Apprentissage multi-capteurs pour le contrôle non destructif de matériaux

Kevin Helvig

THESE DE DOCTORAT

NNT : 2024UPASG079

université PARIS-SACLAY

Apprentissage multi-capteurs pour le contrôle non destructif de matériaux

Multi-sensor learning for material non destructive testing

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Sciences du traitement du signal et des images
Graduate School : Informatique et Science du Numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Traitement de l'information et systèmes (Université Paris-Saclay, ONERA)**, sous la direction de **Pauline TROUVE-PELOUX**, Directrice de recherche, la co-direction de **Christophe PRADERE**, Ingénieur de recherche, et le co-encadrement de **Ludovic GAVERINA**, Ingénieur de recherche

Thèse soutenue à Paris-Saclay, le 27 novembre 2024, par

Kévin HELVIG

Composition du Jury

Membres du jury avec voix délibérative

Yannick LE MAOULT Professeur, IMT Mines Albi-Carmaux	Président & Examinateur
Thomas CORPETTI Directeur de recherche, Université de Rennes	Rapporteur & Examinateur
Olivier AUBRETON Maître de conférence (HDR), Université de Bourgogne Franche-Comté	Rapporteur & Examinateur
Sylvie LE HEGARAT-MASCLE Professeure, Université Paris-Saclay	Examinatrice
Thierry SENTENAC Professeur, IMT Mines Albi-Carmaux	Examinateur

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage multi-capteurs pour le contrôle non destructif de matériaux

Mots clés : apprentissage automatique, fusion de données, contrôle non-destructif, apprentissage profond

Résumé : La détection de défauts dans les structures aéronautiques et spatiales est cruciale, tant pour la fabrication que pour la maintenance. Les méthodes de contrôle non-destructif doivent être rapides, précises, fiables, économiques et de plus en plus automatisées. La complémentarité des différentes techniques d'inspection suggère leur utilisation simultanée pour renforcer la fiabilité des informations ou permettre une détection automatique difficile avec une seule technique.
Dans ce travail de thèse, nous explorons l'utilisation des méthodes d'apprentissage profond et de synthèse d'images pour la détection et la localisation de fissures par thermographie infrarouge laser Flying Spot sur des matériaux métalliques. Notre première contribution se concentre sur la collecte de données sur banc d'essai et la génération d'images synthétiques utilisant des modèles de diffusion pour augmenter la quantité et la diversité d'images ther mographiques disponibles. La deuxième contribution concerne l'utilisation ces jeux de données au travers de protocoles d'entraînement progressif et de transfert de caractéristiques afin d'améliorer la capacité des réseaux de neurones à discriminer et localiser les endommagements sur les images thermiques Flying Spot. Enfin, notre troisième contribution porte sur la construction d'une nouvelle architecture d'apprentissage profond pour la fusion multi-spectrale infrarouge-visible et sa mise en œuvre pour la détection de défauts sur des données couplées thermographiques et visibles. Cette thèse illustre le bénéfice de la fusion multi-spectrale au travers des méthodes d'apprentissage profond, en particulier dans le cadre de la thermographie laser.

Title : Multi-sensor learning for material non destructive testing

Keywords : machine learning, data fusion, non-destructive testing, Deep Learning

Abstract : Defect detection in aeronautical and space structures is crucial for both manufacturing and maintenance. Non-destructive testing methods must be fast, precise, reliable, cost-effective, and increasingly automated. The complementarity of different inspection techniques suggests their simultaneous use to enhance information reliability or enable automatic detection that would be challenging with a single technique.
In this thesis, we explore the use of deep learning and image synthesis methods for the detection and localization of cracks using laser Flying Spot infrared thermography on metallic materials. Our first contribution focuses on data collection on test benches and the generation of synthetic image using diffusion models to increase the quantity and diversity of available thermographic images. The second contribution involves using these datasets with progressive training protocols and feature transfer to enhance the neural networks' ability to discriminate and localize damage in Flying Spot thermal images. Finally, our third contribution proposes the construction of a new deep learning architecture for infrared-visible multi-spectral fusion and its implementation for defect detection on coupled thermographic and visible data. This thesis illustrates the benefit of multi-spectral fusion through deep learning methods, particularly in the context of laser thermography

Il n'est pas nécessaire de construire un labyrinthe quand l'Univers déjà en est un.

L'Aleph,
J. L. Borges

Remerciements

Une maigre épitaphe griffonnée de remerciements et jetée au vague et au sel marin, à la va-vite, pour tous ceux croisés durant ce travail de thèse, comme une brève escale avant de remplir les ballasts pour la rude (?) plongée dans le tas de pages trop dense qui constitue le présent manuscrit.

J'exprime d'abord ma gratitude aux membres du jury : Thomas Corpetti et Olivier Aubreton en tant que rapporteurs, puis Sylvie Le Hégarat-Mascle, Yannick Le Maoult, Thierry Sentenac, Loïc Cudennec en tant qu'examinateurs et membres invités du jury de thèse, pour l'honneur qu'ils m'ont fait en acceptant de soupeser la qualité de ma recherche. Je dédie un remerciement tout particulier pour Yannick Le Maoult et Thierry Sentenac, qui ont été parmi mes enseignants en école d'ingénieur, en contrôle non destructif et en thermographie infrarouge.

Je remercie ensuite chaleureusement ma directrice de thèse, Pauline Trouvé-Peloux, qui a guidé l'ensemble du travail que j'ai mené ici tout en me laissant l'autonomie dont j'avais besoin. Le travail de correction des différentes publications, le soutien pour la préparation des (nombreuses) présentations et la confiance accordée ont réellement été précieux. A mes yeux, il est clair que toute la recherche présentée ici n'aurait pu être abattu sans son aide et sa bienveillance. Je remercie également mon co-directeur, Christophe Pradère, pour son suivi du travail mené malgré la distance et l'éloignement au sujet traité.

Je suis également très reconnaissant envers tous mes co-encadrants : d'abord Baptiste Abeloos pour les échanges sur l'apprentissage machine en général et le volume des réflexions fournis, ce qui fut particulièrement utile lors de l'élaboration de l'architecture de fusion. La micro-divergence sur la source d'inspiration intarissable -à mes yeux- que donne à contempler ce qui grouille et le monde biotique au sens large pour la conception d'une réelle intelligence artificielle -et qui sait, l'émergence d'une véritable conscience ?- fut une vraie source de réflexions intellectuelles. Les arabesques tracées tortueusement dans le lent sillon de l'évolution, sous l'égide du Chaos, sont peut-être aussi riches de concepts que l'immensité des choses mathématiques elles-mêmes... Merci à Jean-Michel Roche et Ludovic Gavérina, qui m'ont d'abord pris en stage au sein du département des matériaux et structures de l'ONERA, m'amenant ensuite presque naturellement à conduire ce travail de thèse, qui de tous me paraissait le plus exaltant, de loin. Merci pour leur aide afin de mener cette recherche, de la jeune pousse de stage jusqu'à sa docte conclusion, éparpillée dans ces pages, notamment sur le plan expérimental.

Il me paraît nécessaire de dédier quelques lignes pour remercier l'ONERA ainsi que

l'Agence de l'Innovation de Défense (AID) en tant d'institutions, pour avoir soutenu et financé mon travail de recherche, et particulièrement la direction scientifique de l'ONERA qui me prima du prix de thèse du domaine "Traitement de l'information et Systèmes", saluant le travail de recherche dépeint par la suite.

Passons ensuite à l'inévitable décompte à la Prévert des pensées pour la masse des doctorants-camarades qui a partagé les années de cette recherche, de près ou de loin. Certains noms se perdront dans le vague, pour sûr. D'abord les "informaticiens" : Salomé, Marie-Ange (*Madame Ange veille sur DTIS-IVA*), Yann (*vais-je la lancer cette boîte de RL pour le jeu vidéo ?*), Marius (*merci pour la "contrebande" de Rocamadour*), Swann (*A. Reynolds est bien le meilleur auteur de Hard-SF*), Clément, Louis, Dao, David, Laura (*j'attends le passage à SAM-2*), Liam (*vive la diffusion*), Adrien, Pol, Solène (*diffusion ++*), Carla, Mélanie, et Quentin. Et chez les "matériciens" aux effectifs moins garnis, qui durent me supporter selon un cycle hebdomadaire seulement, heureusement pour eux ? Loïc, Zoé, Elisa (*merci pour le soutien et les quelques discussions, à l'orée d'une porte*), Charlotte, Anastasia, Stanislas, William, Thomas, Guillaume. Les doctorants plus vieux qui servirent d'ainés : Nathan, Maxime (*Maître-Fromager*), Thomas, Rémy, Alexis, puis Nathan au carré, Lisa, Inès, Achraf, côté matériaux. Je remercie particulièrement Nina, aînée devenue une amie, pour son soutien et son amitié durant la thèse, mais aussi pour avoir supporté un co-bureau tel que moi lorsqu'elle rédigeait son manuscrit, lors de mes quelques passages en terres "matériciennes". J'espère ne pas avoir été si bavard que ça, moi qui suis plus proche de la tombe d'habitude.

Je dédie de nombreux remerciements pour l'ensemble des chercheurs, techniciens et stagiaires des deux départements, qui m'ont souvent soutenus, mais aussi permis de confronter les quelques pièces éparses de connaissances que j'absorbais face à d'autres domaines de recherche, leur donnant bien plus de corps. Là encore trop de noms mais faisons comme on peux : Adrien (*nous verrons si je me lance de manière effective dans le RL, enfin Lunar-lander ne s'écrase plus*), Stéphane, Martial, Aurélien (*merci pour le prêt de la caméra événementielle*), Frédéric (*diffusion forever*), Philippe, Philippe², Elise, Flora, Valentin, Julien, Julien², Alexandre, Pierre, Hélène, Anthelme, Pierre², Yves, Florence, Georges, Bruno, Aurélie, Thibaud, Francoise, Alvérède, Jesus, Céline, Maria, Quentin, Nicolas, Jean-François, Thomas, Myrian, Lina, Alice ...

Je me dois d'accorder une "spéciale dédicace" à Laureen, pour les quelques coups d'oeil utiles sur les articles que j'ai pu produire et pour le soutien sur ces trois années : une soeur d'armes issue des entrailles des Mines d'Albi. Je dédie au passage une pensée pour Zoé, un visage de "matéricienne", ciselée elle-aussi au détour d'une traversée dans cette école.

D'ultimes mots de remerciements sont pour ma mère et pour ma soeur restées sous le soleil Occitan, mais toujours là pour me fournir leur soutien indéfectible à distance. Les voix rassurantes du téléphone et les quelques retours à la maison étaient comme de courtes escales en pleine expédition au long court, solitaire et difficile. J'ai un mot pour ma grand-mère, un roc qui laisse le temps défiler sur elle, imperturbable, et qui finira sans doute par enterrer ce qu'il reste de la famille. Je laisse un merci plus morne qui épilogue avec *Ananké*, pour mon père. Tu as bien joué à me mettre sur cette trajectoire-là par les truchements cinéraires de la Causalité et du Cruel. Toute cette aventure menée durant la thèse s'est révélée un purgatoire cathartique intense et riche, ressuscitant le sens après trois années précédentes plus ternes à cicatriser, sous la lourde coupe lasse du Sérieux et de l'Amer, parfois de l'incompréhension. Se dérober n'est pas une option : il faut bien

essayer de finir par vivre, et nous verrons bien où cette vie mènera.

A ceux mentionnés au-dessus, ainsi qu'à tous ceux qui se retrouvent oubliés en cours de route, par manque de place, par manque de mot ou par défaillance malheureuse du souvenir : merci à tous !

Table des matières

Introduction

Depuis le Big Bang, tout commence à mourir à l'instant même de naître. L'univers n'est qu'un élan vers l'usure et la mort.

Voyez comme on danse,
J. D'Ormesson

Ce manuscrit présente le travail de recherche mené à l'Office National d'Études et de Recherches Aérospatiales (ONERA), au sein du département Traitement de l'Information et Systèmes, unité Mesure, Image, Co-conception (DTIS-MIC), anciennement unité Image Vision Apprentissage (DTIS-IVA). Cette thèse a été menée en collaboration avec le département Matériaux et Structures, unité d'Élaboration et Procédés d'Imagerie et de Contrôle (DMAS-EPIC). Cette recherche porte sur la mise en œuvre des méthodes d'apprentissage automatique pour le contrôle non destructif des matériaux, en particulier par thermographie laser "Flying Spot", et sur la fusion d'informations issues d'instrumentations différentes *via* un apprentissage profond. Nous allons d'abord revenir sur les motivations conduisant à ce travail de recherche, puis sur les différents axes d'investigation explorés au cours de cette thèse. Enfin, nous passerons en revue les différentes contributions scientifiques associées à ces travaux.

Motivations

La maintenance est un processus crucial dans de nombreuses industries telles que l'aéronautique, l'aérospatial, le naval, et le nucléaire. Il est indispensable de disposer de moyens d'inspection à la fois rapides, fiables et peu coûteux, permettant de garantir le bon fonctionnement des systèmes. Or, la détection de défauts ainsi que la caractérisation des propriétés géométriques des pièces métalliques modernes sont des tâches parfois complexes. En effet, les pièces à haute valeur ajoutée présentent des variétés de revêtements ou bien des éléments issus de l'usure et d'éléments géométriques fonctionnels, comme des conduits de refroidissement, de surfaces 3D non planes. Les moyens d'examen et de contrôle non destructifs (CND) consistent en la détection sans contact de défauts variés, sur et dans des composants d'un système. Cela permet d'assurer la viabilité en fonctionnement suivant le caractère critique ou non du défaut détecté. Ces techniques sont de plus en plus présentes dans l'industrie en général, que ce soit dans des secteurs de pointe tels

que l'aéronautique et le secteur spatial ou dans des industries aux échelles de production plus importantes (industrie automobile, électronique). Parmi les défauts pouvant se former lors du cycle de vie d'une pièce mécanique, nous nous concentrons sur la détection de fissures de surface de dimension millimétrique sur les matériaux métalliques, dont la phénoménologie de fissuration simplifiée est illustrée dans la Figure 1.

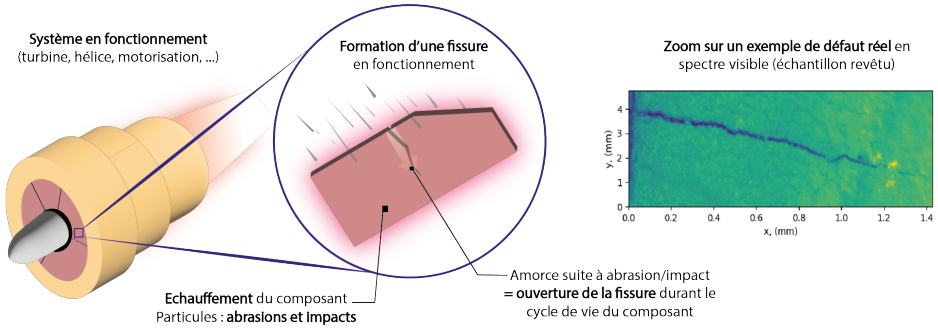


FIGURE 1 – Phénoménologie de fissuration simplifiée et vue rapprochée d'un exemple de fissure en spectre visible sur un échantillon revêtu.

Le contrôle non destructif regroupe un grand nombre de méthodes afin de détecter des défauts et d'en caractériser la criticité. Une première méthode de mise en exergue du défaut est le ressuage, qui repose sur le dépôt d'un agent actif sur la surface, permettant de révéler en optique visible, voire directement à l'œil, la présence ou non d'un endommagement. Mais la dépendance de cette technique à un traitement chimique ou thermique pouvant altérer la pièce ainsi que la potentielle pollution chimique associée ont rendu son usage plus limité aujourd'hui.

L'inspection passive dans le spectre visible est une autre méthode de détection de défauts à mentionner, avec aujourd'hui un grand nombre de capteurs disponibles allant de l'inspection binoculaire jusqu'à l'utilisation de caméras visibles de résolution variée. Ces méthodes sont cependant limitées à l'examen de la surface du matériau étudié, et peuvent de plus être trompées par des éléments comme la rugosité ou des variétés de dépôts sur la surface d'étude qui sont dus à l'usure, par exemple. C'est pourquoi d'autres moyens d'imagerie ont été développés. Ils vont des ultrasons, des rayonnements X à la thermographie infrarouge et aux courants de Foucault. Ils permettent à la fois de compenser les limitations du spectre visible seul et de révéler des endommagements sous la surface du matériau. Le type de défaut détectable dans le matériau peut être la formation de porosités ou de délaminages, ou bien de fissures masquées par l'état de la surface. Les inspections par ultrasons utilisent le balayage d'une onde acoustique aux fréquences ultrasonores pour scanner la surface d'intérêt. La propagation de l'onde dans la matière de la pièce va permettre de détecter des défauts surfaciques et volumiques, fonctionnant de manière analogue aux échographies. Les méthodes tomographiques par rayons X utilisent la transmission des rayonnements pour imager des éléments de surface mais surtout des éléments internes à la pièce, comme des porosités. Ce moyen est généralement lent à mettre en œuvre, le temps de balayer la pièce, et nécessite de plus des traitements de données lourds et exigeants en quantité de données pour imager la pièce. Les méthodes par thermographie reposent sur l'exploitation de la chaleur dégagée par le matériau, soit de manière passive, soit de manière active avec l'adjonction d'une source de chaleur dont

on va pouvoir observer la diffusion sur et/ou dans le matériau inspecté. L'interaction entre le matériau étudié et cette diffusion de chaleur va permettre de trahir la présence d'un endommagement. Les courants de Foucault exploitent l'induction électromagnétique pour repérer un endommagement, et n'étant pas si éloignée des ultrasons dans sa mise en œuvre, avec l'utilisation d'une sonde balayant la surface cible. Cette approche de CND est cependant restreinte aux matériaux ferromagnétiques.

Ces moyens d'imagerie sont généralement difficiles à déployer à grande échelle eu égard à leur complexité de mise en œuvre. Les ultrasons et la thermographie sont les plus employées avec une expertise opérateur plus modérée. Néanmoins, ces méthodes restent coûteuses en temps et en argent et gagneraient à être automatisées pour les déployer plus facilement, tout en réduisant le temps d'examen et de traitement des données collectées.

Pour la détection de défauts millimétriques sur des surfaces métalliques et dans un contexte où l'on cherche à simplifier et accélérer la procédure d'examen, les méthodes par ultrasons sont généralement plus difficiles et longues à mettre en œuvre du fait de la mise en place de sondes, de gel de contact afin de propager le signal acoustique. Au contraire, la propagation de la chaleur produite à distance dans une surface et mesurée par les méthodes thermographiques peut être plus appropriée pour un examen plus rapide, sans démontage ou contact direct avec la pièce à examiner.

Parmi les méthodes de thermographie, on peut distinguer la thermographie Flash, qui consiste en l'utilisation d'une source de chaleur globale échauffant très rapidement l'ensemble de la surface observée. Elle peut être employée autant en transmission qu'en réflexion suivant le type de défaut à détecter et le matériau examiné. Cette technique de contrôle est illustrée dans la Figure 2. Les autres méthodes thermographiques proposent l'utilisation d'un échauffement locale. Parmis elles nous trouvons les thermographies laser et par induction. La thermographie inductive ou par thermo-induction utilise le phénomène magnétique de l'induction pour provoquer un échauffement au niveau du défaut. Elle est néanmoins restreinte aux matériaux qui ont une bonne réponse magnétique : les matériaux métalliques ferro-magnétiques. L'utilisation d'une source laser locale apparaît plus versatile en termes de matériaux observables. La thermographie pulsée consiste cette fois en l'utilisation d'un échauffement dans lequel on introduit une variabilité fréquentielle dans le flux de chaleur transmis à la pièce examinée, ce qui permet de détecter des endommagements.

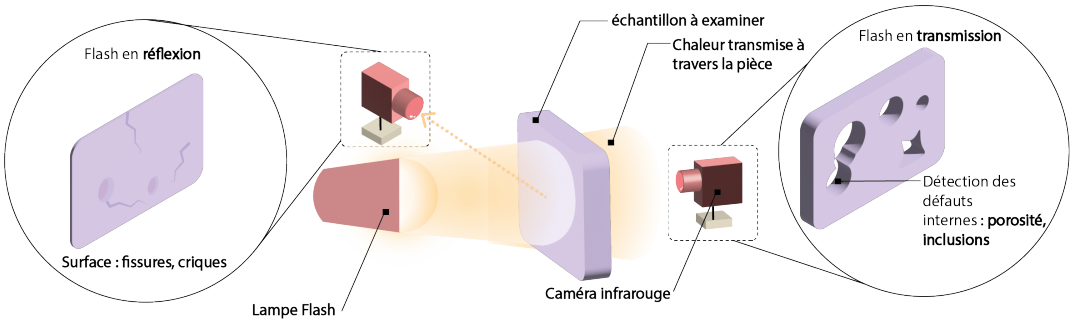


FIGURE 2 – Technique de thermographie Flash illustrée en réflexion et en transmission.

L'utilisation d'une source laser locale est donc plus versatile en termes de matériaux observables. La thermographie pulsée consiste cette fois en l'utilisation d'un échauffement dans lequel on introduit une variabilité fréquentielle dans le flux de chaleur transmis à

la pièce examinée, ce qui permet de détecter des endommagements. La thermographie laser Flying Spot (FST) consiste à balayer la surface à examiner à l'aide d'une source de chaleur locale laser. Ce balayage permet ensuite d'identifier des discontinuités dans la diffusion de la chaleur sur la surface examinée, qui peuvent correspondre à des défauts. Les principes généraux de la technique sont illustrés à la Figure 3, ainsi qu'un exemple de cartographie reconstruite consistant en la somme normalisée des images thermiques d'un enregistrement d'un balayage laser. La fissure est alors caractérisée par la discontinuité dans le signal thermique observé. Cette méthode de CND locale et sans contact direct permet d'envisager des solutions embarquées et automatisées. Comparée à la méthode Flash, elle est plus compacte, ne nécessitant pas de lampe Flash. La FST a fait l'objet de traitements antérieurs à l'ONERA avec la constitution d'un banc d'essai, elle nous est donc apparue pertinente pour le développement de méthodes automatiques pour la détection de fissures de surface.

Cependant, si des travaux existent combinant l'apprentissage automatique avec la thermographie Flash et la thermographie laser pulsée, la littérature actuelle est plus limitée concernant l'automatisation de la thermographie Flying Spot, tant du côté de la robotisation du moyen d'inspection que du côté du traitement des données expérimentales. Nous pouvons l'expliquer par le fait que la FST est une méthode d'examen moins répandue comparée aux approches ultrasons et Flash, avec une absence de données publiques disponibles en quantités importantes. Les approches basées sur l'apprentissage ont pourtant aujourd'hui fait leurs preuves dans le contrôle qualité industriel en spectre visible notamment. Un des enjeux actuels concerne la robustesse de ces méthodes face à de nouveaux échantillons dans un contexte où les données ne sont pas disponibles en grands volumes.

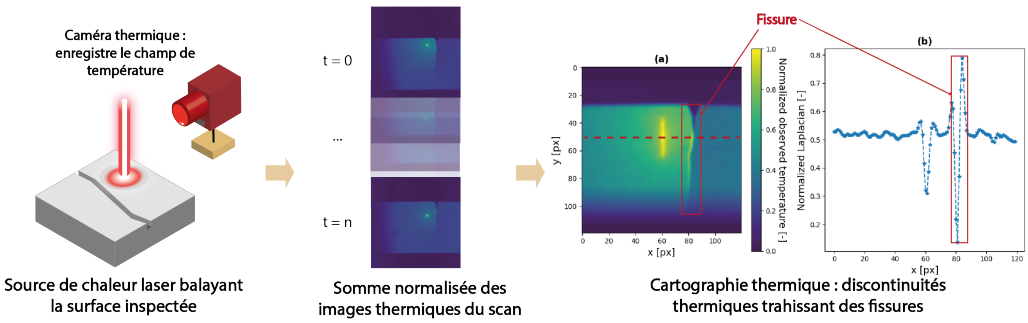


FIGURE 3 – Illustration du principe de la thermographie FST, et reconstruction d'une cartographie thermique à partir d'un balayage d'éprouvette métallique.

Il est commun en contrôle non destructif d'employer différentes instrumentations de manière manuelle, chaque spectre ayant ses informations spécifiques utiles pour déterminer l'état d'usure d'une pièce. L'imagerie visible passive permet par exemple de déterminer les états de surfaces par les textures, les géométries de surface ou de potentiels défauts. Néanmoins, cette instrumentation peut confondre les rayures de surface et les fissurations dommageables pour le fonctionnement de la pièce. De plus, un défaut fin peut être rendu impossible à détecter, ou peut être masqué par la rugosité ou un revêtement de surface. L'estimation complète et précise de la longueur de fissuration peut être complexe dans un tel environnement de détection sans moyens optiques lourds tels qu'un microscope

binoculaire voire des moyens de microscopie électronique.

La FST, quant à elle, permet des détections de fissures et la caractérisation de leur longueur malgré la présence de revêtement grâce au phénomène de propagation de la chaleur à travers la surface du matériau. Elle peut généralement permettre une meilleure estimation de la longueur de l'endommagement que le spectre visible par l'exploitation des phénomènes thermiques. Ce moyen peut aussi être considéré comme peu sensible aux fausses alarmes dues aux rayures de surface superficielles, car celles-ci entraînent une discontinuité dans la propagation de la chaleur qui est au pire minime et au mieux inexistante. Néanmoins, cette modalité d'examen peut être trompée par les revêtements de surface, en particulier sur des surfaces complexes présentant par exemple des régions revêtues et d'autres dénudées, ou bien des géométries de surface usinées (conduits de refroidissement, bords de pièce complexes) : ces éléments sont des sources de fausse alarme potentielle lors du contrôle d'une pièce. Ce moyen d'inspection est de plus limité à l'observation d'un champ dont la dimension est de l'ordre du centimètre : balayer complètement une pièce peut être alors très coûteux en immobilisant du personnel et du matériel.

Dans cette thèse, nous proposons donc de fusionner une inspection en spectre visible et une inspection en FST. La complémentarité de ces deux méthodes d'inspection est illustrée dans la Figure 4 avec une éprouvette observée en spectre visible large champ, où une amorce de fissuration est observable, ainsi qu'une cartographie thermique associée qui révèle l'intégralité de la fissure.

L'exploitation de cette complémentarité de manière autonome grâce à l'apprentissage profond est nouvelle en CND, et demande des investigations, notamment dans la gestion de surfaces non planes et de champs très variables entre le moyen visible et le capteur infrarouge, pouvant nuire à la qualité du réalignement des paires d'images.

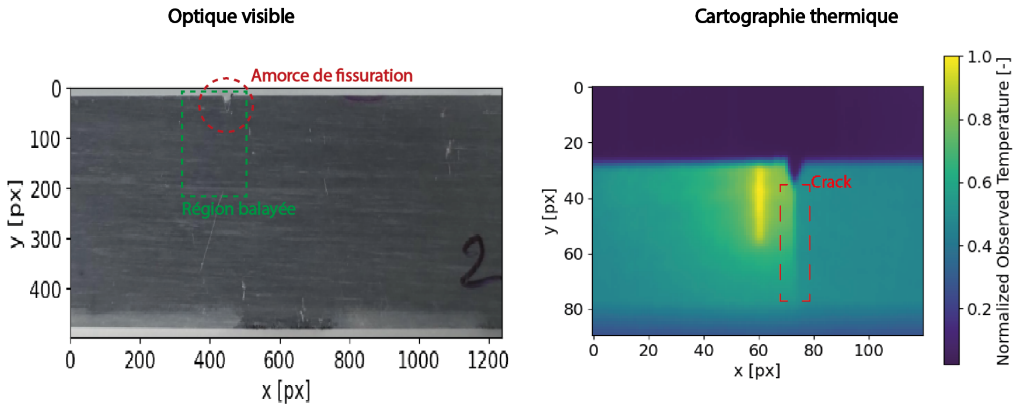


FIGURE 4 – Complémentarité FST-visible avec à gauche l'imagerie visible acquise via une caméra à large champ, à droite un balayage thermique proche de l'amorce de fissure. Le spectre visible ne permet pas d'observer que l'amorce de fissuration, alors que la FST révèle l'intégralité du défaut.

Axes de recherches

L'objectif de ce travail de thèse est donc la mise en œuvre de méthodes d'apprentissage machine appliquées à la thermographie laser Flying Spot afin d'automatiser la détection

de défauts tels que les fissures de surface. Le deuxième objectif est ensuite l'exploitation du couplage entre cette information infrarouge et le spectre visible pour améliorer la détection automatique de fissures. Ce travail a été guidé par quelques questions clés évoquées ici :

— **Comment compenser le manque de données expérimentales afin d'augmenter les capacités de détection et la robustesse à de nouveaux défauts en thermographie laser ?** Le développement de techniques d'apprentissage machine nécessite de disposer de bases de données en volumes assez importants et représentatifs des défauts à inspecter en situation industrielle. Ces bases n'existent pas en FST. Un enjeu a donc été de construire ces bases d'apprentissage tout en assurant tant leur volume d'images que leur diversité, au travers notamment de plages de paramètres expérimentaux étendues.

— **Comment localiser précisément le défaut sur la surface examinée ?** Une première étape est de discriminer les surfaces saines et celles avec défauts. Mais en vue de permettre une caractérisation du défaut, il est nécessaire d'utiliser des architectures de localisation/segmentation de la fissure plus lourdes, plus complexes et plus gourmandes en données que les modèles de classification, et dont l'annotation de localisation peut être coûteuse en expertise humaine.

— **Comment exploiter la complémentarité entre l'information infrarouge et l'information visible par apprentissage dans le contexte du contrôle des matériaux ?** Si l'apprentissage profond pour la fusion de données infrarouges et visibles est largement étudié dans des applications de vision par thermographie passive comme la surveillance, la conduite autonome et la télédétection, la fusion de données automatique par apprentissage reste peu explorée dans le domaine des matériaux et du contrôle non destructif. Cette fusion permettrait pourtant de réduire le nombre de fausses détections ou de détections manquées, ce qui est pertinent pour une application en détection de défaut. L'enjeu est donc de disposer d'une approche de fusion de données apprise robuste à des images provenant de capteurs et de champs dont les résolutions spatiales diffèrent parfois grandement.

Plan du manuscrit

Nous fournissons dans la Figure 5 un diagramme détaillé des travaux qui ont été conduits durant cette thèse.

Le plan chapitre par chapitre est le suivant. La première partie est consacrée à la construction de bases de données en FST. Le Chapitre 1 fournit des éléments de base théoriques en thermographie laser et en FST, puis présente les différentes bases de données collectées ou simulées en Flying Spot sur des jeux d'échantillons métalliques disponibles. Le Chapitre 2 présente la synthèse d'images par modèles de diffusion, qui est la technique d'augmentation de données employée dans le cadre de cette recherche afin de compenser le manque de données potentiel et d'augmenter la diversité statistique au travers du brassage des caractéristiques.

La deuxième partie de ce manuscrit utilise ces différents volumes de données couplés à l'apprentissage profond pour détecter les fissurations. Le Chapitre 3 présente la mise en place d'une méthode d'apprentissage progressif employant de manière hiérarchisée les

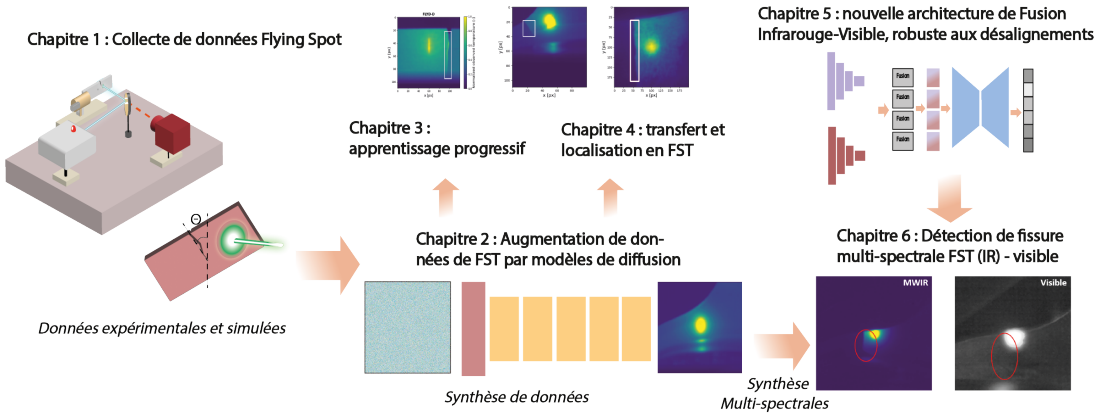


FIGURE 5 – Schéma représentant l'organisation des recherches menées durant la thèse.

données simulées, synthétiques et réelles pour augmenter la capacité des modèles neuronaux à détecter des fissures dans un contexte de manque de données. Nous passons dans le Chapitre 4 de la classification entre images thermiques fissurées ou non à la détection avec localisation de la fissure sur l'image thermographique. Nous exploitons encore dans ce chapitre l'idée de transfert de caractéristiques et d'apprentissage progressif, depuis des échantillons simples disponibles en grand nombre vers des matériaux plus complexes dont le volume de données est plus restreint. Nous y réemployons la synthèse par diffusion sur les échantillons à surface simple pour rehausser encore les capacités de localisation sur l'information thermographique. Le chapitre 4 étudie aussi l'exploitation de l'image thermique individuelle prise au cours d'un balayage, en lieu et place des cartographies thermiques.

La dernière partie est consacrée à la fusion infrarouge-visible. Dans un premier temps, le Chapitre 5 présente notre nouvelle architecture neuronale de fusion infrarouge-visible basée sur les *Transformers* et le mécanisme d'attention. Cette architecture est validée sur des bases de données infrarouge-visible passives de référence dans la communauté de vision par ordinateur et sur des domaines d'application variés : surveillance urbaine, télédétection par satellite, conduite autonome. Le Chapitre 6 clôt le travail conduit avec la mise en application de notre proposition d'architecture de fusion et des méthodes d'augmentation de données basées sur la synthèse d'images, pour le couplage entre FST en spectre infrarouge et modalité visible. Une architecture de la littérature est aussi entraînée afin de fournir un point de comparaison des techniques d'apprentissage machine pour ce problème spécifique de contrôle non destructif avec fusion de données.

L'annexe A dispense des éléments de base de l'apprentissage profond de manière synthétique.

Contributions scientifiques

Ces travaux ont pu être valorisés au travers d'un certain nombre de publications et d'actes associés à des présentations lors de conférences. Ceux-ci sont listés ci-dessous.

Actes de conférence :

— *Towards deep learning fusion of flying spot thermography and visible inspection for*

surface cracks detection on metallic materials, Helvig *et al.*, QIRT 2022 : ce travail fournit une des premières mises en œuvre dans la littérature de l'apprentissage machine sur la donnée FST, et explore les complémentarités entre les informations infrarouge et visible, sans fusion.

— *Laser flying-spot thermography : an open-access dataset for machine learning and deep learning*, Helvig *et al.*, QCAV 2023 : Nous nous focalisons ici sur le déploiement d'une base de données en libre accès pour la thermographie laser flying-spot, palliant cette absence dans la littérature, sur des échantillons simples type éprouvette. Ce travail sera ensuite complété et renforcé dans un article de journal.

— *Détection de fissure sur matériaux métalliques par apprentissage profond et thermographie laser flying-spot : approche par apprentissage progressif*, Helvig *et al.*, Gretsi 2023 : Cet article de conférence fournit une première mise en œuvre de l'apprentissage progressif sur les données de thermographie FST. Ce travail sera ensuite complété et renforcé dans un article de journal.

— *Synthetic visible-IR images pairs generation for multi-spectral NDT using flying spot thermography and deep learning*, Helvig *et al.*, SPIE Thermosense 2024 : Ici les approches d'augmentation de données sont réimplémentées avec des modèles de diffusion text+image tels que Stable Diffusion, d'abord en mono-spectral FST sur nos données en libre-accès, puis un protocole est proposé pour la génération de paires FST-visible synthétiques cohérentes, permettant d'augmenter les performances d'un premier réseau de fusion infrarouge-visible de la littérature.

— *CAFF-DINO : Multi-spectral object detection transformers with cross-attention features fusion*, Helvig *et al.*, IEEE CVPR Workshop on *Perception Beyond the Visible Spectrum* 2024 : Cet acte de conférence de rang A* présente notre architecture de fusion infrarouge-visible développée durant ce travail de thèse qui est basée sur les *Transformers*. Ce modèle atteint voire surpasse les performances de la littérature sur des bases de données de référence en vision par ordinateur pour la fusion IR-visible en vue de détecter des piétons. La robustesse de cette approche aux désalignements systématiques de l'information infrarouge a été de plus étudiée et comparée à une architecture de la littérature.

Articles de revue :

— *Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training*, Helvig *et al.*, Quantitative Infrared Thermography Journal, 2023 : Ce travail développe une méthode de *Curriculum Learning* pour la classification d'images FST, montrant comment une exploitation hiérarchisée et intelligente des données simulées, synthétiques générées par diffusion et enfin réelles pouvait être bénéfique pour les capacités d'un modèle d'apprentissage à identifier des fissures en thermographie.

— *Database for transfer learning in crack detection and localization on metallic materials using flying spot thermography and deep learning*, Helvig *et al.*, Journal of Electronic Imaging, 2024 : Ce travail exploite la base de données open-source développée pour QCAV et propose le transfert d'apprentissage après pré-entraînement sur des acquisitions FST sur pièces simples, où la diversité de défauts est assez élevée mais où la surface est homogène. Le réseau est ensuite transféré vers des matériaux

plus complexes, où la quantité de données est volontairement contrainte. Cette approche de transfert de représentations est analogue à celle développée dans le cadre de l'apprentissage progressif.

Des dépôts Github ont de plus été mis en ligne, fournissant au public une partie du travail développé, sur des données non-propriétaires. Le premier a notamment été déployé pour le travail de conférence QCAV 2023, donnant un accès à notre base open-source : https://github.com/kevinhelvig/FLYD. Un deuxième lien met notre architecture de fusion infrarouge-visible à disposition de la communauté : https://github.com/kevinhelvig/CAFF-DETR. Le dernier dépôt mis en ligne concernera la base de données couplée IR-visible collectée sur éprouvette, FLYD-II, qui est la première base de données de contrôle non destructif multimodale mise en *open-source* à notre connaissance, accessible ici :
https://github.com/kevinhelvig/FLYD-ii.

Première partie

Acquisition et génération de données Flying Spot

Chapitre 1

Collecte de données simulées et expérimentales en thermographie Flying Spot

Sur cet immense tableau d'une nuit céruléenne, la rêverie mathématicienne a écrit des épures. Elles sont toutes fausses, délicieusement fausses, ces constellations !

L'Air et les Songes, G. Bachelard

Sommaire

La thermographie laser *Flying Spot* (FST) est la méthode d'inspection non-destructive étudiée dans le cadre de cette thèse. Cette méthode consiste en l'échauffement local d'une surface par une source laser permettant de détecter des fissures. Bien qu'il y ait eu de nombreuses recherches concernant la mise en place de traitements manuels ajustés par les opérateurs voire semi-automatiques pour la détection de défauts surfaciques, il existe encore peu de développements des méthodes d'apprentissage automatique combinées avec cette approche d'examen dans la littérature. Comme discuté dans l'introduction, cela s'explique en partie par l'absence de travaux de collecte de données en quantité avec cette technique, nécessaires pour les méthodes d'apprentissage machine, a minima pour la validation des performances de détection d'endommagements. La technique demande en effet un certain degré d'expertise et de mise en place de matériel. L'autre raison est le fait que cette technique d'inspection FST est encore peu courante comparativement à d'autres méthodes de contrôle des matériaux comme la thermographie Flash ou l'inspection ultrason.

Dans ce chapitre, nous allons présenter les travaux effectués pour la collecte de bases de données de thermographie FST expérimentales et simulées afin de permettre l'entraînement de systèmes d'apprentissage machine pour la détection de fissures de surface.

Nous allons d'abord présenter la théorie de la FST dans la section 1.1. Une bibliographie focalisée sur les travaux de couplage entre l'inspection thermographique et les méthodes d'apprentissage est ensuite fournie dans la section 1.2. Nous présentons ensuite les bases de données qui ont été constituées dans la section 1.3, en passant en revue les échantillons disponibles, puis la génération d'images simulées par éléments-finis ainsi que les bases de données expérimentales constituées.

1.1 Éléments théoriques

Nous présentons dans cette section quelques bases théoriques concernant la méthode d'examen FST. Nous introduisons d'abord l'intérêt de l'observation de la chaleur au travers du spectre infrarouge puis la technique d'inspection FST sous l'angle théorique. Nous fournissons dans un troisième temps une bibliographie des développements de cette technique d'examen.

1.1.1 Intérêt de la thermographie infrarouge

La mesure par rayonnement infrarouge permet de mettre en exergue des phénomènes d'échauffement liés aux interactions entre les objets à des températures extrêmement variables. La température de tous les objets, au sens d'agitation des particules, induit ainsi un rayonnement mesurable. La plage d'émission occupée par ce rayonnement en fonction de la température interne d'un objet suit la loi de Planck, qui est le plus souvent approximée par la loi de Wien associée au modèle du corps noir. Cette loi décrit l'évolution de la longueur d'onde d'émission d'un corps en fonction de la température réelle de l'objet et est illustrée dans le schéma de principe Figure 1.1. Si les corps extrêmement chauds émettent massivement dans le spectre visible et l'ultraviolet, comme les étoiles bleues/blanches très massives, d'autres corps ont des températures bien plus réduites, dont le rayonnement va se décaler progressivement suivant la température, vers le rouge et l'infrarouge. Les gammes de rayonnement du visible-rouge vers l'infrarouge regroupent des corps encore très chauds comme les naines rouges de classe M à plus de 3.000 K. Les sources de chaleur modérées émettent essentiellement dans le rouge et dans le proche infrarouge, typiques du rayonnement émis par la mise en forme des pièces métalliques en forgeage ou des céramiques techniques en frittage. Nous trouvons ensuite des longueurs d'onde infrarouge beaucoup plus longues correspondant à une température de rayonnement bien plus réduite, pour les tissus vivants biologiques. Pour finir, les corps inertes non-vivants ont des rayonnements dépendant de la température des sources les chauffant et vont émettre dans des longueurs d'onde de plus en plus longues suivant leur température.

Les infrarouges sont classés suivant différentes bandes de travail. Les proches et courts infrarouges (NIR et SWIR, respectivement [0,75 µm, 1,4 µm] et [1,4 µm, 3,0 µm]) concernent des températures d'une à plusieurs centaines de degrés, comme dans les situations d'élaboration de pièces métalliques. Les longueurs d'onde MWIR ([3 µm, 8 µm]) et LWIR ([8 µm, 15 µm]) concernent des températures de rayonnement plus faibles, de quelques dizaines de degrés comme pour les tissus organiques par exemple. Nous pourrions étendre cette étude du lien entre la température d'émission et la longueur d'onde associée au-delà de l'infrarouge, jusqu'à des rayonnements micro-onde pour les températures d'émission les plus faibles (nuages de gaz stellaires ultra-froids, rayonnement cosmologique).

