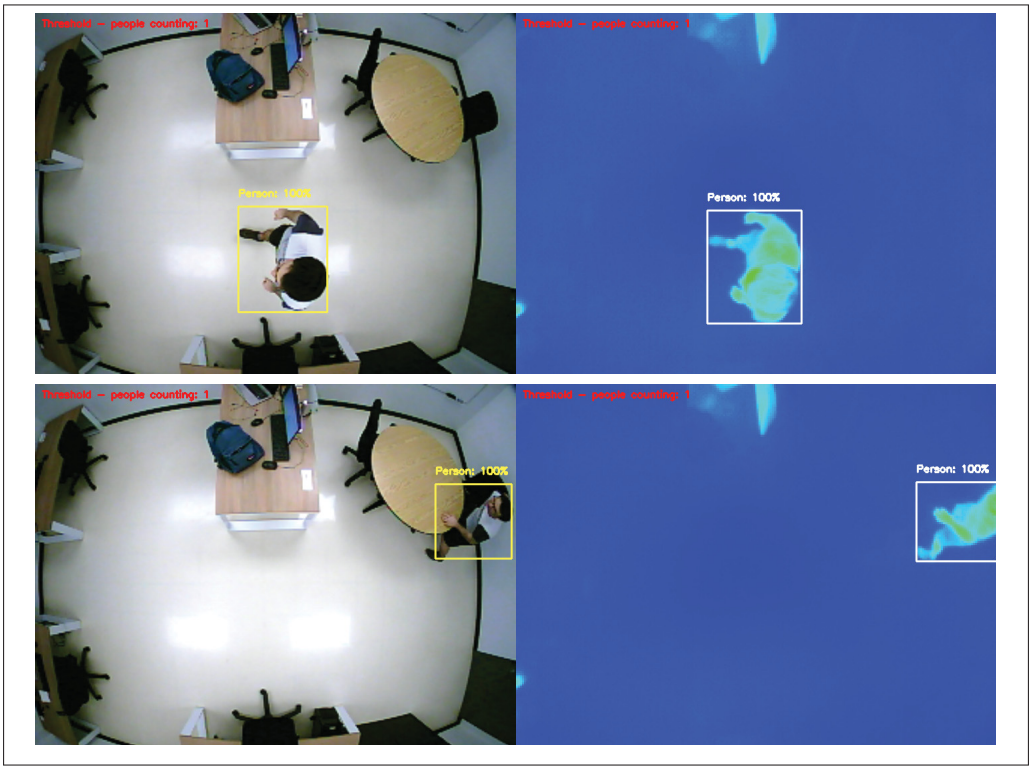


Figure 4.13 Exemple d'images annotées de la base de données Distech-High-IR. RGB à gauche et IR à droite

4.2.3 FLIR ADAS Dataset

La seconde base de données est publique : Group *et al.* (2018). Cette base de données est une référence dans la recherche en infrarouge, de plus, elle a aussi la modalité RGB qui est co-localisée avec l'image IR ce qui permet d'exploiter de nombreuses méthodes. Cette base de données a été extraite de vidéos à 24 images par secondes ce qui nous permet de l'utiliser pour cette étude. Le nombre d'exemples vidéos est de 7498 réparties comme précédemment avec 70% pour l'entraînement, 10% pour la validation et enfin 20% pour le test. Les annotations de boites englobantes sont disponibles pour plusieurs types de classes : personne, vélo, bus, voiture, motocycle, chien, etc. Pour rester en accord avec la première base de données nous avons uniquement considéré les classes : personne, vélo et motocycle.

Figure 4.14 Exemple d'images annotées de la base de données FLIR ADAS Dataset tirées de Group *et al.* (2018)

Dans ce travail nous nous sommes intéressés à la détection. Pour évaluer cette performance, nous avons donc utilisé la métrique de mAP@0.5 :0.95 qui est la métrique standard pour de la détection sur des images en haute résolution. Afin d'évaluer la précision des mesures et de garantir une certaine robustesse, nous avons répété les résultats 3 fois avec des séparations des données différentes.

4.2.4 Résultats

La première expérimentation que nous avons menée concerne l'utilisation du contexte temporel. Comme nous l'avons vu précédemment, la détection d'action nécessite de la temporalité. Dans

le cas de la détection nous soupçonnons une contribution plus faible de la temporalité. Ainsi le tableau 4.2 nous montre la performance de détection en fonction de l'écart entre images en employant la méthode proposée par Corsel *et al.* (2023). Ces résultats présentés sur deux bases de données différentes, nous montrent la supériorité de l'utilisation du contexte temporel. Dans le cas de Corsel *et al.* (2023), l'information temporelle remplace l'information chromatique, ce qui peut détériorer les performances dans certains cas. Dans notre cas, en IR, nous ne perdons pas d'information chromatique. Cependant nous observons une dégradation de la performance en augmentant la durée du clip d'entrée. Intuitivement, il devrait y avoir un compromis entre redondance de l'information dans le cas où l'écart entre images est de 0 et une information trop dé-corrélée avec un écart trop grand. Nos résultats semblent souffrir d'un manque d'images par secondes durant l'acquisition pour trancher. Ainsi nous estimons que l'optimal est en dessous de 80 ms. Pour la suite des expérimentations nous avons donc utilisé un écart entre images de 1 dans les deux bases de données.