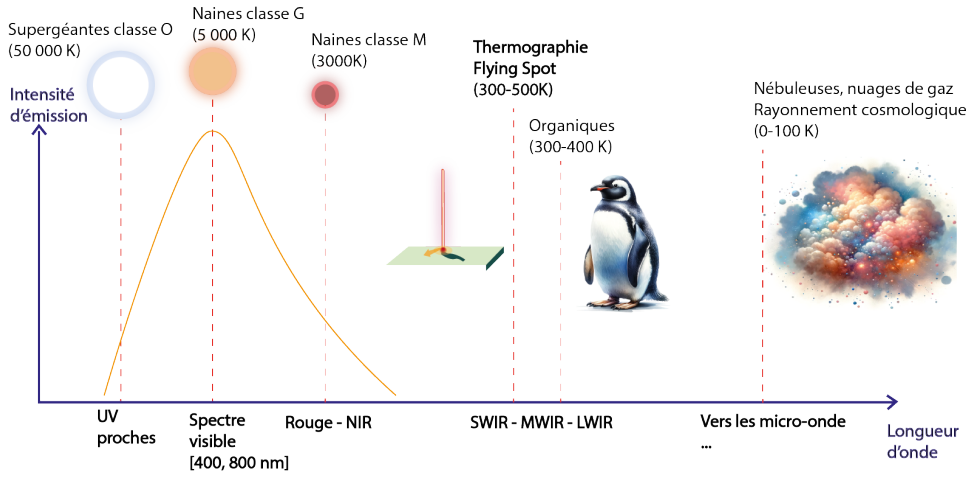


FIGURE 1.1 – Illustration de principe du lien température-longueur d'émission maximale, présentant de manière simplifiée la loi de Planck/de Wien pour différents types de corps afin d'intuiter la longueur de rayonnements de différents corps.

Le rayonnement thermique mesuré en spectre infrarouge peut donc nous informer sur les phénomènes physiques dans des plages de température modérée, de l'ordre du degré jusqu'à plusieurs centaines de degrés, qui sont observables en SWIR-MWIR-LWIR. Il permet, par exemple, l'observation de phénomènes comme les déperditions thermiques par rayonnement ou par convection de manière strictement passive dans un premier temps pour des domaines comme le bâtiment. Il permet aussi la détection d'objets lorsque la modalité visible n'est pas disponible, comme la détection de cibles de nuit, autant dans des contextes urbains que naturels. Néanmoins, il faut tenir compte des perturbations induites par les composantes liées à l'environnement, comme le volume d'air traversé par le rayonnement observé avant d'entrer dans le capteur, en particulier en infrarouge, où le spectre d'absorption de l'eau vaporisée dans l'air joue un grand rôle, suivant la bande passante choisie. L'équation 1.1, dite de la radiosité (ou de la radiométrie), montre comment les différentes contributions de l'environnement affectent la mesure de la luminance et donc de la température.

Par ailleurs, l'estimation de la température réelle à partir de la température observée et de la radiosité est difficile pour un corps donné. En effet, elle est confrontée aux propriétés rayonnantes spécifiques du matériau comme l'émissivité, et à l'absorption du milieu, variable suivant la longueur d'onde observée. Diverses solutions existent pour calculer la température réelle terme à terme à partir de la radiosité, mais elles nécessitent toutes d'employer des hypothèses simplificatrices dépendantes de la technique, et d'estimer les propriétés physiques du matériau étudié.

$$L(\lambda) = \epsilon(\lambda, T)L_{bb}(\lambda, T) + (1 - \epsilon(\lambda, T))L_{env}(\lambda) + \tau(\lambda)(L_{atm}(\lambda) - L_{env}(\lambda)) \tag{1.1}$$

où :

— $L(\lambda)$ est la luminance spectrale mesurée [W m^{-2} sr^{-1} m^{-1}].

— $\epsilon(\lambda, T)$ est l'émissivité de la surface à la température T et à la longueur d'onde λ [unité sans dimension].

— $L_{bb}(\lambda, T)$ est la luminance spectrale du corps noir à la température T [W m^{-2} sr^{-1} m^{-1}].

— $L_{env}(\lambda)$ est la luminance spectrale de l'environnement [W m^{-2} sr^{-1} m^{-1}].

— $\tau(\lambda)$ est la transmittance atmosphérique à la longueur d'onde λ [unité sans dimension].

— $L_{atm}(\lambda)$ est la luminance spectrale de l'atmosphère [W m^{-2} sr^{-1} m^{-1}].

L'observation infrarouge passive donne déjà des indications utiles sur des propriétés données des objets observés, suivant la plage de longueur d'onde d'observation. C'est le cas, par exemple, de la réflexion des surfaces métalliques polies. Cependant, c'est le passage à une thermographie active qui est le plus intéressant dans le contexte de l'étude des matériaux. En effet, pour un corps plutôt conducteur comme un acier ou un alliage d'aluminium qui est très diffusif, le rayonnement incident est souvent limité. Il est alors intéressant d'introduire une source de chaleur afin d'interagir avec la scène observée et d'exacerber la réponse thermique. Par exemple, la thermographie Flash propose d'utiliser la transmission d'un échauffement dans la profondeur d'un matériau, souvent des polymères et composites à matrice polymère, afin de révéler les caractéristiques internes du matériau, comme des délaminages ou des porosités invisibles en optique visible ou en infrarouge passif seul.

A contrario, pour un examen plus local de la surface, on peut utiliser un échauffement ciblé comme une source laser ou provoqué par induction pour révéler des propriétés de surface comme des fissurations, au travers du phénomène transitoire de l'échauffement : c'est l'interaction entre le matériau et le flux de chaleur parcourant la surface qui va alors être observée, en plus des phénomènes thermiques purement passifs. Ceci est le concept-cœur de la thermographie active par source laser. Une approche peut être la mise en place d'un signal thermique entrant dans le matériau présentant une variabilité de nature temporelle et fréquentielle : c'est l'approche par thermographie pulsée. Nous nous intéressons ici à un échauffement local évoluant spatialement sur la surface : c'est l'idée du balayage thermique à la base de la thermographie laser Flying Spot. Cette méthode à l'avantage de pouvoir être mises en oeuvre sur l'ensemble des matériaux métalliques, au contraire de la thermo-induction par exemple qui est réservée aux ferro-magnétiques.

Le retour à la température réelle est généralement difficile car dépendant de propriétés intrinsèques complexes à estimer précisément telles que l'émissivité. Néanmoins, cela n'est pas forcément nécessaire pour l'ensemble des problèmes : ainsi, ce travail de thèse se concentre strictement sur des champs de température observée afin de détecter des structures de surface telles que des fissurations. L'altération qualitative de la diffusion de la chaleur va former des discontinuités dans le flux de chaleur reçu qui vont suffire ici pour révéler la présence d'objets bloquant la diffusion de la chaleur en surface, comme des éléments géométriques fonctionnels, des changements d'alliage ou de revêtement, et surtout des défauts comme des fissures débouchantes.

La méthode d'inspection par thermographie laser Flying Spot repose donc sur le balayage d'un échantillon par une source de chaleur locale : l'apparition de discontinuités thermiques va pouvoir ensuite être interprétée comme étant un endommagement ou non, par un opérateur, humain ou non. La Figure 1.2 fournit une illustration de synthèse de cette méthode, avec un balayage réalisé proche du défaut et la discontinuité thermique induite.

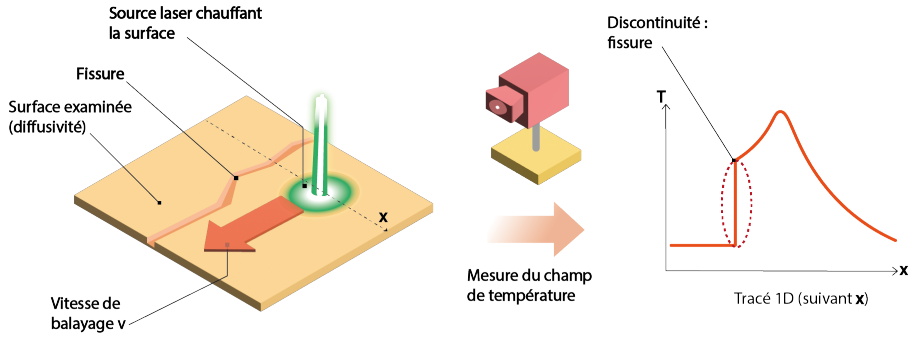


FIGURE 1.2 – Schéma de synthèse sur la méthode FST pour la détection de défauts sur une surface.

L'évolution du champ de température est observé à l'aide d'une caméra infrarouge. L'enregistrement thermique incident est généralement traité par reconstruction : nous sommons et normalisons toutes les images thermiques individuelles prises au cours d'un enregistrement de la température observée dans une zone d'intérêt pré-définie, nous donnant alors des cartographies thermiques reconstruites. Cette reconstruction en cartographies permet généralement de lisser des phénomènes transitoires liés à la diffusion de la chaleur sur la surface. La Figure 1.3 illustre ce passage de l'enregistrement thermique à la cartographie reconstruite. Il faut noter que la cartographie thermique est reconstruite dans le référentiel de la pièce étudiée, nous avons donc pré-estimé les vitesses de déplacement de la pièce afin de recaler ces images depuis le repère caméra vers le repère pièce.

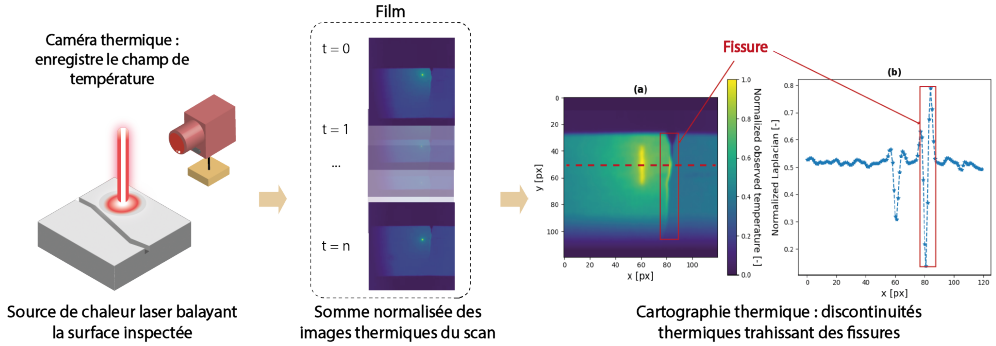


FIGURE 1.3 – Synthèse sur l'enregistrement des images thermiques individuelles puis leur reconstruction en une cartographie thermique.

La Figure 1.4 montre la discontinuité de chaleur liée à une fissure de surface, d'abord sur la cartographie thermique reconstruite acquise sur une éprouvette métallique, puis un tracé 1D du signal après filtrage Laplacien est fourni. Ce type de pré-traitement par cartographie puis application d'un filtre est typique en thermographie, permettant de réduire les dépendances liées à l'état de surface ou aux contours de la pièce étudiée.

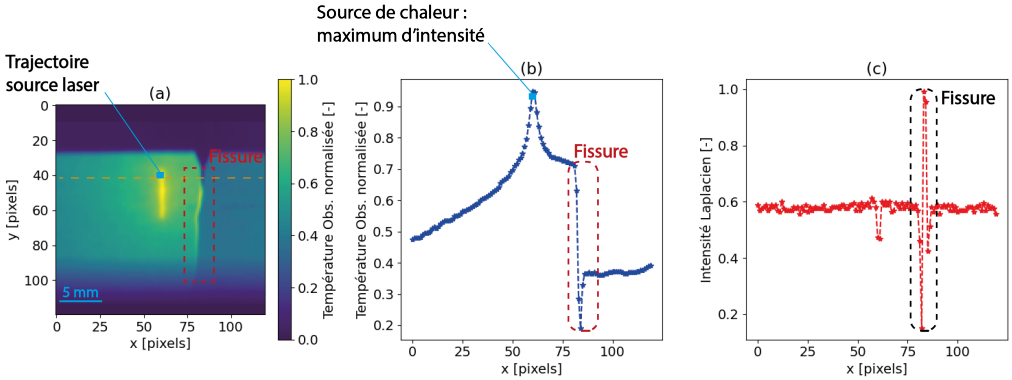


FIGURE 1.4 – (a) montre une cartographie thermique reconstruite. (b) correspond au tracé horizontal de la température observée normalisée [-] suivant la ligne orange. (c) est le tracé 1D après filtrage Laplacien : la discontinuité correspondant à la fissure est alors isolée.

1.1.2 Eléments théoriques sur la thermographie Flying Spot

Nous fournissons maintenant une description théorique de la thermographie laser Flying Spot, modulo quelques hypothèses de simplification. Le transfert de chaleur sur une surface homogène pendant le balayage de la source de chaleur laser peut être modélisé en utilisant l'équation d'advection-diffusion comme suit [1, 2] :

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = \frac{v}{\alpha} \cdot \frac{\partial T}{\partial y}. \tag{1.2}$$

Avec T [K] la température observée, v [mm/s] la vitesse de balayage, et α [mm²/s] la diffusion thermique, qui peut être considérée comme une propriété du matériau.

La densité du flux de chaleur entrant dans la surface peut être modélisée comme une distribution gaussienne correspondant à la région de chauffage par le faisceau laser qui n'est pas parfaitement ponctuel, suivant l'équation ci-dessous :

$$\phi = \frac{A \times P}{S} \times \exp\left(\frac{-(x - x_0)^2 - (y - y_0 - v.t)^2}{R_{\text{spot}}^2}\right). \tag{1.3}$$

Avec A [%] le taux d'absorption, P [W] la puissance, S [mm²] la surface irradiée. Les variables x et y représentent la position spatiale du point laser. x_0 et y_0 correspondent la position de départ du point laser.

Dans la thermographie Flying Spot, il est couramment supposé que la vitesse de balayage reste constante durant l'examen. De plus, notons que la diffusivité thermique du matériau est supposée quasi-constante également avec l'hypothèse des propriétés thermiques de surface homogènes. C'est une hypothèse simplificatrice qui est discutable sur le plan pratique dans l'industrie, par exemple pour des pièces présentant une surface très détériorée ou bien ayant une rugosité variable issue des propriétés intrinsèques du matériau, de l'usure de fonctionnement ou bien du dépôt d'oxydes protecteurs en surface par projection ou par procédés oxydo-réductifs. Dans le cadre de cette thèse, les examens sont réalisés pour de petites variations de température qui sont plus faciles à mesurer dans la

bande MWIR, avec une élévation de la température observée de l'ordre de la centaine de degrés au grand maximum. Des observations en LWIR seraient envisageables mais présentent le risque d'une saturation plus grande des images thermiques sur ces longueurs d'ondes. L'échauffement devrait a priori être plus important pour justifier le passage à une bande SWIR afin de vraiment détecter des discontinuités dans le flux de chaleur.

1.1.3 Le nombre de Péclet

Dans ce contexte, le travail théorique de Krapez [3] propose d'introduire le nombre de Péclet pour configurer les expériences de FST. Ce nombre sans dimension est défini comme le rapport entre le transfert de chaleur convectif et la diffusion thermique. Selon [3], la meilleure détection, comprise comme la discontinuité thermique la plus significative due à la fissure, correspond à un nombre de Péclet de 1. Pour une estimation grossière, la diffusion thermique est de l'ordre de 10^{-7} m^2/s pour l'ensemble des échantillons étudiés. La Figure 1.5 synthétise le nombre de Péclet et ses différentes composantes.

$$Pe = \frac{\text{Flux de chaleur convectif}}{\text{Diffusion de chaleur}} = \frac{v_{spot} \times R_{spot}}{\alpha}. \tag{1.4}$$

Avec v_{spot} [mm/s] : la vitesse de la source de chaleur balayant la surface, R_{spot} [mm] : la taille du point dû à la source de chaleur et α [mm^2/s] : la diffusivité thermique des matériaux.

Ce nombre est utilisé pour fixer les plages de paramètres de travail lors des examens thermographiques par FST. Si ce nombre adimensionnel est utilisé pour sa facilité d'emploi et pour son antériorité d'usage à l'ONERA, des limites sur cet optimum théorique sont soulevées dans la recherche actuelle en thermique des matériaux. D'autres critères existent dans la littérature comme le nombre de Fourier ou le nombre de Biot.

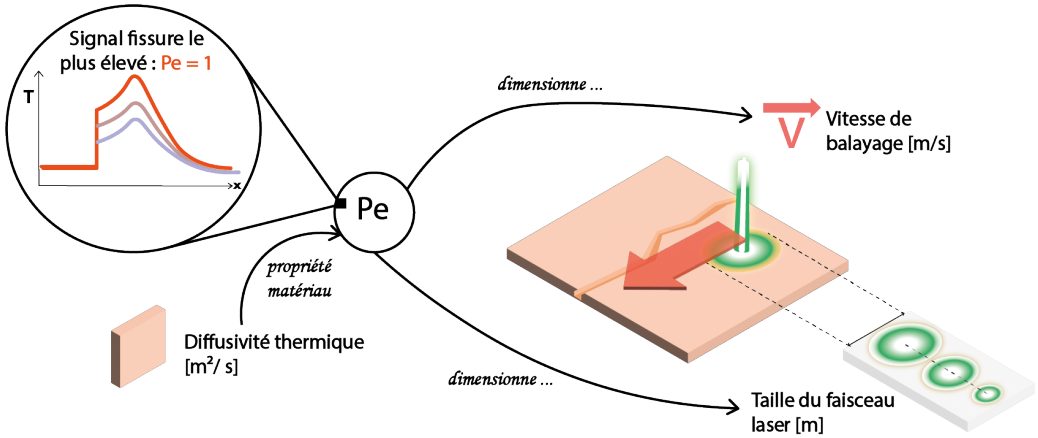


FIGURE 1.5 – Illustration de l'impact du Péclet pour régler les différents paramètres expérimentaux.

1.1.4 État de l'art de la thermogaphie Flying Spot

Afin d'examiner des pièces aéronautiques métalliques, la thermographie infrarouge Flying Spot (FST) peut être utilisée pour fournir des données locales pour une détection et une caractérisation précises des défauts tels que les fissures de surface. Il s'agit, comme nous l'avons vu, d'une technique de contrôle non destructif capable de détecter des défauts *via* le balayage des surfaces à examiner à l'aide d'une source de chaleur locale, ici laser. Les défauts tels que les fissures présentes sur les surfaces des pièces métalliques sont révélés par la perturbation de la propagation de la chaleur mesurée par une caméra infrarouge. La méthode tolère une non-planéité limitée de la surface testée. Cette technique d'examen a été initialement développée pour la détection de fissures dans les pièces d'avions militaires à la fin des années 1960 [4]. Dans [3], J.C. Krapez a réalisé un travail théorique clé sur cette technique, étudiant l'impact des paramètres expérimentaux sur la précision de la détection des fissures. Son approche propose un double passage traversant le défaut de type fissure : la cartographie différentielle entre l'aller et le retour va permettre de minimiser les signaux dus aux géométries de contours tout en maximisant le signal rendu par la cible. C'est l'approche de balayage conventionnelle. Cette technique d'examen par Flying Spot a fait l'objet de recherches importantes depuis lors à l'ONERA. Citons le travail de thèse conduit par T. Maffren [5], qui réalise une étude paramétrique simulée et réelle de l'inspection FST d'échantillons métalliques revêtus en plus de l'étude du comportement thermo-mécanique de ces pièces. Un autre article a quant à lui montré la capacité de détection des micro-craquelures de faïençage sur des échantillons céramiques par thermographie Flying Spot, comparativement à des examens par microscopie électronique [6]. Ces défauts formant un motif en réseau particulièrement difficile à détecter sans moyen de microscopie avancé, consolident l'intérêt de la FST pour la détection de défauts. La méthode par thermographie laser peut aussi fournir des informations sur les propriétés de surface de matériaux comme la diffusivité thermique [7].

En dehors de l'ONERA, nous pouvons citer des travaux plus orientés vers la simulation et la compréhension physique de ce moyen d'examen, en Flying Spot mais aussi en Flying Line [8, 9]. L'idée d'une approche Flying Line est d'utiliser un fibrage pour passer d'un échauffement laser local à une ligne laser permettant de scanner une région beaucoup plus large, bien que souvent à des puissances de chauffe plus limitées. Ces travaux permettent de mieux appréhender l'influence de chaque paramètre d'inspection sur le signal produit par une fissure dans le cadre de l'approche en double passage conventionnelle [3]. Un autre travail explore lui aussi l'influence des différents paramètres d'examen sur le signal thermique produit par la fissure en constituant une étude paramétrique très diversifiée de la FST [1]. Cette approche a été proposée pour le balayage de régions globales de pièces en contexte d'automatisation industrielle comme dans [10]. Citons enfin un autre article relativement récent étudiant en détail la technique thermographique et ses paramètres de fonctionnement comme la distance de balayage optimale pour maximiser la réponse [11]. La littérature récente montre que cette technique d'examen est encore en investigation et en recherche active.

En dehors de l'examen de surfaces industrielles, notons aussi l'emploi de la FST dans d'autres contextes, comme la conservation et la restauration d'œuvres d'art : elle y permet de déceler la formation de craquelures malgré la couche de peinture de surface [12].

1.2 Traitements des données thermographiques

Plusieurs travaux ont proposé un traitement semi-automatique des données de FST par un filtre de Canny pour la reconstruction de cartographies en vue de détecter visuellement des fissures sur des aubes de turbine [13]. D'autres travaux proposent des méthodes d'automatisation de la détection des défauts par filtrage, mais ceux-ci sont difficiles à généraliser depuis des surfaces homogènes simples vers l'examen de pièces réelles complexes [14].

A contrario, les méthodes d'apprentissage profond peuvent fournir des méthodes de détection automatiques plus génériques et robustes, capables d'extraire des informations contextuelles des données. Des exemples d'utilisation de l'apprentissage profond sont disponibles dans la littérature pour d'autres méthodes d'examen que la FST, comme pour l'inspection thermique passive pour la surveillance des processus de fabrication additive [15], ou depuis la caractérisation automatique des matériaux [16]. Au démarrage de la thèse, il existait peu de travaux explorant le couplage entre l'apprentissage profond et la thermographie laser Flying Spot, à l'exception de la publication [17] qui proposait d'utiliser des architectures complexes de réseaux neuronaux récurrents (RNN) basées sur l'apprentissage profond pour gérer les caractéristiques temporelles lors d'un échauffement laser. Les RNN sont des réseaux neuronaux spécialisés dans le traitement des données séquentielles, comme les séries temporelles [18]. Ce type d'architecture peut néanmoins être difficile à entraîner et à annoter par des experts. Un autre travail propose une approche plus précise pour la super-résolution des données FST, c'est-à-dire la synthèse d'une image à haute résolution spatiale permettant de mieux identifier un défaut [19].

Nous constatons que le nombre de travaux sur l'association de l'apprentissage profond et de la thermographie laser Flying Spot est ainsi relativement limité. Cela s'explique par la difficulté à produire des données sur des expérimentations contraignantes à la fois en termes d'expertise humaine, de matériel et d'échantillons disponibles. L'autre raison à ce manque dans la recherche est aussi tout simplement le caractère encore rare de ce moyen d'examen.

C'est pourquoi nous avons travaillé à la constitution d'ensembles de données permettant l'entraînement de modèles d'apprentissage machine. La suite du chapitre présente les différents aspects associés à cette collecte de données pour la thermographie Flying Spot, depuis la simulation jusqu'aux données expérimentales.

1.3 Constitution des bases de données

Cette section présente l'agrégation des différentes bases de données expérimentales et simulées à partir des moyens d'essais de l'ONERA. Les sous-sections 1.3.1 et 1.3.2 discutent les choix expérimentaux réalisés, d'un côté notre proposition de thermographie FST en simple passage et de l'autre l'extension de la plage des paramètres de travail autour du Péclet optimal de 1. La sous-section 1.3.4 présente les échantillons métalliques disponibles pour cette étude. La sous-section 1.3.5 décrit la formation des volumes de données simulées par éléments finis. Enfin, les sous-sections 1.3.6 et 1.3.7 présentent les bases de données réelles acquises, sur échantillons d'essais mécaniques simples puis sur échantillons complexes revêtus.

1.3.1 Intérêt d'une thermographie simple-passage

L'approche la plus courante en thermographie laser Flying Spot est la thermographie dite en double passage, telle que décrite dans l'article originel de Krapez [3]. Celle-ci consiste en un double balayage, aller puis retour, idéalement orthogonal au défaut. On peut ensuite procéder à la soustraction des deux cartographies correspondantes. Cela permet de réduire le signal produit par les contours et les hétérogénéités de surface, tout en maximisant la réponse thermique liée au défaut traversé. Ce traitement simplifie la détection d'un potentiel endommagement en maximisant la réponse due au défaut tout en minimisant la réponse associée aux autres éléments de surface.

Cette approche, si elle a ses avantages dans un contexte de détection et d'identification manuelle par opérateur humain du défaut, n'est pas forcément idéale dans le cadre d'une détection automatique. En effet, de nombreux a priori sont à fournir par l'opérateur : un minimum d'information sur la localisation et l'orientation du défaut est nécessaire pour réaliser le double passage. De plus, la différence entre les deux cartographies pose des problèmes de ré-alignement au pixel près, qui sont en général ajustés manuellement afin que la différence ne produise pas d'artefacts superflus. Suivant les moyens d'acquisition, ces ajustements peuvent être très coûteux en temps.

Dans un contexte d'automatisation, nous proposons donc de procéder à une thermographie laser Flying Spot en simple-passage. Notre approche est agnostique à l'orientation du défaut, ce qui réduit en partie la dépendance à des informations fournies par l'opérateur humain. Plus précisément, nous réalisons des balayages ne traversant pas la fissure avec des orientations variables. Le fait de ne faire qu'un seul balayage aller divise par deux le temps d'examen. De plus, cette approche non traversante peut permettre de suivre l'ouverture de fissuration et de mieux caractériser sa longueur. L'inconvénient de l'approche est l'augmentation des perturbations dues aux géométries de surface et autres sources d'artefacts produits notamment par les contours de la pièce, qui vont présenter un signal du même ordre de grandeur que celui induit par la fissure.

La Figure 1.6 illustre une cartographie obtenue en double passage suivant l'approche conventionnelle [3] tandis que la Figure 1.7 fournit une cartographie en simple passage. Les deux images correspondent au même échantillon. Nous voyons que la première carte donne un signal fort pour le défaut mais au prix d'une phase de pré-traitement manuel et d'ajustements des cartographies aller et retour coûteux en temps. La deuxième cartographie est obtenue directement mais présente un signal lié au contour de la pièce plus important.

Pour résoudre cette difficulté nous proposons l'emploi de l'apprentissage profond, qui s'ajuste généralement bien à ce type de contexte.

1.3.2 Extensions des paramètres d'acquisition

Outre l'acquisition en un seul passage, nous proposons également un mode d'acquisition dit élargi, dans lequel les paramètres de l'expérimentation, tels que la puissance du laser et sa vitesse, sont plus éloignés de l'optimum théorique indiqué par le nombre de Péclet. De plus, les balayages sont effectués sur une distance assez large par rapport au défaut, entre 0 et approximativement un centimètre de la cible. Cela permet d'augmenter la variabilité des données. Par ailleurs, cela s'inscrit dans notre volonté de limiter les contraintes expérimentales et l'ajout d'a priori opérateurs. Enfin, des acquisitions loin de

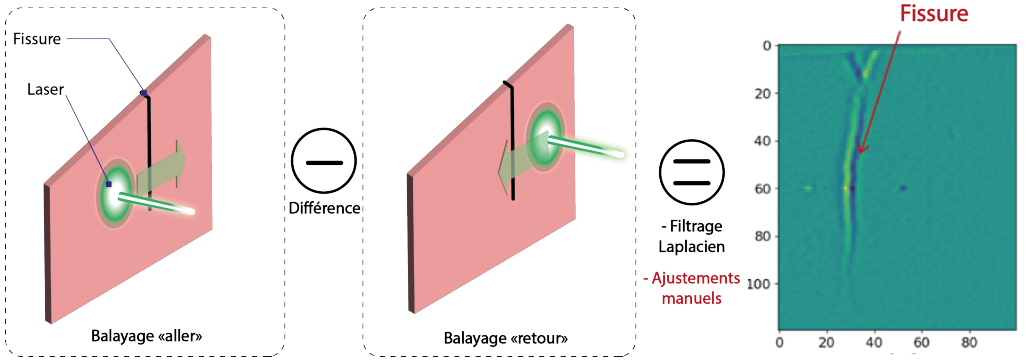


FIGURE 1.6 – Filtrage Laplacien appliqué à la soustraction des cartographies thermiques reconstruites aller et retour, pour un double-balayage traversant le défaut.

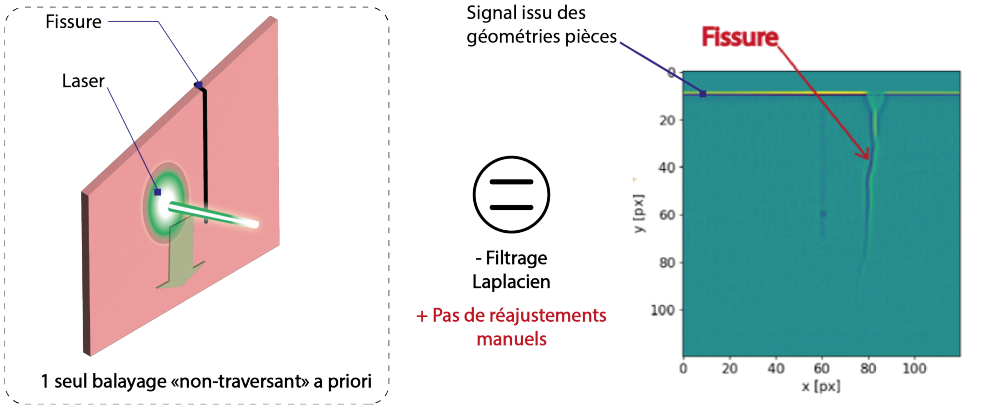


FIGURE 1.7 – Filtrage Laplacien appliqué directement à une cartographie thermique reconstruite, pour un balayage non-traversant.

la fissure ouvrent la voie pour une étude de la capacité de détection en fonction de la distance à la fissure.

Pour cela, nous allons grandement jouer sur les paramètres d'acquisition exprimés dans le nombre de Péclet. Ainsi, si la plage du Péclet reste autour du point de fonctionnement de 1, nous allons avoir des variations dans la vitesse de balayage, dans la taille de faisceau, en plus des variations de propriétés thermiques de surface liées aux matériaux étudiés (rugosité, diffusivité de la surface du matériau). Nous jouons particulièrement sur la distance entre la fissure et la trajectoire de balayage pour nos inspections. Cela permet de reproduire des signaux dus à l'endommagement d'une grande variabilité, par exemple qui seraient extrêmement ténus, voire de mener des pré-détections exploitant l'information infrarouge passive pure pour des fissurations très loin du faisceau, i.e. au-delà du centimètre. Ces choix permettent d'entraîner des modèles d'apprentissage robustes à une grande variété de réponses thermiques dues au défaut sur le jeu d'échantillons disponible ici, ce qui se justifie bien avec notre proposition de système de détection automatique polyvalent minimisant l'influence d'a priori humains dans l'inspection. Nous pouvons ainsi penser à un contexte purement exploratoire, où la position de l'endommagement serait

complètement inconnue, ou bien où les propriétés des matériaux seraient peu/mal connues (pièce après vieillissement en environnement inconnu, diffusivité de surface inconnue). Le protocole général d'acquisition est illustré avec la Figure 1.8, avec les deux régimes de données de thermographie laser : l'un passif à moyenne et grande distance du défaut pour une pré-détection du défaut à partir de l'amorce et du contexte image. L'autre régime peut être qualifié d'actif, proche du défaut où la discontinuité thermique est la plus facile à mesurer et permettant une bonne caractérisation du défaut.

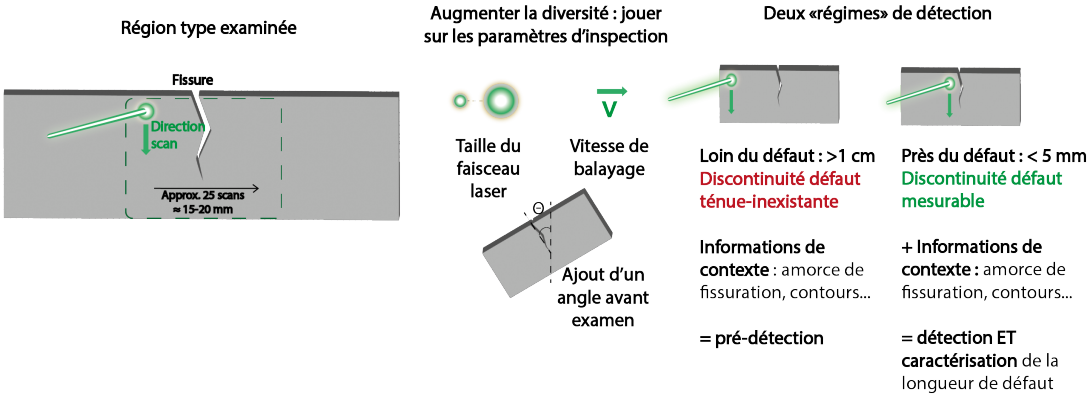


FIGURE 1.8 – Principes généraux des protocoles d'acquisition en vue de maximiser la quantité et la diversité des données disponibles.

En résumé, toutes les acquisitions présentées dans cette thèse sont des acquisitions à un seul passage ne traversant pas l'endommagement observé. L'acquisition de données a été réalisée pour constituer un détecteur de thermographie laser minimisant les a priori des opérateurs grâce à l'apprentissage machine, capable d'exploiter à la fois l'information passive à longue distance du dommage ainsi que les phénomènes thermiques liés au balayage laser et à la discontinuité de chaleur à moindre distance de la cible.

1.3.3 Présentation du banc d'essai Flying Spot

Le banc de FST de l'ONERA est illustré Figure 1.9. Le laser travaille à une puissance fixée entre 0,5 et 3 W. La longueur d'onde est de 532 nm. Une lentille dichroïque permet de renvoyer la source de chaleur vers la surface de l'échantillon, qui est fixé sur un support motorisé permettant sa translation dans les 3 directions. La lentille laisse passer le rayonnement infrarouge vers la caméra infrarouge MWIR refroidie mesurant le champ thermique dans une plage infrarouge entre 3 et 5 microns, ce qui permet de suivre l'évolution de l'échauffement sur la surface observée. Ce choix de bande de travail permet d'étudier des échauffements relativement faibles, de l'ordre de la dizaine de degrés, sur des échantillons métalliques en environnement fermé. Les vitesses de balayage disponibles sont comprises entre 0,5 et 2,5 mm/s. La taille du faisceau laser est réglable et des faisceaux larges tendent à fournir de meilleures détections loin de l'endommagement, dans un contexte où l'on limite les a priori sur l'emplacement précis de la fissure-cible. Ce paramètre est néanmoins difficile à estimer précisément : le *tempo* de la diffusion de la chaleur dépend en effet des propriétés de surface, pouvant rendre la mesure précise du rayon de

la Gaussienne du laser assez complexe. Nous avons la plage suivante pour l'estimation de la taille de spot, entre 0,5 et 1,5 mm.

Deux objectifs optiques sont utilisés dans cette recherche : un objectif de 25 mm équipé d'un auto-focus. C'est ce premier qui est le plus utilisé car facile à mettre en œuvre. Sa résolution spatiale est de 200 microns. Le deuxième objectif est un objectif dit G1 250 mm Microscope : il ne dispose pas d'un auto-focus et son champ d'observation est très réduit, compliquant alors son utilisation.

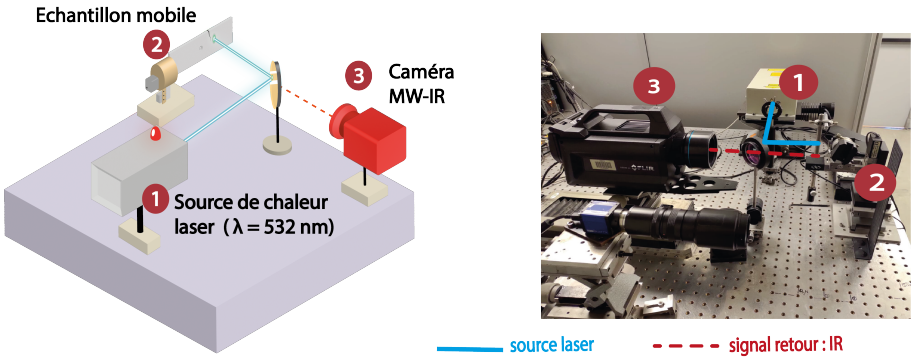


FIGURE 1.9 – Illustration du banc d'essai FST de l'ONERA.

1.3.4 Échantillons disponibles

Tous les échantillons examinés durant la thèse sont des échantillons métalliques, présentant ou non un revêtement. Ils sont présentés par degré de complexité croissant et non par ordre de disponibilité chronologique. Des campagnes de récupération et de collecte de rebuts d'essais mécaniques ont été nécessaires.

Échantillons simples : éprouvettes de fatigue. Le premier grand jeu d'échantillons est constitué d'éprouvettes d'essais mécaniques fournies par des groupes de recherche en matériaux de l'ONERA, notamment d'essais de fatigue/traction n'allant pas jusqu'à rupture. Les échantillons sont constitués en matériaux variables pré-entaillées puis fissurés de façon variable suivant l'essai mécanique conduit. Ils sont généralement constitués en acier ou alliages similaires. La réponse au transfert thermique est ainsi assez proche de la réponse des super-alliages série AM. Si les échantillons ne présentent pas d'oxydation protectrice, ils présentent une certaine variété d'états de surface plus ou moins rugueux. Ce premier lot d'échantillons présente ainsi une diversité assez grande de faciès de fissuration mécaniques, allant de la simple ouverture, très difficile à repérer, jusqu'à la fissuration très profonde et proche de la rupture. Il est à noter que ce jeu d'échantillons n'a été disponible qu'en fin de première moitié de thèse.

La Table 1.1 réunit l'ensemble des échantillons pour cette catégorie de pièce, avec les longueurs de défauts associées. Il faut noter que la numérotation est héritée des essais mécaniques d'origine et ne suit donc pas la longueur de fissuration. Un échantillon sans fissure est dénommé 11. Celui-ci nous permet de produire des données de contrôle, chaque coin fournissant une région d'examen. Cette pièce présente de plus certains coins oxydés et d'autres dépolis, ce qui ajoutera à la diversité du groupe de contrôle. Ces coins sont distingués par des lettres allant de 11A à 11D.

Échantillon	1	2	3	4	5	6	7	8	9	10
Longueur de fissuration mesurée [mm]	6	8	7	13	9	15	11	8	17	23

TABLE 1.1 – Longueurs de fissurations mesurées à l'aide d'une caméra visible sur 10 éprouvettes issues d'essais mécaniques.

La Figure 1.10 montre quelques exemples de ces échantillons, avec la Figure 1.11, illustrant la diversité de fissuration disponible sur une surface homogène et aux contours géométriques simples. Sur l'éprouvette 2 le défaut est ainsi pratiquement invisible à l'oeil nu, en dehors de l'indentation d'amorçage. L'éprouvette 4 est quant à elle nettement fissurée.

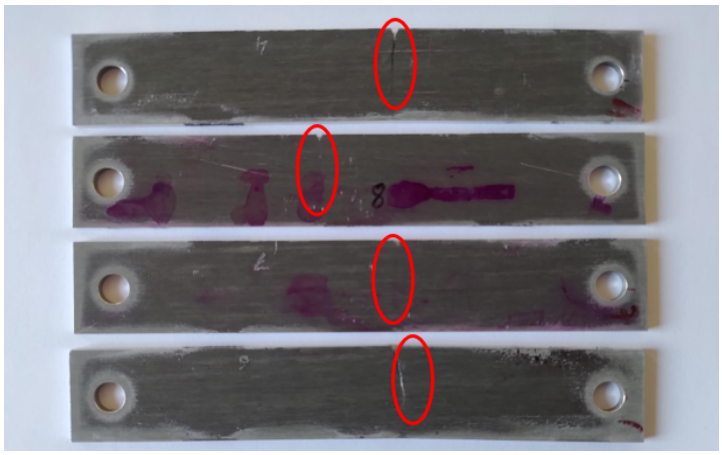


FIGURE 1.10 – Exemples d'échantillons de la base FLYD constituée d'échantillons d'essais de fatigue.

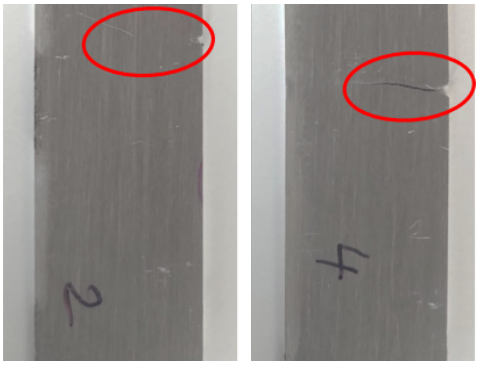


FIGURE 1.11 – Vue rapprochée des échantillons 2 et 4. Ces deux exemples illustrent la diversité des faciès de fissuration.

Les échantillons sont distribués de la manière suivante : les pièces 1 à 5 puis 8, 9, 10, et 11A à 11 C nous donnent le jeu de pièces d'entraînement, le jeu de test sera constitué des échantillons 6 et 7 ainsi que 11D.

Échantillons complexes et revêtus. Grâce à la collaboration entre l'ONERA et un acteur industriel, nous disposons d'échantillons d'application industrielle. Ce sont des pièces fonctionnelles présentant différents endommagements suite à l'usure de fonctionnement. Le matériau est un superalliage monocristallin de la série AM : un type d'alliage métallique particulier typique des pièces devant supporter des régimes de fonctionnement à haute température. Un dépôt de revêtement thermique par projection, ici une barrière thermique lamellaire en zircone yttriée, est de plus réalisé afin d'améliorer encore le comportement thermo-mécanique de cet organe en fonctionnement, particulièrement pour des régimes à températures et à sollicitations mécaniques très élevées. Nous disposons ici de 5 échantillons présentant une fissure et de 5 échantillons sains.

L'endommagement est ici formé suite à l'usure de la pièce mécanique en fonctionnement : l'abrasion ou l'impact lié à des particules dans un milieu de fonctionnement extrême sur le plan thermique et mécanique va fragiliser le composant, amorçant alors une fissuration qui va pouvoir s'ouvrir au cours du temps jusqu'à devenir critique et menacer le bon fonctionnement du système. Sur le plan de l'examen, cette surface particulière va présenter différents états des régions partiellement revêtues par la projection de la barrière

thermique. D'autres régions vont présenter un état de surface beaucoup plus rugueux, ce qui est dû à l'hétérogénéité du dépôt et à son propre état de surface après consolidation et fonctionnement, ce qui rend la diffusion de chaleur sur cette partie de la surface complexe à appréhender. Si l'amorce de la fissuration est détectable en spectre visible, le revêtement déposé cache généralement la fermeture de la fissure observée. Plus de détails sur les aspects matériaux sont fournis dans un des travaux de thèse antérieurs conduits à l'ONERA et travaillant sur des échantillons comparables [5]. La Figure 1.12 fournit un exemple d'acquisition en spectre visible et en cartographie thermique reconstruite de ce type d'échantillons. Les fissurations sont de l'ordre du millimètre en termes de longueur avec une plage large allant de 1 à 7 mm, mesurée en microscopie optique visible.

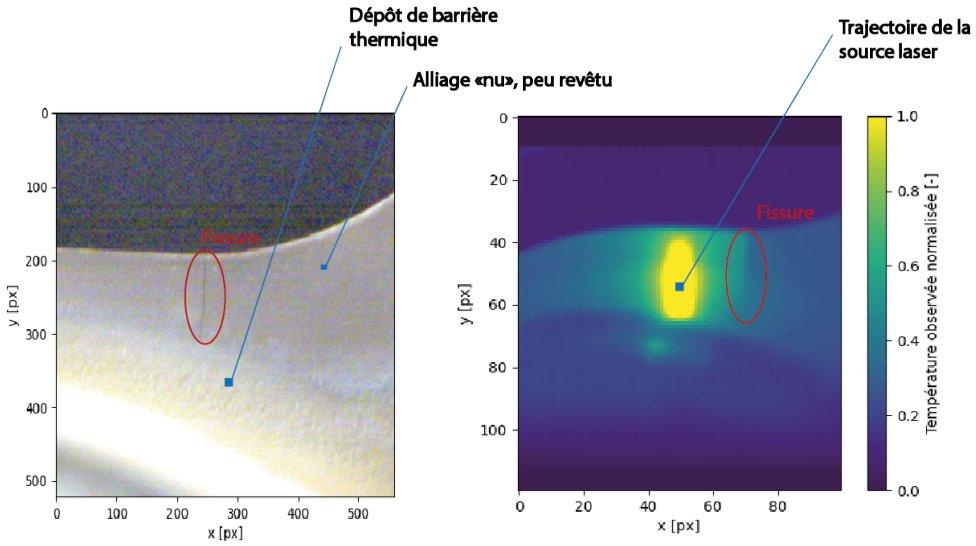


FIGURE 1.12 – À gauche, prise de vue haute définition en spectre visible d'un échantillon métallique complexe. À droite, cartographie thermique reconstruite sur un échantillon similaire.

Une phase importante de ce travail de thèse a été focalisée sur la collecte de bases de données, un élément indispensable dans l'apprentissage machine pour une évaluation statistique propre des performances des algorithmes. Les différentes bases de données compilées durant la thèse sont regroupées et décrites ici. Pour chaque famille d'échantillons, la plage des paramètres d'inspection est élargie comme décrit plus haut, de manière à augmenter la quantité de données sur une base d'échantillons limitée. Nous jouons aussi sur la distance avec le défaut ciblé, l'éloignement pouvant aller jusqu'au centimètre, ce qui amène l'objet au bord du champ pour une inspection FST. Cela permet notamment de rendre l'algorithme d'apprentissage sensible à des défauts dont la réponse est bien plus réduite, ou bien dans le cas d'un balayage exploratoire de la surface sans pré-localisation a priori par l'opérateur humain.

1.3.5 Données simulées par logiciel éléments-finis

Les éléments théoriques vus dans les sections 1.1 et 1.1.3 ont été introduits dans un simulateur par éléments finis. Cette simulation permet de comprendre la physique

Paramètre	Valeur
Longueur de fissure [mm]	5 ou 10
Orientation [-]	0 à 90
Vitesse de scan [mm/s]	0.5 à 2.5
Taille de spot [mm]	0.5 à 1.5
Distance défaut-source [mm]	0 à 10
Diffusivité Thermique [m^2/s]	$7.1e^{-7}$
Maillage	Triangulaire

TABLE 1.2 – Paramètres de base utilisés pour toutes les données simulées par éléments-finis.

de la FST. Nous l'utilisons pour produire des données de simulation. Pour cela, nous avons utilisé le logiciel COMSOL pour simuler des scènes de thermographie laser Flying Spot [20].

Le cadre de ce modèle est basé sur les éléments finis : cette approche par interpolations d'estimations locales d'un problème physique permet de modéliser les phénomènes complexes sans solution exacte évidente, comme dans la diffusion de chaleur en présence d'un défaut en 2D. La première simulation, déjà implémentée à l'ONERA avant le début de la thèse, correspond à l'implémentation de la loi d'advection, ce qui permet ensuite de reconstruire l'évolution du champ de température observée. Cette simulation est simple et ne présente que la trajectoire de l'échauffement ainsi qu'un défaut, qui est modélisé par une fente d'air parfaitement rectiligne. Les paramètres de la source de chaleur laser sont sa puissance, son diamètre de faisceau, et les informations définissant la trajectoire de balayage (point d'origine, longueur de scan). Les simulations suivent les variations de paramètres décrites dans la Table 1.2.

Des limites ont été identifiées pour cette première simulation, telles que la simplicité de la représentation du défaut. De même, si l'on peut modéliser partiellement des phénomènes comme la rugosité ou les propriétés de surface dans des paramètres globaux comme la diffusivité, cela ne rend pas réellement compte des variabilités de surface locales ou de phénomènes de diffusion comme l'apparition de *Speckles* dus à la rugosité de surface locale. C'est pourquoi nous avons implémenté une amélioration de cette simulation en ajoutant des contours et des régions d'air fictives à la surface observée, illustrées respectivement dans les Figures 1.13 et 1.14.

Ces deux blocs de données synthétiques forment alors la base FEM 1, ne présentant pas de contours, et la base FEM 2, qui présente un contour simple rectiligne.

Les principes de production de ces bases simulées sont illustrés dans la Figure 1.15. Nous y retrouvons les deux types de simulation (homogène puis avec contours simplifiés). Un récapitulatif des paramètres de la simulation est également fourni.

1.3.6 Acquisitions sur les échantillons simples

La base de données FLYD correspond à la base de données collectée sur les éprouvettes d'essai mécaniques. Cette base présente des acquisitions en FST simple passage pour des fissures assez diverses, mais sur les surfaces simples typiques des éprouvettes d'essais mécaniques, aux propriétés de surface homogènes et sans contours ou éléments

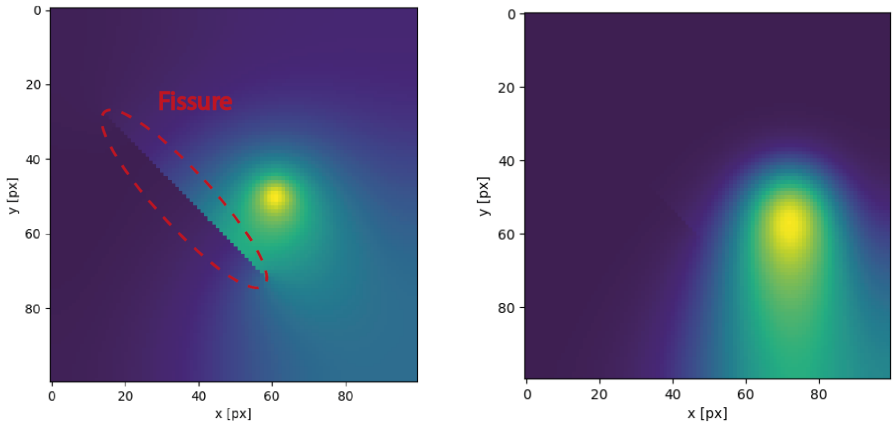


FIGURE 1.13 – Exemples d'images thermiques simulées par éléments finis, base FEM 1.

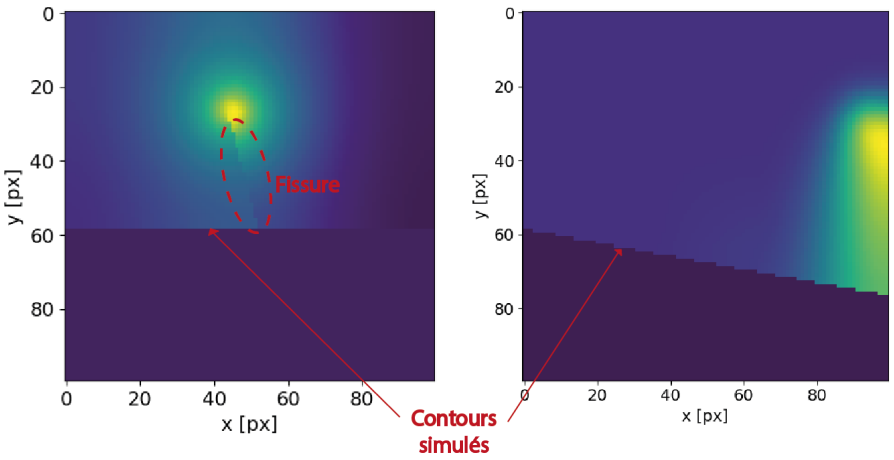


FIGURE 1.14 – Exemples d'images thermiques simulées par éléments finis, base FEM 2.

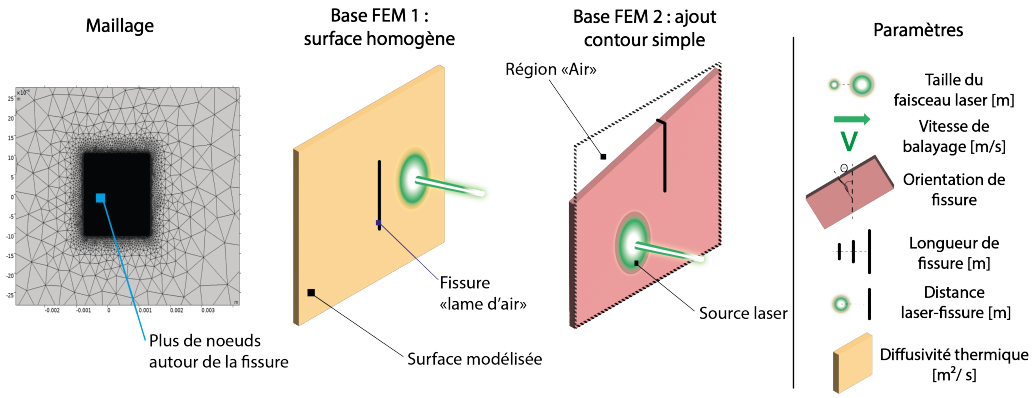


FIGURE 1.15 – Principes généraux des bases de données simulées et récapitulatif des paramètres qui sont modifiés pour la synthèse.

de surface complexes. Les paramètres d'examen sont élargis de manière à maximiser le nombre d'acquisitions réalisées, fournissant alors une base d'enregistrements thermiques

relativement large. Les balayages sont ainsi effectués pour deux vitesses d'examen, 0,5 mm/s et 1,5 mm/s. La taille du spot est fixe à 1,5 mm. Le protocole d'acquisition est illustré dans la Figure 1.16. Nous jouons à la fois sur la position de départ du scan et sur la région examinée pour augmenter encore le volume de données disponible et la diversité des caractéristiques produites.

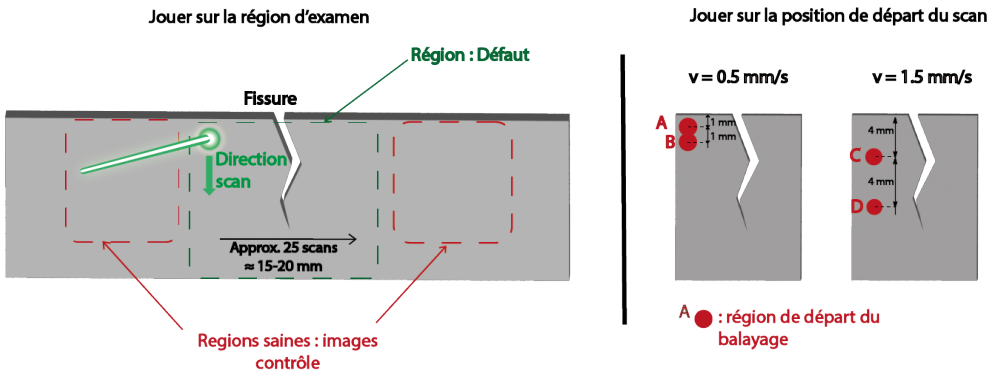


FIGURE 1.16 – Illustration du protocole d'acquisition appliqué pour constituer la base de données FLYD.

Ces enregistrements sont illustrés eux-aussi Figures 1.17 et 1.18, qui fournissent des exemples d'images thermiques individuelles, au cours d'un enregistrement.

Cette base de données a été mise en ligne en libre accès pour contribuer à pallier le manque de données FST disponibles pour la communauté.

La base de données nommée **FLYD-D** pour FLYD-Detection correspond à la base FLYD avec des annotations de localisation d'objet, c'est-à-dire une boîte englobante encadrant la fissure. Ces annotations sont produites avec le logiciel Label-Studio [21]. La Figure 1.19 illustre cette base de données avec un panel de détections de fissures automatiques réalisées sur balayages thermiques reconstruits.

La base de données **FLYD-Frames** est constituée en sélectionnant de manière aléatoire les films thermiques afin d'isoler des images thermiques individuelles en lieu et place de cartographies thermiques reconstruites. Le nombre de données est restreint à 1000 images thermiques individuelles dans cette base de données afin de rester sur un volume similaire à FLYD. Elles sont disponibles en très grandes quantités : pour donner un ordre de grandeur, nous passerions de quelques milliers d'enregistrements thermiques à plusieurs dizaines de millions d'images thermiques individuelles, sur l'ensemble du travail de thèse et même en éliminant les déplacements sub-pixelliques. Ces images thermiques individuelles sont intéressantes car elles peuvent être plus difficiles à interpréter pour un modèle d'apprentissage. Les cartographies thermiques présentent l'intérêt de mitiger certains phénomènes transitoires perturbateurs liés aux variations de l'état de surface par exemple. La Figure 1.20 illustre cette base de données avec un panel de détections de fissures automatiques réalisées sur ces images thermiques individuelles.

1.3.7 Acquisitions sur les échantillons complexes

Les bases de données **DAWN** ont été acquises à l'aide d'un objectif 25 mm sur les échantillons revêtus fournis par l'industriel. Ces bases sont constituées à partir de 3 échan-

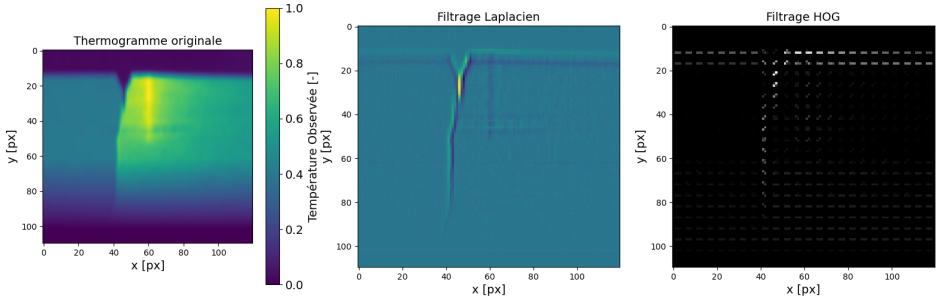


FIGURE 1.17 – Exemple de cartographie thermique reconstruite présentant un défaut, acquise sur un échantillon d'essai mécanique. De gauche à droite, cartographie, filtrat Laplacien et histogrammes de gradients orientés (HOG) associés.

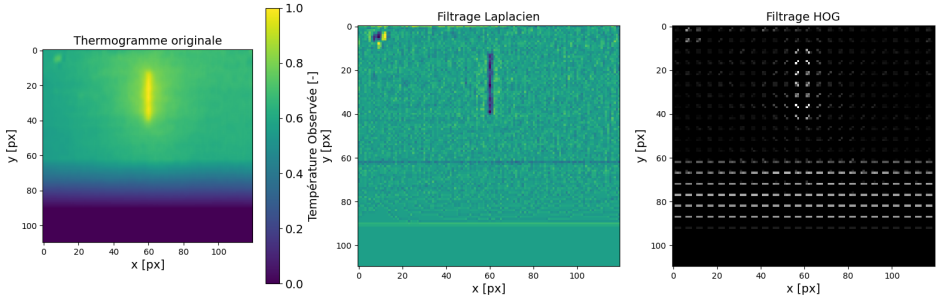


FIGURE 1.18 – Exemple de cartographie thermique reconstruite saine, acquise sur un échantillon d'essai mécanique. De gauche à droite, cartographie, filtrat Laplacien et histogrammes de gradients orientés (HOG) associés.

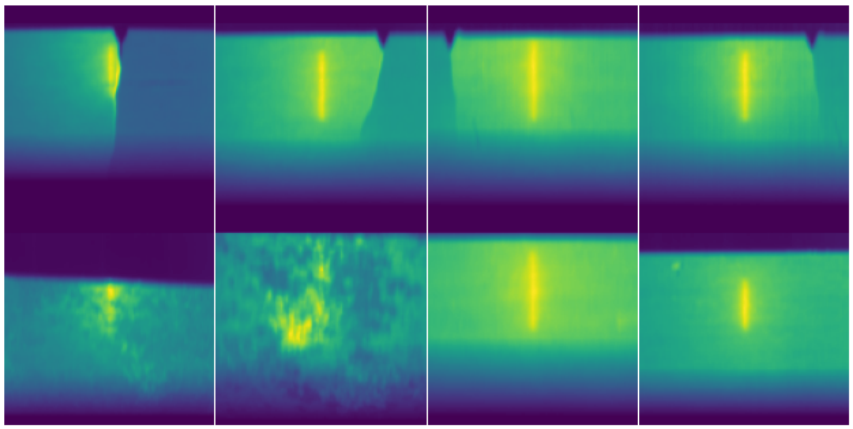


FIGURE 1.19 – Mosaïque de cartographies thermiques constituant la base de données FLYD. Les positifs avec fissures sont sur la première ligne, les négatifs en dessous.

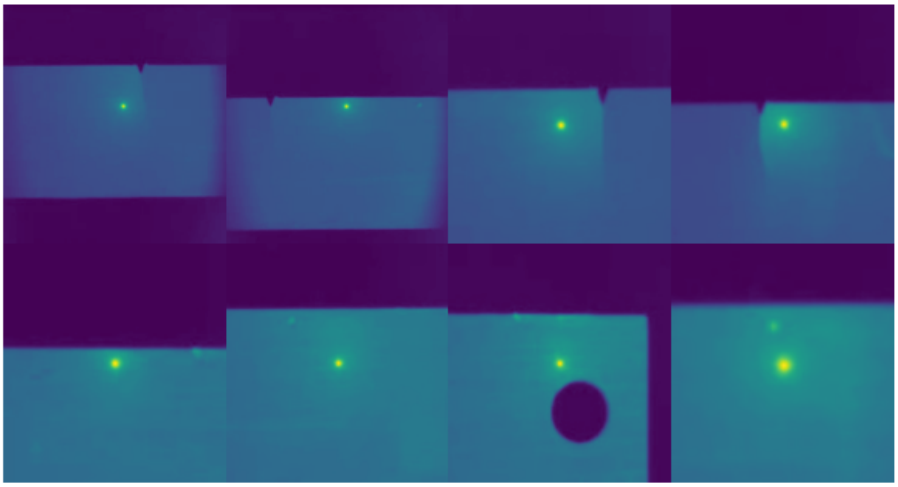


FIGURE 1.20 – Mosaïque d'images thermiques individuelles constituant la base de données FLYD-Frames. Nous pouvons voir la position du spot laser en temps réel, en lieu et place d'une trajectoire de la source chauffante.

tillons présentant une fissure et 3 sains. Les films utilisés sont obtenus à partir de balayages aux paramètres élargis par rapport à l'optimum théorique, comme pour FLYD. De plus un angle aléatoire compris entre 0 et 45 degrés est ajouté lors de certains balayages permettant d'augmenter encore la diversité dans les enregistrements thermiques. La vitesse de balayage couvre une plage de 0.5 à 2.5 mm/s (pas de 0.5 mm/s). Les balayages peuvent ensuite être synthétisés en cartographies thermiques reconstruites pour l'entraînement de modèle. Ces enregistrements sont illustrés Figures 1.21 et 1.22, qui montrent chacune une cartographie thermique reconstruite. L'une présente un défaut et l'autre est saine. On constate que le défaut n'est pas très difficile à identifier à l'oeil, néanmoins des sources de perturbations potentielles sont identifiées. Les artefacts liés par exemple à la réflexion non-souhaitée de la source de chaleur sur une surface non parfaitement plane, la discontinuité entre le revêtement protecteur et le matériau plus dénudé ... Un premier jeu très large de données est constitué contenant approximativement 900 cartographies thermiques et appelé **DAWN-J**. Les cartographies peu éclairées, avec trop d'artefact sont expurgées pour former deux jeux de données. Le deuxième jeu de données est appelé **DAWN-E** et contient 600 reconstructions en cartographie thermiques, avec une bonne variété d'angles. La base de données désignée par **DAWN** dans le reste de l'étude est un raffinement supplémentaire de cette base, via un tri aléatoire et une restriction plus fine du Peclet autour de 1, contenant 300 cartographies thermiques reconstruites.

DAWN est utilisée pour l'entraînement classique d'architectures d'apprentissage automatique. DAWN-E sera surtout utilisé pour la génération de synthétiques en utilisant les modèles de diffusion. DAWN-J sera utilisé pour évaluer les performances des systèmes d'apprentissage face à des acquisitions dégradées.

La Figure 1.23 illustre cette base de données avec un panel des différentes cartographies

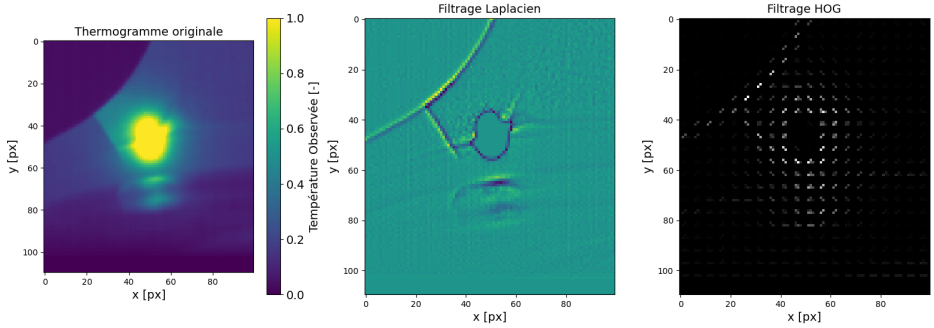


FIGURE 1.21 – Exemple de cartographie thermique reconstruite présentant un défaut, acquise sur un échantillon complexe. De gauche à droite, cartographie, filtrat Laplacien et Histogrammes de gradients orientés (HOG) associés.

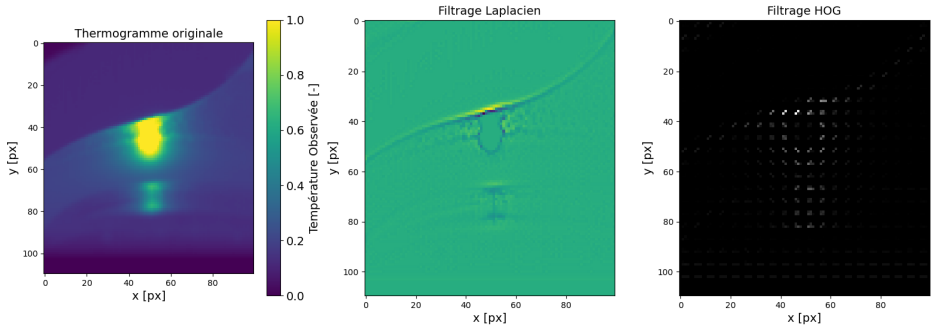


FIGURE 1.22 – Exemple de cartographie thermique reconstruite ne présentant pas de défaut, acquise sur un échantillon complexe. De gauche à droite, cartographie, filtrat Laplacien et Histogrammes de gradients orientés (HOG) associés.

thermiques reconstruites. Nous pouvons y voir la variété en angle d'inspection par rapport à la fissure ainsi que les phénomènes pouvant perturber l'apprentissage comme les contours complexes, artefacts, la saturation et la forme "altérée", ovoïde de la trajectoire du spot laser de chauffe qui est en grande partie due à l'état de surface au niveau de la barrière thermique.

Une autre base de données a été acquise à l'objectif 250 mm G1 microscopique et est dénommée **DAWN-G1**. Cette fois les acquisitions sont réalisées strictement parallèles au défaut examiné. Un volume d'enregistrements contenant plusieurs centaines de films thermiques est constitué. Les problématiques liées à la reconstruction des cartographies pour des balayages dont la dimension d'inspection est largement supérieure au champ image sont ici évitées en se restreignant à l'étude des images thermiques individuelles. Si les volumes de données produits ici sont assez importants en image-par-image, cette base est par la suite sous-échantillonnée à approximativement 200 images thermiques individuelles de manière aléatoire, en vue de contraindre fortement la quantité de données. Elle sera justement employée au chapitre 4 pour évaluer le transfert des apprentissages depuis des échantillons génériques vers des bases complexes et restreintes en quantité de données. Ces enregistrements sont illustrés eux-aussi Figures 1.24 et 1.25, qui fournissent des exemples d'images thermiques individuelles, au cours d'un enregistrement. Nous constatons en particulier grâce aux caractéristiques filtrées, que l'utilisation de l'objectif G1 fait apparaître

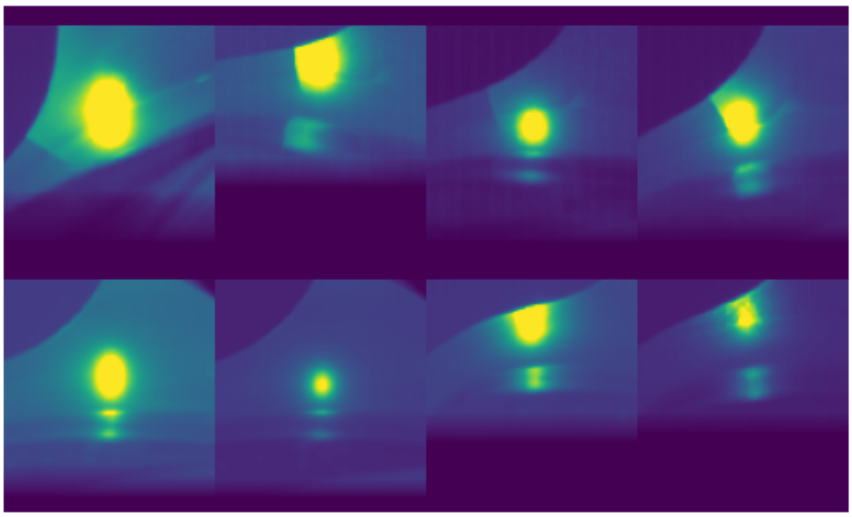


FIGURE 1.23 – Mosaïque de cartographies thermiques constituant la base de données DAWN. Les positifs avec fissures sont sur la première ligne, les négatifs en dessous.

énormément de détails de surface qui bruitent les cartographies comme les rugosités et le dépôt d'oxydes en surface, ce qui peut potentiellement mettre en difficulté une méthode d'apprentissage machine.

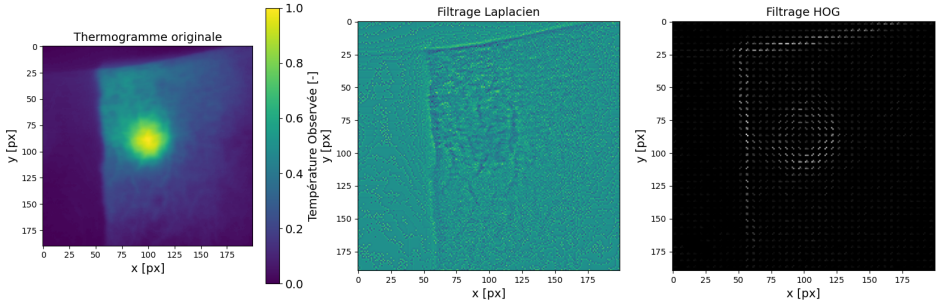


FIGURE 1.24 – Exemple d'image thermique individuelle présentant un défaut, acquise sur un échantillon complexe. De gauche à droite, cartographie, filtrat Laplacien et Histogrammes de gradients orientés (HOG) associés.

La Figure 1.26 illustre la base de données DAWN-G1 avec un panel des différentes images thermiques individuelles. Elle illustre la différenciation avec les cartographies précédentes tant au niveau du phénomène thermique observé que le champ image, grâce au changement d'optique, avec l'état de surface bien plus marqué et varié.

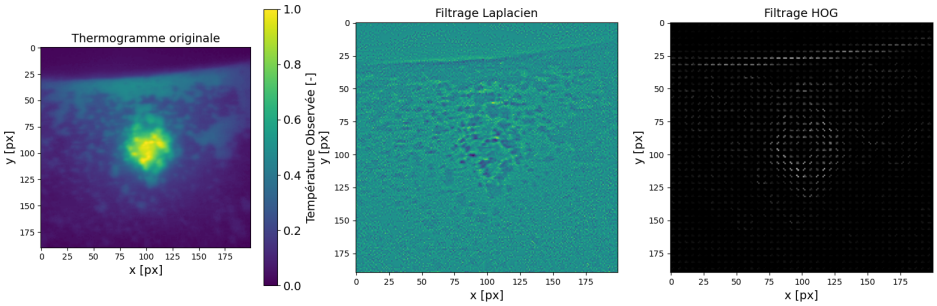


FIGURE 1.25 – Exemple d'image thermique individuelle ne présentant pas de défaut, acquise sur un échantillon complexe. De gauche à droite, cartographie, filtrat Laplacien et Histogrammes de gradients orientés (HOG) associés.

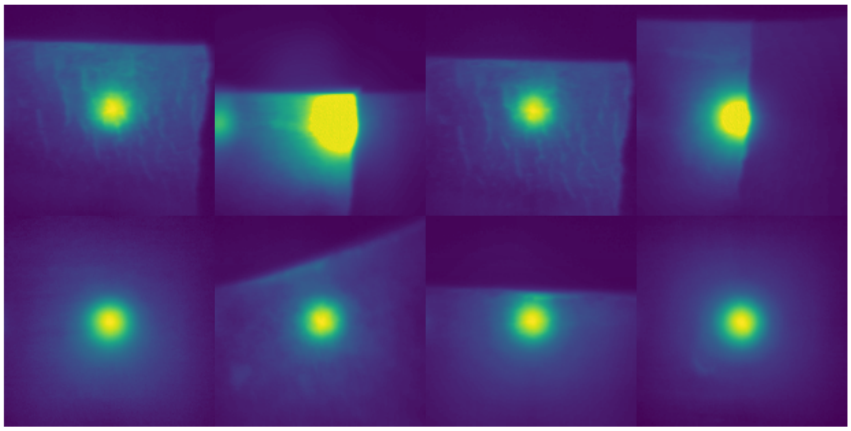


FIGURE 1.26 – Mosaïque d'images thermiques individuelles constituant la base de données DAWN-G1. Les positifs avec fissures sont sur la première ligne, les négatifs en dessous..

1.4 Synthèse des différentes bases de données simulées et expérimentales

Nous fournissons la Table 1.3 comme point de repère permettant d'identifier les différents volume de données constitués, leur type, leur usage, ainsi qu'un renvoi vers les chapitres de déploiement.

1.5 Conclusion

Nous avons donc présenté, dans ce premier chapitre, la théorie de la méthode d'examen non-destructif par thermographie laser Flying-Spot, ainsi que les moyens expérimentaux et simulés disponibles. Nous avons décrit les données collectées sur le plan simulé comme

Base	Source	# images	Section associée	Utilisation	Chapitres de déploiement
FEM-1	Simulation	28,975	Section 1.3.5	pré-entraînement	Chap. 3
FEM-2	Simulation	6,000	Section 1.3.5	pré-entraînement	Chap. 3
FLYD	Experimentation	1200	Section 1.3.6	pré-entraînement, détection	Chap. 3, 4
FLYD-Frames	Experimentation	800	Section 1.3.6	pré-entraînement, synthèse d'images	Chap. 3, 4
DAWN-J	Experimentation	900	Section 1.3.7	étude d'ablation	Chap. 3
DAWN-E	Experimentation	600	Section 1.3.7	synthèse d'images	Chap. 2
DAWN	Experimentation	300	Section 1.3.7	détection	Chap. 3, 4
DAWN-G1	Experimentation	200	Section 1.3.7	détection	Chap. 4

TABLE 1.3 – Résumé des différentes bases de données expérimentales et simulées constituées. Nous distinguons les données simples obtenues par simulations, des données réelles obtenues expérimentalement sur des échantillons du type éprouvette et sur les échantillons complexes revêtus.

sur le plan expérimental en justifiant les choix qui ont été pris afin de constituer des détecteurs de fissures par thermographie robustes aux conditions de surface et sans a priori des opérateurs. Néanmoins, nous pouvons noter que la quantité d'échantillons disponibles, et donc de faciès de fissuration, reste limitée et peu diversifiée en termes d'orientation ou de longueur de fissure notamment, concernant les échantillons complexes d'application. Cela pose souvent des problématiques dans les performances de généralisation pour les modèles neuronaux i.e. de détection de l'endommagement sur de nouveaux échantillons ou de nouvelles pièces, ainsi que dans la stabilité générale de l'apprentissage des modèles avec le risque d'un phénomène de sur-apprentissage.

L'intelligence artificielle générative, qui utilise un modèle d'apprentissage pour produire de nouvelles données synthétiques à partir des données d'entrée, offre potentiellement une solution pour résoudre ce déficit en diversité des données. Ces méthodes sont explorées dans le chapitre suivant.

Chapitre 2

Génération d'images par modèles de diffusion pour l'augmentation de données en FST

La réalité peut se permettre d'être improbable, pas la fiction.

L'autre moi-même, A. R. Damasio

Sommaire

Comme nous avons pu le voir dans le chapitre précédent, la collecte de données expérimentales diversifiées peut vite être restreinte en FST. Nous sommes confrontés à la fois à des contraintes sur le nombre d'échantillons disponibles et sur la quantité d'enregistrements de balayage réalisables pour une surface d'examen donnée. Nous avons aussi constaté que la simulation par éléments finis fournissait des images extrêmement simples qui manquent de réalisme, ce modèle étant plus orienté vers la simulation des comportements thermophysiques purs que vers la diversité des générations d'images thermiques produites. Des limites logicielles empêchent donc de produire en grande quantité et facilement des scènes plus complexes avec cet outil de simulation.

La génération d'images synthétiques permet d'augmenter significativement la quantité totale d'images d'entraînement disponibles, ce qui est particulièrement intéressant dans notre contexte où la donnée peut être rare avec des échantillons peu diversifiés et dont l'examen peut être difficile à réaliser. L'approche d'augmentation de données par la synthèse d'images est courante en vision par ordinateur et notamment dans le milieu médical [22, 23]. En mélangeant les propriétés de la distribution des images d'entraînement, elle permet ainsi de générer de nouvelles images spécifiques et différenciées par rapport à la distribution d'entrée. Cet apport de données peut augmenter tant les performances des modèles de détection d'objets que leur robustesse face à de nouvelles données inconnues, comme de nouvelles pièces à inspecter absentes de l'entraînement d'origine.

Nous proposons dans ce chapitre de nous concentrer sur la synthèse de données et plus particulièrement sur la génération d'images au travers des modèles de diffusion par débruitage. Dans la section 2.1, nous présentons un état des lieux bibliographique des modèles génératifs, des premiers travaux jusqu'aux méthodes de diffusion les plus récentes.

Nous traitons ensuite en détail le fonctionnement des modèles de diffusion par débruitage pour la synthèse d'images, puis les employons sur les données de FST dans la section 2.2. Nous abordons enfin l'utilisation de ces modèles avec un couplage texte+image pour la génération d'images FST dans la section 2.3.

Ce chapitre est le premier à aborder des notions d'apprentissage profond, dont les fondements sont présentés en annexes A.1 et A.2. L'augmentation de données est abordée en annexe A.3.

2.1 Bibliographie des modèles génératifs

Cette section dresse un état des lieux de la génération profonde d'images pour l'augmentation de données, c'est-à-dire l'agrandissement artificiel du volume de données. L'accent est notamment mis sur des modèles génératifs récents basés sur la diffusion par débruitage. Nous évoquons ensuite quelques exemples de déploiements de ces approches présents dans la littérature. Enfin, nous discutons des métriques employées pour mesurer la qualité de la synthèse d'images. Ces modèles nous permettent ensuite de synthétiser de manière artificielle des données supplémentaires pour l'apprentissage profond appliqué en FST.

2.1.1 Historique de l'apprentissage profond pour la synthèse d'images

La génération d'images consiste en l'utilisation d'un système d'apprentissage apprenant à reproduire la distribution d'une base de données, permettant alors de produire de nouvelles images de manière artificielle. Un schéma de synthèse sur l'intérêt de l'approche générative pour l'augmentation de données est fourni dans la Figure 2.1. C'est aujourd'hui une approche courante proposée en apprentissage automatique moderne afin d'augmenter les performances de modèles sur des tâches de vision par ordinateur. Cette technique permet à la fois une augmentation brute des performances pour détecter ou localiser un objet par exemple, mais aussi des gains en généralisation face à de nouveaux lots d'images dont les propriétés diffèrent de la distribution de données d'entraînement.

Les gains permis par ces modèles de synthèse d'images peuvent s'expliquer au travers de deux hypothèses qui ne sont pas encore tranchées dans la littérature actuelle de l'apprentissage profond. La première est l'augmentation de la variabilité des données d'entraînement, au travers du mélange des propriétés des images comme les contours ou l'orientation et la forme des objets observés. La deuxième hypothèse pourrait être associée à la distillation de connaissances, où l'utilisation de modèles génératifs massifs pour entraîner un modèle spécialisé est une approche courante [24–27]. Le modèle de synthèse d'images formerait durant l'apprentissage des caractéristiques riches liées à la distribution des propriétés des données d'entrée. La synthèse de données sert alors d'intermédiaire pour transférer ces connaissances à un modèle spécialisé dans une tâche précise comme la détection d'objets.

Les méthodes génératives permettent aussi de manière plus secondaire l'identification qualitative des caractéristiques apprises par le modèle, en extrayant les formes et contours générés par l'espace latent du modèle.

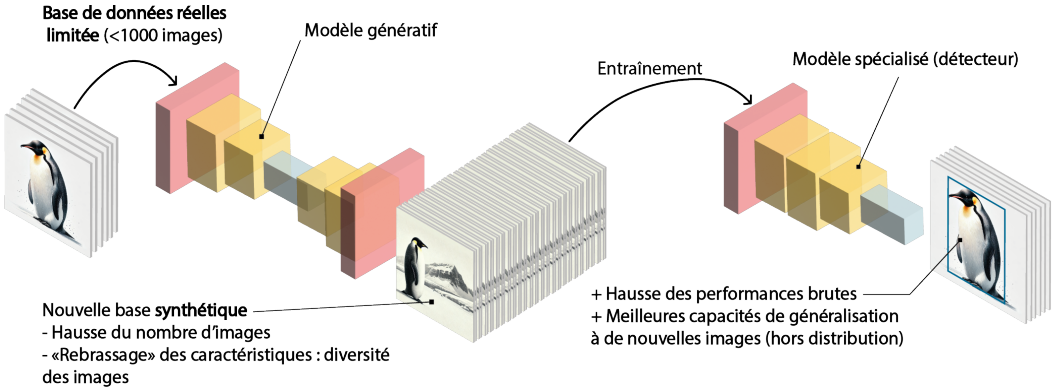


FIGURE 2.1 – Synthèse résumant l'intérêt de l'IA générative dans le cadre de l'entraînement d'un modèle spécialisé comme un détecteur d'objet en contexte de données limitées.

Les modèles auto-encodeurs sont les premières architectures de synthèse de données par apprentissage. Leur principe date de la fin des années 1980 avec comme premières applications la synthèse de séquences 1D pour la reconstruction de signaux [28–30]. Dans cette architecture, les caractéristiques des données d'entrée sont d'abord comprimées sur le plan dimensionnel par apprentissage grâce à une structure neuronale appelée encodeur sous une représentation vectorielle appelée espace latent. Une deuxième structure neuronale va procéder à la décompression de cet espace pour générer l'image synthétique : c'est le décodeur, qui est couramment construit en symétrie par rapport à l'encodeur sur le plan neuronal. Le vecteur latent noté $\mathbf{u}$ sert de représentation des caractéristiques de la distribution des images fournies en entrée. Cette architecture est un composant fondamental des autres approches génératives.

Une fois l'entraînement réalisé, on peut partir d'un bruit et de cet espace latent pour générer de nouvelles images conservant les caractéristiques de la distribution d'origine.

Toute une diversité de VAE existent, suivant la manière de contraindre l'espace latent pour qu'il tende à suivre une distribution gaussienne, afin générer de nouvelles images tels que : *Beta-VAE* [31], *Conditional VAE* [32], etc. Les auto-encodeurs sont généralement simples à implémenter et ont un nombre relativement limité de paramètres. Néanmoins, la diversité et la qualité de la synthèse, et donc son bénéfice sur les capacités de généralisation du modèle spécialisé, ne sont pas garanties. Il est de plus souvent nécessaire d'élaborer une architecture spécifique pour arriver à reproduire la distribution associée à un problème donné.

Ensuite, les réseaux de neurones génératifs antagonistes (GANs) ont été développés : cette approche d'apprentissage est applicable à une diversité d'architectures et repose sur la mise en place d'un couple générateur-discriminateur. Le générateur peut être basé sur un auto-encodeur et synthétise une image que le discriminateur évalue ensuite comme étant une image réelle ou artificielle. Les deux modèles sont co-entraînés afin de maximiser la vraisemblance des synthèses. Cette opposition permet, selon les problèmes et le succès de l'apprentissage, d'obtenir des images de synthèse cohérentes avec la distribution réelle. Néanmoins, cette approche peut vite être coûteuse en termes de calcul. De plus, le couple

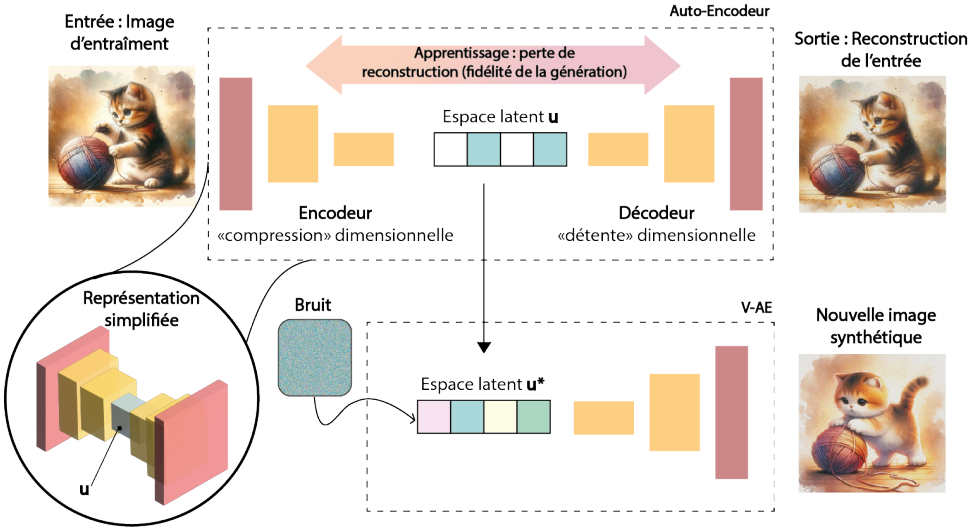


FIGURE 2.2 – Fonctionnement d'un auto-encodeur. Dans la phase d'apprentissage le modèle apprend à reproduire l'entrée. Puis lors de la phase de synthèse l'auto-encodeur variationnel synthétise de nouvelles images basées sur les caractéristiques d'origine altérées d'un bruit.

binaire entre discriminateur et générateur rend l'apprentissage de l'ensemble instable : ainsi l'architecture est très sensible aux explosions de gradient et aux changements de mode de comportement de la fonction de coût, ce qui entraîne une divergence de cette dernière au cours de l'apprentissage. L'architecture nécessite alors des ajustements fins des hyper-paramètres durant l'entraînement antagoniste pour arriver à une convergence du modèle, ce qui complique l'utilisation de ces approches de synthèse d'images. Une architecture de référence est Cycle-GAN, pensée pour le transfert de style, c'est-à-dire l'injection de caractéristiques d'un domaine A vers un domaine B [33]. Une partie de la recherche s'est longtemps focalisée sur la stabilisation de l'apprentissage des GANs au travers de la régularisation des fonctions de coût apprises [34, 35].