Tableau 4.2 Performance de la méthode TYOLO en fonction de l'écart entre images. Le modèle YOLOv5n a été utilisé sur 3 répétitions.

Base de données	écart entre images	durée du clip (ms)	mAP@0.5 :0.95 ↑
Distech-High-IR	0	0	0.694 ± 0.007
	1	400	0.700 ± 0.001
	2	800	0.697 ± 0.010
FLIR ADAS Dataset	0	0	0.210 ± 0.012
	1	83	0.258 ± 0.019
	3	250	0.244 ± 0.015
	5	416	0.246 ± 0.017
	12	1000	0.215 ± 0.016

Nous avons étudié l'ajout de contexte jusqu'à 1 seconde de temps. Dans le cas de Distech-High-IR cela représente jusqu'à 2 images de delta. Et dans le cas de FLIR ADAS Dataset un écart maximum de 12 images. Sur les deux bases de données, l'écart optimal est de 1 image. Cela montre que l'information temporelle nécessaire à la détection est de très court terme.

Notre seconde expérimentation (voir tableau 4.3) concerne les augmentations de données que nous avons proposées. Afin d'évaluer la performance de nos augmentations de données, nous avons évalué son effet sur plusieurs détecteurs. Notre augmentation de données permet d'améliorer les résultats sur les deux bases de données étudiées et sur différents détecteurs. Ainsi notre contribution permet une amélioration de la performance de 17%. De plus, l'application du modèle proposé par Corsel *et al.* (2023) nous permet d'améliorer de 23% la performance du détecteur en utilisant le contexte temporel : sans utiliser plus d'annotations ni augmenter le temps de calcul en mode inférence.

Tableau 4.3 Performance des augmentations de données sur différents détecteurs

Modèle	augmentation temporelle	mAP@0.5 :0.95 ↑	
		Distech-High-IR	FLIR ADAS
Yolov5		0.700 ± 0.001	0.258 ± 0.019
	flou	0.691 ± 0.016	0.227 ± 0.017
	suppression de régions	0.702 ± 0.003	0.249 ± 0.020
	bruit	0.693 ± 0.007	0.274 ± 0.023
	(notre) ajustement	0.704 ± 0.002	0.331 ± 0.028
	(notre) pré-entraînement	0.707 ± 0.004	0.294 ± 0.016
	(notre) pré-entraînement + ajustement	**0.710** ± 0.004	**0.349** ± 0.027
SSD		0.611 ± 0.001	0.211 ± 0.027
	flou	0.624 ± 0.001	0.227 ± 0.007
	suppression de régions	0.627 ± 0.003	0.231 ± 0.029
	bruit	0.621 ± 0.001	0.237 ± 0.022
	(notre) ajustement	0.630 ± 0.002	**0.244** ± 0.028
	(notre) pré-entraînement	0.617 ± 0.002	0.232 ± 0.023
	(notre) pré-entraînement + ajustement	**0.632** ± 0.003	0.239 ± 0.012
FCOS		0.521 ± 0.005	0.159 ± 0.026
	flou	0.525 ± 0.006	0.150 ± 0.017
	suppression de régions	0.549 ± 0.014	0.174 ± 0.021
	bruit	0.514 ± 0.023	0.165 ± 0.016
	(notre) ajustement	0.536 ± 0.017	0.169 ± 0.013
	(notre) pré-entraînement	0.533 ± 0.008	0.172 ± 0.018
	(notre) pré-entraînement + ajustement	**0.556** ± 0.013	**0.176** ± 0.017

Dans le cadre de la base de donnée Distech-High-IR nous pouvons voir que la performance globale est bonne, ce qui est aussi le cas visuellement sur les détections proposées par le modèle.

Nous avons remarqué que l'ajout de contexte temporel dans le modèle lui permet un gain dans les scores de confiance dans les cas d'une détection sur le bord de l'image. Par exemple en présence d'un pied uniquement dans l'image 4a figure 4.17. Initialement, le pied n'est pas détecté mais en ajoutant du contexte avec l'image précédente, qu'on peut voir sur le canal rouge, il est détecté en 4b. Bien que l'image 2a/b soit quasiment statique, nous pouvons voir que le peu de mouvements entre les deux personnes à gauche permet au modèle utilisant le contexte, de délimiter proprement les deux personnes, ce qui est un résultat surprenant compte tenu de la complexité de la situation. Les résultats sur la base de données FLIR ADAS Dataset suivent le même schéma. La performance est globalement améliorée par l'ajout de contexte dans le modèle. Notre augmentation de données améliore de 7% absolue la performance et de 12% par rapport au modèle sans contexte temporel.

La suppression de régions proposée par Duran-Vega *et al.* (2021) améliore la performance dans le cas de Distech-High-IR contrairement à la base de données FLIR ADAS. Cette différence nous montre bien que les augmentations de données ne sont pas automatiques et ont des hyperparamètres à ajuster durant les itérations d'optimisation. La base de donnée de Distech-High-IR semblerait avoir plus d'occlusions que la base de données FLIR ADAS Dataset. Cette augmentation de données a été proposée pour les améliorer. En ce qui concerne l'insertion de bruit dans le flux proposé par Duran-Vega *et al.* (2021), il améliore la performance sur FLIR ADAS Dataset et la réduit sur Distech-High-IR. Ce bruit a été initialement introduit pour rendre la détection plus robuste aux effets de trainée rencontrés lors de la capture vidéo avec de longues expositions. Comme le FLIR ADAS Dataset ne présente pas de vidéo fixe, ces effets sont accentués, comparé au Distech-High-IR, qui lui est statique, ce qui peut expliquer les résultats quant à cette augmentation de données. Notre augmentation de données présente un seul hyperparamètre s qui influence l'amplitude des transformations appliquées en pourcentage de l'image de base. Nous avons étudié son effet sur les bases de données en figure 4.15. Notre augmentation de données présente un plateau de performance entre 5% à 10% ainsi nous préconisons une utilisation dans cette plage. Nos expérimentations nous ont mené vers une sélection de $s = 5\%$ basée sur l'ensemble de validation.