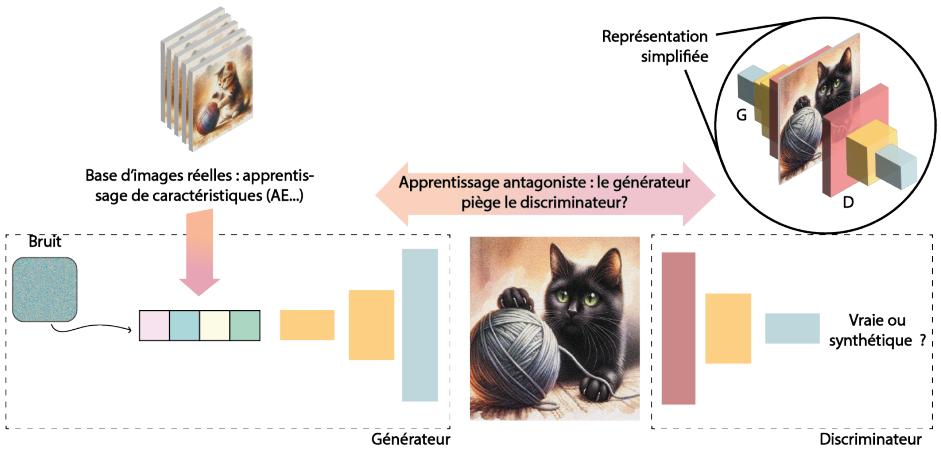


FIGURE 2.3 – Illustration du fonctionnement d'un apprentissage antagoniste par GAN. Le générateur va générer une image qui va être évaluée par le discriminateur comme réelle ou fausse, synthétique.

Les modèles de diffusion quant à eux sont des architectures de génération d'images qui appartiennent à la famille des modèles basés sur la modélisation des phénomènes d'écoulements, où une transformation générique est appliquée de manière itérative à une entrée afin de l'altérer progressivement jusqu'au résultat souhaité [36, 37]. Une première proposition d'architecture de diffusion fut élaborée par Sohl-Dickstein [38] : la distribution d'entrée est convertie progressivement en un bruit aléatoire. Puis la transformation est réversible, ce qui permet de générer de nouvelles images synthétiques. Cette architecture d'apprentissage présente une similarité forte entre processus de transformation thermodynamique simple et modèle de diffusion. Ainsi, la diffusion peut être assimilée à la détente progressive et réversible d'un gaz froid dans un grand volume. Cette première architecture est néanmoins trop coûteuse sur le plan calculatoire, et peu optimale. Ces travaux ont mené ensuite au modèle de diffusion probabilistique par débruitage (DDPM) par Ho et al., apprenant à transformer l'entrée en bruit gaussien [39]. Un bruit aléatoire couplé à une hypothèse de réversibilité permet ensuite de générer des images de synthèse, se passant d'apprentissage antagoniste. Le composant réalisant ce bruitage est généralement confié à un modèle U-net de débruitage, qui est une architecture type auto-encodeur [40]. Afin d'accélérer l'apprentissage, les modèles de diffusion implicite par débruitage (DDIM) ont été élaborés, reposant notamment sur des raccourcis dans la transformation de débruitage exploitée [41]. Ils sont associés aux modèles de diffusion latents (LDM) qui proposent de diffuser sur un espace latent comprimé plutôt que sur la distribution brute des images, et de procéder a posteriori à une sur-résolution de cet espace [42]. Le caractère progressif de la transformation appliquée rend généralement l'architecture très générique, généralisable à de nombreux problèmes sans adaptation architecturale coûteuse en temps.

Aujourd'hui, on couple ces modèles de diffusion ayant fait leurs preuves avec des modèles de langage, comme DALL-E et Stable Diffusion [41].

La mise en cascade de processus de diffusion en partant d'une diffusion sur un espace latent très petit puis en sur-échantillonnant vers des espaces dimensionnels de plus en plus grands, permet encore des progrès dans la qualité des images synthétiques produites. Nous pouvons mentionner le modèle récent Stable-Cascade qui est basé sur l'architecture de diffusion en cascades Wuerstchen [43]. Pour finir, les modèles texte-images que l'on peut qualifier de modèles de fondation sont des modèles agnostiques à la tâche ou au type de données des problèmes, ils sont considérés comme hautement versatiles.

Les modèles de diffusion sont aujourd'hui considérés comme une approche très performante pour la synthèse d'images cohérentes et diversifiées. Néanmoins, le processus de génération est toujours contraint, notamment par la latence et le coût calculatoire inhérents au débruitage. D'autre part, pour la synthèse couplée au texte, la capacité du réseau de langage à conditionner efficacement la génération est encore perfectible.

Les perspectives les plus explorées dans la littérature actuelle sont donc l'utilisation d'un conditionnement textuel de plus en plus riche sémantiquement et pertinent du point de vue de la compréhension, comme dans les modèles de diffusion les plus récents : DeepFLOYD-IF, SD XL, DALL-E 3 [44–46]. D'autre part, l'essor récent des modèles de consistance comme successeurs et suites naturelles des modèles de diffusion, pouvant potentiellement largement réduire la latence dans le mécanisme de la diffusion sans perte de qualité de synthèse, commence à ouvrir des voies vers la réduction drastique du besoin calculatoire de ces modèles lors de l'apprentissage [47, 48].

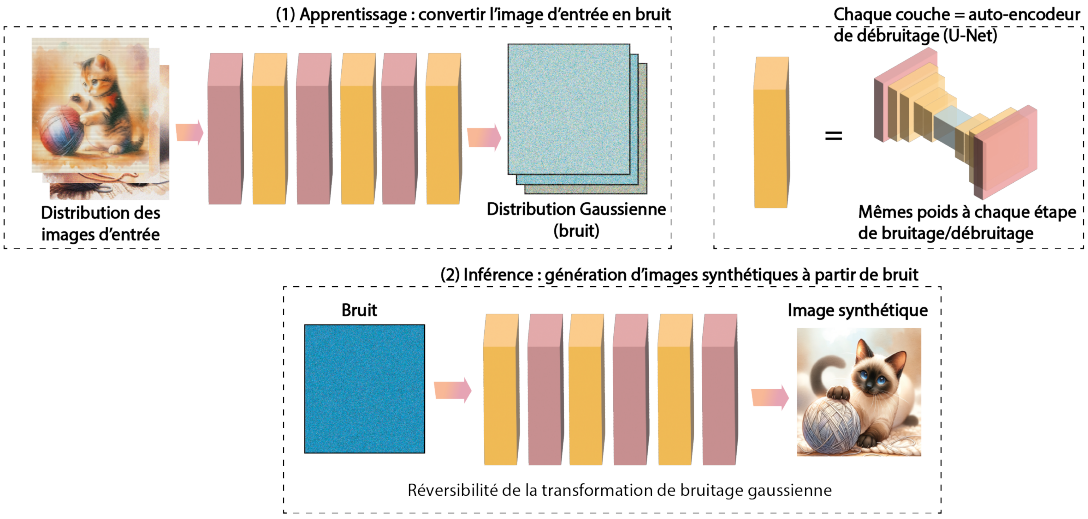


FIGURE 2.4 – Fonctionnement d'un modèle de diffusion : lors de l'apprentissage le modèle apprend à convertir des images en bruit gaussien. En phase d'inférence ce modèle va être inversé et permettre la génération d'images synthétiques à partir d'un vecteur de bruit aléatoire de départ.

2.1.2 Exemples d'application analogues au travail de thèse

Nous pouvons pointer ici quelques travaux d'applications de la synthèse d'images pour l'augmentation de données, fournissant des analogies pour guider le travail de thèse, tant sur le fait de traiter des données non conventionnelles que sur l'application au contexte de CND. Il n'y a pas de travaux comparables à l'heure actuelle en thermographie.

La synthèse d'images pour augmenter le volume de données est une approche commune dans le domaine médical, où la radio-imagerie présente des contraintes de production et de diversité de données similaires à celles rencontrées en thermographie. De plus, dans ce domaine d'application, il y a une problématique de respect de la vie privée et les données synthétiques servent alors d'intermédiaires pour éviter la fuite de données privées sensibles tout en permettant l'entraînement de modèles. Mentionnons l'utilisation d'un GAN, puis notablement l'utilisation d'un modèle de diffusion pour augmenter le bassin total de données disponibles pour les radiographies thoraciques [23]. Du côté du contrôle qualité en optique visible, un certain nombre de travaux existent qui exploitent des modèles génératifs pour l'augmentation de données ou pour le pré-entraînement, en particulier du côté des approches dites de détection d'anomalies [49]. Nous pouvons relever par exemple l'utilisation d'un *Beta-VAE* [31] pour le contrôle des joints de soudure en spectre visible, ou bien l'utilisation d'un GAN pour une architecture généraliste de contrôle de défauts évaluée sur la base de données de référence en contrôle qualité optique MVTEC [50].

2.1.3 Métriques de synthèse d'images

La qualité de génération d'images fournie par les modèles mentionnés précédemment peut être évaluée grâce à différentes métriques. Néanmoins, la plupart reposent sur le principe d'une mesure de distance entre les distributions réelles et synthétiques. La distance de Fréchet calculée avec le modèle Inception (FID) est une mesure classique dans

le contexte de la génération d'images, évaluant la qualité des images générées [51]. Elle mesure la similarité entre les distributions d'images synthétiques et réelles. La FID implique d'apprendre la distribution des deux ensembles d'images en utilisant le réseau de neurones Inception [52] pour fournir un espace latent comprimé des caractéristiques de chaque distribution d'images. Chacune des deux distributions voit ensuite sa moyenne et sa covariance estimées. Le score FID, estimant une distance entre ces distributions, associe des valeurs plus faibles à une plus grande similarité et à une meilleure qualité de génération d'images. Cette métrique est généralement considérée comme une bonne mesure du photoréalisme des images générées par les modèles de diffusion. On peut donner une taxonomie empirique des différentes valeurs du FID :

— **Pour** $0 \leq$ **FID** < 20 : la génération est très bonne, c'est-à-dire que les distributions réelles et synthétiques sont très similaires. Les anomalies et artefacts de synthèse sont très limités.

— **Pour** $20 \leq$ **FID** < 50 : la génération est considérée comme bonne, mais on observe un début d'effets de bord tels que des distorsions et artefacts de synthèse en nombre limité : blobs, altérations de contours et de textures.

— **Pour** $50 \leq$ **FID** < 100 : la génération est considérée comme moyenne. La distribution d'entrée est assez bien répliquée dans l'ensemble, mais les effets de bord, artefacts de génération, etc., sont bien identifiables sur les images.

— **Pour FID** ≥ 100 : la génération est considérée comme mauvaise, avec des artefacts de génération en quantité et des caractéristiques de la distribution d'entrée qui sont mal comprises et reproduites par le modèle de génération.

Cette échelle est seulement indicative, le facteur essentiel restant le gain de performance permis par l'utilisation de ces données de synthèse. De plus, les scores comme le FID sont assez sensibles au pré-traitement des images, qui peut altérer négativement l'évaluation de la métrique : les modèles de diffusion ont souvent des résolutions spatiales de sortie deux à trois fois plus grandes que les images d'origine, ce qui peut dégrader artificiellement la valeur de la mesure de qualité de synthèse.

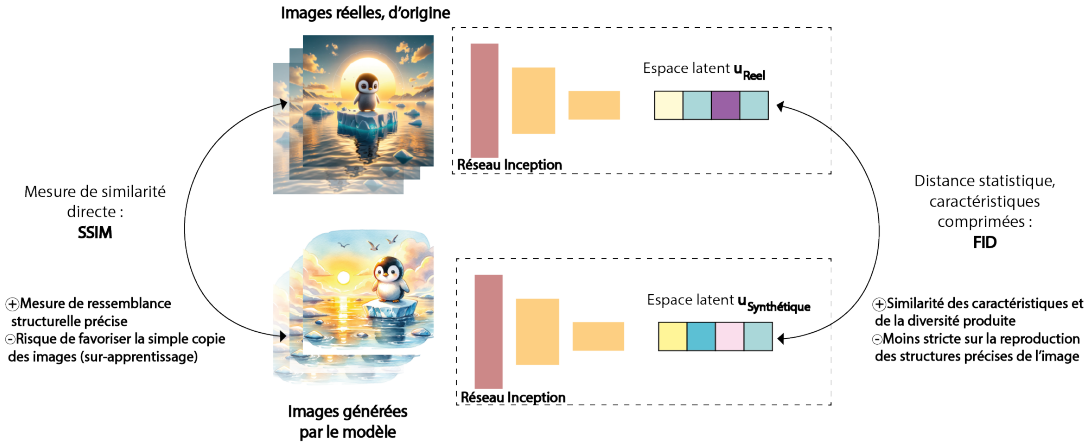


FIGURE 2.5 – Comparaison des deux métriques SSIM et FID, vulgarisation de leur place dans l'évaluation de la génération d'images par rapport aux images réelles d'entraînement.

Si la FID est souvent privilégiée dans la littérature généraliste de synthèse d'images moderne, notons l'existence d'autres métriques pour l'IA générative. Par exemple, le *Structu-*

ral Similarity Index Measure (SSIM) revient à une mesure directe de la similarité entre les images de chaque distribution, sans compression dimensionnelle en un espace latent [53]. La figure 2.5 fournit une représentation vulgarisée de ces deux métriques et de leur rôle dans la mesure de la qualité de synthèse d'images. D'autres métriques existent dans la littérature mais sont moins courantes ou bien focalisées sur des tâches spécifiques comme le débruitage de données.

2.2 Modèles de diffusion par débruitage pour la FST

Nous présentons ici de manière détaillée le fonctionnement des modèles de diffusion par débruitage, puis leur mise en application sur les données de FST, en particulier les données collectées sur des échantillons complexes revêtus.

2.2.1 Éléments théoriques

Proposés initialement par Ho et al. [39], les DDPM fonctionnent en inversant un processus de diffusion pour transformer progressivement du bruit en données structurées. Mathématiquement, le processus de diffusion, considéré comme stochastique, est décrit par une série d'opérations où à chaque étape t, les données x_t sont progressivement bruitées jusqu'à obtenir un signal aléatoire à l'étape finale T (en général Gaussien). Le processus de diffusion peut être formalisé comme suit :

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{2.1}$$

où q correspond à la distribution de probabilité du processus de diffusion, $\mathcal{N}$ est la distribution normale, β_t est un paramètre de bruit contrôlant l'intensité de la diffusion à l'étape t, et $\mathbf{I}$ est la matrice identité.

Pour générer de nouvelles données, le DDPM suppose la réversibilité de ce processus itératif : il apprend alors à prédire le bruit ϵ qui a été ajouté à chaque étape, et l'utilise pour progressivement débruiter le vecteur de bruit de départ jusqu'à produire une image synthétique. Cela est exprimé par l'équation de prédiction du bruit suivante :

$$\epsilon_\theta(x_t, t) = \epsilon - \nabla_{x_t} \log p_\theta(x_{t-1}|x_t), \tag{2.2}$$

où ϵ_θ est le bruit prédit par le modèle à l'étape t, et p_θ est la distribution apprise par le modèle pour inverser le processus de diffusion.

L'algorithme de Langevin joue un rôle crucial dans ce processus de génération, en particulier dans l'implémentation DDPM de Ho et al. [39]. Il permet de générer des échantillons à partir d'une distribution de probabilité complexe guidée par le logarithme du gradient de cette distribution, ajoutant de plus un terme de bruit à chaque itération (bruitage de la distribution). Pour un état donné x_t, l'itération de Langevin se formule comme suit :

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z, \tag{2.3}$$

où α_t est le pas de temps, $\sigma_t = \sqrt{\beta_t}$, et $z \sim \mathcal{N}(0, \mathbf{I})$ est un bruit gaussien, permettant ainsi une exploration efficace de l'espace formé par la distribution des données. Cela

contribue généralement à l'amélioration de la qualité des images synthétiques produites par la diffusion.

La Figure 2.6 fournit une illustration de l'altération progressive de la diffusion d'entrée vers un bruit gaussien, montrant le passage de la distribution complexe d'entrée à une distribution normale centrée.

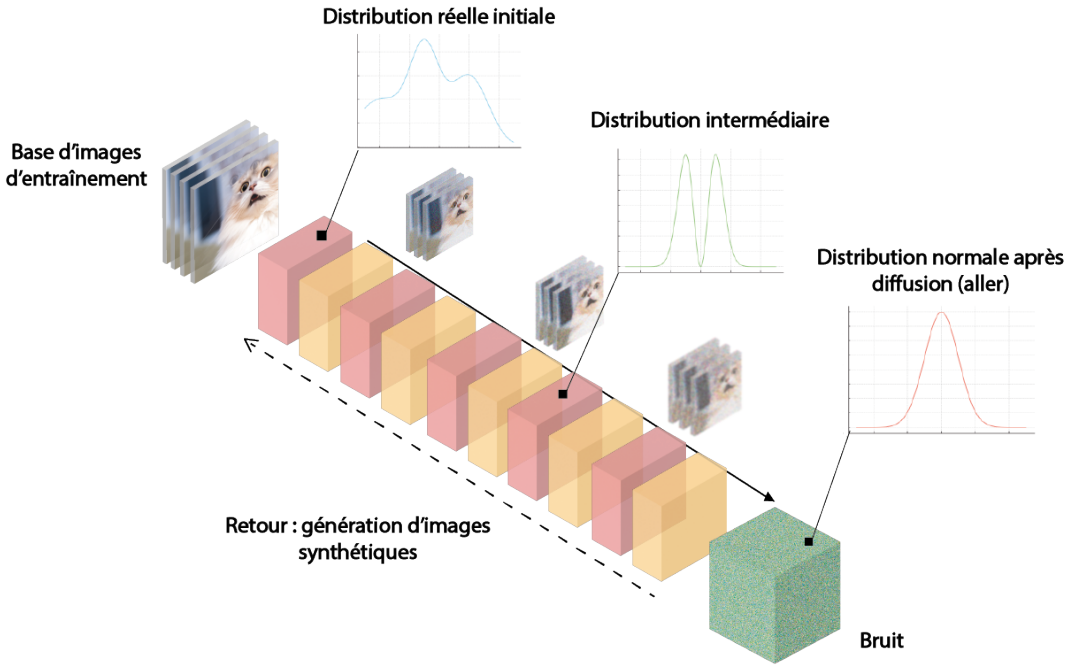


FIGURE 2.6 – Illustration détaillée du processus de diffusion dans un DDPM. Le schéma montre l'altération progressive de la distribution d'entrée et son impact sur les images d'apprentissage est mis en valeur.

Algorithm 1 Entraînement du DDPM (aller)

1: Initialiser modèle de paramètres θ
2: **repeat**
3: Échantillonner $x_0 \sim q(x_0)$
4: Échantillonner $t \sim \text{Uniform}(\{1, \ldots, T\})$
5: Échantillonner $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
6: Prendre le gradient de la descente sur $\nabla_\theta \left\| \epsilon - \epsilon_\theta \left(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t} \epsilon, t \right) \right\|^2$
7: **until** Convergence

Algorithm 2 Synthèse par DDPM (retour)

1: Échantillonner $x_T \sim \mathcal{N}(0, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: Échantillonner $z \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $z = 0$
4: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z$
5: **end for**
6: **Return** x_0

La différence principale entre les DDPM et les *Denoising Diffusion Implicit Models* (DDIM) repose sur leur méthode de génération à partir du bruit. Les DDPM effectuent une série d'étapes de débruitage pour convertir le bruit en images directement sur l'espace des images originales. Les DDIM, en revanche, accélèrent le processus de diffusion inverse et permettent une reconstruction plus rapide des images sans compromettre la qualité. Cela est permis grâce à l'utilisation de simplifications dans la diffusion inverse, permettant de se

passer d'un débruitage complet. Il y a aussi une pré-compression latente des images dans les modèles dits de diffusion latente (LDM) [42]. Cela rend alors les DDIM particulièrement adaptés pour des applications nécessitant une génération d'image relativement rapide.

2.2.2 Application aux échantillons complexes

Deux DDPM sont entraînés pour augmenter le volume de données disponibles sur les échantillons complexes. Ils sont entraînés spécifiquement sur chaque classe. Nos modèles apprennent à convertir les cartes thermiques en un bruit gaussien. Puis ce réseau nous permet de générer de nouvelles cartes synthétiques, par réversibilité. Le nombre d'étapes de débruitage est maintenu à la valeur de 1000, standard pour ce modèle. Elle donne empiriquement le meilleur rapport entre qualité de synthèse et vitesse de génération.

Les modèles ont ici été entraîné jusqu'à l'obtention d'une synthèse la plus satisfaisante possible pour un expert, sur le plan qualitatif. Le temps d'entraînement a pris approximativement une semaine par classe sur une carte graphique RTX A5000.

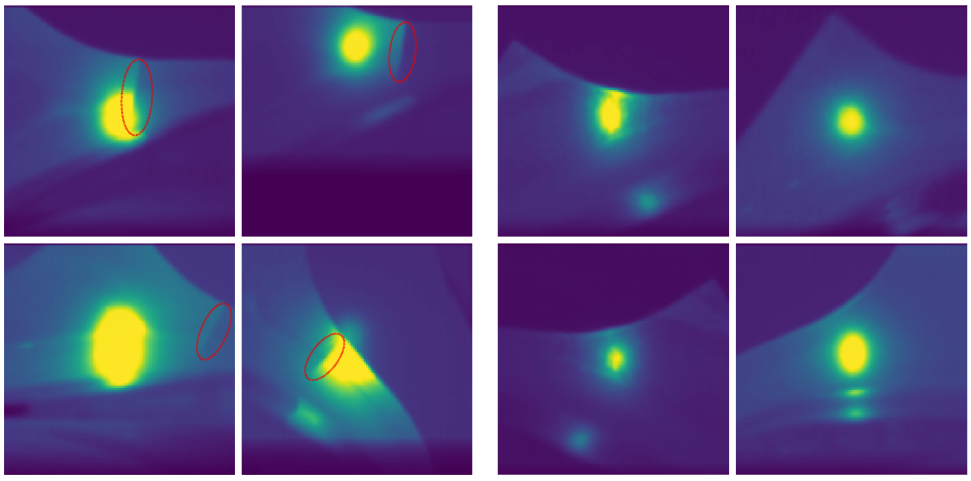


FIGURE 2.7 – Exemples d'images générées par DDPM à partir des thermogrammes d'échantillons revêtus. À gauche, les positifs (fissures entourées en rouge), à droite, des exemples du groupe contrôle.

Une base de données synthétiques est ainsi constituée, contenant approximativement 8 000 cartographies reconstruites artificielles. Cette base est nommée DAWN-Synth dans la suite de ce manuscrit et sera notamment utilisée dans le chapitre 3. Le FID obtenu est de 24, indiquant une bonne génération d'images : les images sont relativement proches des cartographies thermiques de la base d'entraînement, tout en présentant de légères variations d'orientation ou de longueur de défaut notamment. À l'œil, les données sont très difficiles à distinguer des images réelles d'origine.

2.3 Modèles de diffusion guidés par le texte : Stable Diffusion

Les modèles texte-images se différencient particulièrement des DDPM vus précédemment par l'ajout d'un guidage sémantique texte, permettant de mieux contraindre la

génération. Dans cette section, nous utilisons la génération d'images par modèle texte-image. Ces volumes de données synthétiques sont employés pour les travaux des chapitres 4 et 6.

2.3.1 Principes

À partir de l'architecture DDIM de base, un conditionneur est ajouté permettant d'injecter aux différents stades du débruitage un vecteur latent correspondant aux caractéristiques textuelles souhaitées. Cela permet ainsi de guider la sémantique générale de la scène produite par le débruitage, guidant l'apparition de caractéristiques précises, telles qu'un style particulier ou une forme donnée. L'approche utilisée pour ce conditionneur est généralement basée sur le modèle CLIP (*Contrastive Language-Image Pre-Training*) [54]. Elle est devenue classique dans les architectures de l'état de l'art afin de produire un couplage texte-image (architectures COCA, DALL-E [46, 55]). Pour simplifier, cette méthode utilise un apprentissage contrastif qui permet la formation d'associations entre les vecteurs appris côté texte et côté image. Cet espace de couplage des modalités permet de former des vecteurs multi-modaux spécifiques associant la sémantique du texte et les caractéristiques images associées. Plus de détails sur les méthodes d'apprentissage contrastif sont présents ici [56].

Ces modèles, que l'on peut qualifier de **modèles de fondation**, sont pensés pour être facilement ré-ajustables (voir annexe A.5), en particulier dans des cas où la quantité de données est réduite. Nous pouvons ici mentionner la méthode *Dreambooth*, qui promet un ajustement du modèle pour synthétiser de nouvelles images avec moins de 5 images de départ, et permet au modèle d'absorber de nouvelles informations sémantiques texte-image sans ré-entraînement complet, avec l'encodeur texte gelé notamment.

Nous proposons ici d'utiliser un modèle Stable Diffusion. Un DDIM est couplé à une description de texte par un mécanisme d'attention croisée qui sera détaillé au chapitre 5. Le lecteur peut aussi se référer à l'annexe A.6 sur le mécanisme d'attention standard. Cette architecture, synthétisée dans le schéma Figure 2.8, présente le bénéfice d'être en libre accès, utilisable et ré-ajustable sur une machine locale disposant d'une carte graphique adaptée et disposant d'une très large communauté en ligne, en particulier grâce à *Hugging Face* et à la librairie associée *Diffusers* [57].

2.3.2 Génération d'images synthétiques sur les échantillons simples

Un modèle Stable Diffusion est ajusté grâce à la méthode *Dreambooth* sur des images individuelles des échantillons simples du dataset FLYD. Nous avons entraîné deux modèles pour chaque classe (fissuré ou sain). La base d'entraînement originelle contient un total de 400 images. Le texte de guidage utilisé était : *"A thermal frame from a laser scan of a metallic part { presenting a ; without a } surface crack."* Le nombre de pas d'inférence (i.e. le nombre d'étapes de débruitage) est fixé à 50 itérations après quelques essais empiriques. Un coefficient de congruence à la description de texte, qui permet de jouer sur l'influence de l'information sémantique fournie par le texte lors de la diffusion, est réglé aussi empiriquement à 10, améliorant légèrement la qualité de la synthèse. Sa valeur standard est généralement à 7,5 pour Stable Diffusion.

4000 images thermiques ont ainsi été produites afin de procéder à de l'augmentation

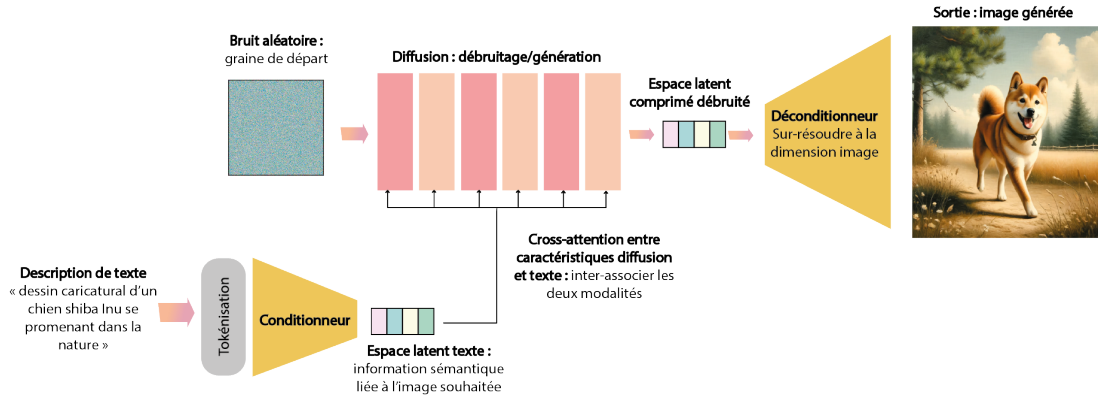


FIGURE 2.8 – Schéma de principe : architecture Stable Diffusion. Un bruit aléatoire passe dans un réseau de diffusion dont la génération est contrainte par un modèle de langage.

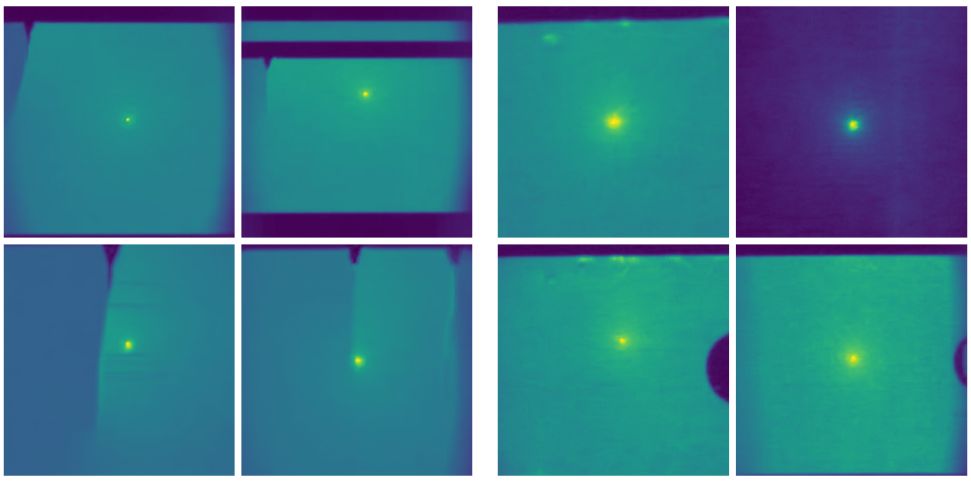


FIGURE 2.9 – Exemples d'images générées par SD à partir des images thermiques individuelles sur des échantillons d'essais de fatigue, FLYD-frames. À gauche, les positifs, à droite, des exemples du groupe contrôle.

de données. Elles sont utilisées dans le Chapitre 4. La mesure de FID donne une valeur de 60. Néanmoins, à l'œil, les images semblent assez bien générées : des propriétés des images d'origine sont bien reproduites comme les bords de pièce ou la trajectoire du laser. Le score indique sans doute un moindre photoréalisme des images générées : cela pourrait être dû à l'utilisation d'un modèle entraîné pour la génération "artistique", moins adapté pour la production d'images photoréalistes. Cette synthèse d'image est néanmoins bien plus rapide que la génération par DDPM, passant de deux semaines à quelques heures sur une carte graphique RTX A5000.

Une piste d'amélioration de cette synthèse qui aurait pu être plus explorée est la modification de l'entrée textuelle associée. Néanmoins, les quelques essais empiriques n'ont pas permis d'améliorations significatives de la synthèse sur le plan qualitatif. Ce générateur semble moins qualitatif que le précédent, mais nous l'utiliserons pour le pré-apprentissage en tâche de localisation de fissures sur les échantillons-éprouvettes.

2.3.3 Génération d'images à partir des échantillons complexes

Un modèle Stable Diffusion généraliste a été réajusté pour les échantillons complexes, afin d'avoir un point de comparaison au moins qualitatif avec la génération DDPM sur le même type d'images. Ce modèle permet aussi d'évaluer qualitativement la perte de performance de génération imputable au passage d'un DDPM à un DDIM, bien qu'il y ait toujours un gain notable dans la vitesse de génération de ces données. Le texte de génération est le même que celui des éprouvettes. Nous avons généré ainsi un volume de 2 000 images synthétiques. L'entraînement a duré approximativement deux heures par classe sur une carte graphique RTX A5000, ce qui est bien plus rapide que pour les DDPM.

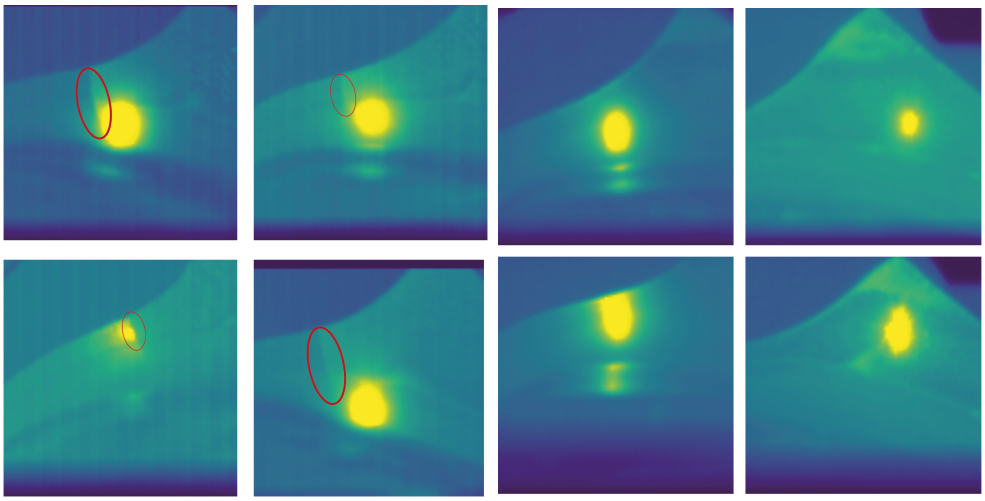


FIGURE 2.10 – Exemples d'images générées par Stable Diffusion à partir des images thermiques individuelles sur des échantillons complexes (DAWN-Synths). À gauche, les positifs (fissures entourées en rouge), à droite, des exemples du groupe contrôle.

La Figure 2.10 fournit des exemples de cette génération d'images avec Stable Diffusion pour les échantillons complexes. Qualitativement, la génération semble respecter les éléments principaux, néanmoins les images semblent notamment plus bruitées et plus floues, avec une trajectoire laser moins nette et plus exotique comparativement à la génération via DDPM (voir Figure 2.7). L'image globale semble aussi plus "chaude" que celle obtenue via DDPM. Quantitativement, on obtient un FID de 55, quantifiant cette dégradation comparative au DDPM, mais il faut aussi noter l'accélération dans la vitesse de synthèse d'image. Ce volume de données synthétiques n'est pas réemployé dans le reste du travail de thèse.

Néanmoins, il a l'avantage de synthétiser rapidement des images, après un apprentissage rapide par réajustement de modèle. Nous verrons de plus dans le chapitre 5 que les données qu'il synthétise sont suffisantes pour permettre l'augmentation des performances de détection de fissures en FST pour des modèles de localisation d'objet.

2.4 Conclusion

Ce chapitre a passé en revue les différentes méthodes de génération d'images par apprentissage pour l'augmentation de données employées dans le cadre de ce travail de

thèse. Nous avons aussi étudié l'évaluation de cette synthèse, tant qualitativement que quantitativement à l'aide de la mesure du FID. Les différentes bases constituées ont été réemployées tout au long de ce manuscrit.

Des perspectives existent pour améliorer ce travail de génération, d'abord sur le plan de la méthodologie ou du choix du modèle de synthèse. Nous pourrions ainsi nous tourner vers le déploiement de modèles plus avancés que ceux disponibles lors du travail de thèse : un passage à Stable Diffusion XL ou Wuerstschen [43, 45], qui montrent des capacités de synthèse supérieures et une réajustabilité à de nouvelles données accrue, est envisageable suivant la capacité GPU disponible. En parallèle, une exploration plus poussée de l'utilisation du texte grâce à un protocole de sélection de l'entrée textuelle plus strict serait intéressante : elle pourrait améliorer, au moins marginalement, la qualité de la synthèse d'images pour les modèles type Stable Diffusion.

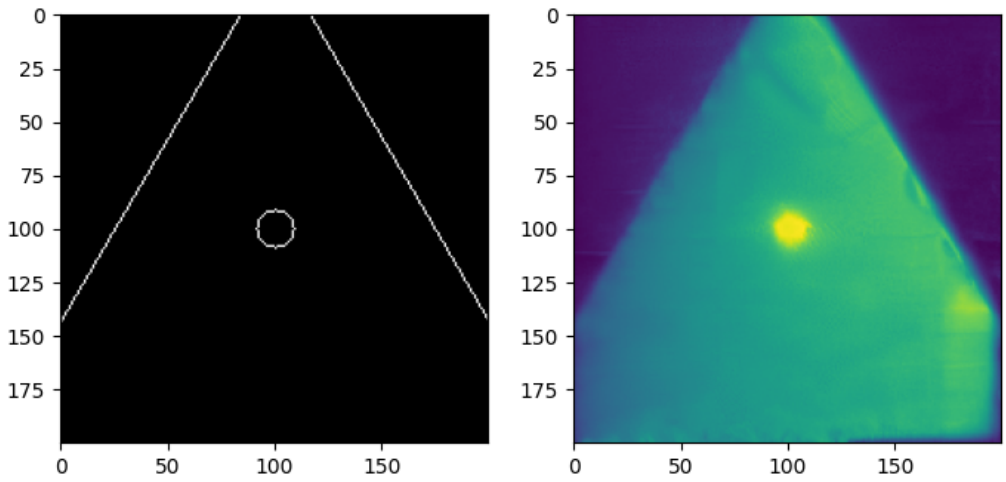


FIGURE 2.11 – Exemple de tentative de synthèse par forme libre, en spectre infrarouge. À gauche, nous trouvons le contour généré via un code aléatoire, qui guide un modèle type Stable Diffusion. À droite, une des variations qualitativement correctes est montrée.

La génération de formes libres, comme illustrée avec un contour triangulaire généré aléatoirement sur la Figure 2.11, pour guider les modèles de diffusion par contours, tant pour un diffuseur en spectre visible qu'infrarouge, permettrait de générer des scènes encore plus variées, et par là même d'obtenir un modèle robuste à des échantillons radicalement différents et inconnus de ceux disponibles à l'avenir. Néanmoins le taux d'échec de la synthèse est de plus de 50 % sur des expériences préliminaires avec cette forme simple : un grand nombre de générations sont constatées avec des contours ou des surfaces mal compris par le modèle.

Sur le plan de l'évaluation de la qualité de la synthèse d'images, nous nous sommes focalisés exclusivement sur le FID qui évalue la qualité de la génération uniquement du point de vue des propriétés des données. Ainsi, la construction d'une métrique d'évaluation de la qualité de synthèse d'images allant au-delà de ces propriétés et prenant en compte des éléments liés au phénomène physique associé à la diffusion de la chaleur serait un travail méthodologique pertinent. Cela pourrait même mener à l'élaboration d'une fonction de perte pour réajuster efficacement ces modèles aux données de FST. Nous pourrions nous

baser sur les travaux réalisés dans la télédétection radar-SAR comme travail analogue de départ pour construire une telle métrique [58].

Le prochain chapitre verra une première application de ces bases de données synthétiques en association avec des données simulées et expérimentales afin d'augmenter les performances de détection de fissures en FST pour des modèles neuronaux de classification.

Deuxième partie

Détection et localisation automatique de fissures sur données thermographiques

Chapitre 3

Apprentissage progressif pour la classification de fissures apprise en thermographie laser

Le jeu est donc, sous deux formes essentielles d'exercice sensorimoteur et de symbolisme, une assimilation du réel à l'activité propre, fournissant à celle-ci son alimentation nécessaire et transformant le réel en fonction des besoins multiples du moi.

Psychologie et pédagogie, J. Piaget.

Sommaire

L'approche dite d'apprentissage progressif ou *curriculum learning* en anglais est issue du transfert du jeu dans l'enfance compris sous l'angle de la psychologie du développement vers les systèmes d'apprentissage machine. Un réseau de neurones peut ainsi être entraîné progressivement, des données simples vers les plus complexes : ce type de protocole permet, suivant la tâche réalisée, de faciliter l'apprentissage et d'augmenter les performances. Nous pouvons remarquer que les données simulées, puis synthétiques et réelles, collectées dans les chapitres 1 et 2 peuvent être triées suivant leur degré de complexité. Il y a ainsi une hiérarchie dans les caractéristiques présentes sur chaque type d'image thermique : les cartographies reconstruites simulées présentent une surface homogène avec une fissure extrêmement simple et rectiligne. Les données synthétisées par diffusion apportent une granularité comparable aux données réelles, avec notamment des artefacts et des contours complexes, tout en étant disponibles en grande quantité. Les images obtenues expérimentalement sont quant à elles disponibles en quantités restreintes et serviront à réajuster le modèle.

Dans ce chapitre, nous développons un premier travail d'exploitation des données simu-

lées, synthétiques, puis expérimentales à l'aide d'une méthode d'apprentissage progressif pour les modèles de classification afin d'identifier les cartographies thermiques reconstruites avec ou sans fissure. Une étude bibliographique sur cette approche d'entraînement et ses cas d'applications est fournie dans la Section 3.1. Notre protocole qui exploite chacun des types de données est présenté dans la Section 3.2. Puis, les résultats des expérimentations d'apprentissage réalisées sont donnés dans la Section 3.3. Enfin, une étude d'ablations est décrite dans la Section 3.4, étudiant l'influence de l'approche proposée sur les performances de détection de fissures en condition expérimentale standard et en condition dégradée, ainsi que la part de chaque type de données, simulées, synthétiques ou réelles, dans les scores obtenus.

La méthode élaborée ici est appliquée au problème de la détection de fissures sur les échantillons métalliques complexes, sur des cartographies thermiques reconstruites. Elle a fait l'objet d'une présentation à la conférence GRETSI 2023 et d'un article de journal publié dans la revue QIRT en 2023 également.

Les notions d'entraînement de modèles et de tâche de classification sont abordées plus en détail dans les annexes A.1 et A.2. Le concept de transfert d'apprentissage est brièvement décrit dans l'annexe A.5.

3.1 État de l'art sur l'apprentissage progressif

L'apprentissage progressif est inspiré de l'étude de l'apprentissage en psychologie cognitive et comportementale, notamment via le jeu et son rôle dans le développement sensorimoteur des mammifères et des humains. Le concept de modelage (*shaping*) en psychologie, c'est-à-dire la mise en exergue d'une progressivité dans la réalisation d'une tâche, consiste d'abord à apprendre à traiter une version très simplifiée du problème, puis à complexifier la tâche progressivement jusqu'au problème réel final. La progressivité facilite largement l'apprentissage de nouvelles compétences et le conditionnement moteur, en particulier dans le développement de l'enfant. Par exemple, l'apprentissage progressif de la marche chez l'enfant où l'individu passe progressivement de la marche à quatre pattes jusqu'à la marche à deux pattes. Nous pouvons aussi mentionner l'apprentissage du langage par imitation qui suit également une forme de progressivité, de la simple imitation du langage parental jusqu'à l'apprentissage de la lecture et de l'écriture.