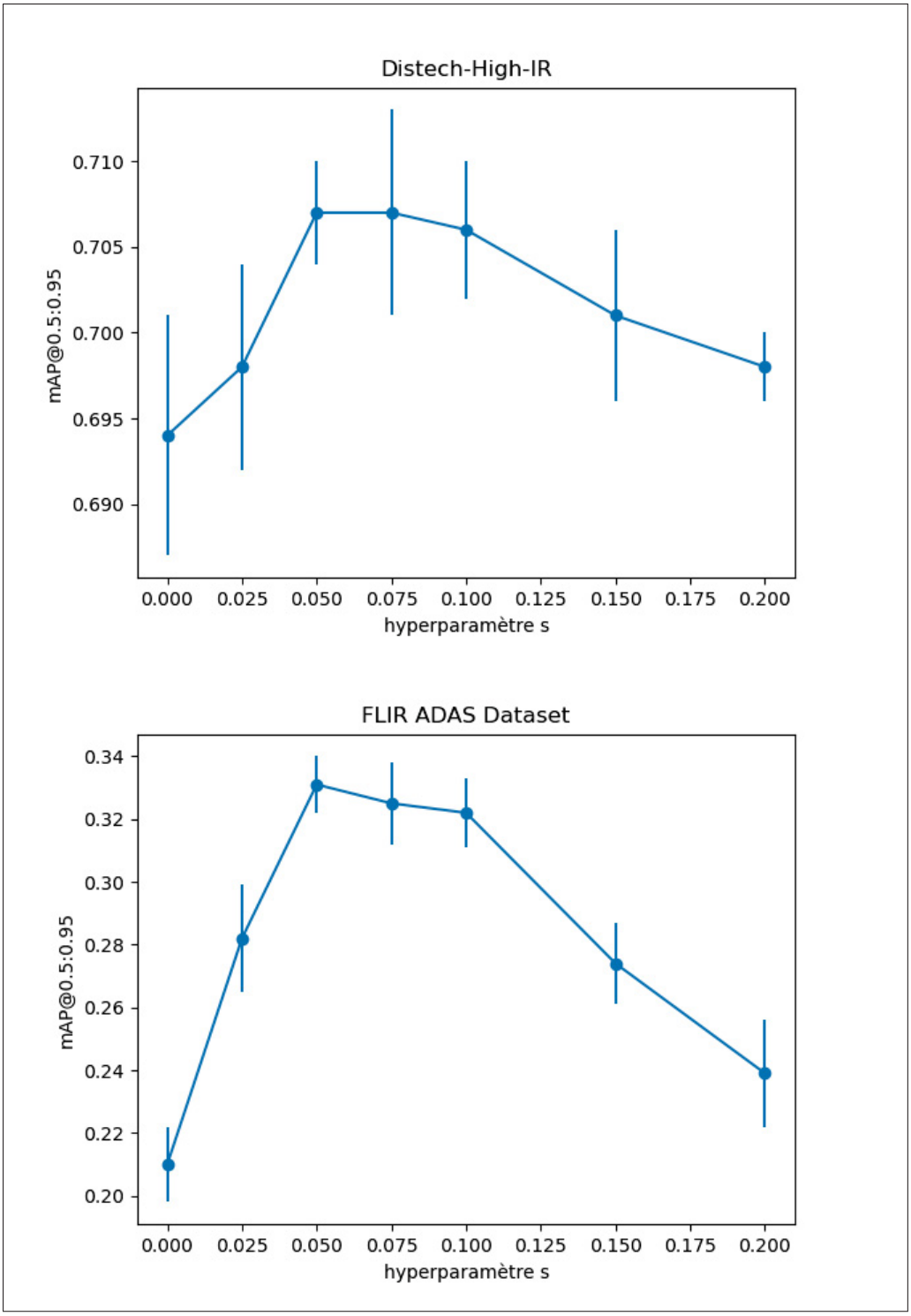


Figure 4.15 Effets de l'hyperparamètre *s* sur les deux bases de données

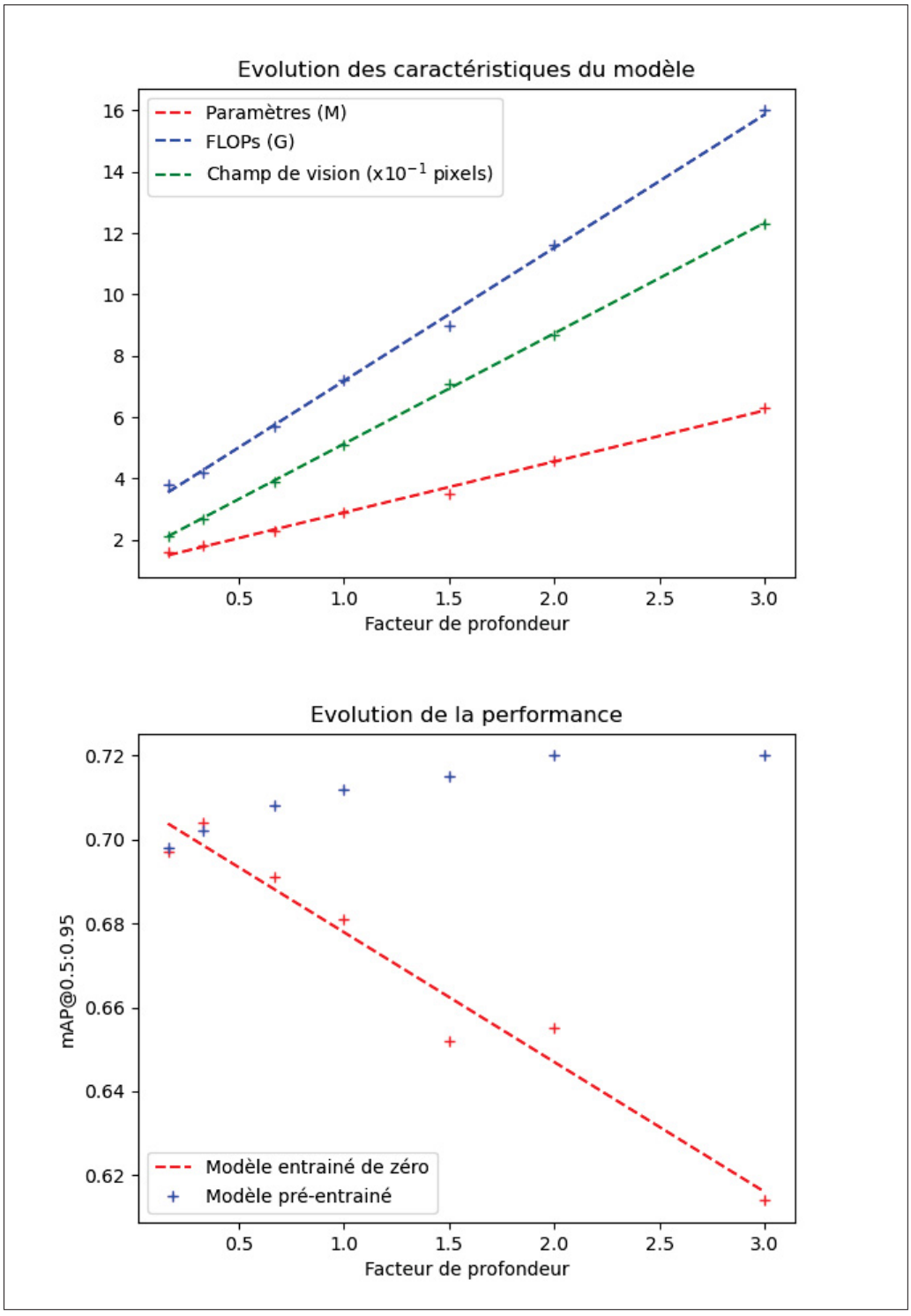


Figure 4.16 Effets de la profondeur du modèle

Nous avons aussi étudié l'architecture du modèle dans le cadre de l'application de la méthode proposée par Corsel *et al.* (2023). Intuitivement l'information temporelle ajoutée au niveau des canaux de couleurs, rajoute l'épaisseur à l'objet dû au décalage des couleurs (voir figure 4.12). Ce qui est par ailleurs utilisé par l'auteur pour détecter de petits objets. Dans notre cas les objets sont de taille normale, ce qui nous a encouragé à étudier le champ de vision utile pour chaque cellule de détection. Nous avons utilisé le modèle Yolov5 (car celui-ci est notre borne supérieure) en changeant les paramètres de profondeur du modèle pour en étudier les effets. Les résultats obtenus (voir figure 4.16) semblent contredire notre intuition. Les modèles avec le moins de champ de vision performent mieux. Cependant le nombre de paramètres augmente aussi avec le champ de vision, ce qui accroît la difficulté à entraîner le modèle. Les courbes d'apprentissage du modèle nous montrent que le modèle converge très vite vers une solution pertinente. Cela peut être dû à une tâche trop simple à réaliser ou encore une base de données trop petite. Dans ce cas, les modèles les plus petits sont de dimension plus adaptée et performent mieux que les modèles plus profonds qui sont partiellement entraînés (un apprentissage plus long les guide vers un sur-apprentissage). Lorsque les modèles sont pré-entraînés sur VOC (par Everingham, Van Gool, Williams, Winn & Zisserman (2010)), cela inverse la tendance car les modèles sont entièrement pré-entraînés. Les modèles étant plus grands, ils peuvent retenir plus d'informations concernant le pré-apprentissage, ce qui les rend plus performants. Bien que ces observations soient intéressantes, nous avons changé notre étude pour réduire le nombre de variables. Ainsi nous avons utilisé Yolov5, version nano qui est la plus petite, en ajoutant de la dilatation dans l'opération de convolution à plusieurs niveaux pour en augmenter le champ de vision. Cette opération a pour effet d'élargir le champ réceptif de la convolution sans en changer le nombre de paramètres, ni d'ajouter d'opération mathématique. Afin de ne pas trop perturber la manière de fonctionner du modèle, nous avons réparti les valeurs de dilatation sur plusieurs couches de convolutions en ajoutant au maximum 1 pixel de dilatation par couche.

Les résultats de cette étude (voir tableau 4.4) nous permettent d'affirmer, dans notre cas, que le champ de vision n'a pas d'effet sur les performances du modèle. Que le modèle soit pré-entraîné ou non, sa performance semble être indépendante de la dilatation ajoutée. Ainsi le gain de

Tableau 4.4 Dilatation des convolutions dans le modèle Yolov5 nano sur la base de donnée Distech-High-IR

Dilatation ajoutée	champ de vision (pixels)	mAP@0.5 :0.95 ↑	
		entraîné de zéro	pré-entraîné
0	21	0.652 ± 0.017	0.700 ± 0.010
+1	25	0.651 ± 0.015	0.698 ± 0.012
+2	29	0.653 ± 0.003	0.700 ± 0.009
+3	33	0.644 ± 0.013	0.694 ± 0.011
+4	37	0.646 ± 0.004	0.696 ± 0.006

performance que nous observons par les améliorations proposées par la méthode de Corsel *et al.* (2023) ne fait pas intervenir la corrélation entre pixels séparés de plus de 21 pixels de distance. Nos observations précédentes sur l'amélioration des détections sur le bord des images ainsi que les images pratiquement statiques vont aussi dans ce sens. Ainsi nous avons détourné l'utilisation de la méthode proposée par Corsel *et al.* (2023) qui semble avoir beaucoup d'avantages sur une application infrarouge. Non seulement le coup d'annotation n'est pas augmenté, mais l'architecture ainsi que le coût de calcul restent constants.

4.2.5 Conclusion

Dans cette partie nous avons étudié l'effet du contexte temporel sur la détection des personnes en infrarouge en haute résolution. Nous avons proposé 2 augmentations de données qui permettent d'améliorer les performances des méthodes de détection en utilisant le contexte temporel. La première permet d'utiliser des données classiques pour un pré-entraînement visant une application utilisant le contexte temporel. Bien que l'augmentation de données ne soit pas complexe, elle permet de mettre en lien des tâches (détection avec et sans contexte temporel) qui étaient auparavant séparées d'un grand changement de distribution. De cette manière nous pouvons amplifier les avantages du pré-entraînement. Nous avons aussi présenté une seconde augmentation de données permettant d'augmenter la généralisation du modèle durant la phase d'ajustement de la détection. Dans cette étude a également permis de quantifier le gain qu'introduit le contexte temporel dans le cadre de la détection. De plus, nous avons comparé