Cette approche, qui est développée en particulier dans la théorie du conditionnement de Pavlov [59], est apparue pertinente pour les systèmes d'apprentissage machine. Le principe clé de cette idée d'apprentissage progressif est illustré par une expérience de pensée typique dans la Figure 3.1. Ici, un système robotique passe par des environnements synthétiques simplifiés pour faciliter l'apprentissage du déplacement dans un environnement réel complexe. L'approche consiste en la mise en place d'un apprentissage progressif échelonné, partant de l'introduction de données génériques et simples pour aller ensuite vers des données aux propriétés plus spécifiques et complexes jusqu'aux données d'application finales [60]. Cette approche est efficace en cas de données limitées et a déjà été largement explorée pour la robotique et pour le traitement du langage. Notons en robotique un travail portant sur l'utilisation d'environnements simplifiés obtenus par simulation qui accélèrent l'apprentissage et permettent d'augmenter les capacités de navigation en situation réelle pour des systèmes robotiques [61]. D'autre part, l'apprentissage progressif a montré une capacité impressionnante pour amener un système d'apprentissage à une

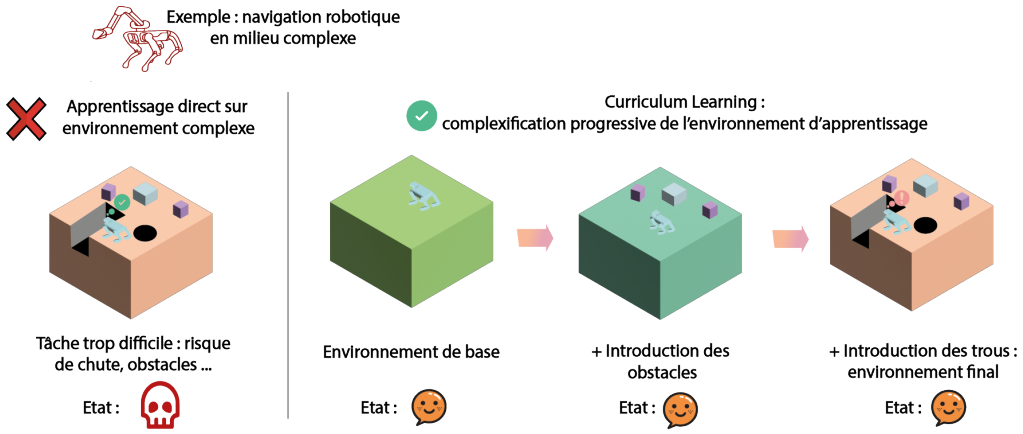


FIGURE 3.1 – Schéma illustrant l'apprentissage progressif dans un contexte de navigation robotique.

meilleure compréhension sémantique en traitement du langage naturel [62]. Les approches d'apprentissage progressif reposent sur l'utilisation de données simplifiées généralement issues de la simulation numérique, qui sont ensuite hiérarchisées suivant leur difficulté. Il y a peu de mise en œuvre de cette approche du côté de la vision par ordinateur, néanmoins une application peut être mentionnée comme exemple sur un problème de classification d'images [63], où des gains de performance notables sont obtenus sur la base de données de référence CIFAR-10 [64]. Une autre utilisation de l'apprentissage progressif particulièrement intéressante a été proposée dans le domaine médical : le score associé à la sortie fournie est utilisé pour hiérarchiser les données [65]. Néanmoins, l'utilisation de cette mesure n'est pas toujours un indicateur réel de la difficulté des caractéristiques de l'image d'entrée.

En l'état de nos connaissances, l'apprentissage progressif n'avait jusque-là pas été mis en œuvre dans le cadre du contrôle qualité et encore moins avec des données d'imagerie CND par thermographie active. Dans ce contexte, nous proposons d'abord d'entraîner le modèle pour la détection de fissures rectilignes simplifiées, comme des indentations, pour commencer à apprendre à détecter cet objet. Puis, nous proposons de passer à des fissures plus réalistes dans des contextes de détection de plus en plus difficiles contenant des artefacts de réflexion, de bruit ou bien de revêtements perturbant la diffusion de la chaleur et obtenues par synthèse d'image. Une dernière étape de réajustement sur les données réelles finalise l'apprentissage.

3.2 Protocole d'entraînement progressif

Cette idée d'apprentissage progressif semble pertinente vis-à-vis des données et des moyens disponibles dans le cadre de notre recherche. Nous avons des données produites par simulations très simples centrées sur le comportement thermique de la surface, et des images synthétiques par diffusion. Enfin, nous avons un volume relativement limité de cartographies thermiques sur échantillons assez complexes. Nous développons dans cette section notre protocole d'entraînement par apprentissage progressif appliqué pour la détection des fissurations de surface sur des cartographies thermiques reconstruites obtenues expérimentalement sur les échantillons complexes revêtus. Cette proposition

d'apprentissage progressif vise à former des caractéristiques de manière progressive en vue de faciliter l'apprentissage de systèmes machines et d'augmenter les performances de détection de fissures par rapport à un entraînement direct.

3.2.1 Procédure d'apprentissage

La Figure 3.2 décrit notre protocole d'entraînement basé sur l'apprentissage progressif, fournissant un exemple d'image d'entraînement pour chaque étape. Le protocole d'entraînement commence par un apprentissage de caractéristiques génériques sur des données simulées par éléments finis. Ces données très simples présentent une surface 2D complètement homogène, avec un échauffement localisé et une fissure rectiligne modélisée par une lame d'air. Cette première étape permet de focaliser l'apprentissage de représentation sur des formes simplifiées, faciles à appréhender, des aspects thermiques tels que la diffusion de chaleur sur la surface et comment celle-ci peut être altérée par la présence de défauts (étape 1). Nous avons un deuxième échelon d'apprentissage sur des données simulées où, cette fois, le réseau est exposé à la présence de contours rectilignes simplifiés (étape 2). Nous cherchons ici à faire distinguer au modèle d'apprentissage une différence de contraste thermique assimilée à un défaut réel et la discontinuité bien plus importante qui traduit le contour de la pièce.

Ensuite, le réseau est ré-entraîné sur des images synthétiques produites par nos modèles de diffusion (étape 3). Cette étape nous permet de finaliser l'apprentissage en passant des caractéristiques génériques à leurs pendants plus spécifiques, associés à la diffusion de chaleur sur une pièce complexe. Un ajustement du modèle est finalement réalisé sur un échantillon limité de données réelles (étape 4), ce qui conclut l'apprentissage.

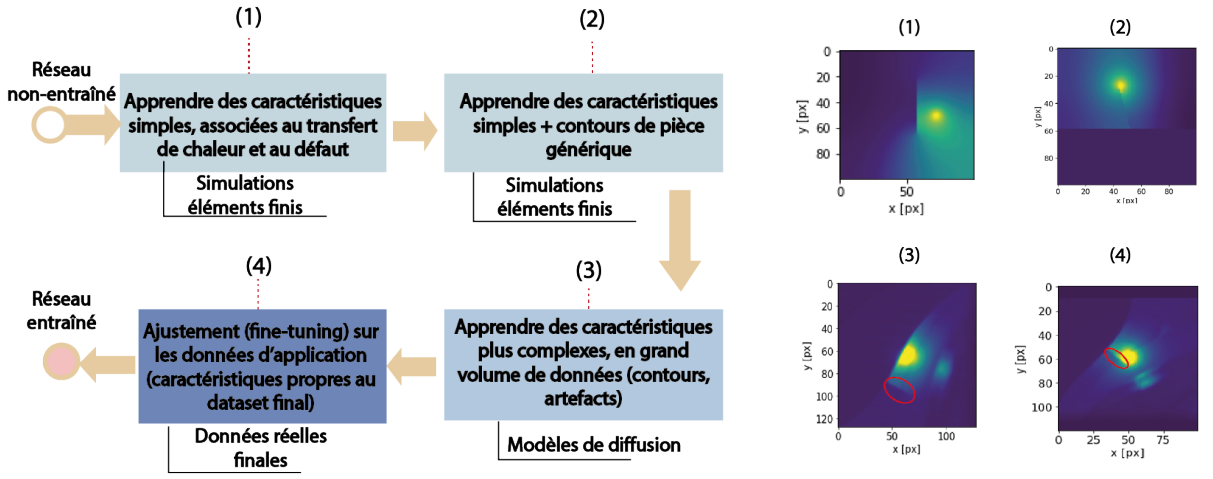


FIGURE 3.2 – À gauche, schéma-bloc de la procédure d'apprentissage progressif. À droite, exemples illustrant les données utilisées pour chaque étape de l'apprentissage.

Des augmentations de données typiques telles que les inversions horizontales et verticales, ainsi que des rotations aléatoires sont appliquées pour l'entraînement de toutes les architectures, et cela à tous les échelons. C'est une stratégie commune afin d'augmenter artificiellement les volumes de données limités. Pour toutes les architectures, les poids sont initialisés aléatoirement.

3.2.2 Bases de données utilisées

Les données utilisées pour notre apprentissage progressif sont décrites plus en détail dans le chapitre 1 (données réelles et simulées), ainsi que dans le chapitre 2 pour les données synthétiques générées par modèle de diffusion par débruitage (DDPM). Nous revenons ci-dessous sur leur rôle dans la perspective de la formation de caractéristiques et de l'apprentissage progressif :

— Un premier ensemble d'entraînement a été produit par la simulation par éléments-finis et présente uniquement le champ de température simulé sur la surface examinée sans aucun contour ou structure. La surface est de plus complètement homogène. Cela permet de focaliser l'apprentissage sur la caractéristique de classification la plus importante : la fissure, et comment celle-ci altère la diffusion de la chaleur, pour une grande diversité de paramètres. Cet ensemble de données est désigné par FEM1 et contient 13 000 images thermiques sans fissures et 15 000 avec fissures.

— Un deuxième jeu de données simulées FEM a été constitué avec cette fois l'adjonction d'un contour extrêmement simple rectiligne séparant l'environnement simulé de l'air du reste de la surface métallique. Une petite rotation est ajoutée à ce bord simulé augmentant la variabilité des contours. Cet ensemble de données est nommé FEM2. La variabilité due à l'orientation et à l'emplacement de la surface métallique simulée n'est pas incorporée dans ce second ensemble de données de simulation : elle sera introduite par des augmentations de données communes pendant l'entraînement telles que la rotation ou les retournements miroir. Ce second ensemble de données simulées contient 6 000 images thermiques. Il s'agit d'un ensemble équilibré entre les images avec fissures et sans fissures.

— 20 000 images synthétiques générées par modèles de diffusion ont permis d'agrandir le jeu de données réelles. Ces données introduisent le modèle d'apprentissage aux contours réels complexes, aux artefacts et reflets, aux variabilités du dépôt de surface, le tout sur un volume de données largement augmenté.

— Un petit jeu de données réelles est utilisé pour raffiner l'apprentissage, afin de maximiser la performance sur le jeu d'images finales.

La Table 3.1 fournit un récapitulatif des bases de données employées, dans l'ordre, avec le type et le volume de données associés. Des renvois vers les chapitres décrivant mieux ces données sont de plus ajoutés. Les bases DAWN-b et DAWN-J sont strictement utilisées pour l'étude d'ablations seulement. Rappelons que l'ensemble des bases de données hors-ablation est équilibré, c'est-à-dire contenant quasiment autant de cartographies avec et sans fissure.

La Figure 3.3 fournit une illustration conceptuelle supplémentaire permettant de mettre en valeur les différents types de données vues par le modèle à chaque étape d'apprentissage, des propriétés simples des images simulées aux contours complexes et artefacts des données réelles.

3.3 Expérimentations

Nous présentons ici les modèles neuronaux sur lesquels l'approche d'apprentissage progressif proposée a été appliquée. Nous nous concentrons dans cette partie sur une tâche

Base	Source	# images	Paramètres d'examen	Chapitre
FEM-1	Simulation	28 975	plage élargie	Chap. 1
FEM-2	Simulation	6 000	plage élargie	Chap. 1
DAWN-E	Expérimentation	600	plage élargie	Chap. 1
DAWN-Synths	Diffusion	20 000	-	Chap. 2
DAWN	Expérimentation	330	paramètres optimaux	Chap. 1
DAWN-B	Expérimentation	21	paramètres optimaux	-
DAWN-J	Expérimentation	950	plage dégradée	-

TABLE 3.1 – Récapitulatif des différentes bases de données employées. L'origine, le nombre d'images et une indication des paramètres d'examen sont ajoutés.

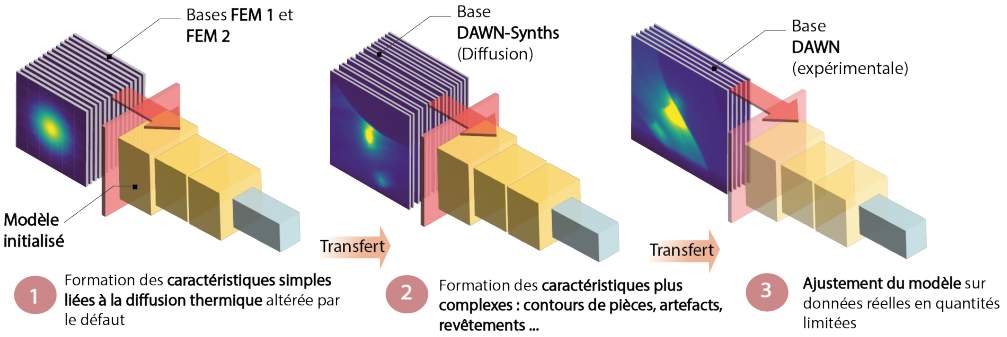


FIGURE 3.3 – Apprentissage progressif pour la FST, de la donnée simulée des bases FEM 1 et 2 aux données générées par diffusion puis aux données réelles.

de classification binaire d'images, c'est-à-dire la capacité d'un modèle d'apprentissage à identifier les images thermiques reconstruites comme présentant une fissure ou non. Nous fournissons ensuite les résultats obtenus par l'apprentissage proposé face à un entraînement direct, autrement dit un entraînement effectué directement sur les données réelles . Enfin, une étude permettant d'identifier les régions des images impactant la prise de décision des modèles est conduite dans un objectif d'explicabilité des modèles neuronaux.

3.3.1 Architectures

Nous testons diverses architectures d'apprentissage profond pour la classification binaire des cartographies thermiques. Premièrement, nous évaluons un réseau convolutif, VGG 13 [66]. Puis nous comparons ce modèle avec des architectures basées sur le mécanisme d'attention appelées les *Transformers*. Ce mécanisme, et plus précisément l'auto-attention, permet un apprentissage hiérarchisé associant les différentes régions de l'image d'entrée [67,68]. Le mécanisme d'attention est décrit en détail dans l'annexe A.6. Nous retenons deux réseaux de l'état de l'art basés sur le mécanisme d'attention, les réseaux Swin et CaiT [69, 70]. Nous testons également une méthode issue de l'apprentissage machine traditionnel, constituée d'un filtre à histogrammes de gradients orientés (HOG) comme extracteur de caractéristiques, et un séparateur à vastes marges (SVM) pour réaliser la classification [71]. Elle fournit un repère de performances pour les méthodes classiques.

3.3.2 Métriques de classification

Pour quantifier la capacité des modèles d'apprentissage à distinguer les images présentant ou non un défaut, nous avons évalué les métriques suivantes : l'*accuracy*, le F1-score, la précision et enfin le rappel [72].

L'*accuracy* représente la proportion d'échantillons correctement prédits sur le nombre total d'échantillons. C'est une estimation globale des performances de détection du modèle, sans différenciation entre détections manquées et fausses détections. Cette métrique peut néanmoins fournir une mesure peu fiable dans les cas de gros déséquilibre entre les différentes classes dans les données. Les bases de données sont équilibrées dans notre cas, donc cette métrique reste pertinente dans notre contexte d'étude.

La précision quantifie la proportion de prédictions de fissures par le modèle qui sont correctement classées comme présentant un défaut parmi toutes les prédictions avec défaut proposées par le modèle. Le rappel (ou sensibilité) en revanche estime la proportion de prédictions correctes de fissures parmi toutes les instances présentant réellement un défaut. Pour simplifier, la précision fournit une quantification des fausses détections de fissures, tandis que le rappel estime les détections de défauts manquées. Une précision élevée indique un modèle qui confond peu d'images saines comme endommagées, tandis qu'un bon rappel indique un détecteur qui rate peu de fissures. Le rappel peut être assimilé à la probabilité de détection employée dans le contrôle des matériaux.

Le score F1, ou *f-score*, qui combine la précision et le rappel en une seule métrique, est particulièrement utile. Il est calculé en prenant la moyenne harmonique de la précision et du rappel, fournissant des aperçus précieux sur les capacités prédictives du modèle et sa performance à travers différentes classes de données. C'est une métrique globale qui est comparable à l'*accuracy* mais est moins sensible aux dichotomies dans la distribution statistique des données. Les équations correspondant à chaque métrique sont données ci-dessous :

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} = \frac{\text{Classifications correctes}}{\text{Total}}$$

$$\text{Precision} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{3.1}$$

$$\text{rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{3.2}$$

$$\text{Score F1} = \frac{2 \times \text{Precision} \times \text{rappel}}{\text{Precision} + \text{rappel}} \tag{3.3}$$

avec les faux positifs (FP) correspondant à des fausses alarmes, tandis que les faux négatifs (FN) correspondent à des détections manquées. Les vrais positifs (VP) sont des images thermiques de fissures vraies, tandis que le vrai négatif (VN) correspond à des images thermiques sans fissure vraies. Une illustration est fournie Figure 3.4 pour ces concepts, et les différents éléments qu'ils capturent, dans la capacité d'un modèle à détecter les objets-cibles.

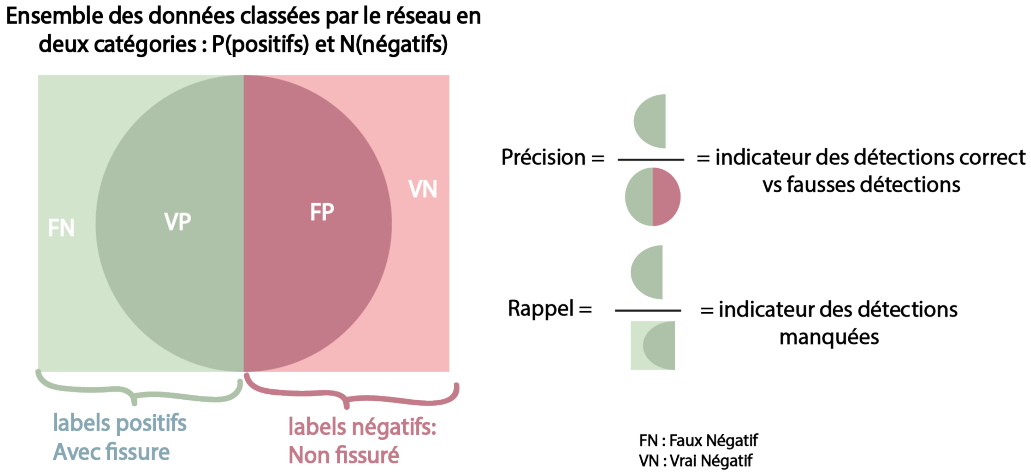


FIGURE 3.4 – Illustration du rappel et de la précision pour un problème de classification binaire.

3.3.3 Impact de l'apprentissage progressif

La Table 3.2 fournit l'évaluation des métriques de performances pour les différentes architectures entraînées, d'abord dans le cas d'un entraînement direct i.e. seulement sur les données réelles de la base DAWN. Puis les scores après passage dans notre protocole d'apprentissage progressif sont fournis. Les scores sont élevés pour l'ensemble des architectures étudiées, surpassant la méthode de référence d'au moins 5 % en *accuracy*. Nous pouvons ensuite voir des gains dans les métriques pour tous les modèles sélectionnés suite à l'apprentissage progressif, par exemple avec le modèle CaiT qui passe de 90 à 99 % d'*accuracy*. Si la capacité à identifier correctement les images est globalement améliorée comme montré par l'*accuracy* et le F1-score, nous pouvons aussi constater grâce à notre méthode d'apprentissage un rééquilibrage entre les fausses détections et les détections manquées i.e. un écart rappel-précision plus réduit.

Méthode	Architecture	*accuracy* [%]	F1	Précision	Rappel
Référence	HOG+SVM	83	0.83	0.82	0.83
Entraînement direct	VGG13	92	0.92	0.95	0.90
	Swin	87	0.86	0.92	0.86
	CaiT	90	0.91	0.93	0.89
Apprentissage progressif	VGG13	97	0.97	0.96	0.99
	Swin	96	0.96	0.94	0.98
	CaiT	99	0.98	0.98	0.99

TABLE 3.2 – Performances de classification binaire obtenues en entraînement direct ou avec la méthode d'apprentissage progressif.

3.3.4 Explicabilité de la classification : GRAD-CAM

Nous avons réalisé une étude permettant d'identifier les régions sur l'image d'entrée qui ont influencé la prise de décision du modèle. Cette étude se base sur le calcul des

gradients au sein du modèle, réalisée grâce à la librairie *pytorch-Grad-CAM* pour cette mesure de gradient [73].

Cette mesure de gradient est calculée par rapport à la prédiction du modèle, puis propagée jusqu'aux pixels de l'image d'entrée. La carte obtenue est appelée carte d'activation. Cette méthode permet donc de tirer des association entre les éléments vus par le modèle d'apprentissage sur l'image et la prise de décision. Notons que cette étude ne donne pas d'information sur les critères de décision, mais indique les régions impactant la prise de décision par le modèle.

Nous comparons les activations neuronales entre le modèle VGG et le modèle CaiT afin d'identifier des potentielles divergences de comportement entre les modèles basés sur les convolutions et ceux basés sur les *Transformers*. Le gradient est estimé à partir de la sortie souhaitée pour le modèle : nous nous concentrons ici sur l'activation pour la classe des images fissurées. Dans le cadre de cette étude et pour les deux modèles c'est l'avant-dernière couche d'extraction de caractéristiques avant la prise de décision par le modèle qui est choisie pour effectuer la mesure de l'activation, car formant les représentations les plus abstraites et profondes du modèle [74]. Notons que nous intéressons ici à l'étude des régions qui activent le plus la couche étudiée dans le modèle, et non sur l'intensité absolue de cette activation.

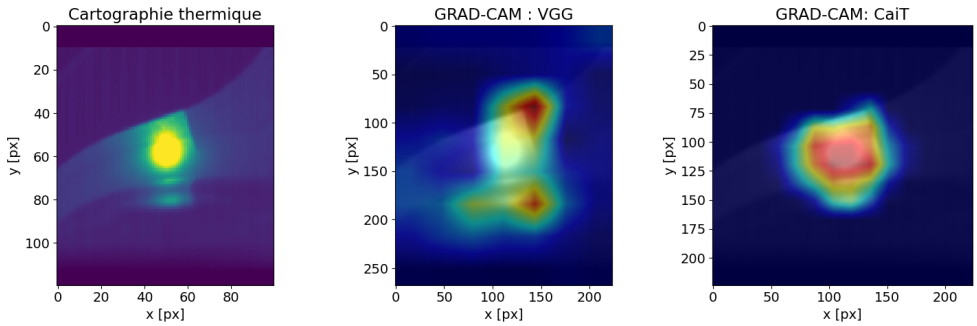


FIGURE 3.5 – Carte d'activation obtenue par méthode GRAD-CAM pour les deux modèles VGG et CaiT, après un apprentissage progressif. Exemple avec fissure.

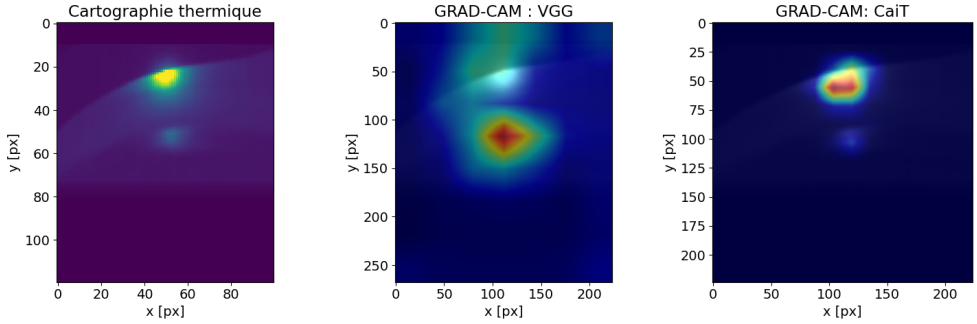


FIGURE 3.6 – Carte d'activation obtenue par méthode GRAD-CAM pour les deux modèles VGG et CaiT, après un apprentissage progressif. Exemple sans fissure.

***CaiT* versus VGG.** Après observation de l'ensemble des cartes d'activations sur les données de test, nous présentons des exemples de carte d'activité avec et sans apprentis-

sage progressif dans les Figures 3.5 et 3.6 afin d'illustrer l'activité typique mesurée pour une cartographie reconstruite avec puis sans fissure. Pour les deux modèles étudiés nous observons une association entre la zone du lobe de l'échauffement thermique, la région de la fissure sur l'image, et l'activité neuronale mesurée. Nous remarquons sur les Figures que CaiT présente une activité forte sur la trajectoire de la source laser sur la surface, et moindre sur la fissure elle-même. Au contraire du VGG qui a une activité concentrée sur la fissure et moindre sur la zone d'échauffement. Une hypothèse pour expliquer ces différences d'activation neuronale pourrait être que CaiT, utilisant le mécanisme d'attention, se concentre bien plus sur le contexte, et en particulier sur l'observation de la source de la chaleur pour sa prise de décision. Le modèle VGG, basé sur les convolutions (opération plus locale), semble se restreindre à la géométrie de la fissure ou à des éléments comme les amorces de fissures et les contours.

La Figure 3.7 fournit des exemples de cartes d'activation supplémentaires, afin d'appuyer le discours sur les différences d'activation entre VGG et CaiT.

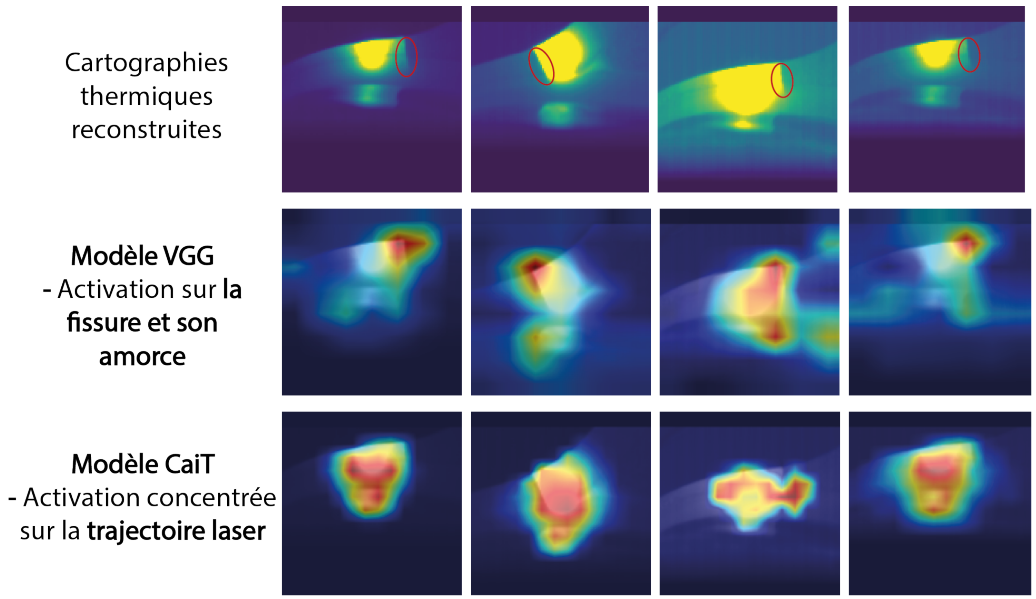


FIGURE 3.7 – Exemples supplémentaires de cartes d'activation obtenues sur des cartographies thermiques reconstruites présentant une fissure. Les activations pour VGG sont en haut, celles pour CaiT sont en bas. La fissure est entourée en rouge sur chaque cartographie reconstruite.

Apprentissage progressif versus apprentissage direct. Nous fournissons un exemple de carte d'activation sur la Figure 3.8 avec et sans apprentissage progressif pour le modèle VGG. L'étude de l'ensemble des cartes d'activation de la base de données nous montre que l'apprentissage direct induit une activation importante autour du halo de chauffe et de l'endommagement, tandis que l'apprentissage progressif induit une activité plus localisée sur la fissure. Il est difficile d'expliquer ce changement dans l'activation entre les deux types d'entraînement, néanmoins nous pourrions tendre à dire que l'apprentissage progressif permettrait au modèle de se focaliser plus facilement sur la fissure.

Pour le modèles basé sur les *Transformer*, peu de différences sont constatées : pour les deux types d'entraînement l'activation est concentrée autour de la trajectoire de chauffage laser. Nous constatons seulement sur quelques exemples une activation sur une région un

peu élargie par rapport à la zone de chauffage après l'apprentissage progressif, comme sur l'exemple fourni par la Figure 3.9.

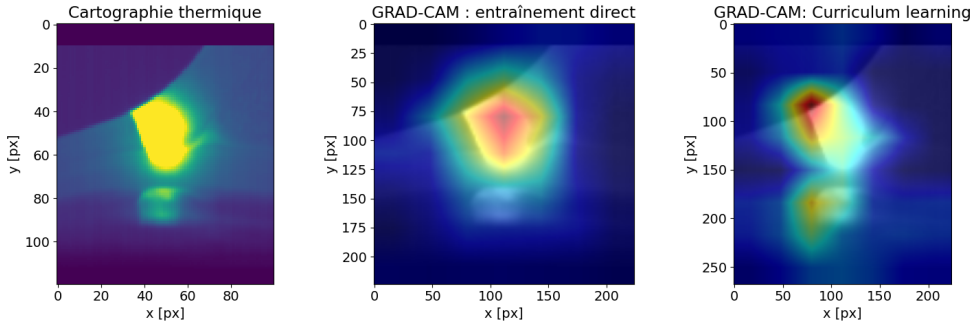


FIGURE 3.8 – Carte d'activation fournie par le modèle VGG, en entraînement direct et après apprentissage progressif. Autre exemple avec fissure.

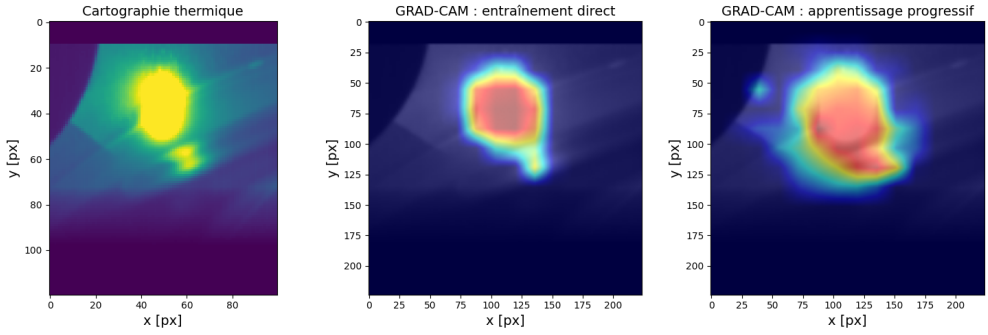


FIGURE 3.9 – Carte d'activation fournie par le modèle CaiT, en entraînement direct et après apprentissage progressif. Autre exemple avec fissure.

3.4 Etude d'ablations

Nous procédons ensuite à une étude d'ablations qui consistent à altérer des éléments de l'apprentissage pour en estimer l'impact relatif sur les métriques. La première ablation consiste ici à retirer une étape de notre protocole d'entraînement ce qui permet d'en déduire son influence relative sur l'apprentissage global de la tâche demandée suivant le modèle de détection étudié. Nous avons ensuite évaluer la robustesse de notre méthode d'entraînement face à différents changements dans les données expérimentales. Nous étudions d'abord l'apprentissage sur des échantillons complètement inconnus, puis sur des reconstructions thermiques dégradées où le défaut et les différents éléments structuraux comme les contours de pièce sont plus difficiles à identifier. Ces expérimentations sont réalisées sur les architectures VGG13, Swin et CaiT. Une étude de l'influence du volume d'images synthétiques sur l'apprentissage est enfin menée en fin de section.

3.4.1 Impact relatif des différents volumes de données

Nous étudions l'impact de chaque type de données sur les performances de détection avec notre apprentissage progressif. La Table 3.3 fournit les performances de classification obtenues par chaque modèle, après le retrait de chacune des étapes de l'apprentissage progressif proposé.

Nous constatons comme point marquant la perte de 36 % de performances lorsque les données produites par diffusion sont retirées de l'apprentissage et sans rajustement de modèle. Cela appuie le fait que les données simulées par élements-finis sont trop simplistes pour permettre directement de bonnes performances sur données réelles. Au contraire, les données synthétiques par diffusion permettent d'obtenir des performances proches de la procédure d'apprentissage complète, même sans réajustement sur les données réelles. Enfin le retrait de l'ajustement du modèle sur les données expérimentales finale ne diminue que de 3 % l'*accuracy* : cet ajustement apporte donc finalement assez peu sur la performance de détection. Cela témoigne aussi de la pertinence de la synthèse d'images par diffusion : leur apport suffit pour permettre au réseau d'augmenter de manière notable les performances de détection de fissures sur les données FST.

Ablation	Données réelles ?	VGG13			Swin			CaiT		
		Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.
Pas d'étapes 1,2 (FEM)	Oui	96	0.97	0.97	98	0.97	0.99	95	0.93	0.97
	Non	96	0.94	0.99	96	0.94	0.99	94	0.91	0.99
Pas d'étape 3 (diff.)	Oui	61	0.75	0.67	66	0.70	0.65	70	0.72	0.65
	Non	55	0.55	0.99	57	0.56	0.83	55	0.54	0.94
Pas de données réelles		97	0.94	0.99	94	0.89	0.99	97	0.94	0.99

TABLE 3.3 – Performances obtenues par les modèles de détection sur la base DAWN lors de l'étude d'ablations.

La conclusion de cette ablation est que le rôle des images simulées par éléments-finis est moins déterminant particulièrement comparé à celui de la synthèse d'images par diffusion, pour notre protocole d'apprentissage progressif.

3.4.2 Généralisation aux échantillons inconnus

Nous cherchons à mesurer l'influence de notre méthode d'apprentissage progressif sur les capacités de détection dans le cas de nouveaux échantillons. Pour cela deux échantillons ne faisant pas partie de la base d'entraînement DAWN sont exploités pour évaluer la capacité de généralisation. L'échantillon 4 est pratiquement identique aux autres échantillons, tandis que le numéro 2 présente une marque d'inclusion sur sa surface qui est absente du jeu d'échantillons de départ. Les modèles CaiT et VGG13 sont entraînés soit par apprentissage direct soit par apprentissage progressif.

Le Tableau 3.4 regroupe les performances obtenues en *accuracy* de test pour 2 échantillons non vus qui sont notés 2 et 4 à la fois en entraînement direct et en apprentissage progressif. Pour l'échantillon 4 ce tableau montre que le processus d'entraînement basé sur l'apprentissage progressif augmente la robustesse face aux deux échantillons inconnus, comparé à l'entraînement direct. La baisse de performance est plus marquée pour l'échantillon 2. Cela peut être expliqué par une surface anormale en comparaison avec

d'autres échantillons, comme le montre la figure 3.10. Cette limite de revêtement atypique pourrait perturber la classification de l'image par le modèle. Dans notre étude, les *Transformers* semblent également avoir une meilleure capacité de généralisation que les réseaux convolutifs plus classiques ce qui est cohérent avec les travaux récents sur ce type d'architecture [75].

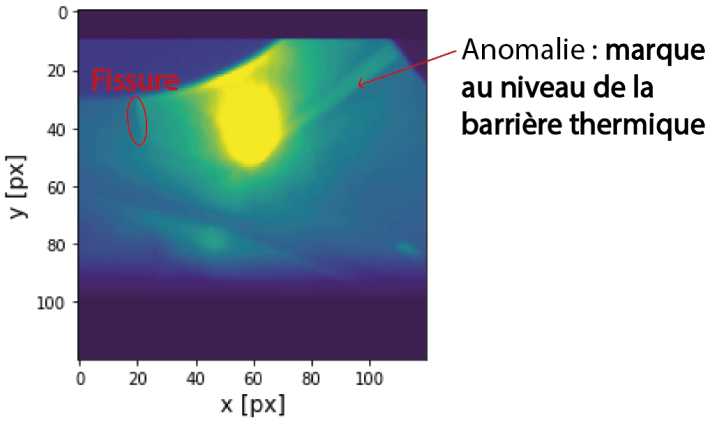


FIGURE 3.10 – Exemple d'image thermique de l'échantillon 2. La spécificité de la surface examinée peut perturber la détection.

Ech.	#image	Acc. (CaiT, direct)	Acc.(CaiT, progress.)	Acc.(VGG13, direct)	Acc.(VGG13, progress.)
2	11	59	82	41	77
4	10	90	96	78	90

TABLE 3.4 – Accuracy obtenues en test sur le jeu d'échantillons inconnus, correspondant aux échantillons non vus lors de l'entraînement. Le nombre d'images par échantillon est indiqué.

Globalement nous constatons une diminution des performances sur les nouveaux échantillons. Néanmoins l'approche par apprentissage progressif permet de mitiger la baisse de performance sur ces nouvelles données.

3.4.3 Conditions expérimentales dégradées

Nous évaluons ici la capacité d'apprentissage des modèles grâce à notre approche dans un contexte d'acquisitions FST dégradées i.e. présentant par exemple un éclairement thermique limité, une saturation très importante de la source laser masquant les contours des pièces. Une base de données dégradées a ainsi été constituée à partir du banc d'essai, notée DAWN-J, illustrés avec la Figure 3.11 montrant un exemple de cartographie reconstruite après rehaussement de contraste. Elle regroupe tant les acquisitions prises à réglages optimales que d'autres obtenues en condition dégradées : fissures loin de l'échauffement, et échauffement beaucoup plus limité (longueur de balayage réduite et vitesse de scan élevée). Nous obtenons ainsi les scores suivants, présentés dans la Table 3.5 lorsque DAWN est remplacée par cette base DAWN-J.

Architecture	*accuracy* (direct)	*accuracy* (progressif)
VGG13	pas de convergence	69
Swin	73	74
CaiT	72	78

TABLE 3.5 – Scores obtenus sur la base DAWN-J pour les architectures sélectionnées, en entraînement direct et en apprentissage progressif.

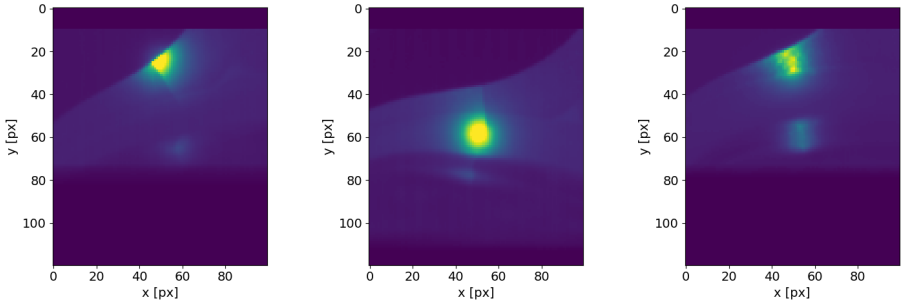


FIGURE 3.11 – Exemples d'images dites en condition dégradée. Un réhaussement de contraste CLAHE est appliqué ici, permettant de mieux voir les contours. L'échauffement thermique est plus réduit dans l'ensemble : il y a moins d'information contextuelle pour guider l'apprentissage.

Nous pouvons voir que ces données dégradées et plus contrastées perturbent fortement l'apprentissage direct, avec notamment une des architecture d'apprentissage incapable d'apprendre (VGG). Cependant nous constatons que notre méthode d'apprentissage permet un gain net dans la performance de discrimination des cartographies thermiques. Nous pourrions expliquer ces gains par une meilleure robustesse des représentations formées durant l'apprentissage progressif qui résistent mieux à l'introduction de données bruités ou dégradées. Cela ouvre des perspectives pour le redéploiement par ajustement du modèle en particulier sur des données où les paramètres physiques seraient radicalement différents.

3.4.4 Impact du volume d'images générées par diffusion

Nous cherchons à évaluer l'impact de la quantité de cartographies synthétiques sur les gains de performance obtenues grâce à notre proposition d'apprentissage progressif. Pour cela nous limitons ici la quantité de données utilisées lors de la phase d'apprentissage par diffusion, avec plusieurs sous-échantillons de la base synthétique. Cela permet d'identifier un seuil relatif à partir duquel l'apport de nouveaux volumes de données synthétiques de ce type n'a peu/plus d'impact sur la capacité d'un modèle à mieux détecter la fissure sur les cartographies expérimentales. Pour réduire le coût calculatoire et temporel de cette étude, l'expérience est conduite directement sur un réseau initialisé aléatoirement, sans apprentissage sur données de simulation FEM. Nous limitons aussi le nombre d'époques à 10, pour le pré-apprentissage sur le lot d'images diffusées comme pour l'ajustement sur les données réelles, permettant de produire rapidement cette expérience mais expliquant un décalage potentiel avec les scores généraux précédents à volume de données équivalents.

Le graphique 3.12 fournit l'évolution de l'*accuracy* sur données réelles finales pour dif-

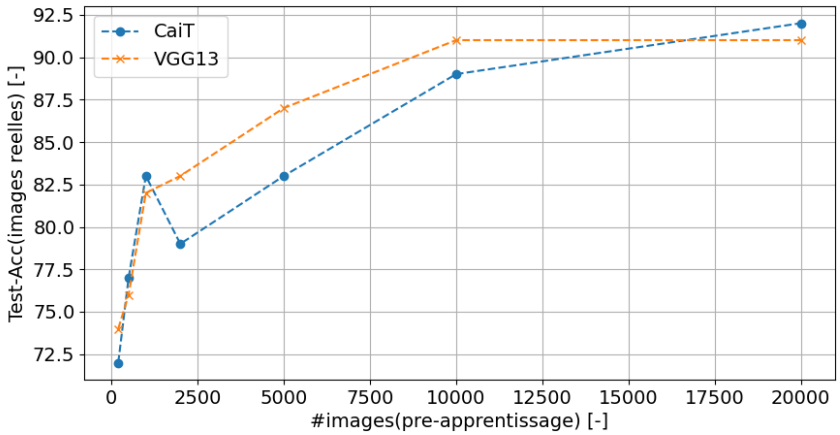


FIGURE 3.12 – Évolution de l'*accuracy* sur données réelles en fonction de la taille d'échantillons d'images synthétiques pour pré-apprentissage pour les architectures VGG13 et CaiT.

férentes tailles de sous-échantillons de la base de pré-entraînement synthétique. L'*accuracy* semble augmenter continuellement suivant la taille de la base de pré-apprentissage tant pour CaiT que pour VGG, bien que pour la dernière architecture on tend à identifier un plateau après 10.000 à 15.000 cartes synthétiques.