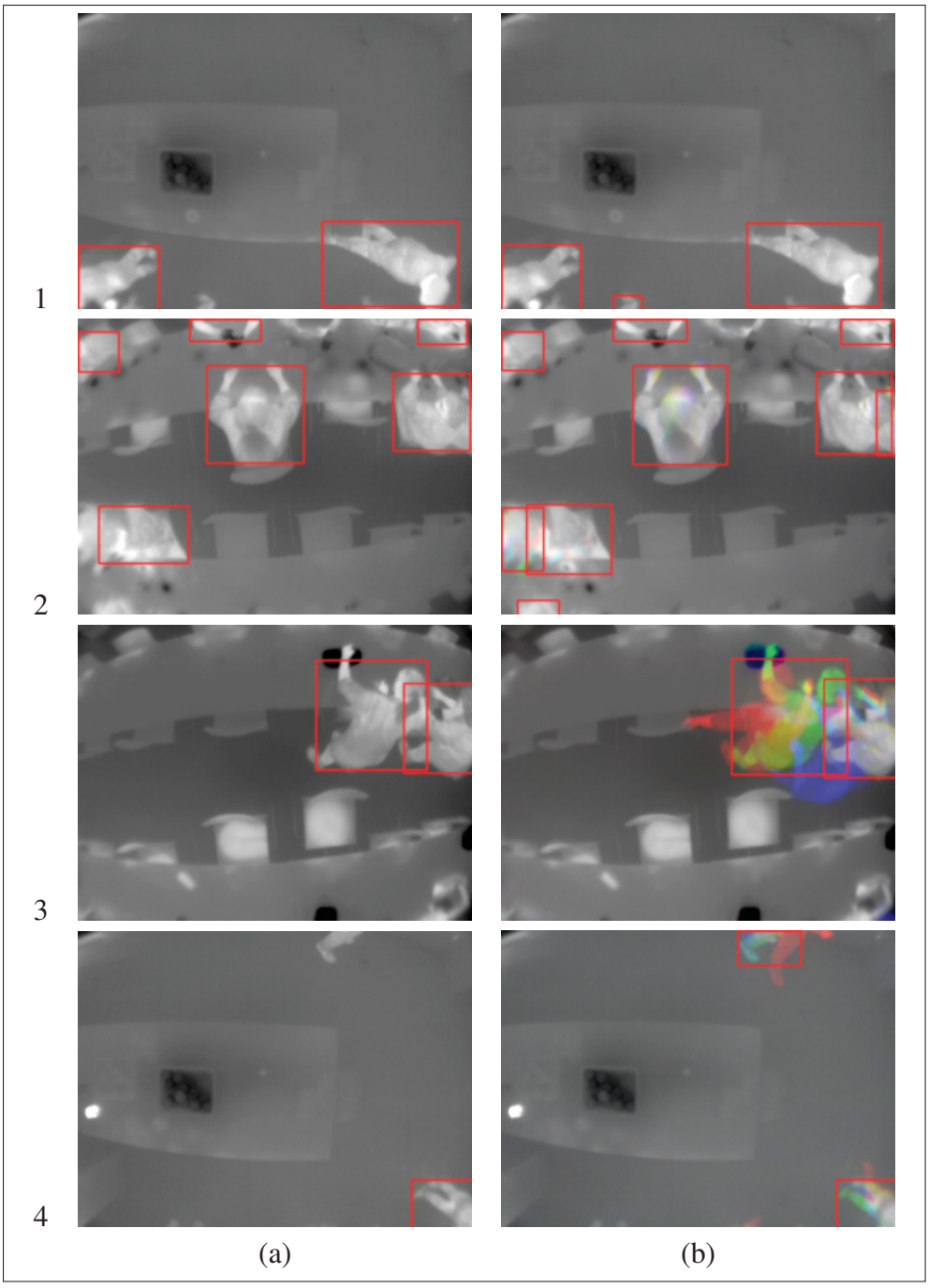


Figure 4.17 Exemple de détection sur l'ensemble de test, en (a) l'image IR en (b) en utilisant le contexte temporel. Le point de fonctionnement a été placé au maximum de score F1 sur l'ensemble de validation dans les deux cas

expérimentalement la temporalité optimale pour une bonne détection à celle de la classification d'actions. Celles-ci jouent sur des ordres de grandeurs différents pour ces deux tâches. Cela explique l'utilisation de méthodes et d'architectures différentes dans ces deux cas. Enfin nous avons étudié le champ de vision nécessaire au bon fonctionnement de la méthode proposée par Corsel *et al.* (2023) dans le but de l'agrandir : ce qui semble n'avoir aucun effet dans notre cas.