3.4.5 Conclusions des ablations

Les ablations conduites ici nous permettent de tirer quelques directions pour l'utilisation des données artificielles dans la suite du travail. D'abord, la mesure de l'influence de chaque type de données montre la place limitée des données de simulation dans la performance de détection des modèles entraînés par apprentissage progressif, comme indiqué dans la section 3.4.1. Ensuite, nous avons pu voir que notre procédure d'apprentissage permettait une meilleure capacité de généralisation du modèle de détection face à de nouveaux échantillons, comme décrit dans la section 3.4.2. Nous avons aussi évalué la robustesse de la méthode d'apprentissage développée face à des conditions d'examen difficiles, dans la section 3.4.3. L'impact de la quantité de données de synthèse générées sur la performance a été évalué dans la section 3.4.4. Cette étude tend à indiquer la présence d'un seuil au-delà duquel le nombre d'images synthétiques n'a plus d'impact sur la performance.

3.5 Conclusion

Nous avons donc vu que l'approche d'apprentissage progressif proposée permet d'augmenter la capacité des modèles d'apprentissage à mieux distinguer les cartographies thermiques avec et sans fissures. Les scores de détection ainsi que l'étude d'ablation conduite tendent à montrer l'efficacité de l'utilisation des données synthétiques produites par diffusion pour l'augmentation des performances de détection, mais aussi l'intérêt limité des données simulées par Comsol comparativement aux données synthétisées par diffusion.

Une autre approche pourrait être le pré-apprentissage sur une base de pré-entraînement sur des données réelles, qui pourrait être plus adaptée pour pré-apprendre des caractéristiques utiles pour la détection de la fissure dans un environnement simplifié : une surface homogène par exemple, avec une certaine variété de défauts plus faciles à détecter sur la surface. Cette approche sera investiguée notamment dans le chapitre suivant.

Plusieurs idées pourraient améliorer notre proposition d'apprentissage progressif.

Une première idée serait d'ajouter des modules de simulations plus complexes ou de génération par diffusion intermédiaire simplifiant certaines structures. Cependant, ces ajouts risqueraient d'alourdir sensiblement le nombre d'étapes d'entraînement. Le gain en performance vis à vis de cet alourdissement sera à évaluer. Une autre approche serait de mieux redistribuer les données d'entraînement, sans a priori humain en s'inspirant de [65], qui utilise le score de détection comme mesure de la difficulté de la tâche sur les données.. Néanmoins, la mesure de score associé à la classe n'est pas la plus fiable. L'exploration d'autres mesures utilisées en apprentissage incrémental ou en apprentissage actif pourrait alors permettre de développer des stades d'entraînement supplémentaires pour un apprentissage progressif entre les différentes données, qui soient plus pertinents que la seule séparation entre données simulées, puis synthétiques, et enfin réelles, pour les performances de détection de fissures par FST.

Enfin, l'utilisation de générateurs d'images pour produire la base de diffusion peut être considérée comme un *proxy* pour transférer les représentations apprises par le modèle génératif vers un modèle spécialisé en détection, par exemple. Cela peut revenir à une forme d'apprentissage maître-étudiant (*Teacher-Student* en anglais) inclus dans les techniques dites de distillation de connaissances (*Knowledge Distillation*) [76,77]. Suivant un certain travail de construction architecturale, nous pourrions alors envisager la récupération des caractéristiques latentes apprises par les modèles génératifs, plutôt que de procéder à un transfert des caractéristiques par l'intermédiaire d'images synthétiques qui restent coûteuses en temps de calcul pour être produites en grandes quantités. Le transfert de caractéristiques des modèles de diffusion est assez complexe et encore peu étudié, néanmoins nous notons quelques travaux en détection d'anomalies donnant des premières pistes de recherche [78–80].

Cette étude portait sur la détection de la fissure. Nous allons désormais nous focaliser sur la tâche de localisation précise du défaut sur les données thermographiques. Dans le prochain chapitre, nous nous intéresserons à des architectures de détection d'objets plus grandes et plus complexes à mettre en œuvre et verrons comment une approche de transfert d'apprentissage peut être appliquée à ces modèles, au travers de l'utilisation de données collectées en grande quantité sur des échantillons très simples mais présentant une grande diversité de fissures. Nous allons aussi nous intéresser à l'exploitation des images thermiques individuelles en lieu et place des cartographies thermiques reconstruites : cela permettrait à la fois de réaliser une détection temps-réel mais aussi d'augmenter considérablement le volume de données disponible.

Chapitre 4

Localisation de fissures par pré-apprentissage depuis des surfaces simples

Il y a, entre le système cérébral humain et son environnement, une incertitude fondamentale qui ne peut être comblée : la biologie de la connaissance nous montre, en effet, qu'il n'y a aucun dispositif, dans le cerveau humain, qui permette de distinguer la perception de l'hallucination, le réel de l'imaginaire.

Introduction à la pensée complexe,
E. Morin.

Sommaire

Le laboratoire des matériaux de l'ONERA réalise en continu des essais mécaniques de traction et de fatigue, dont nous avons pu récupérer les rebuts. L'examen FST de ces échantillons obtenus au cours de la thèse nous a permis de construire une base de données élargie. Cette base présente des enregistrements thermographiques avec une grande variété de fissurations sur une surface néanmoins assez simple : les contours des pièces sont rectilignes et la surface est homogène, sans géométries particulières. L'apprentissage sur ces données collectées peut s'inscrire dans une approche d'apprentissage progressif où l'on apprend d'abord les caractéristiques de l'objet cible, puis la robustesse aux artefacts et sources de bruit, par transfert des apprentissages. Contrairement au chapitre 3, l'objectif ici est de localiser la fissure.

Nous présentons d'abord brièvement l'idée de transfert d'apprentissage depuis une surface simple dans la section 4.1. Puis nous évaluons les performances de classification

pour ces échantillons simples en préambule, dans la section 4.2. Ensuite, nous nous intéressons dans la section 4.3 à la tâche de localisation du défaut à travers une approche de transfert des représentations formées sur la base d'échantillons simples appelée FLYD vers des échantillons plus complexes dont les bases sont volontairement réduites et plus complexes afin d'augmenter la difficulté de l'apprentissage, issues des acquisitions sur les échantillons complexes de la base de données DAWN. Nous expérimentons aussi la détection apprise image par image, c'est-à-dire directement sur l'image thermique individuelle, qui permettrait de tendre à une détection en temps-réel du défaut lors de l'examen. Nous évaluons dans la section 4.4 les bénéfices de performance de localisation de fissures sur ces bases FLYD couplées avec la synthèse d'images par les modèles Stable Diffusion. Nous nous intéressons enfin à l'impact de la longueur de fissuration sur la localisation par un modèle dans la section 4.5.

Le travail présenté ici a fait l'objet d'une communication lors de la conférence *Quality Control by Artificial Vision* (QCAV 2023), focalisée sur la constitution de la base de données FLYD, ainsi que d'un article de revue, dans le *Journal of Electronic Imaging*, concentré sur la localisation d'objet et le transfert d'apprentissage. Le lien GitHub suivant donne accès en ligne aux différentes bases de données FLYD, ainsi que quelques résultats et architectures : https://github.com/kevinhelvig/FLYD.

Les notions d'entraînement de modèles pour la vision par ordinateur sont abordées plus en détail dans les annexes A.1 et A.2. L'idée de transfert d'apprentissages est synthétisée dans l'annexe A.5. Le lecteur peut aussi se référer à l'annexe A.6 sur les *Transformers* et le mécanisme d'attention.

4.1 Apprendre et transférer des caractéristiques génériques

La théorie des représentations présentée brièvement en annexe A.5 peut servir de socle de compréhension pour le concept de transfert d'apprentissages d'un problème à un autre. Nous pouvons voir l'apprentissage comme l'extraction de propriétés associées aux données qui sont utiles à la résolution d'une tâche. Ces caractéristiques abstraites se révèlent plus ou moins ré-adaptables suivant la profondeur du réseau utilisé, et donc ré-applicables à un autre problème. Nous proposons d'utiliser cette approche de transfert des caractéristiques à partir des données de la base FLYD d'échantillons simples. Cet apprentissage sur des surfaces simples et homogènes, mais avec une plus large diversité de fissures, permet de former des représentations abstraites propres à une variété large de défauts avec des formes et des orientations de fissurations diversifiées, tout en étant dans le cas d'une scène simple où la détection est facilitée. Les caractéristiques formées par un modèle sur ces données peuvent servir de premier modèle d'initialisation transférable à des problèmes spécifiques où le régime de données est plus restreint et où la détection de l'endommagement est plus difficile, typiquement sur des pièces spécifiques où la quantité de données de FST est plus limitée, et avec des échantillons présentant un revêtement ou une géométrie de pièce plus complexe.

Cette approche est analogue à l'apprentissage progressif décrit dans le chapitre 3 et qui exploite la progression dans la complexité des données vues.

Une autre question est de savoir s'il est possible de transférer des représentations

formées par le réseau depuis les cartographies thermiques reconstruites vers les images thermiques individuelles lors de l'enregistrement.

L'annexe A.5 fournit quelques détails techniques et illustrations didactiques supplémentaires sur le transfert d'apprentissage.

4.2 Classification sur des surfaces simples

Nous présentons ici les résultats de l'entraînement d'un groupe de modèles de classification binaire sur la base FLYD. Ce travail fournit un point de comparaison avec les performances obtenues sur les échantillons complexes dans le Chapitre 3.3.3. Une étude d'activation neuronale pour l'explicabilité grâce à l'outil GRAD-CAM est aussi menée [74].

4.2.1 Base de données

Les échantillons constituant la base de données FLYD sont les échantillons d'essais mécaniques de fatigue essentiellement (voir Chapitre 1 pour plus de détails). Plusieurs jeux ont été constitués. FLYD-C est une base de données de classification binaire entre cartographies thermiques reconstruites avec et sans fissures, c'est-à-dire identifier si oui ou non l'image présente une fissure. Elle contient approximativement 900 cartes thermiques pour l'entraînement et 200 pour l'évaluation des performances. FLYD-D est une base de données où les fissures sont localisées par boîtes englobantes sur ces cartographies thermiques reconstruites. Les annotations par boîtes englobantes sont produites manuellement dans les deux formats les plus courants de la littérature : le format COCO où la boîte est indiquée par le quadruplet (x, y, w, h). x et y sont les coordonnées du premier point de la boîte, tandis que h et w désignent la hauteur et la largeur de la boîte. Les annotations ont aussi été converties au format YOLO : quadruplet (x1, y1, x2, y2) où M1(x1, y1) désigne le coin en haut à gauche et M2(x2, y2) désigne le coin en bas à droite.

Une troisième base de données est nommée FLYD-Frames. Elle est cette fois constituée d'images thermiques individuelles sélectionnées aléatoirement à partir des films thermiques. Ce jeu d'images contient approximativement 800 images thermiques annotées pour la localisation. La Figure 4.1 et la Figure 4.2 présentent des cartes thermiques issues de FLYD-D et de FLYD-Frames, respectivement, avec les annotations de localisation de la fissure sur la surface grâce à des boîtes englobantes.

4.2.2 Performances de classification

Les performances de classification sont évaluées sur la base de données FLYD-C afin de valider la conformité de cette base de données FST pour cette tâche de détection simple.

Le déploiement de ces modèles est permis par la librairie *timm/pytorch image models*, associée à *Hugging Face* [81]. Nous évaluons à la fois des architectures basées sur les convolutions et les Transformers. La référence d'apprentissage machine traditionnelle est à nouveau le descripteur HOG comme extracteur de caractéristiques et le SVM comme classifieur. Tous les modèles présentés ici sont entraînés pendant 200 époques, avec optimisation ADAM et une entropie croisée comme fonction de coût d'apprentissage. Un ensemble d'augmentations de données est appliqué de manière aléatoire sur les images d'entraînement : rotations, fenêtrages de taille variable, perspectives, miroirs horizontaux

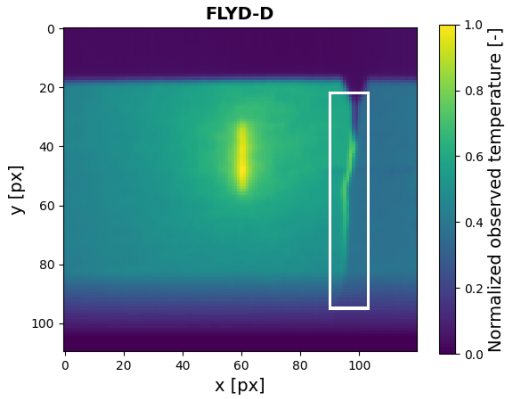


FIGURE 4.1 – Exemple de cartographie thermique reconstruite annotée, base de données FLYD-D.

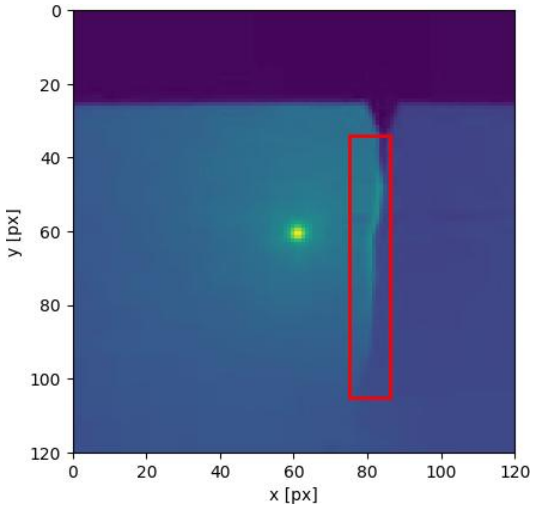


FIGURE 4.2 – Exemple d'image thermique individuelle annotée, base de données FLYD-frames.

et verticaux. Le lecteur est invité à se référer aux annexes pour plus de détails théoriques sur ces points de l'entraînement de réseaux.

Architecture	Initialisation	*Test-accuracy*	F1	*Précision*	*Rappel*
HOG+SVM [71]	-	79.1	0.79	0.78	0.79
VGG13 [66]	Init. Aléat.	83.9	0.840	0.910	0.781
	Init. Imagenet	90.2	0.902	0.985	0.832
VGG16 [66]	Init. Aléat.	75.5	0.713	0.977	0.561
	Init. Imagenet	81.1	0.795	0.963	0.677
ConvNext [82]	Init. Aléat.	-	-	-	-
	Init. Imagenet	98.6	0.990	0.987	0.994
ViT-B [68]	Init. Aléat.	86.7	0.881	0.850	0.916
	Init. Imagenet	98.6	0.987	0.975	0.999
ViT-L [68]	Init. Aléat.	84.3	0.862	0.819	0.910
	Init. Imagenet	99.0	0.990	0.981	1.00
Swin [83]	Init. Aléat.	-	-	-	-
	Init. Imagenet	98.9	0.990	0.987	0.993
CaiT [70]	Init. Aléat.	-	-	-	-
	Init. Imagenet	**98.9**	**0.990**	**0.981**	**0.999**
CoAtNet [84]	Init. Aléat.	74.1	0.736	0.742	0.735
	Init. Imagenet	98.3	0.982	0.983	0.982
Big-Transfer [85]	Init. Imagenet	98.6	0.986	0.987	0.985

TABLE 4.1 – Performances de différentes approches de classification binaires sur FLYD-C. On distingue le premier groupe des architectures basées convolution [66, 82], et le groupe des architectures basées sur le mécanisme d'attention [68, 70, 83–85].

La Table 4.1 montre les scores de classification obtenus pour différentes architectures de classification binaire. Chaque architecture est entraînée après avoir été initialisée aléa-

toirement et après pré-entraînement sur la base Imagenet [86]. Nous notons une borne inférieure pour l'*accuracy* de 74.1 % en initialisation aléatoire, et de 81.1 % en initialisation Image-net [86] pour les modèles neuronaux. Les performances obtenues sont plus faibles que celles obtenues à la section 3.3.3 pour les mêmes architectures de détection (VGG, Swin, CaiT) en initialisation aléatoire. Cette différence pourrait s'expliquer par la plus grande diversité en fissures. Néanmoins les performances sont bien plus élevées après pré-entraînement sur Image-Net et avec des modèles plus avancés de la littérature.

Les résultats montrent aussi la pertinence d'un transfert d'apprentissage depuis des bases grand-public généralistes vers la donnée thermographique, malgré la grande différence entre données visibles généralistes et celles issues de l'imagerie infrarouge Flying Spot. Si l'information FST a ses propriétés propres, un ajustement vis-à-vis du spectre visible (contours, textures) est suffisant pour adapter l'extraction des caractéristiques à ce type de données et atteindre des performances élevées. Cette ré-adaptabilité des caractéristiques semble plus prégnante pour des architectures d'apprentissage modernes du type des *Transformers*.

4.2.3 Étude d'explicabilité GRAD-CAM

Une étude de l'activité des gradients a été conduite comme au chapitre précédent afin d'investiguer les éléments des cartographies thermiques impactant la décision du modèle sur la base FLYD. La librairie *pytorch-Grad-CAM* est employée [73]. L'activité neuronale est à nouveau mesurée sur la dernière couche neuronale de deux modèles. Le premier est le modèle convolutif VGG13. L'autre est encore une fois CaiT comme modèle dérivé des *Transformers* basés sur le mécanisme d'attention. Nous étudions dans les deux cas les modèles pré-entraînés sur Imagenet, et l'activation est mesurée par rapport à la dernière couche du modèle avant la couche de classification. Nous rappelons ici que nous nous concentrons sur l'étude des régions qui activent le plus la couche étudiée dans le modèle, et non sur l'intensité de cette activation.

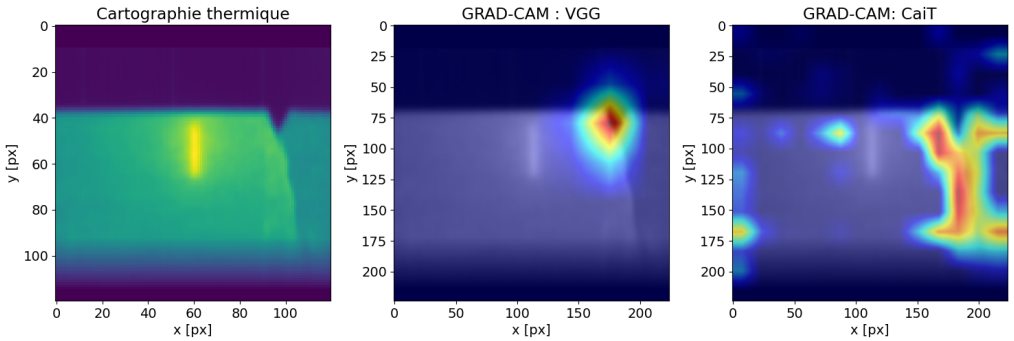


FIGURE 4.3 – Carte d'activation fournie par les deux modèles VGG pré-entraîné et CaiT pré-entraîné pour un exemple avec fissure.

Nous avons observé l'activation typique des modèles. Elle est illustrée par un exemple dans la Figures 4.3, pour un cas avec fissure. Nous constatons une activité des gradients différente de ceux obtenus dans le chapitre précédent à la section 3.3.4. Les modèles laissent de côté l'activité autour de la source d'échauffement sur cette base de données : cela est assez attendu vis-à-vis des caractéristiques de ces cartographies, où le lobe d'échauffement

produit par le laser est bien moins important et saturé que pour les échantillons complexes. Nous observons ensuite que le modèle VGG présente une activation forte au niveau de l'amorce de fissuration : le modèle semble utiliser cette structure géométrique pour sa prise de décision, en délaissant d'autres éléments de l'image comme l'intégralité de la longueur de fissuration. Au contraire, nous observons que le modèle basé sur l'attention présente sur la majorité des cartographies une activation neuronale équilibrée sur toute la longueur de fissuration, ainsi que sur des structures géométriques comme les contours et des différences de contraste thermique en bordure d'image.

Une dernière étude d'activation porte sur la comparaison entre le modèle initialisé aléatoirement puis pré-entraîné sur Image-net. La Figure 4.4 fournit un exemple de carte d'activation sur une cartographie avec fissure. Nous pouvons constater que le modèle pré-entraîné limite son activité à l'élément le plus saillant dans la détection de la fissure, ici l'amorce. Le réseau initialisé de manière aléatoire tend à venir suivre l'ensemble de la géométrie de la fissure, ainsi que la région entre la trajectoire de chauffe et le défaut, pour valider ou non sa décision de classification, comme illustré sur la carte d'activité de l'exemple.

Nous rappelons que cette étude d'activation neuronale permet d'identifier des corrélations entre le choix d'un réseau et des régions des images d'entrée mais pas de conclure sur le lien de causalité entre cette activation et le choix réalisé par le modèle.

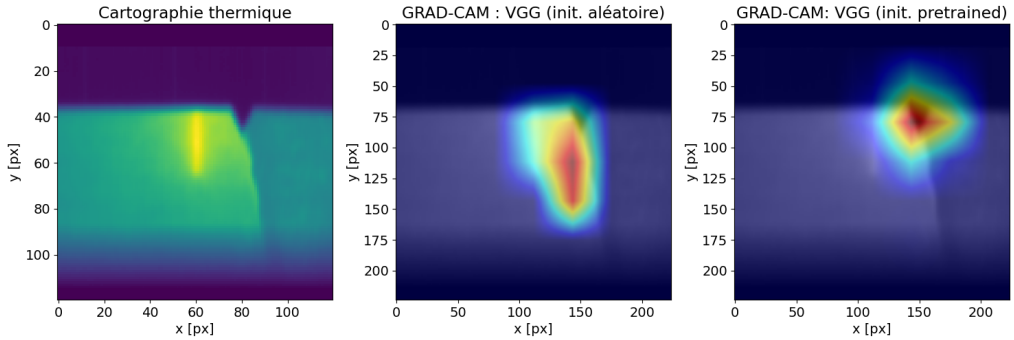


FIGURE 4.4 – Carte d'activation fournie par le modèle VGG, après entraînement, dans le cas d'une initialisation aléatoire versus d'un pré-entraînement Imagenet [86], exemple avec fissure.

Des exemples supplémentaires de cartes d'activation obtenus sur modèles pré-entraînés sont fournis dans la Figure 4.5, afin d'appuyer le discours sur les différences d'activation entre VGG et CaiT.

4.3 Localisation de fissures par transferts d'apprentissages

Nous proposons de localiser la fissure, en procédant au transfert les caractéristiques formées lors de l'apprentissage des éprouvettes dans l'apprentissage sur les échantillons métalliques complexes d'application. Comme au chapitre 3 nous retrouvons l'idée d'une progressivité dans la difficulté de la tâche, en allant d'une tâche simple d'apprentissage

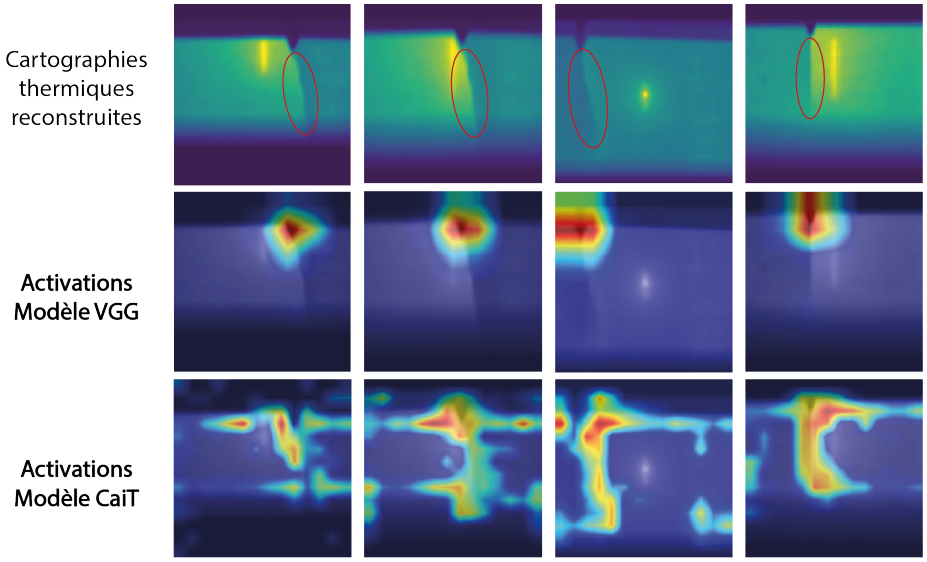


FIGURE 4.5 – Exemples supplémentaires de cartes d'activation obtenues sur des cartographies thermiques reconstruites présentant une fissure. Les activations pour VGG sont en haut, celles pour CaiT sont en bas. L'endommagement est entouré en rouge sur chaque cartographie reconstruite.

des caractéristiques neuronales clés associées à la fissure dans un contexte ou elle est facile à identifier, vers une tâche plus complexe qui est la détection de fissure avec présence d'un revêtement parcellaire, de contours de pièces plus complexes... Le travail de cette section utilise la base FLYD-D pour apprendre ces caractéristiques simples. Les bases A et B sont des bases plus complexes vers lesquels transférer les apprentissages. Elles correspondent ici respectivement à des sous-échantillons restreints à 150 images d'entraînement/réajustement de modèle plus 50 images d'évaluation de métrique. La base A correspond à une sélection/restriction aléatoire sur la base de données DAWN, qui est constitué de cartographies. La base B correspond à DAWN-G1 qui passe aux image thermiques individuelles acquises avec l'objectif microscope de 250 mm. Le lecteur peut se référer au Chapitre 1 pour plus de détails sur ces bases. La Figure 4.6 illustre l'idée générale de transfert qui va être appliquée ici, depuis un apprentissage sur FLYD vers les bases restreintes acquises sur échantillons complexes.

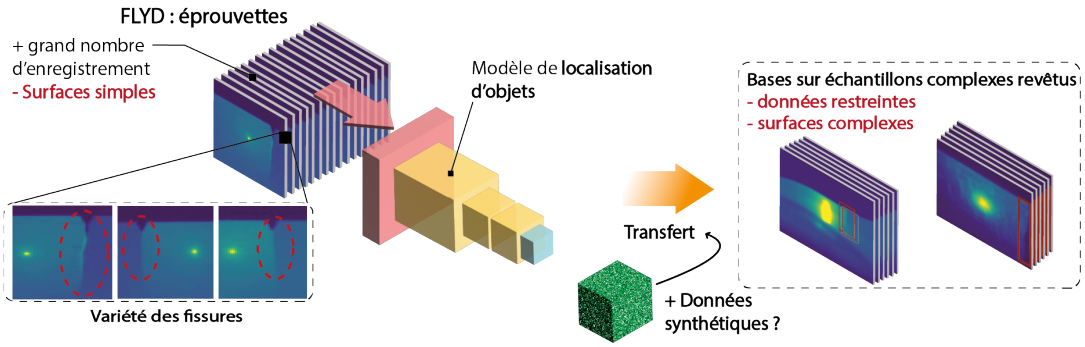


FIGURE 4.6 – Concept de transfert de modèle après un apprentissage sur éprouvettes vers des bases restreintes.

4.3.1 Architectures et métriques

Les modèles de localisation par apprentissage utilisés ici sont les suivants. Les comportements des architectures YOLO (*You Only Look Once*) -v5 et -v8 sont étudiés [87, 88]. L'étude porte également sur l'architecture YOLO-v9, sortie très récemment et utilisant un nouveau mécanisme de gradients lors de l'apprentissage [89]. Un premier *Transformer* pour la détection appelé *Detection TRansformer*(DETR) avec Resnet-50 comme extracteur de caractéristiques [90, 91] est également étudié. Nous testons aussi les modèles RT-DETR [92], qui est un modèle type DETR optimisé en terme de nombre de paramètres ce qui réduit le temps d'inférence, ainsi que Deformable-DETR [93] qui utilise une attention déformable permettant de mieux capturer des objets de faible dimension et d'accélérer la vitesse de l'apprentissage. La notion d'attention déformable est présentée de manière synthétique dans l'annexe A.8. La Table 4.2 fournit un résumé des architectures utilisées avec le nombre de paramètres associés. Les paramètres des modèles de localisation sont initialisés à partir des poids pré-entraînés disponibles en ligne, en lieu et place d'une initialisation aléatoire.

Architecture	Détails	# param
YOLO-v5	Modèle taille M	25M
YOLO-v8	Modèle taille M	25M
YOLO-v9	Modèle taille C	26M
DETR	Extracteur Resnet-50	41M
Deformable-DETR	Extracteur Resnet-50	41M
RT-DETR	Extracteur Resnet-50	33M

TABLE 4.2 – Architectures de localisation d'objets déployées dans ce chapitre.

La *Mean Average Precision* (mAP) est la mesure standard dans la littérature de traitement d'images actuelle pour la détection et localisation d'objet. La mAP est la métrique conventionnelle pour évaluer la performance des tâches de localisation sur les images et les séquences vidéos. Dans le cas présent ici où un seul objet est détecté, la mAP équivaut simplement à la (AP) car il n'y a qu'une seule classe d'objet à détecter.

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c = AP \tag{4.1}$$

Où C représente le nombre total de classes, qui est de 1 dans ce cas puisqu'il n'y a que la classe de défaut.

La Précision Moyenne (AP) est typiquement calculée comme la moyenne des valeurs de précision à un ensemble de niveaux de rappel. Une approche commune est de calculer la précision à onze niveaux de rappel discrets $[0, 0.1, 0.2, \ldots, 1]$:

$$AP = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}} p_{\text{interp}}(r) \tag{4.2}$$

où $p_{\text{interp}}(r)$ est la précision maximale à n'importe quel niveau de rappel r' tel que $r' \geq r$.

La précision à la i-ème détection est définie comme suit :

$$p(r_i) = \frac{\text{Vrai Positif}_{(r_i)}}{\text{Vrai Positif}_{(r_i)} + \text{Faux Positif}_{(r_i)}} \tag{4.3}$$

Les formules fournissent une méthode structurée pour évaluer la précision et la fiabilité de la tâche de détection de défauts, facilitant une mesure standard de la performance de classification et de localisation de l'objet-cible.

Différentes variations de la mAP peuvent donc être évaluées, variant selon le seuil d'intersection sur union (IoU) donnant des métriques de performance plus ou moins strictes. Comme illustré dans l'équation suivante, l'IoU établit un seuil de chevauchement entre les boîtes englobantes et la vérité terrain. Ce seuil détermine si la détection proposée par une architecture est acceptée ou rejetée en fonction du pourcentage de chevauchement obtenu par les propositions de boîtes du réseau. Les mAP les plus communes dans la littérature de vision par ordinateur sont la mAP50, la mAP75 et enfin la mAP qui est aussi notée mAP50-95.

$$IoU = \frac{\text{Aire de l'Intersection}}{\text{Aire de l'Union}}. \tag{4.4}$$

La $mAP50$ est la métrique avec le seuil d'acceptation d'IoU de seulement 50%, le plus large parmi les mAPs typiques de la littérature. Tandis que la mAP est calculée à plusieurs seuils d'IoU, allant de 0.5 à 0.95, par paliers de 0.05. C'est la métrique la plus stricte utilisée dans cette étude, exigeant à la fois un excellent rappel d'objet et une excellente correspondance avec la vérité terrain. Pour toutes les mAP, le plus proche de 1 est le meilleur. On considère généralement dans la littérature une valeur supérieure à 0.5 comme acceptable.

La Figure 4.7 fournit une illustration du fonctionnement de cette métrique d'évaluation sur un objet trivial, et centrée sur cette idée d'IoU comme seuil critique pour valider ou non une détection.

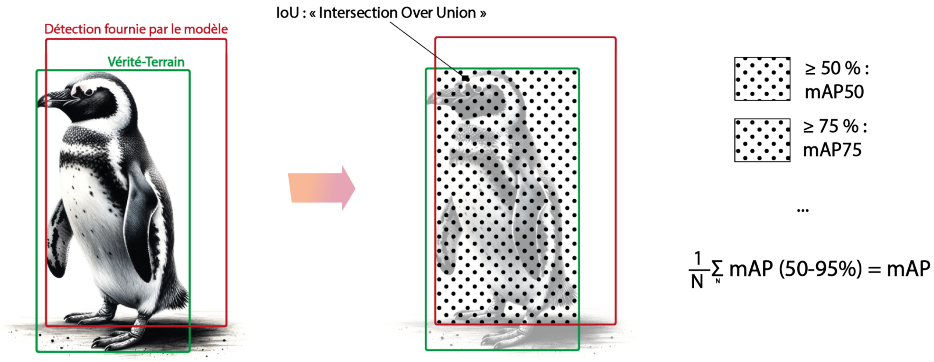


FIGURE 4.7 – Illustration simplifiée du fonctionnement de l'intersection sur union (IoU) et de son lien avec la mAP, en tant que seuil minimum d'aire de recoupement pour accepter une détection.

4.3.2 Performances de localisation sur les échantillons simples

Les architectures sélectionnées sont entraînées à l'aide des annotations dédiées de la base FLYD-D puis évaluées sur le jeu de test associé. Les scores de performance obtenus

sont présentés dans le Tableau 4.3 : toutes les architectures atteignent une mAP50 supérieur à 0,9, ce qui correspond à une localisation à seuil de 50 % du défaut très élevée. Ces scores de performance relativement élevés peuvent être expliqués par une complexité de surface limitée sur les échantillons considérant les bords des pièces ou les propriétés de surface, même s'il existe une diversité dans la longueur des fissures ou l'orientation des défauts.

YOLO-v8 et YOLO-v9 offrent les meilleures performances de localisation précise de l'objet-cible avec une mAP de 0,61. La Figure 4.8 présente un exemple de lot de détections effectuées en utilisant le modèle YOLO-v8 sur le sous-ensemble de test FLYD-D. Les architectures DETR sont globalement moins performantes, excepté le modèle RT-DETR.

Données	Architecture	mAP50 (↑)	mAP (↑)
FLYD-D	YOLO-v5	0.96	0.59
	YOLO-v8	0.96	**0.61**
	YOLO-v9	**0.99**	**0.61**
	DETR	0.92	0.48
	Deformable DETR	0.93	0.49
	RT-DETR	**0.99**	0.60

TABLE 4.3 – Performances de localisation des architectures entraînées sur la base de données FLYD.

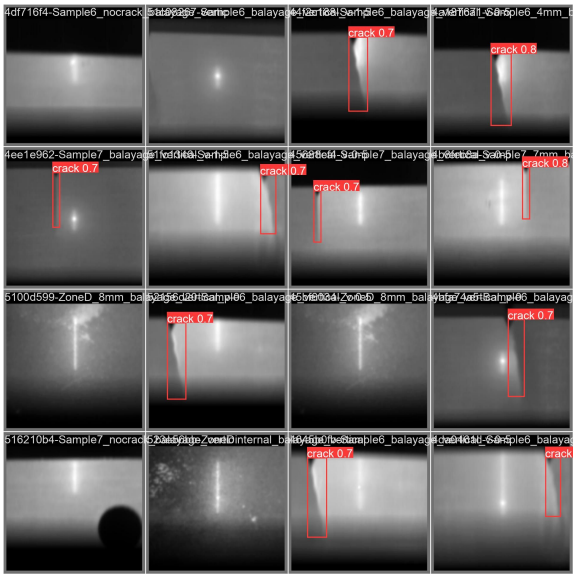


FIGURE 4.8 – Exemples de détection de fissures réalisées par le modèle YOLO-v8 sur les données de pré-apprentissage sur la base de données FLYD.

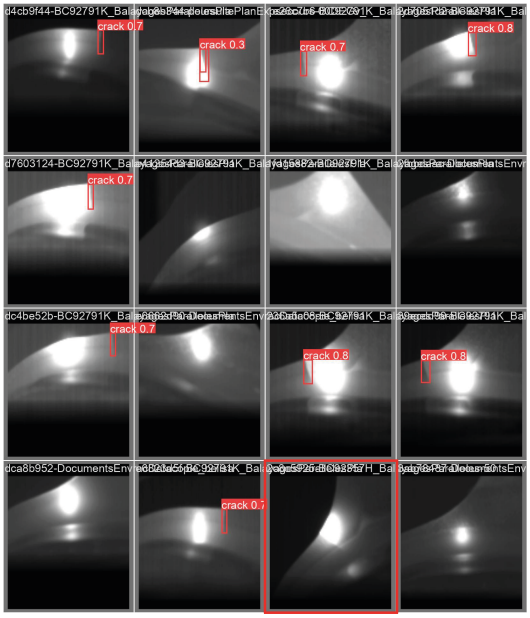


FIGURE 4.9 – Exemples de détection de fissures réalisées par le modèle YOLO-v5, après pré-apprentissage sur FLYD, sur la base de données A. Les cartographies avec détection manquée sont encadrées en rouge.

4.3.3 Performances de localisation sur les échantillons complexes

Pré-entraînement	Base	Architecture	mAP50	mAP
Entraînement direct	Base de données A	YOLO-v5	0.97	0.64
		YOLO-v8	0.97	0.63
		YOLO-v9	0.98	0.68
		DETR	0.89	0.52
		Deformable DETR	0.96	0.63
		RT-DETR	**0.99**	**0.69**
	Base de données B	YOLO-v5	0.75	0.30
		YOLO-v8	0.58	0.32
		YOLO-v9	0.91	0.36
		DETR	0.89	0.48
		Deformable DETR	0.95	**0.64**
		RT-DETR	**0.99**	0.60

TABLE 4.4 – Performances de localisation des architectures entraînées sur les bases de données A et B sans pré-entraînement FLYD avec une initialisation par apprentissage sur les bases COCO/Imagenet [86, 94].

Pré-entraînement	Base	Architecture	mAP50	mAP
Pré-entraînement FLYD-D	Base de données A	YOLO-v5	0.97	**0.71**
		YOLO-v8	**0.98**	0.69
		YOLO-v9	**0.98**	0.70
		DETR	0.92	0.57
		Deformable DETR	0.96	0.63
		RT-DETR	0.99	0.69
	Base de données B	YOLO-v5	0.79	0.30
		YOLO-v8	0.96	0.37
		YOLO-v9	0.94	0.37
		DETR	0.96	0.59
		Deformable DETR	0.96	0.64
		RT-DETR	**0.99**	**0.61**

TABLE 4.5 – Performances de localisation des architectures entraînées sur les bases de données A et B après pré-entraînement sur la base FLYD.

Les Tableaux 4.4 et 4.5 présentent les scores de mAP pour l'entraînement direct avec les jeux de données A et B ou avec un pré-entraînement sur le jeu de données FLYD-D. L'entraînement direct correspond à un modèle pré-entraîné sur COCO/Imagenet [86, 94].

Résultats sur la base de données A. Les performances sont élevées sans pré-entraînement FLYD, comme le montre le Tableau 4.4 pour l'entraînement direct. Néanmoins le Tableau 4.5 qui donne les scores après pré-apprentissage sur FLYD montre une augmentation jusqu'à + 7 % en mAP au maximum. Le pré-entraînement contribue à

augmenter le score de localisation pour toutes les métriques mesurées. Nous notons néanmoins que les architectures les plus récentes basées sur les *Transformers* comme RT-DETR et Deformable-DETR bénéficient peu du pré-apprentissage. La Figure 4.9 donne des exemples de détections réalisées en utilisant le modèle YOLO-v5 sur le sous-ensemble de test du jeu de données A. Une seule image présente une détection manquée, encadrée en rouge. RT-DETR avec un pré-entraînement sur FLYD-D est la meilleure architecture, avec une mAP50 de 0,99 et un mAP de 0,69.

Résultats sur la base de données B. Les scores de performance sur cette base sont réduits par rapport aux deux jeux de données précédents en entraînement direct, comme montré dans la Table 4.4. Une augmentation de performance avec le pré-entraînement à partir de FLYD est observée allant jusqu'à +11 % dans le Tableau 4.5. Cela confirme l'intérêt du jeu de données de pré-entraînement sur des données plus éloignées expérimentalement de FLYD. Le modèle Deformable-DETR surpasse les autres modèles : la mAP est supérieure à 0,48 aussi bien pour le modèle pré-entraîné que pour l'entraînement direct. Ici le modèle YOLO-v8 montre les bénéfices les plus importants issus du pré-entraînement FLYD : la mAP50 est augmentée jusqu'à 0,38. Le pré-entraînement est bénéfique pour la plupart des architectures, particulièrement pour la détection précise du défaut mesurée par la mAP. Les Figures 4.10 et 4.11 illustrent des exemples de détection sur le jeu de données B en utilisant l'architecture YOLO-v8 pour l'entraînement direct et pour le pré-entraînement avec FLYD-D. Les exemples illustrent qualitativement les avantages de l'augmentation de la mAP pour le modèle pré-entraîné, avec une réduction des détections manquées pour plusieurs images thermiques et une hausse de la confiance du modèle dans ses détections.

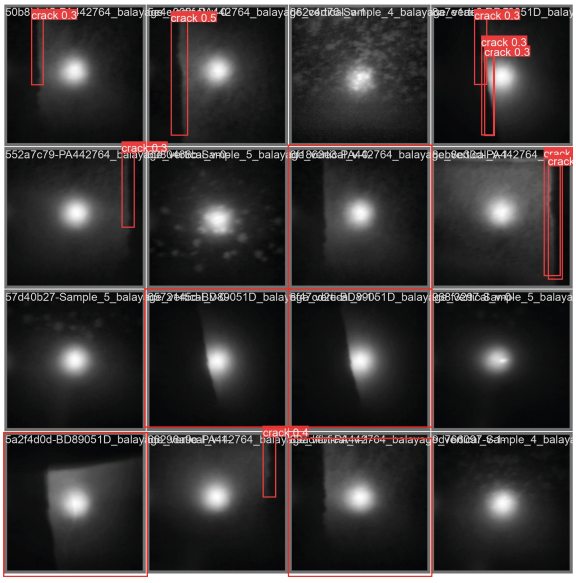


FIGURE 4.10 – Exemples de détection de fissures réalisées par le modèle YOLO-v8, en entraînement direct (base de données B, échantillon de test). Les images thermiques individuelles avec détection manquées sont encadrées en rouge.

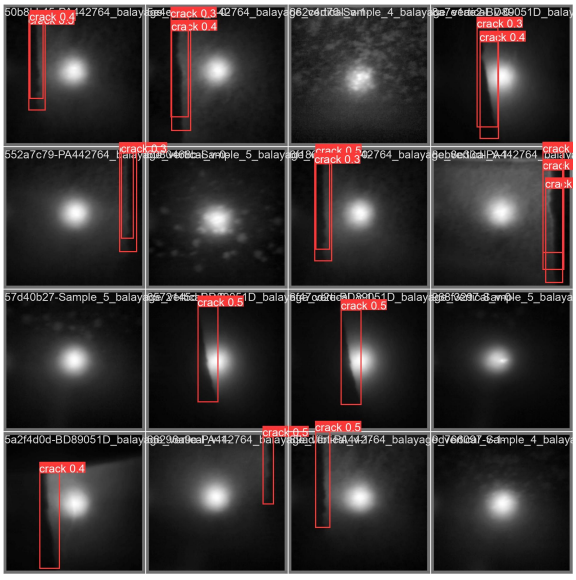


FIGURE 4.11 – Exemples de détection de fissures réalisées par le modèle YOLO-v8, après pré-apprentissage sur FLYD (base de données B, échantillon de test). Les images thermiques individuelles avec détection manquées sont encadrées en rouge.