CONCLUSION ET RECOMMANDATIONS

Dans ce mémoire, nous avons étudié la détection des personnes en infrarouge. Bien que ce domaine ne soit pas très développé dans la littérature, il présente beaucoup de défis et d'avantages, que ce soit pour développer une solution de détection garantissant l'anonymat des individus, ou encore pour une utilisation de la vision dans l'obscurité. L'utilisation des capteurs infrarouges ont un futur prometteur dans le domaine des bâtiments intelligents et permettent non seulement de renseigner de la température de la pièce, mais aussi de réaliser le comptage du nombre de personnes présentes. De cette manière, une modulation dynamique de la climatisation, flux d'air ou encore de la température, peuvent réduire la consommation globale du bâtiment. De plus, une estimation du taux d'occupation peut permettre d'optimiser la place utilisée (par exemple dans le cadre de bureaux partagés) ou encore une détection d'intrusion. Pour finir, une détection des actions peut permettre des applications à grande échelle dans les hôpitaux ou maisons de retraite afin de prévenir rapidement d'une chute. Nous avons étudié le niveau de supervision nécessaire pour réaliser une détection en basse résolution, ce qui a donné lieu à une publication. Par la même occasion, nous avons contribué à l'élaboration d'une base de données publique pour encourager la recherche dans ce domaine et en augmenter la reproductibilité. Nous avons aussi amélioré les techniques actuelles de reconnaissance d'actions, utilisant les images infrarouges très basses résolutions, en proposant une évaluation sur la base de données citée précédemment. Pour finir, nous avons étudié l'utilisation du contexte temporel pour la détection afin d'obtenir des méthodes plus robustes et utilisables dans l'industrie. Dans ce cadre, nous avons proposé 2 augmentations de données permettant d'améliorer les résultats des méthodes de détection utilisant un contexte temporel.

Dans le cadre de la détection d'actions, nos augmentions de données peuvent aussi être appliquées pour améliorer les résultats. L'ajout de contexte temporel est également un axe à développer dans le cas de la basse résolution avec la base de données FIR-Image-Action-Dataset. Pour finir, la plupart des défis dans ce domaine sont induits par un changement de distribution qui pourrait

donc être étudié plus en profondeur : par exemple en utilisant des méthodes d'adaptation de domaine.

BIBLIOGRAPHIE

Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. (2019). Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. *Brainlesion : Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries : 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pp. 161–169.

Bergmann, P., Fauser, M., Sattlegger, D. & Steger, C. (2020). Uninformed students : Student-teacher anomaly detection with discriminative latent embeddings. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4183–4192.

Buolamwini, J. & Gebru, T. (2018). Gender shades : Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, pp. 77–91.

Cao, J., Sun, L., Odoom, M. G., Luan, F. & Song, X. (2016). Counting people by using a single camera without calibration. *2016 Chinese control and decision conference (CCDC)*, pp. 2048–2051.

Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. (2018). Grad-cam++ : Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847.

Chen, Y., Zhou, X. S. & Huang, T. S. (2001). One-class SVM for learning in image retrieval. *Proceedings 2001 international conference on image processing (Cat. No. 01CH37205)*, 1, 34–37.

Chen, Z., Wang, Y. & Liu, H. (2018). Unobtrusive sensor-based occupancy facing direction detection and tracking using advanced machine learning algorithms. *IEEE Sensors Journal*, 18(15), 6360–6368.

Corsel, C. W., van Lier, M., Kampmeijer, L., Boehrer, N. & Bakker, E. M. (2023). Exploiting Temporal Context for Tiny Object Detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 79–89.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words : Transformers for image recognition at scale. *arXiv preprint arXiv :2010.11929*, –.

Dubail, T., Guerrero Peña, F. A., Medeiros, H. R., Aminbeidokhti, M., Granger, E. & Pedersoli, M. (2022). Privacy-preserving person detection using low-resolution infrared cameras. *European Conference on Computer Vision*, pp. 689–702.

Duran-Vega, M. A., Gonzalez-Mendoza, M., Chang, L. & Suarez-Ramirez, C. D. (2021). TYOLOV5 : a temporal Yolov5 detector based on quasi-recurrent neural networks for real-time handgun detection in video. *arXiv preprint arXiv :2111.08867*, 1–16.

E. Santos, A. C. S. & Pedrini, H. (2019). Spatio-temporal Video Autoencoder for Human Action Recognition. *VISIGRAPP (5 : VISAPP)*, pp. 114–123.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88, 303–338.

Gao, C., Li, P., Zhang, Y., Liu, J. & Wang, L. (2016). People counting based on head detection combining Adaboost and CNN in crowded surveillance environment. *Neurocomputing*, 208, 108–116.

Girshick, R. (2015). Fast R-CNN.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative Adversarial Networks.

Group, F. et al. (2018). Flir thermal dataset for algorithm training.

Gutoski, M., Aquino, N. M. R., Ribeiro, M., Lazzaretti, E. & Lopes, H. S. (2017). Detection of video anomalies using convolutional autoencoders and one-class support vector machines. *XIII Brazilian congress on computational intelligence*, 2017, 1–12.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, P., Griffin, C., Kacprzyk, K., Joosen, A., Collyer, M., Shtedritski, A. & Asano, Y. M. (2021). Privacy-preserving Object Detection. *arXiv preprint arXiv :2103.06587*, 1–10.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Ji, S., Xu, W., Yang, M. & Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 221–231.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M. & Wei, Y. (2021). Layercam : Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30, 5875–5888.

Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X. et al. (2017). T-cnn : Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), 2896–2907.

Karayaneva, Y., Baker, S., Tan, B. & Jing, Y. (2018). Use of low-resolution infrared pixel array for passive human motion movement and recognition. *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*, pp. 1–2.

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 1–14.

Lee, D.-S., Hull, J. J. & Erol, B. (2003). A Bayesian framework for Gaussian mixture background modeling. *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)*, 3, III–973.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft coco : Common objects in context. *European conference on computer vision*, pp. 740–755.

Liu, M. & Zhu, M. (2018). Mobile video object detection with temporally-aware feature maps. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5686–5695.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. & Berg, A. C. (2016). SSD : Single Shot MultiBox Detector. *arXiv :1512.02325 [cs]*, 9905, 21–37.

Loshchilov, I. & Hutter, F. (2019). Decoupled Weight Decay Regularization.

Marszalek, M., Laptev, I. & Schmid, C. (2009). Actions in context. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2929–2936.

Morshed, M. G., Sultana, T., Alam, A. & Lee, Y.-K. (2023). Human Action Recognition : A Taxonomy-Based Survey, Updates, and Opportunities. *Sensors*, 23(4), 2182.

Oksuz, K., Cam, B. C., Kalkan, S. & Akbas, E. (2021). One metric to measure them all : Localisation recall precision (lrp) for evaluating visual detection tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 9446–9463.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62-66.

Radon, J. (1917). On the determination of functions from their integrals along certain manifolds. *Mathematisch-Physische Klasse*, 69, 262–277.

Redmon, J. & Farhadi, A. (2018). YOLOv3 : An Incremental Improvement.

Ren, S., He, K., Girshick, R. & Sun, J. (2016). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks.

Saul, L. K. & Roweis, S. T. (2000). An introduction to locally linear embedding. *unpublished. Available at : http ://www. cs. toronto. edu/˜ roweis/lle/publications. html*, –.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M. & Lockhart, J. W. (2020). Diagnosing gender bias in image recognition systems. *Socius*, 6, 2378023120967171.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017). Grad-cam : Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K. & Woo, W.-c. (2015). Convolutional LSTM network : A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 1–9.

Shi, Y., Tian, Y., Wang, Y. & Huang, T. (2017). Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimedia*, 19(7), 1510–1520.

Song, J., Yang, Z., Zhang, Q., Fang, T., Hu, G., Han, J. & Chen, C. (2018). Human action recognition with 3D convolution skip-connections and RNNs. *Neural Information Processing : 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I 25*, pp. 319–331.

Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F. & Wu, Y. (2021). Rethinking counting and localization in crowds : A purely point-based framework. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3365–3374.

Sun, K., Zhao, Q. & Zou, J. (2020). A review of building occupancy measurement systems. *Energy and Buildings*, 216, 109965.

Tao, L., Volonakis, T., Tan, B., Jing, Y., Chetty, K. & Smith, M. (2018). Home activity monitoring using low resolution infrared sensor. *arXiv preprint arXiv :1811.05416*, 1–8.

Tao, L., Volonakis, T., Tan, B., Zhang, Z., Jing, Y. & Smith, M. (2019). 3D convolutional neural network for home monitoring using low resolution thermal-sensor array. *3rd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2019)*, pp. 1–6.

Tateno, S., Zhu, Y. & Meng, F. (2019). Hand gesture recognition system for in-car device control based on infrared array sensor. *2019 58th Annual conference of the society of instrument and control engineers of Japan (SICE)*, pp. 701–706.

Tateno, S., Meng, F., Qian, R. & Hachiya, Y. (2020a). Privacy-preserved fall detection method with three-dimensional convolutional neural network using low-resolution infrared array sensor. *Sensors*, 20(20), 5957.

Tateno, S., Meng, F., Qian, R. & Li, T. (2020b). Human motion detection based on low resolution infrared array sensor. *2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp. 1016–1021.

Ullah, A., Muhammad, K., Haq, I. U. & Baik, S. W. (2019). Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. *Future Generation Computer Systems*, 96, 386–397.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 1–27.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, –.

Vishwakarma, D. K., Kapoor, R. & Dhiman, A. (2016). A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. *Robotics and Autonomous Systems*, 77, 25–38.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P. & Hu, X. (2020). Score-CAM : Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25.

Wang, X., Zhang, S., Qing, Z., Shao, Y., Zuo, Z., Gao, C. & Sang, N. (2021). Oadtr : Online action detection with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7565–7575.

Yu, S., Chen, X., Sun, W. & Xie, D. (2008). A robust method for detecting and counting people. *2008 International conference on audio, language and image processing*, pp. 1545–1549.

Zhang, H. (2020). *FIR-Image-Action-Dataset*. Repéré à https://github.com/visiongo-kr/FIR-Image-Action-Dataset#fir-image-action-dataset.

Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. (2017). mixup : Beyond empirical risk minimization. *arXiv preprint arXiv :1710.09412*, 1–13.

Zheng, H., Zhong, X., Huang, W., Jiang, K., Liu, W. & Wang, Z. (2022). Visible-Infrared Person Re-Identification : A Comprehensive Survey and a New Setting. *Electronics*, 11, 1–18.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, 2, 28–31.