Cette étude montre qu'il est possible de détecter le défaut directement sur image thermique individuelle. Une diminution des scores est observée entre les jeux de données A et B, ainsi qu'avec ou sans pré-entraînement sur FLYD-D. Le modèle DETR cependant permet d'atténuer la réduction de performance, en particulier avec le pré-entraînement sur FLYD-D.

Ces expériences ont mis en évidence l'intérêt du pré-entraînement en utilisant le jeu de données FLYD-D pour la détection et la localisation de fissures d'abord sur un jeu de données présentant des paramètres expérimentaux assez similaires à ceux de FLYD-D (jeu de données A). Le gain de performance par rapport à l'entraînement direct est d'autant plus marqué sur un jeu de données présentant des paramètres plus différenciées comme le jeu de données B. Cela confirme la capacité de transfert des apprentissage depuis le jeu de données de pré-entraînement FST proposé. En ce qui concerne la comparaison des performances des architectures sélectionnées, RT-DETR est le modèle avec les meilleures performances sur le jeu de données B en considérant à la fois l'entraînement direct et le pré-entraînement sur FLYD. Tandis que les architectures basées sur YOLO après pré-entraînement sur le jeu de données FLYD-D sont les plus performantes sur le jeu de données A, considérant la mAP. Ce jeu de données présente à la fois plus de variété dans les fissures et une quantité accrue de données ainsi que des défauts plus petits et plus fins à cause de la focale utilisée. Cela pourrait avantager les modèles de la famille YOLO tandis que les modèles DETR sont reconnus dans la littérature pour avoir plus de difficultés avec les petits objets qui sont plus représentés dans FLYD et dans le jeu de données A [95].

4.4 Pré-entraînement sur images générées par diffusion

Dans cette section nous comparons l'influence d'un pré-apprentissage sur des données générées par Stable Diffusion par rapport au transfert de l'apprentissage depuis les données expérimentales obtenues sur les échantillons simples, vis-à-vis performances de localisation obtenues précédemment sur la base de données A et la base de données B. Cela nous permet d'évaluer le gain de performance en transfert de données synthétiques génériques face à un transfert depuis les acquisitions sur le jeu d'imageries réelles de FLYD. En effet le jeu de données synthétiques peut permettre de maximiser la qualité et la diversité des représentations construites par le réseau de neurones pour une tâche donnée en rebrassant la diversité des propriétés des fissures déjà présente dans FLYD. La synthèse de données par diffusion est présentée dans la section 2.3.2.

4.4.1 Auto-annotation interactive pour la localisation

Nous mettons cette fois en place une approche d'auto-annotation interactive et complètement supervisée sur ce jeu de données synthétiques plutôt qu'une phase d'annotation manuelle très coûteuse en temps. Un réseau de localisation est d'abord entraîné sur FLYD : il fournit une première localisation de fissures sur les images thermiques individuelles à annoter. Un opérateur humain corrige ensuite les propositions de localisation soumises par le réseau. Il ajuste ainsi les boites englobantes et supprime les fausses détections potentielles. Cela permet de partiellement automatiser le travail d'annotation tout en gardant

un humain dans la boucle de production des données. Nous n'avions pas mis en place cette méthodologie pour les données réelles, celles-ci étant bien moins nombreuses, ni pour les données du chapitre 3 utilisées strictement pour la classification sans notion de localisation.

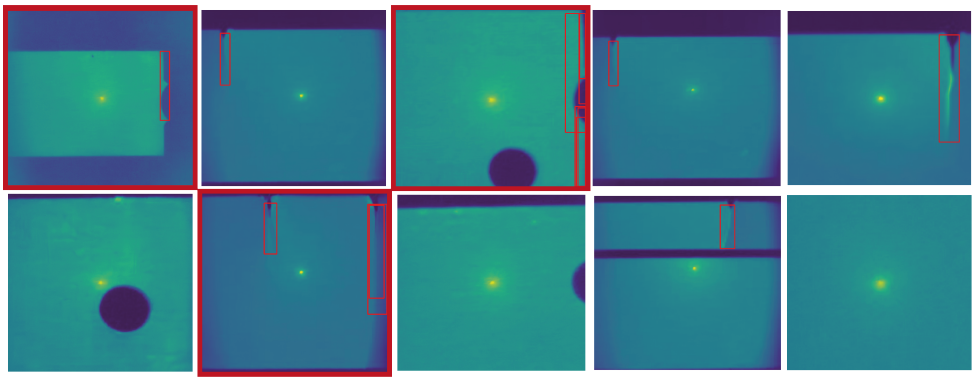


FIGURE 4.12 – Exemples de génération d'images à partir des images thermiques individuelles de FLYD après la première phase d'auto-annotation. On peut constater des cas de faux négatif qui seront corrigés a posteriori par l'opérateur humain.

Cette procédure d'annotation est appliquée sur FLYD-Frames et sur FLYD-Synths. Cela permet notamment de réaliser un pré-apprentissage de localisation supervisé sur un grand volume d'images de synthèse, qui tend à fournir de meilleures performances comparativement à une approche d'apprentissage auto/non supervisée sans annotations à nombre de données égal.

4.4.2 Performances de localisation sur échantillons simples et complexes

Performances sur FLYD-Frames. La Table 4.6 présente les performances de localisation sur des données mono-spectrales de la base FLYD-Frames, en comparant les cas avec et sans pré-entraînement sur des données synthétiques. Les performances obtenues sont supérieures à celles rapportées la section précédente. Les résultats mettent en évidence une augmentation de mAP jusqu'à 21 %, lors de l'utilisation d'un pré-entraînement avec des données synthétiques, tant sur FLYD-Frames que sur FLYD-D, pour les trois modèles testés. Si c'est un accroissement de performance plus réduit, de l'ordre de 1 % pour les deux premiers modèles, l'architecture YOLO-v9 présente des gains de performances plus notables de 21 %, le manque de données de départ affectant grandement les performances de détection de fissures. La mesure de localisation de fissure à seuil large (mAP50) ainsi que celle de localisation précise du défaut (mAP) sont toutes deux améliorées. Les figures 4.13 et 4.14 montrent des exemples de détections sans pré-entraînement sur FLYD-Synths, puis avec ce pré-entraînement. On constate sur le plan qualitatif une augmentation générale de la confiance de détection du modèle indiquée sur chaque imagette d'exemple ainsi qu'une suppression de certains cas de double détection. Un comportement similaire a été observé pour l'ensemble des détecteurs évalués.

Transfert vers les bases de données A et B. Nous nous intéressons aux transferts depuis FLYD-Synths vers les autres bases de données A et B, de volume de données

Entraînement	Modèle	mAP50 (↑)	mAP (↑)
Sans pré-entraînement sur données synthétiques	YOLO-v5	**0.94**	**0.44**
	YOLO-v8	0.93	0.42
	YOLO-v9	0.75	0.32
Avec pré-entraînement sur données synthétiques	YOLO-v5	0.95	0.46
	YOLO-v8	0.94	0.47
	YOLO-v9	**0.96**	**0.53**

TABLE 4.6 – Comparaison des performances de localisation sur la base de données FLYD-Frames (mAP50, mAP) sans et avec pré-entraînement sur des données synthétiques.

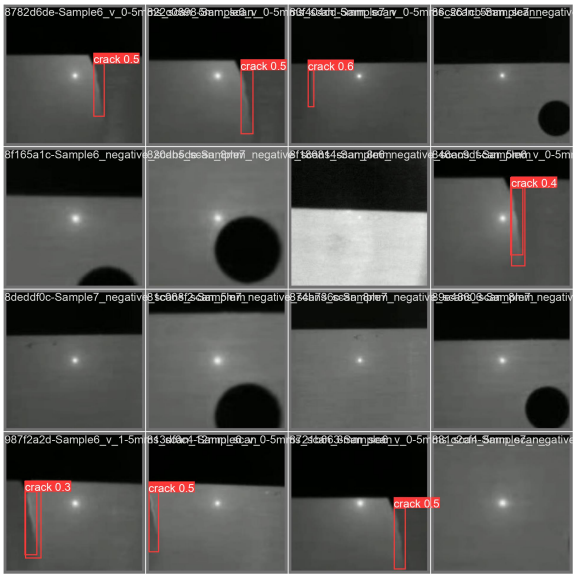


FIGURE 4.13 – Exemples de détection de fissures réalisées par le modèle YOLO-v8 sur la base de données FLYD-Frames, en entraînement direct seulement.

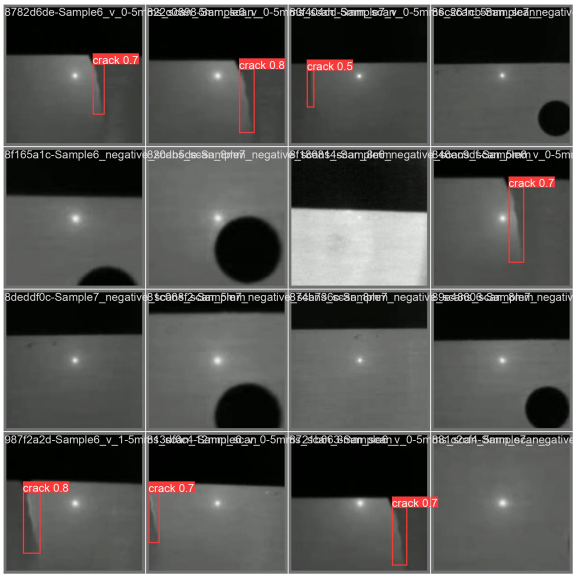


FIGURE 4.14 – Exemples de détection de fissures réalisées par le modèle YOLO-v8 sur la base de données FLYD-Frames, après pré-apprentissage sur FLYD-Synths.

restreint. Les modèles ont ainsi d'abord été entraînés sur FLYD-Synths avant d'être réajustés sur les bases-cibles. Nous évaluons aussi les performances obtenues sur la base de données FLYD-D avec le passage de la carte thermique individuelle à la cartographie. Les deux types de données permettraient donc d'alimenter des modèles tant pour de la détection image-par-image que sur cartographie reconstruite, augmentant alors énormément la quantité de données et la diversité disponible.

4.5 Impact des propriétés de la fissure sur la détection/localisation

Il peut être intéressant de dresser une évaluation de l'impact des propriétés de la fissure sur la détection par un système de détection, comme de la longueur de fissuration. Un échantillon dédié à cette étude a été indenté en plusieurs points, avec des longueurs des

Pré-entraînement	Data	Architecture	mAP50	mAP
Pre-entraînement FLYD-Synths	Base de données A	YOLO-v5	0.97	0.69
		YOLO-v8	**0.98**	**0.72**
		YOLO-v9	0.96	0.71
	Base de données B	YOLO-v5	0.98	0.41
		YOLO-v8	**0.99**	**0.46**
		YOLO-v9	0.83	0.36
	FLYD-D	YOLO-v5	0.99	0.59
		YOLO-v8	**0.99**	**0.65**
		YOLO-v9	0.99	0.62

TABLE 4.7 – Performances de localisation des architectures entraînées sur les bases de données A, B et FLYD-D, après pré-apprentissage sur les données générées par diffusion de la base FLYD-Synths.

défauts maîtrisées. Chaque fissure a ensuite été balayée au Flying Spot. Cette étude se concentre sur le modèle YOLO-v8 après entraînement sur données synthétiques, qui a fourni les meilleures performances sur l'étude développée en section 4.4 et sur les données des bases FLYD.

Nous nous concentrons dans cette partie sur l'évolution de la mAP, qui mesure à la fois la classification de l'image comme sain ou fissuré mais aussi la qualité de la localisation fournie pour la fissure par rapport à un expert humain.

4.5.1 Protocole expérimental

Nous isolons d'abord les performances de localisation du réseau sélectionné dans la base de données FLYD pour les échantillons simples 1 à 8, dont les longueurs de fissuration ont été estimées. Cette expérimentation se fait donc uniquement sur les cartographies thermiques reconstruites de FLYD.

En plus de l'extraction des performances sur FLYD, une éprouvette aux dimensions similaires à FLYD est utilisée, afin d'évaluer les performances de généralisation sans ré-entraînement en fonction de la longueur du défaut. L'échantillon est indenté à l'usinage electro-pulse pour des longueurs d'indentation allant de 3 à 16 mm. L'éprouvette en aluminium est dépolie et sablée afin de maximiser la rugosité de surface. Cela permet de procéder à un balayage propre limitant les phénomènes de réflexion spéculaire, ou bien une diffusion trop rapide de l'échauffement sur la surface du matériau due aux propriétés de conduction de l'aluminium. La Figure 4.15 fournit une illustration de l'échantillon après sablage. Nous pouvons remarquer que la plupart des indentations sont assez visibles à l'oeil nue : c'est l'impact de la longueur de fissuration sur le comportement des modèles d'apprentissage qui nous intéresse ici.

Nous produisons ainsi entre 10 et 15 films thermiques par fissure, partant d'une distance de 1 cm et se rapprochant progressivement de la fissure ciblée. La Figure 4.16 illustre une image thermique typique acquise durant un de ces balayages avec des pré-détections issues du modèle YOLO-v8 entraîné sur FLYD-Synths. Les balayages sont réalisés parallèlement à la fissure à une distance de 0 à environ 1 cm de par et d'autre de celle-ci. Cela permet de moyenner les performances de détection obtenues par rapport à la dis-

tance à l'endommagement inspecté. La vitesse de balayage et la taille de faisceau suivent les paramètres d'inspection suivis pour la base de données FLYD. Nous mesurons les performances de localisation sur les cartographies thermiques reconstruites.

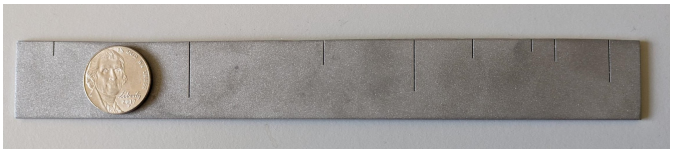


FIGURE 4.15 – Prise de vue de l'échantillon métallique indenté. La plupart des indentations sont visibles à l'oeil nue ainsi que la rugosité ajoutée par le sablage.

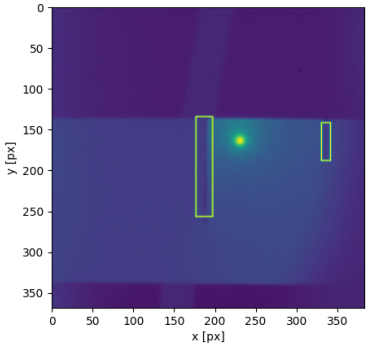


FIGURE 4.16 – Image thermique individuelle acquise durant un balayage avec pré-détection du défaut, modèle YOLO-v8 entraîné sur FLYD-Synths.

4.5.2 Evolution de la performance de détection en fonction de la longueur de fissure

Résultats sur FLYD. Les résultats de détection avec le modèle en fonction de la longueur de fissuration sont fournis dans la Figure 4.17. Un exemple de localisation est fourni sur l'échantillon 7 dans la Figure 4.18. La fissure de cet échantillon est relativement longue mais présente un faciès de fissuration peu ouvert.

Nous constatons que la mAP50 décrit un comportement relativement uniforme en fonction de la longueur de fissure. Le défaut le plus petit semble être moins bien détecté, mais il est de plus constaté une diminution de performance pour les défauts au-delà de 13 mm. On n'observe donc pas, au sein d'un certain régime de bon fonctionnement, d'influence de la taille de fissure sur la qualité de la détection à seuil de 50 %. Le comportement de la mAP est lui plus erratique et difficile à interpréter. Le pic de mAP à 0.59 pour l'échantillon 7 de longueur de 13 mm pourrait s'expliquer par sa présence dans le jeu de test de FLYD. C'est le processus de sélection du modèle, réalisé sur la mAP évaluée sur les échantillons de test, qui peut aboutir à une maximisation de la performance sur cet échantillon, expliquant alors ce pic. Cette hétérogénéité dans la distribution des performances de localisation avec le critère le plus strict de la mAP est difficile à expliquer, mais pourrait s'expliquer par la diversité de défauts. Une hypothèse serait que le modèle pourrait valoriser durant son apprentissage une localisation moyenne de la fissure correspondant à une bonne mAP50, délaissant en partie la précision de la localisation ce qui est associée à une mAP plus faible.

Échantillon métallique indenté. Les résultats de détection avec le modèle sélectionné en fonction de la longueur de fissuration sont fournis dans le graphe Figure 4.19.

Il est à noter que les défauts de type indentation sont totalement absents de l'ensemble des bases de données d'entraînement utilisées pour faire apprendre au modèle.

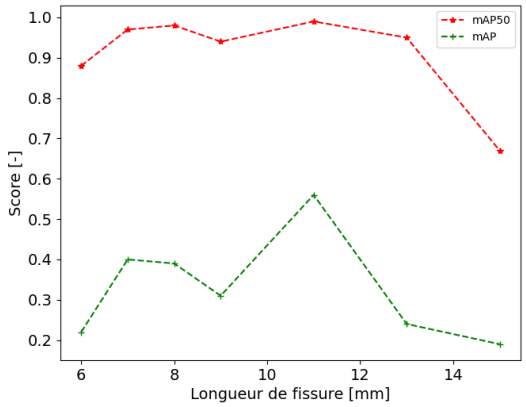


FIGURE 4.17 – Performance de détection en fonction de la longueur de fissuration sur la base de données FLYD, estimée pour les échantillons 1 à 8.

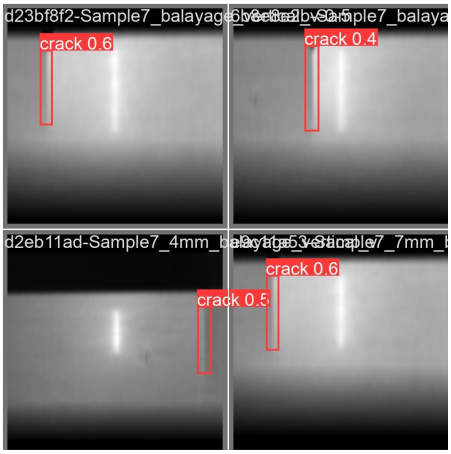


FIGURE 4.18 – Exemples de détection sur l'échantillon simple 7 de la base FLYD.

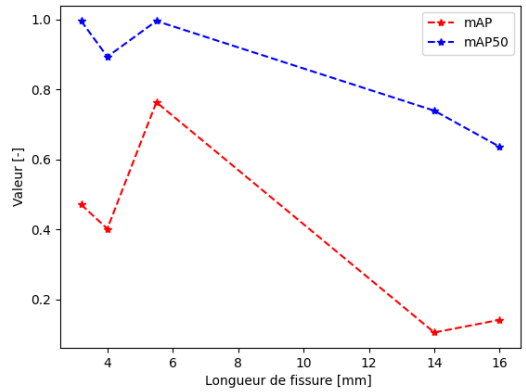


FIGURE 4.19 – Performance de détection en fonction de la longueur de fissuration. Les scores sont obtenus sans ré-apprentissage sur ce type de fissure "indentation".

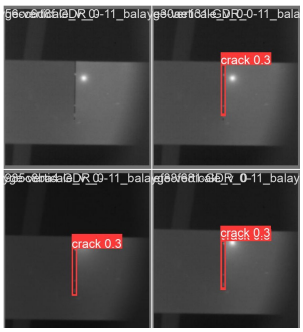


FIGURE 4.20 – Exemple de détection pour le défaut de 14 mm où un abaissement de la mAP est détecté : on constate que l'endommagement est tout de même bien détecté sur les trois quart des images. .

Cela pourrait expliquer la baisse de performance : les défauts type indentations sont trop nets comparativement à la base de données de FLYD par exemple où l'ouverture des fissures est plus réaliste et donc moindre. Ils pourraient alors être confondus avec des discontinuités de chaleur comparables aux contours de pièce. Un exemple de cas d'exemples de détection et de non-détection est fourni dans la Figure 4.20. De plus, la base d'entraînement de FLYD contient essentiellement des défauts dont la dimension est inférieure à 10 mm avec une longueur moyenne de fissure estimée à 8,5 mm, ce qui pourrait avoir un impact fort sur les capacité de généralisation pour des fissuration plus longues. Ce phénomène de performance moindre sur des défauts longs pourrait être supprimé par l'ajout de fissures de longueur supérieures ce qui peut réhausser les performances de localisation de ce type d'endommagement. Si les modèles entraînés sur FLYD gardent une détection

à seuil de 50 % correcte, leur capacité de généralisation est limitée concernant la qualité de la localisation pour ce type de pseudo-défaut.

4.5.3 Conclusion sur l'impact des propriétés de la fissure

Cette section nous a montré que si les modèles fournissent des bonnes performances de détection et localisation globales en particulier sur FLYD pour la mAP50, la précision la plus stricte de la localisation en fonction de la longueur de fissure est plus hétérogène, que ce soit sur la base de données d'entraînement FLYD ou bien pour la généralisation à l'échantillon indenté.

A l'avenir une prochaine version de la base de données FLYD devrait tenir compte de ce phénomène, en introduisant des fissurations supplémentaires et de forme plus diversifiée, et notamment de ce type de pseudo-fissure indentée durant l'entraînement, ce qui n'a pas été fait dans le cadre de ce travail de thèse. Une autre approche pourrait être envisagée en prospective du côté de l'apprentissage actif et interactif : nous pourrions ré-entraîner nos modèles en focalisant l'apprentissage sur les défauts qui fournissent les performances les plus faibles, au travers d'un mécanisme de pondération par exemple.

4.6 Conclusion

Ce chapitre a donc présenté les premiers résultats à notre connaissance de localisation de fissure par FST couplé à l'apprentissage profond. Pour cela nous avons utilisé le pré-entraînement sur des surfaces simples pour la thermographie laser infrarouge, soit sur des données expérimentales soit sur des données synthétiques. Nous avons notamment vu l'utilisation de la base de pré-entraînement FLYD afin de former des caractéristiques sur une grande diversité de fissures dans un contexte de détection facile qui sont transférées vers des problèmes plus complexes de FST où les données sont en quantité restreinte. Nous montrons aussi ici que la base de données permet des gains dans la qualité de la localisation de fissures sur les données Flying Spot sur les surfaces revêtues plus complexes et au volume de données volontairement restreint par transfert des apprentissages. L'introduction de données synthétiques produites par diffusion à partir de ces surfaces génériques pour la localisation a aussi été étudiée. Elle montre des gains de performances équivalents voir supérieurs à l'utilisation de la base de pré-entraînement réelle. La base de données FLYD a été mise à disposition de la communauté et pourrait être étendue à long terme, avec l'introduction potentielle au long cours d'échantillons supplémentaires, dans le cadre de prochaines thèses ou des prochaines campagnes d'inspection FST. Couplé à de la synthèse d'images nous pourrions encore augmenter les capacités de transfert et de généralisation des modèles entraînés sur FLYD, dans le cadre d'un déploiement industriel par exemple. Le mécanisme d'auto-annotation fait partie de cette perspective d'évolution et de maintenance à long terme en permettant de corriger et d'agrandir continuellement le volume de données face à de nouveaux échantillons. Il faudrait de plus que les prochaines collectes veillent à diversifier autant que faire ce peut les longueurs et les faciès de fissuration présents.

Le travail présenté ici peut néanmoins être amélioré. Ainsi l'introduction dans la fonction de perte de la loi de diffusion de chaleur ou d'une loi suffisamment fidèle telle que l'advection-diffusion permettrait d'augmenter la robustesse à de nouvelles scènes et donc

les performances de localisation avec une quantité donnée équivalente ou moindre. C'est l'approche des réseaux de neurones informés par la physique (*Physics Informed Neural Networks* en anglais, PINN). Ils pourraient par exemple être intégrée dans le processus de débruitage d'un modèle de diffusion via la composition d'une fonction de coût couplant physique et statistique, ce qui augmenterait le réalisme tant statistique que physique des images synthétiques.

Sur l'axe de la préparation des données, le système d'auto-annotation actuel totalement supervisé pourrait être encore amélioré et automatisé. Il devrait alors être remplacé à plus ou moins court terme par une approche de pré-annotation par apprentissage actif [96,97]. Néanmoins l'aspect interactif de ces annotations devrait être conservé. Le processus de correction humaine réduit le risque d'apparition de biais comme la propagation dans les auto-annotations d'un faux positif systématique sur un artefact de reconstruction de la cartographie thermique. Un tel système d'annotation évolutif pourrait être couplé tout aussi bien avec un système d'apprentissage ensembliste [98,99] assemblant plusieurs sous-modèles de détection spécialisés : dans un type de pièce, un type de défaut, voire un spectre d'inspection et sa physique [100]. Nous pourrions alors imaginer développer et combiner plusieurs réseaux spécialisés suivant le type de pièce au travers d'une *Mixture d'Experts* (MoE), qui consiste en un mécanisme de *switch* appris pour sélectionner le modèle le plus adapté suivant la donnée rencontrée [101–104]. Cela pourrait faciliter le ré-entraînement d'un des modèles experts voire l'adjonction d'un expert supplémentaire tout en améliorant grandement la capacité de généralisation à de nouvelles pièces à examiner une fois le système déployé sur banc d'essai [105,106].

Les premiers chapitres étaient focalisés sur l'exploitation de données FST seules pour la détection de l'endommagement. Si des performances élevées sont atteignables comme nous avons pu le voir, des non-détections peuvent toujours se produire. Celles-ci semblent souvent associées à une non-compréhension par le modèle des sources de fausses alarmes comme les contours des pièces et éléments de géométrie de surface. L'introduction d'une modalité supplémentaire où des caractéristiques de changement de texture sont plus faciles à identifier permettraient de mieux comprendre les sources de biais en présence, comme c'est la cas du spectre visible. Nous proposons donc d'étudier dans la partie suivante le couplage avec l'imagerie visible, pour voir si il est possible de bénéficier efficacement de la complémentarité entre les deux moyens d'imagerie pour détecter les endommagements par apprentissage. Le prochain chapitre sera ainsi dédié à une nouvelle architecture de fusion plus robuste aux problématiques de désalignements entre les modalités qui sont plus dures à recaler proprement dans le contexte du CND, où les champs examinés sont parfois très différents.

Troisième partie

Fusion multi-spectrale

Chapitre 5

Méthode de fusion infrarouge-visible basée sur l'attention croisée

Le crâne est notre caverne, et les représentations mentales sont des ombres. L'information dans une représentation interne, c'est tout ce que nous pouvons savoir du monde extérieur.

Comment fonctionne l'esprit,
S. Pinker.

Sommaire

La détection d'objets au travers de plusieurs modalités n'est pas un problème simple : la manière de combiner les informations de chaque spectre joue un rôle important dans la construction d'un modèle efficace vis-à-vis des performances de localisation d'objets, mais aussi du temps de calcul et de la mémoire exploitée. C'est un problème encore ouvert pour lequel des gains en capacité de détection et en efficience mémoire sont toujours possibles. Ce champ de recherche s'est élargi, avec le développement récent du mécanisme d'attention et des architectures basées sur les *Transformers*. Un autre enjeu est la robustesse de l'apprentissage à des recalages imparfaits voire à l'absence complète de recalage inter-modalités, qui est aussi peu étudiée dans la littérature alors qu'elle apparaît critique pour nombre d'applications, comme notamment la détection de défauts sur des pièces 3D dans un contexte industriel. Ainsi, l'étalonnage des deux moyens d'imagerie peut être difficile à obtenir, car il nécessite une mire dont les structures sont visibles dans les deux modalités. De plus, cet étalonnage peut être affecté par des différences de champ observé dans chaque modalité. Un décalage au cours du cycle de vie du système d'acquisition multi-modal peut aussi survenir.

Ce chapitre traite des développements architecturaux conduits durant la thèse, en vue de la conception d'une nouvelle architecture de fusion. Après un point bibliographique consacré à la fusion de données infrarouge-visible dans la section 5.1 et les motivations pour un développement architectural dans la section 5.2, nous présentons notre modèle de fusion CAFF-DINO dans la section 5.3. Cette proposition est basée sur une fusion des cartes de caractéristiques infrarouges et visibles par une opération d'attention croisée qui dérive de l'auto-attention. Ensuite, après avoir présenté les bases de données infrarouge-visible publiques de référence dans la communauté de vision par ordinateur dans la section 5.4, la capacité de localisation d'objets multi-spectrales de notre modèle est comparée aux approches de la littérature actuelles dans la section 5.5. La robustesse de notre architecture aux désalignements est aussi étudiée, avec une étude d'ablations permettant de valider le choix des éléments de l'architecture tout en étudiant le degré de modularité de notre modèle.

Les concepts clés d'apprentissage profond utilisés dans ce chapitre sont définis plus en détails en annexes, en particulier l'annexe A.6 traitant des *Transformers* et du mécanisme d'attention.

Le travail développé dans ce chapitre a fait l'objet d'un acte de conférence publié dans le *Workshop on Perception Beyond the Visible Spectrum (PBVS)* associé à la conférence *Computer Vision and Pattern Recognition* (CVPR-2024). Le modèle développé ici est de plus disponible en accès libre suivant ce répertoire : https://github.com/kevinhelvig/CAFF-DETR.

5.1 Bibliographie sur la fusion multi-spectrale par apprentissage

Nous dressons d'abord un état de l'art sur la fusion par apprentissage des données infrarouge-visible pour la détection d'objets. Nous décrivons dans une deuxième partie les principales opérations sur la fusion de caractéristiques. Enfin, nous décrivons l'état de l'art des *Transformers* pour la détection d'objets mono-spectrale.

5.1.1 Fusion infrarouge-visible pour la détection d'objets

Des premiers travaux sont apparus dès le milieu des années 2010, développant des architectures d'apprentissage profond pour la fusion multi-spectrale, en particulier en infrarouge-visible pour la conduite autonome [107].

Une certaine diversité d'architectures de fusion a été développées depuis. Nous pouvons classer les différentes stratégies de fusion de l'information de chaque spectre à partir de la localisation de cette fusion, autrement dit l'endroit dans le modèle où les deux caractéristiques de chaque spectre sont assemblées et combinées. Le découpage entre les différentes méthodes est illustré par des schémas de principe dans la Figure 5.1 afin d'identifier cette localisation de la fusion d'informations.

L'approche par fusion précoce peut être vue comme la plus intuitive et la plus simple à mettre en œuvre. Elle est aussi appelée fusion à l'entrée et consiste en une concaténation directe des images de chaque spectre. La couche neuronale qui permet l'injection d'informations visuelles dans le modèle est simplement adaptée pour passer des 3 canaux

RGB habituels à du 4 canaux ou à du 6 canaux, suivant la dimensionalité de la modalité ajoutée. Une telle architecture est généralement facile à mettre en œuvre par simple concaténation et convolution, mais ce type de modèle montre des performances non optimales de manière générale par rapport aux architectures de fusion plus tardives. Suivant la complexité de l'architecture, il est possible que le modèle n'arrive pas à extraire les informations pertinentes propres à chaque modalité efficacement dans le cadre de cette fusion [108].

La deuxième approche proposée est la fusion tardive dans laquelle les informations propres à chaque spectre sont combinées très en aval dans la structure du réseau. La fusion par score en est un représentant extrême : un système de poids permet de combiner les localisations fournies par deux modèles dans chaque spectre, sans aucune combinaison d'information [109]. La fusion des caractéristiques de chaque spectre n'est donc pas réalisée, ou bien seulement de manière très tardive suivant l'architecture. Elle peut être vue comme analogue aux méthodes d'ensemble dans sa topologie, avec un modèle spécifique à chaque modalité et un mécanisme de poids permettant de sélectionner la meilleure sous-architecture suivant l'information d'entrée. Cette approche semble fournir les meilleures performances de détection multi-spectrales. Néanmoins, elle amène à un dédoublement des structures de traitement et, suivant l'échelle de l'architecture proposée, elle est vite coûteuse en temps de calcul et en mémoire [108, 110, 111].

Le dernier type de fusion de l'état de l'art peut être appelé fusion des caractéristiques ou fusion en milieu de réseau [108]. Cette approche est intermédiaire aux deux précédentes et propose de combiner des caractéristiques mono-modalité après une extraction de caractéristiques séparée, à un certain niveau d'abstraction mais toujours en amont du module de décision. Cela permet d'extraire des informations spécifiques dans chaque modalité, avant leur combinaison et leur mise en relation plus en aval du modèle.

Les fusions de données visuelles en milieu de réseau ou tardives dans les modèles neuronaux trouvent de plus un écho en neurologie. Ainsi, le traitement de la perception visuelle de chaque œil est distribué à travers les nerfs optiques et des sous-structures associées comme le *Chiasma optique* et le *Tractus optique* avant que l'information ne soit finalement fusionnée et traitée dans le cortex visuel, à l'arrière du crâne [112, 113].

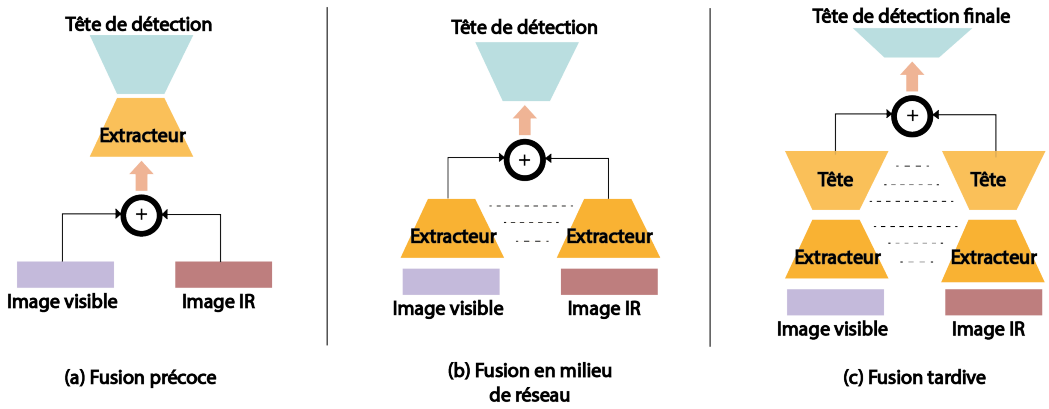


FIGURE 5.1 – Illustration des grandes familles de fusion typiques. Les traits en pointillés indiquent les possibilités de connexions résiduelles, cela pour différentes profondeurs d'abstraction des réseaux.

Cette dernière famille de fusion présenterait l'avantage de réduire le coût calculatoire global, à condition toutefois que l'opération de fusion soit suffisamment expressive : c'est-à-dire qu'elle soit efficace dans sa capacité à extraire et à mettre en relation les informations de chaque modalité pour que la tête du réseau puisse identifier les objets à détecter. Nous pouvons mentionner l'approche *Cyclic Fuse and Refine* [114] qui propose le principe général de ces fusions à l'extracteur de caractéristiques, comme illustré à la Figure 5.2 : une couche de convolution extrait des caractéristiques fusionnées par concaténation de chaque modalité, à différentes profondeurs d'abstraction. Cette approche va former un résidu fusionné à partir de l'information des deux spectres, qui est ensuite traité par chaque couche d'un extracteur de caractéristiques qui est généralement identique à un réseau de neurones mono-spectral. Des approches plus poussées ont été par la suite développées pour enrichir ces caractéristiques d'association infrarouge-visible fusionnées. Nous trouvons notamment le brassage (*shuffling* en anglais) [115] des caractéristiques mono-modales entre elles, à la fois dans la profondeur de ces caractéristiques ainsi que sur le plan spatial, employant une opération d'attention pour extraire des inter-relations entre chaque spectre. Ce mécanisme de brassage peut permettre de forcer le réseau à combiner l'information de chaque spectre plutôt que de ne baser sa décision que sur l'extraction de caractéristiques issues d'une seule modalité [115]. D'autres équipes proposent l'utilisation d'opérations d'auto-attention ou assimilées pour combiner les caractéristiques de chaque spectre à différents étages des extracteurs de caractéristiques [116, 117]. Cependant, ces approches sont généralement conçues et testées pour un unique détecteur et requièrent de redévelopper la tête de détection déployée ou l'extracteur de caractéristiques.

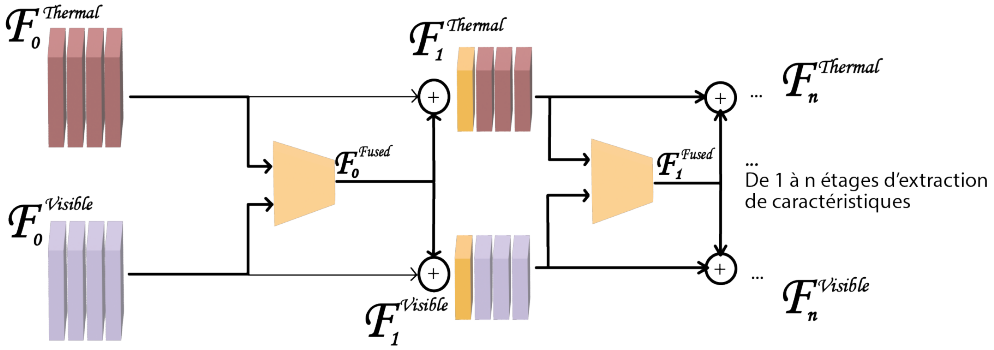


FIGURE 5.2 – Illustration de l'approche *Cyclic Fuse and Refine* [114]. Pour chaque profondeur d'abstraction allant de 1 à n, un module de fusion extrait des caractéristiques fusionnées F_i^{Fused} qui sont ensuite réinjectées dans les voies de traitement mono-spectrales.

Notons que ces différents travaux se focalisent principalement sur des données alignées.

5.1.2 Opérations pour la fusion de caractéristiques

Plusieurs opérations de base permettent de combiner des vecteurs de caractéristiques spécifiques à chaque spectre ou modalité.

La concaténation des caractéristiques de chaque modalité est la méthode la plus simple pour combiner les informations. On peut ensuite appliquer une couche de convolution, permettant de former des caractéristiques fusionnées multi-modales. Cependant, bien que

cette opération soit relativement peu coûteuse en termes de calcul, elle est limitée dans sa capacité à extraire des informations associatives entre les modalités en raison de la taille du noyau, de la densité des filtres et de leurs interrelations. Malgré cela, cette opération reste extrêmement utile, permettant d'assembler facilement l'information de chaque modalité avant de procéder à d'autres opérations comme de l'interpolation pour la fusion et la compression spatiale des cartes de caractéristiques, ou bien l'utilisation de couches de convolution pour procéder à leur fusion.

L'opérateur "+" est utilisé dans la suite de ce manuscrit pour formaliser cette opération, par simplicité d'usage. Pour illustrer, voici comment l'opération de concaténation peut être formulée pour des vecteurs colonnes :

$$\text{Concat}(F^{(1)}, F^{(2)}) = F^{(1)} + F^{(2)} = \begin{bmatrix} F^{(1)} \\ F^{(2)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ . \\ . \\ x_n^{(2)} \end{bmatrix} \tag{5.1}$$

Où $F^{(1)}$ et $F^{(2)}$ représentent respectivement les caractéristiques de la modalité 1 et de la modalité 2. Cette opération empile simplement les vecteurs colonnes $F^{(1)}$ et $F^{(2)}$ pour former une matrice où les caractéristiques des deux modalités sont concaténées verticalement.

Une autre approche pour combiner des informations multi-modales consiste à réaliser le produit des caractéristiques. Elle joue un rôle crucial dans le couplage vision-langage, démontrant une certaine efficacité même avec des volumes de données limités, grâce à l'intégration d'attention [118]. Un exemple de fusion multi-modale par produit des caractéristiques de chaque modalité a notamment été développé pour combiner la vision et le langage : c'est la méthode CLIP (*Contrastive Language-Image Pre-training*) [54], où une fonction de similarité consistant en un produit scalaire entre l'information associée à une donnée visuelle I et celle associée à un texte T est calculée par :

$$\text{CLIP}(F^{(Img)}, F^{(text)}) = \sigma\left(\frac{1}{\sqrt{d}} \cdot \text{Sim}\left(\frac{W_I F^{(Img)}}{\|W_I F^{(Img)}\|_2}, \frac{W_T F^{(text)}}{\|W_T F^{(text)}\|_2}\right)\right) \tag{5.2}$$

où $F^{(Img)}$ et $F^{(text)}$ représentent respectivement les caractéristiques extraites depuis l'image et depuis le texte, W_I et W_T sont les poids appris pour l'image et le texte, $\text{Sim}(x, y)$ est une mesure de similarité (par exemple par produit scalaire), σ est une fonction d'activation (par exemple, une fonction softmax), et d est la dimension des vecteurs d'entrée. Cette approche illustre comment les caractéristiques peuvent être utilisées pour évaluer la compatibilité entre images et textes dans des tâches de vision par ordinateur avancées. Cette opération de multiplication est également employée pour guider les modèles de génération d'images par diffusion, comme le conditionnement de *Stable Diffusion* et DALL-E [41,46]. Cependant, cette opération reste coûteuse en termes de calcul, générant des cartes de caractéristiques couplées de grande taille, ce qui implique d'avoir à disposition des moyens de calcul importants pour être adaptée à de nouvelles modalités de fusion.

D'autres opérations d'inter-corrélation plus explicites peuvent être utilisées, telles que l'introduction d'une mesure de distance vectorielle ou de corrélation entre les caractéristiques des différentes modalités, ce qui peut ensuite informer le modèle d'apprentissage.

On peut notamment utiliser la Similarité Cosinus, formalisée dans l'équation (5.3). Cette approche implique le calcul d'un produit scalaire normalisé entre les caractéristiques de chaque modalité, permettant de mesurer l'écart entre leurs espaces dimensionnels abstraits respectifs. Cette opération est couramment employée pour évaluer la similarité entre des vecteurs de grande dimension, comme dans l'approche d'intelligence artificielle alternative du calcul hyper-dimensionnel [119]. En revanche, des mesures de distance ou de magnitude vectorielle plus communes, comme les régressions ridge et lasso, ne fournissent pas une information de comparaison entre les modalités qui soit aussi expressive [120, 121].

$$\text{Cos-sim}(F^{(1)}, F^{(2)}) = \frac{F^{(1)} \cdot F^{(2)}}{\|F^{(1)}\| \times \|F^{(2)}\|} \tag{5.3}$$

Où $F^{(1)}$ et $F^{(2)}$ représentent respectivement les caractéristiques de la modalité 1 et de la modalité 2. Cette équation montre comment le produit scalaire normalisé mesure la similarité cosinus entre les vecteurs de caractéristiques, fournissant une indication du degré de désalignement ou de similarité entre les modalités.

Nous pouvons enfin mentionner le mécanisme d'attention et l'un de ses dérivés, l'attention croisée. Ces deux opérations sont capables d'extraire des inter-relations multi-échelles entre différentes régions de l'entrée. L'attention classique, ou auto-attention, met en relation l'ensemble des régions au sein d'une même modalité, ce qui peut entraîner l'extraction d'informations redondantes dans le cadre d'une information multi-modale. En revanche, l'attention croisée va calculer l'attention de la modalité observée par rapport à la modalité cible, ce qui met explicitement en rapport la région observée dans une modalité que l'on peut qualifier de référence par rapport à celle observée dans la modalité cible. Elle peut être considérée comme une opération de corrélation des caractéristiques de chaque modalité.

Nous nous focalisons ici sur le formalisme de l'auto-attention et de l'attention croisée, afin d'identifier leurs éléments de différence. L'attention et les *Transformers* sont décrits et illustrés dans l'annexe A.6. L'attention classique peut être définie par :

$$\text{Self-Attn}(Q^{(1)}, K^{(1)}, V^{(1)}) = \text{softmax}\left(\frac{Q^{(1)}(K^{(1)})^T}{\sqrt{d}}\right) V^{(1)} \tag{5.4}$$

Où $Q^{(1)}$, $K^{(1)}$ et $V^{(1)}$ représentent respectivement les requêtes, les clés et les valeurs extraites pour la modalité 1 seulement. Cette opération montre comment l'attention classique calcule un poids d'attention pour chaque élément de la modalité en fonction de sa similarité avec tous les autres éléments, puis combine les valeurs correspondantes en fonction de ces poids. L'auto-attention est alors assimilable à une opération d'inter-corrélation entre les différents éléments de la modalité 1.

L'attention croisée peut être définie par :

$$\text{Cross-Attn}(Q^{(1)}, K^{(2)}, V^{(2)}) = \text{softmax}\left(\frac{Q^{(1)}(K^{(2)})^T}{\sqrt{d_k}}\right) V^{(2)} \tag{5.5}$$

Où $Q^{(1)}$ représente les requêtes de la modalité 1, $K^{(2)}$ les clés de la modalité 2 et $V^{(2)}$ les valeurs de la modalité 2. Cette opération montre comment l'attention croisée calcule un poids d'attention pour chaque élément de la modalité 1 ($Q^{(1)}$) en fonction de sa similarité avec chaque élément de la modalité 2 ($K^{(2)}$), puis combine les valeurs correspondantes de

la modalité 2 ($V^{(2)}$) en fonction de ces poids. Dans un contexte de fusion multi-modale, l'attention croisée peut être interprétée comme une opération d'indexation ou d'inter-corrélation des caractéristiques de la modalité (2) par rapport à la modalité (1) qui sert de référence.

Les opérations de fusion classiques sont illustrées à la Figure 5.3. L'attention croisée peut être assimilée à une mise en relation des caractéristiques de chaque spectre, comme illustré dans la Figure 5.4.

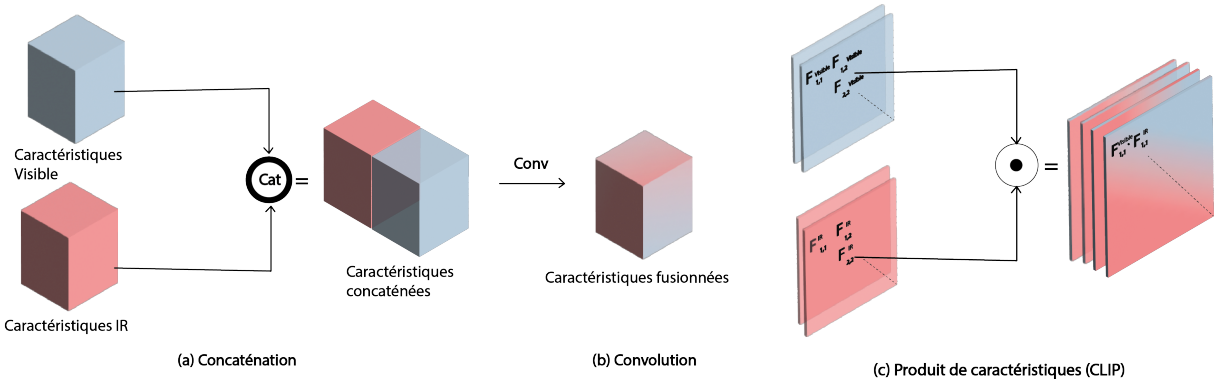


FIGURE 5.3 – Illustration d'opérations sur les caractéristiques mono-spectrales présentes dans la littérature.

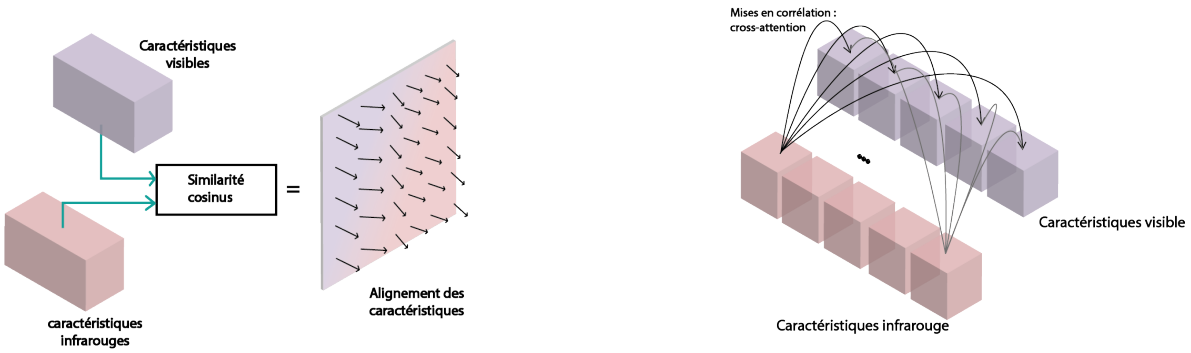


FIGURE 5.4 – Illustration de la similarité cosinus et de l'attention croisée.

5.1.3 Évolution des *Transformers* de détection

Les *Transformers* de détection sont, à notre connaissance, encore peu étudiés dans la littérature de fusion multi-spectrale. Pourtant, ceux-ci sont aujourd'hui largement utilisés dans l'état de l'art de la détection d'objets en spectre visible.

Les premières solutions de détection d'objets par apprentissage machine performantes émergent dès les années 2000 avec le développement des cascades de Haar, reposant sur l'extraction de caractéristiques définies par l'humain, couplée à un système de classification par apprentissage machine pour la détection de visages [122]. Ce domaine de recherche a ensuite connu une forte progression dans la capacité des approches à détecter efficacement

des objets grâce aux recherches en apprentissage profond pour la vision. Nous pouvons noter notamment l'élaboration des R-CNN et Fast(er)-RCNN, qui sont les réseaux de neurones les plus connus permettant une localisation d'objets par apprentissage profond [123]. Néanmoins, ces architectures reposent souvent sur la mise en place d'une série importante de modules et de sous-modules, comme des réseaux de proposition de région (RPN), sous-traitant la tâche de localisation en plusieurs étapes intermédiaires, mais alourdissant l'ensemble et compliquant leur entraînement.

Au contraire, *DEtection TRansformer* (DETR) [124] propose l'utilisation d'un *Transformer* monolithique qui utilise le système de requêtes pour remplacer le RPN. À partir d'une extraction de caractéristiques, le modèle génère des requêtes associées à des propositions de localisation d'objet par mécanisme d'attention, qui sont ensuite filtrées grâce à différentes techniques comme la suppression des non-maximums dans le score de confiance. C'est une simplification importante par rapport aux Fast(er)-RCNN : en effet, l'architecture DETR est plus minimaliste avec seulement deux modules de traitement neuronaux principaux. Un extracteur de caractéristiques classique y pré-traite l'information d'entrée afin d'en extraire des caractéristiques utiles de l'image d'entrée. De l'autre, un détecteur consistant en un *Transformer* calcule de manière itérative une équation d'attention afin de fournir en sortie une matrice des détections et localisations de cibles associées. L'efficience de ce modèle est permise notamment par le mécanisme d'attention, dont le pouvoir expressif [125] serait supérieur aux couches de convolution plus conventionnelles.

L'architecture Deformable-DETR est le deuxième grand chaînon de cette lignée [93]. Ce modèle, qui emploie l'extraction de caractéristiques à différents degrés d'abstraction, introduit l'attention déformable, moins rigide que l'attention conventionnelle et surtout moins dépendante de la taille de fenêtrage, qui a tendance à affecter la granularité des détections réalisées. Le modèle, en plus de présenter un entraînement accéléré, s'améliore dans la localisation d'objets de faibles dimensions spatiales, une limite du DETR d'origine. Ensuite, une autre architecture a été proposée, H-Deformable DETR [126], qui améliore le mécanisme de requête à l'aide d'une mise en correspondance entre les requêtes fournies dans le *Transformer* de détection et la vérité terrain, lors de l'apprentissage. Une annexe expliquant la déformabilité pour les réseaux de neurones est fournie A.8.

DAB-DETR [127] repose aussi sur un système de correction de requêtes, à l'aide de l'introduction d'un mécanisme de requête dynamique des détections potentielles, qui sont mises à jour à la manière d'un pré-traitement à travers chaque couche d'attention. Malgré l'alourdissement calculatoire, le modèle permet tout de même une amélioration significative des performances obtenues.

DINO-DETR, appelé DINO par abus de langage, est l'un des derniers *Transformers* de détection ayant émergé dans la littérature [128]. Il s'agit d'une amélioration de la mise en correspondance des requêtes par rapport à H-Deformable DETR. Il déploie aussi un apprentissage contrastif par débruitage et une palette étendue d'augmentations de données. Cette architecture permet d'atteindre des performances état-de-l'art sur une bonne part des problèmes de localisation et en particulier sur la base de données COCO [94].

5.2 Motivations pour la recherche de nouvelles architectures

L'état de l'art indique d'une part que le choix d'une fusion de caractéristiques est prometteuse, et d'autre part que l'utilisation du mécanisme d'attention pour la fusion multi-spectres est encore peu explorée. La capacité de ce type d'opération à faire des corrélations entre différentes régions des images d'entrée offre une solution pour traiter des images partiellement désalignées. L'effet du désalignement entre chaque spectre sur les performances est lui aussi peu étudié dans la littérature. En effet, la recherche est principalement tirée par les applications en conduite autonome et en surveillance, où les paires d'images sont généralement recalées. Néanmoins, ce recalage est plus difficile à obtenir en contexte de contrôle non destructif où le champ d'observation propre à chaque moyen d'imagerie peut être très différent, avec des résolutions parfois très variables rendant l'appariement précis difficile. Une robustesse à ces phénomènes serait bénéfique pour l'ensemble des problèmes de vision multi-modale car permettant de simplifier en partie les protocoles d'acquisition de données multi-spectres.

Nous proposons donc d'élaborer une nouvelle architecture de fusion basée sur les *Transformers* de détection et utilisant la fusion des caractéristiques par opération d'attention croisée pour combiner les informations de chaque modalité.

5.3 CAFF-DINO

Cette section présente la constitution de l'architecture CAFF-DINO. Cette architecture procède d'abord à l'extraction de caractéristiques dans chaque spectre séparément. Ensuite, le modèle proposé emploie un module de fusion des caractéristiques à chaque niveau de l'extraction, qui est basé sur l'attention croisée. Les cartes de caractéristiques fusionnées ainsi produites viennent alimenter un *Transformer* de détection, basé sur DINO [128] dans notre proposition de modèle finale.

5.3.1 Composition générale du modèle

Le fonctionnement général de l'architecture que nous avons développée est illustré à la Figure 5.5. Notre modèle reprend la structure typique des *Transformers* de détection, avec des extracteurs de caractéristiques séparés pour chaque modalité. Puis, une tête de détection consistant en un *Transformer* de détection réalise la localisation d'objets à partir des caractéristiques fusionnées issues des images d'entrée. Un module de fusion est intercalé entre chaque extracteur : pour chaque degré d'abstraction de l'extraction de caractéristiques, le module fusionne les caractéristiques de chaque modalité et les corrèle entre elles, par une opération d'attention croisée. Ce module permet d'estimer une nouvelle carte de caractéristiques fusionnée en sortie de module, qui est transmise au *Transformer* de détection.

L'architecture globale est construite pour permettre de réduire le coût en ré-entraînement au travers d'une conception modulaire : on peut réutiliser des extracteurs et têtes déjà pré-entraînés, ce qui réduit le coût en calcul et en données lors de l'entraînement, qui est de surcroît réalisé avec l'extraction de caractéristiques gelée pour les deux spectres. Il

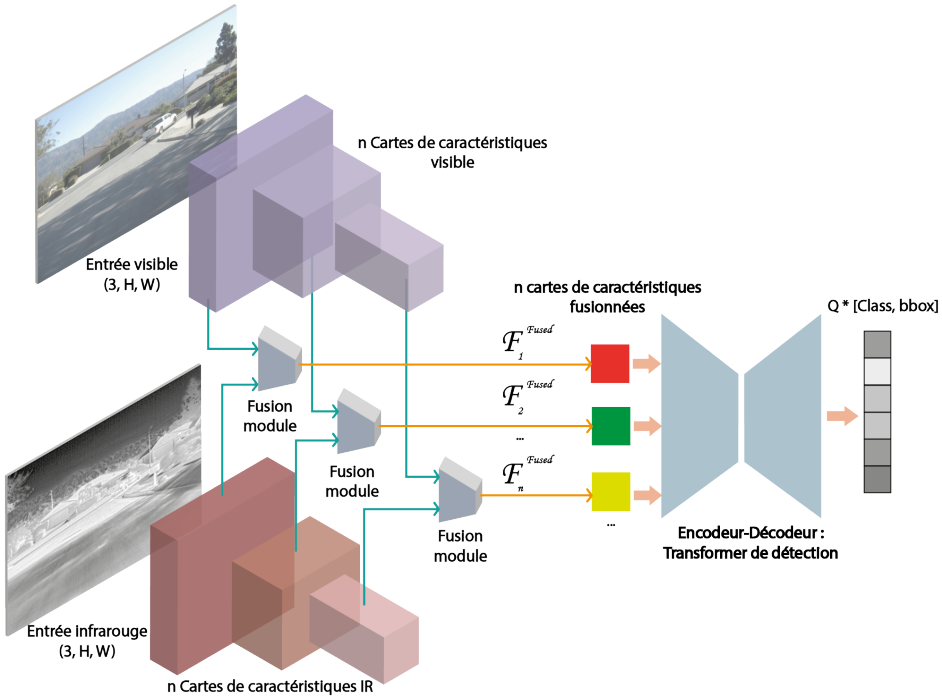


FIGURE 5.5 – Illustration de CAFF-DINO pour la fusion d'images infrarouge-visible. L'architecture est composée de deux extracteurs de caractéristiques mono-spectraux, combinés à l'aide de 4 modules de fusion pour la tête DINO. Cette tête basée sur les *Transformers* de détection réalise la localisation des objets de l'image à partir des deux modalités fusionnées.

n'y a pas de réinjection des caractéristiques fusionnées dans chaque extracteur de caractéristiques mono-spectral qui restent donc strictement indépendants, contrairement aux architectures basées sur *Cyclic Fuse and Refine* [114], ce qui joue encore à augmenter la modularité. L'architecture est ainsi facile à adapter à de nouveaux *Transformers* de détection ou à de nouveaux extracteurs de caractéristiques, suivant les futurs progrès architecturaux en apprentissage profond pour la vision par ordinateur.

5.3.2 Module de fusion proposé

Divers travaux dans la littérature ont déjà proposé l'utilisation de l'auto-attention pour fusionner des caractéristiques [116, 117]. *A contrario*, nous proposons l'emploi de l'attention croisée qui est une opération permettant d'extraire directement une information de corrélation du spectre infrarouge par rapport au spectre visible. Le module de fusion est nommé CAFF pour *Cross-Attention Features Fusion*, il est illustré par la Figure 5.6 identifiant chaque traitement appliqué aux caractéristiques mono-spectrales.

L'idée derrière notre méthode de fusion est de forcer l'extraction d'associations significatives et de corrélations entre les modalités d'entrée. Au lieu d'un travail précédent utilisant l'opération d'auto-attention [117], l'opération proposée ici est l'attention croisée, développée pour la vision dans [129] : cette opération est directement axée sur l'extraction d'informations associatives entre les deux modalités. Pour chaque étape de l'extraction

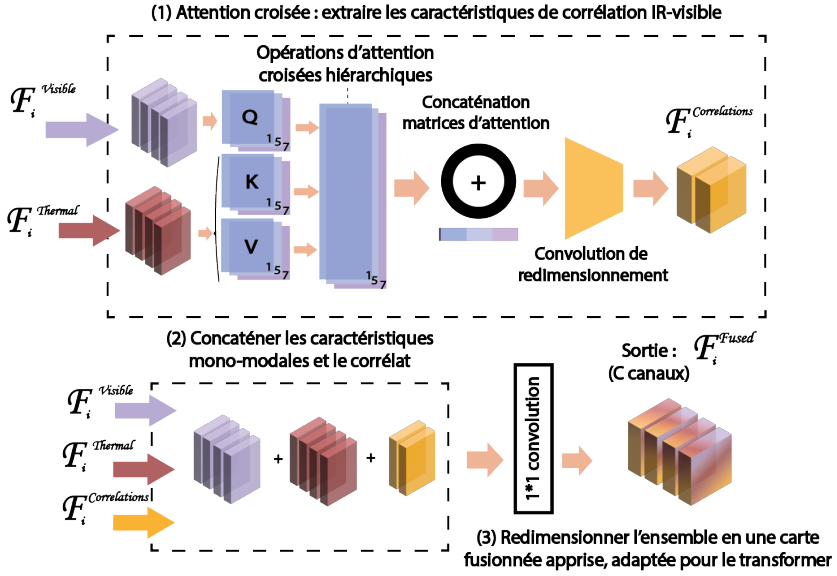


FIGURE 5.6 – Illustration du module de fusion des caractéristiques CAFF appliqué à la fusion IR-visible. À chaque niveau d'extraction, les caractéristiques mono-spectrales sont mises en corrélation en utilisant des opérations d'attentions croisées hiérarchiques (1). Les cartes de caractéristiques IR, visibles et de corrélation sont ensuite concaténées (2). Ces caractéristiques sont finalement fusionnées avec des opérations de convolution 1*1 et adaptées à la tête de détection (3).

de caractéristiques, un module nommé **Fusion de Caractéristiques par Attention Croisée** ou *Cross-Attention based Features Fusion* **(CAFF)** effectue une attention croisée hiérarchique d'une modalité vers l'autre. Les opérations effectuées sur les cartes de caractéristiques, à chaque niveau d'abstraction, sont illustrées dans la Figure 5.6. Comme montré dans le schéma, l'attention croisée est réalisée avec plusieurs tailles de noyau, afin de permettre une extraction de caractéristiques corrélées à différentes échelles spatiales, d'où le terme hiérarchique. Ensuite, les informations de corrélation obtenues sont combinées avec les caractéristiques mono-modales.

Pour n cartes de caractéristiques extraites par spectre (ici thermiques et visibles), la mise en œuvre proposée est formalisée comme suit. Tout d'abord, une attention croisée est calculée, mettant en corrélation les informations extraites de la modalité thermique vers le spectre visible qui sert de référence, comme décrit dans l'équation (5.6). L'opération est appliquée sur les caractéristiques $F_i^{\text{thermique}}$ et F_i^{visible}. Les matrices de requêtes, de clés et de valeurs sont définies en utilisant plusieurs couches de convolution avec différentes tailles de noyaux (attention hiérarchique), augmentant la capacité du modèle à extraire des corrélations multi-niveaux entre les caractéristiques de chaque modalité.

$$\text{CrossAttn}_i^k = \text{softmax}\left(\frac{Q_i^k(F_i^{\text{visible}}).K_i^k(F_i^{\text{thermal}})^T}{\sqrt{d_k}}\right) V_i^k(F_i^{\text{thermal}}),$$
$$\forall i \in \{1,\ldots,n\}, \quad \forall k, \text{taille de noyau} \tag{5.6}$$

Les matrices d'attention croisée obtenues sont empilées et compressées à l'aide d'une

couche de convolution de profondeur C imposée par le choix du *Transformers* de détection de la littérature. Pour enrichir les caractéristiques de corrélation extraites, plusieurs blocs d'auto-attention sont appliqués sur la carte d'attention croisée fusionnée. Les informations de corrélation sont à nouveau compressées en utilisant une couche de convolution unitaire, dans une profondeur de 64 canaux et définies comme $\mathrm{F}_i^{\text{Corrélations}}$. Les cartes de caractéristiques visibles et thermiques sont concaténées avec $\mathrm{F}_i^{\text{Corrélations}}$ pour former $\mathrm{F}_i^{\text{Stacked}}$, dans une profondeur de 2*C+64 canaux comme décrit dans l'Equation (5.7).

$$\mathrm{F}_i^{\text{Stacked}} = F_i^{\text{Visible}} + F_i^{\text{Thermal}} + \mathrm{F}_i^{\text{Corrélations}}, \quad \forall i \in \{1, \ldots, n\} \tag{5.7}$$

Une convolution 1*1 est appliquée sur les caractéristiques concaténées, ce qui compresse une dernière fois la dimensionalité du vecteur de caractéristiques fusionnées à la forme attendue par l'encodeur des têtes de *Transformers* dans l'équation 5.9.

$$\mathrm{F}_i^{\text{Fused}} = \mathrm{Conv}_{2C+64\to C}^{k=1}(\mathrm{F}_i^{\text{Stacked}}), \tag{5.8}$$

$$\forall i \in \{1, \ldots, n\} \tag{5.9}$$

Le module de fusion proposé est optimisé pour la structure de base Swin-Large et le détecteur DINO, et est nommé CAFF. Ce choix est justifié dans la Section 5.5.4 qui présente une comparaison des têtes et des structures de base. Le Tableau 5.1 décrit le module de fusion, le type de structure de base, la taille des noyaux de convolution utilisés dans les attentions croisées pour calculer les clés, les requêtes et les valeurs, ainsi que le nombre de blocs d'attention supplémentaires. Le module CAFF correspond au module optimisé pour un extracteur de caractéristiques Swin-Large, tandis que CAFF* est optimisé pour un extracteur Resnet50, et est comparé à CAFF dans l'étude d'ablation.

Fusion	Extracteurs	Noyaux pour attn croisée	# (bloc d'attn)
CAFF	Swins	$k = \{1, 5, 7\}$	2
CAFF*	Resnets	$k = 1$	3

TABLE 5.1 – Hyper-paramètres associés aux deux modules de fusion des caractéristiques : CAFF et CAFF*.

5.4 Bases de données de référence

Plusieurs bases de données font référence pour la comparaison de performance des architectures de fusion infrarouge-visible dans la communauté de la détection d'objets multi-spectrale. Nous décrivons ici leurs principales caractéristiques.

5.4.1 LLVIP pour la surveillance en milieu urbain

LLVIP est une base de fusion infrarouge-LWIR visible pour la surveillance en milieu urbain. Les paires d'images sont découpées puis recalées, limitant le décalage sub-pixel.

Les acquisitions, essentiellement de nuit et dans le même type d'environnement de ville, sont favorables à l'infrarouge. La quantité de données disponible est relativement importante comparée aux autres bases multi-spectrales (12.000 paires pour l'entraînement, 4.000 paires pour le test). La base de données est mono-classe, réduisant la difficulté du problème.

5.4.2 FLIR pour la conduite autonome

La base de données FLIR-ADAS a été développée par l'entreprise FLIR Teledyne pour l'entraînement d'architectures de localisation multi-spectrale dans un contexte de conduite autonome. Les images sont acquises depuis un véhicule dans des contextes variés : campagne, ville, et jour comme nuit. Néanmoins, la base d'origine est plutôt construite pour un entraînement mono-spectral. Il contient ainsi un nombre conséquent d'images non appairées ou partiellement recalées.

FLIR-aligned est ainsi une version épurée et purgée de la base de données mise en ligne par Teledyne : c'est la base FLIR la plus largement utilisée par la communauté. Cette version épurée présente un jeu d'entraînement limité à 4000 images sélectionnées car ayant une paire associée et un recalage correct au niveau du pixel de l'image visible par rapport à l'infrarouge. Le volume de données est ainsi plus faible, ce qui représente une première contrainte d'apprentissage par rapport à LLVIP. La base présente aussi un problème de détection multi-classes en lieu et place d'une tâche de détection mono-classe : le modèle apprend à y détecter les classes "personne", "vélo" et "voiture", ce qui est plus diversifié que LLVIP en termes de propriétés d'objets à identifier. 1000 paires d'images sont dédiées au test.

5.4.3 VEDAI pour la surveillance aérienne et spatiale

La base de données VEDAI est une base de données pour la fusion en contexte de surveillance aérienne et spatiale, fournissant des images recalées infrarouges et visibles et visant à la détection de cibles de type véhicule (détection des classes voiture et camion). Le nombre de paires d'images total est de plus de 1100 : 900 pour l'entraînement, 200 pour le test. Cela est très restreint comparé aux bases précédentes. Ce volume de données raréfié peut avoir un impact sur l'apprentissage d'une architecture basée sur les *Transformers*, qui présentent généralement un nombre important de paramètres à apprendre. De plus, la cible est particulièrement difficile à repérer, notamment pour un *Transformer* (objet petit sur l'image). Plusieurs jeux d'annotations contradictoires semblent être présents en ligne pour cette base. Dans le cadre de nos expérimentations, nous nous sommes donc concentrés sur les annotations trouvées ici : (lien vers les annotations employées https://universe.roboflow.com/uni-project-o9mo5/vedai-tsw2j). Cette base nous permet d'évaluer les performances de notre approche en dehors de l'environnement urbain des deux bases précédentes.

La Figure 5.7 montre des paires visible-infrarouge issues de chaque base de données présentée ci-dessous, permettant d'en apprécier les ressemblances et les propriétés spécifiques.

FIGURE 5.7 – Paires Visible-IR pour les bases de données LLVIP, FLIR-aligned et VEDAI.

5.5 Expérimentations

Une série d'expérimentations a été conduite afin de valider l'architecture précédemment élaborée sur les bases de données de référence mentionnées précédemment (section 5.5.2). Nous évaluons ensuite la robustesse de l'architecture au désalignement systématique d'une modalité (Section 5.5.3). Enfin, une étude d'ablations nous permet de montrer le degré d'interchangeabilité des têtes et des extracteurs de caractéristiques disponibles dans la littérature, et soutient de plus le choix de la composition SWIN + CAFF + DINO comme meilleur modèle (Section 5.5.4).

5.5.1 Mise en œuvre

Les différents composants d'origine mono-spectrale que sont la tête de détection et les extracteurs de caractéristiques ont été initialisés grâce à des poids pré-entraînés sur la base COCO accessibles en ligne [94]. Le module de fusion a quant à lui été initialisé aléatoirement. Afin de réduire le nombre de paramètres total, qui est très important, et le temps d'inférence, les extracteurs de caractéristiques mono-spectraux ont été gelés durant l'entraînement : il n'y a pas de propagation de gradient, ce qui permet de réduire le risque

de sur-paramétrisation, notamment sur des bases où la quantité de données est plus limitée. Les hyper-paramètres ont été définis suivant les spécifications de l'architecture mono-spectrale d'origine, sauf contre-indication indiquée par les auteurs [128]. La métrique de mesure de la qualité de la localisation d'objets est ici à nouveau la mAP et ses variantes standards que sont la mAP50, la mAP75 (voir Chapitre 4, section 4.3.1 pour plus de détails sur cette mesure). C'est la librairie PyCocoTool, standard pour l'évaluation des *Transformers* de détection sous Pytorch, qui a été utilisée pour l'estimation des métriques.

5.5.2 Performances de localisation sur les bases de données de référence

Notre architecture est confrontée aux autres modèles de la littérature sur les bases de données LLVIP, FLIR-aligned et VEDAI. Nous évaluons de plus le visible seul, l'infrarouge seul et la fusion des deux.

Base	Modality	Extracteur	Détecteur	mAP50 (↑)	mAP75 (↑)	mAP (↑)
LLVIP	Visible	CSPD53	YOLOv5 [117]	90.8	51.9	50.0
	IR	CSPD53	YOLOv5 [117]	94.6	72.2	61.9
	Vis+IR	-	HalfWay [114]	91.4	60.1	55.1
	Vis+IR	-	ProbEn [130]	93.4	50.2	51.5
	Vis+IR	-	GAFF [116]	94.0	60.2	55.8
	Vis+IR	-	CSSA [115]	94.3	66.6	59.2
	Vis+IR	CFB	CFT-YOLO-v5 [117]	97.5	72.9	63.6
LLVIP **(our)**	Visible	Swin-Large	DINO	91.3	59.8	54.4
	IR	Swin-Large	DINO	97.3	79.0	67.5
	Vis+IR	Swin-Large-CAFF	DINO	**98.1**	**79.0**	**68.5**

TABLE 5.2 – Comparaison de performance de détection entre l'architecture CAFF-DINO (swin-Large-CAFF+DINO) et les approches de la littérature, sur la base de données LLVIP.

Base de données LLVIP. Les résultats de notre architecture sur cette base sont fournis dans le Tableau 5.2. Les performances de notre modèle surpassent celles de la littérature : la mAP augmente de 4.9 % comparé au meilleur modèle de l'état de l'art [117]. Néanmoins, la fusion semble un peu moins intéressante vue sous l'angle de la comparaison IR seul contre visible et IR combinés, avec une hausse de mAP de 1 %, ce qui peut être dû à une distribution des données principalement favorable à la modalité infrarouge avec des scènes capturées de nuit, en milieu urbain, et face à un problème mono-classe de détection de piétons. Deux exemples de détection multi-spectrale réalisées par notre architecture sont fournis dans la Figure 5.8. Le score de confiance est défini ici par le *logit* associé à la détection d'un objet donné i.e l'intensité de la réponse du neurone de sortie associé à la classe détectée. Il est seuillé à 50 % pour l'ensemble des exemples de ce chapitre.

Base de données FLIR-aligned. Les résultats sur cette base sont comparés aux modèles de la littérature dans le Tableau 5.3. Les performances générales sont plus réduites sur cette tâche, qui est, comme discuté auparavant, plus difficile. Les performances de notre approche de fusion sont bien supérieures à celles référencées dans la littérature. Le bénéfice de la fusion est bien plus prégnant dans cette base : en effet, la base FLIR-aligned contient une plus grande variété entre les images urbaines et celles en campagne, avec un mélange entre acquisitions de jour et de nuit, augmentant la nécessité pour le

Base	Modalité	Extracteur	Détecteur	mAP50 (↑)	mAP75 (↑)	mAP (↑)
FLIR-aligned	Visible	Resnet-50	Faster-RCNN	64.9	21.1	28.9
	IR	Resnet-50	Faster-RCNN	74.4	32.5	37.6
	Visible	CSPD53	YOLO-v5 [117]	67.8	25.9	31.8
	IR	CSPD53	YOLO-v5 [117]	73.9	35.7	39.5
	Vis+IR	ResNet18	GAFF [116]	72.9	32.9	37.5
	Vis+IR	-	ProbEn [130]	75.5	31.8	37.9
	Vis+IR	CFB	CFT-YOLO-v5 [117]	78.7	35.5	40.2
	Vis+IR	-	CSSA [115]	79.2	37.4	41.3
	Vis+IR	-	ICA-Fusion [131]	79.2	36.9	41.4
FLIR-aligned **(our)**	Visible	Swin-Large	DINO	75.6	33.5	39.2
	IR	Swin-Large	DINO	77.2	41.3	43.6
	Vis+IR	Swin-Large-CAFF	DINO	**85.5**	**51.6**	**50.5**

TABLE 5.3 – Comparaison de performance de détection entre l'architecture CAFF-DINO (swin-Large-CAFF+DINO) et les approches de la littérature, sur la base de données FLIR-aligned.

FIGURE 5.8 – Paires visible-infrarouge issue de la base LLVIP. La vérité-terrain est encadrée en rouge, tandis que les détections réalisées par CAFF-DINO sont en bleu. Le seuil de confiance a été fixé à 50 %.

modèle d'exploiter les caractéristiques de chaque modalité. Deux exemples de détection multi-spectrale réalisées par notre architecture sont fournis dans la Figure 5.9.

Base de données VEDAI. Les résultats de l'approche CAFF-DINO sont décrits dans le Tableau 5.4, accompagnés des modèles de la littérature. À première vue, si CAFF-DINO surclasse à nouveau les autres approches, mono ou multi-modales, le bénéfice de la fusion est ici plus mitigé : si la mAP est augmentée de 0.9 %, les autres métriques sont légèrement diminuées par rapport à l'infrarouge seul. Plusieurs explications sont plausibles ici : d'une part, la raréfaction du volume de données rend la convergence du modèle plus difficile. Une explication peut venir des caractéristiques formées depuis la base COCO

Base	Modalité	Extracteur	Détecteur	mAP50 (↑)	mAP75 (↑)	mAP (↑)
VEDAI	Visible	CSPD53	YOLOv5 [117]	74.3	46.9	46.2
	IR	CSPD53	YOLOv5 [117]	74.0	46.8	46.1
	Vis+IR	-	Early Fusion [132]	-	-	44.0
	Vis+IR	-	Mid-Fusion [132]	-	-	44.6
	Vis+IR	CFB	CFT-YOLO-v5 [117]	85.3	65.9	56.0
VEDAI **(our)**	Visible	Swin-Large	DINO	86.5	69.4	55.7
	IR	Swin-Large	DINO	81.2	64.0	52.2
	Vis+IR	Swin-Large-CAFF	DINO	86.4	68.2	**56.6**

TABLE 5.4 – Comparaison entre l'architecture CAFF-DINO (swin-Large-CAFF+DINO) et les approches de la littérature, sur la base de données VEDAI.

utilisée en pré-entraînement de nos extracteurs, qui sont peut-être moins pertinentes pour des tâches de fusion de données acquises en télédétection par satellite.

Nous avions tenté un entraînement complet sur cette base, avec les extracteurs dégelés, néanmoins nous n'avions pas observé de performances supérieures à l'approche proposée. Cela peut indiquer une sur-paramétrisation du modèle actuel. Une solution serait de réaliser un pré-apprentissage mono-spectral des extracteurs sur une base satellitaire accessible et riche en données, semblable à COCO en ordre de grandeur, soit plus de 10.000 images.

Deux exemples de détection multi-spectrale réalisées par notre architecture sont fournis dans la Figure 5.10.

5.5.3 Robustesse de la localisation au désalignement systématique entre les deux modalités

Dans cette partie, un désalignement systématique a été induit sur l'image infrarouge à la fois durant l'apprentissage et durant l'inférence pour l'évaluation des performances. Cette expérimentation a été conduite sur les bases de données LLVIP et FLIR-aligned. Notre approche a été ici comparée avec l'architecture CFT-YOLO v5 [117], qui présente l'intérêt d'utiliser une opération de fusion relativement proche, l'auto-attention, et dont le code est récent et accessible en ligne : https://github.com/DocF/multispectral-object-detection.

Les tableaux 5.5 et 5.6 montrent la mAP évaluée à chaque décalage, ainsi que la diminution relative du mAP associée à ce décalage systématique, pour les bases LLVIP et FLIR-aligned. Nous constatons que la diminution relative de la mAP est généralement moins importante pour CAFF-DINO que pour l'architecture de la littérature [117], soulignant la robustesse de l'architecture proposée.

L'approche CFT-YOLO [117] pourrait avoir plus de mal à compenser le désalignement, particulièrement dans les données FLIR qui contiennent des changements d'environnement plus difficiles à gérer, nécessitant de réaliser une fusion précise des deux spectres.

Les deux modèles étudiés ici convergent également vers les performances obtenues dans le spectre non décalé (le spectre visible) pour le plus grand désalignement de 200 pixels, car plus d'informations sont perdues dans l'IR en raison du désalignement systématique.

FIGURE 5.9 – Paires visible-infrarouge issue de la base FLIR-aligned. La vérité-terrain est encadrée en rouge, tandis que les détections réalisées par CAFF-DINO sont en bleu. Le seuil de confiance a été fixé à 50 %.

Modèle	10px	50px	100px	200px
mAP(CFT-YOLO-v5)	51.3 (-19 %)	50.3 (-21 %)	49.2 (-23 %)	51.8 (-19 %)
mAP(CAFF-DINO)	57.4 (-16 %)	59.7 (-12 %)	57.0 (-17 %)	53.9 (-21 %)

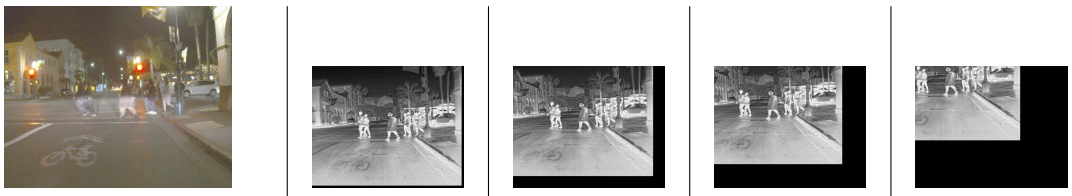
TABLE 5.5 – Illustration du mAP et de la diminution relative du score après désalignement sur des données LLVIP. La diminution relative du score est calculée comme la différence relative entre la mAP sur les données alignées et les données désalignées. Le désalignement est indiqué en pixels (px).

5.5.4 Ablations

Une série d'expériences a été conduite sur des données alignées afin de valider les choix de composition architecturale. Cela permet aussi de confirmer partiellement l'aspect générique de l'architecture proposée, comme l'interchangeabilité des extracteurs ou des têtes de détection. Nous comparons aussi différentes opérations de fusion à l'attention croisée proposée.

Comparaison de plusieurs têtes de détection. La performance de plusieurs têtes est évaluée sur LLVIP et FLIR afin d'estimer le bénéfice comparatif de notre approche

FIGURE 5.10 – Paires visible-infrarouge issue de la base de données VEDAI. La vérité-terrain est encadrée en rouge, tandis que les détections réalisées par CAFF-DINO sont en bleu. Le seuil de confiance a été fixé à 50 %.



Modèle	10px	50px	100px	200px
mAP(CFT-YOLO-v5)	34.7 (-13 %)	28.7 (-28 %)	28.7 (-28 %)	29.0 (-27 %)
mAP(CAFF-DINO)	49.8 (-1 %)	37.2 (-26 %)	39.4 (-22 %)	43.7 (-13 %)

TABLE 5.6 – Illustration du score de détection (mAP) et de la diminution relative du score après désalignement sur des données FLIR-aligned. La diminution relative du score est calculée comme la différence relative entre la mAP sur les données alignées et les données désalignées. Le désalignement est indiqué en pixels (px).

de fusion avec les différents détecteurs basés sur l'attention élaborés dans la littérature. Les détecteurs évalués ici sont DETR [124], Lite-DETR [133], Deformable-DETR [93], H-Deformable-DETR [126] et DINO [128]. L'expérience est menée avec l'extraction de caractéristiques Resnet-50 (CAFF*), en raison de la plus grande disponibilité de poids

pré-entraînés en libre accès avec cette structure de base par rapport à Swin-Large. Le Tableau 5.7 montre les résultats de cette expérience. Nous pouvons observer de cette ablation que le mécanisme de fusion proposé permet d'avoir une architecture convergente sur l'ensemble des têtes proposées, tendant à soutenir l'aspect modulaire de la technique de fusion. La performance de détection sur ce jeu de données suit généralement la chronologie des progrès techniques réalisés sur les *Transformers* de détection, du DETR original à DINO et aux architectures modernes. Enfin, les avantages de l'utilisation de têtes de *Transformers* de détection plus avancés tels que DINO sont confirmés.

Base	Extracteur	Tête	mAP75	mAP
LLVIP	Resnet-50-CAFF*	DETR	65.5	58.9
		Deformable-DETR	69.4	61.1
		H-Deformable-DETR	75.5	65.0
		Lite-DINO	77.7	65.7
		DINO	**78.2**	**67.0**
FLIR-a	Resnet-50-CAFF*	DETR	20.6	15.7
		Deformable-DETR	36.4	**38.3**
		H-Deformable-DETR	33.7	34.6
		Lite-DINO	33.2	36.8
		DINO	**40.6**	37.9

TABLE 5.7 – Comparaison de différentes têtes de détection basées sur les *Transformers* disponibles dans la littérature associées avec le module CAFF* sur LLVIP et FLIR-aligned.

Comparaison de différents extracteurs de caractéristiques. L'expérience est ici menée avec Deformable-DETR, qui fournit des structures de base pré-entraînées publiques Resnet50, Swin-Tiny et Swin-Large associées à cette architecture. Les performances obtenues avec DINO sont également mesurées, pour une comparaison entre CAFF et CAFF* (seuls les poids Resnet50 et Swin-Large sont disponibles).

Base	Tête	Extracteur	mAP75	mAP
LLVIP	Deformable-DETR	Resnet-50-CAFF*	69.4	61.1
		Swin-tiny-CAFF*	**76.2**	64.6
		Swin-Large-CAFF*	75.7	**64.9**
LLVIP	DINO	Resnet-50-CAFF*	76.3	66.3
		Swin-Large-CAFF*	78.1	67.6
		Resnet-50-CAFF	74.7	65.1
		Swin-Large-CAFF	**79.0**	**68.5**

TABLE 5.8 – Comparaison de différents extracteurs couplés à nos modules de fusion CAFF* et CAFF, associés avec deformable-DETR et DINO sur LLVIP.

La table 5.8 met en évidence l'avantage d'utiliser des extracteurs de caractéristiques plus expressifs et riches, tels que ceux basés sur le mécanisme d'attention, comparativement à Resnet qui est basé sur les convolutions uniquement. L'écart de performance avec l'extracteur de caractéristiques Swin est plus réduit, les deux structures de base offrant des performances comparables. Ici Swin-Large a été privilégié dans le modèle CAFF-DINO en raison de la disponibilité de ses poids pré-entraînés sur COCO. Les scores utilisant DINO mettent également en avant le plus grand bénéfice de CAFF sur la fusion basée sur un

extracteur de base Swin, et respectivement CAFF* sur une base Resnet.

Comparaison d'approches alternatives de fusion de caractéristiques. Deux modules de fusion alternatifs sont proposés comme points de comparaison avec la fusion CAFF. La fusion par **concaténation de caractéristiques uniques (concat)** consiste à concaténer les cartes de caractéristiques des deux spectres, à chaque niveau d'abstraction. Ce module de fusion est illustré dans la Figure 5.11. Le processus de cette fusion de caractéristiques peut être formalisé comme suit, pour n extractions de caractéristiques réalisées par chaque structure de base mono-spectre :

Les cartes de caractéristiques de chaque spectre, nommées respectivement F_i^{Thermal} et F_i^{Visible}, sont empilées, donnant un vecteur de caractéristiques d'une profondeur de 2*C (équation 5.10). C est la profondeur de la carte de caractéristiques attendue par l'encodeur du *Transformer*.

$$\begin{aligned} \mathrm{F}_i^{\text{Stacked}} &= F_i^{\text{Visible}} + F_i^{\text{Thermal}}, \\ &\forall i \in \{1, \ldots, n\} \end{aligned} \tag{5.10}$$

Une couche de convolution 1*1 compresse une dernière fois ce vecteur de caractéristiques concaténées à la forme attendue par l'encodeur d'origine du *Transformers* de détection, montré dans l'équation 5.11.

$$\begin{aligned} \mathrm{F}_i^{\text{Fused}} &= \mathrm{Conv}_{2C \to C}^{k=1}(\mathrm{F}_i^{\text{Stacked}}), \\ &\forall i \in \{1, \ldots, n\} \end{aligned} \tag{5.11}$$

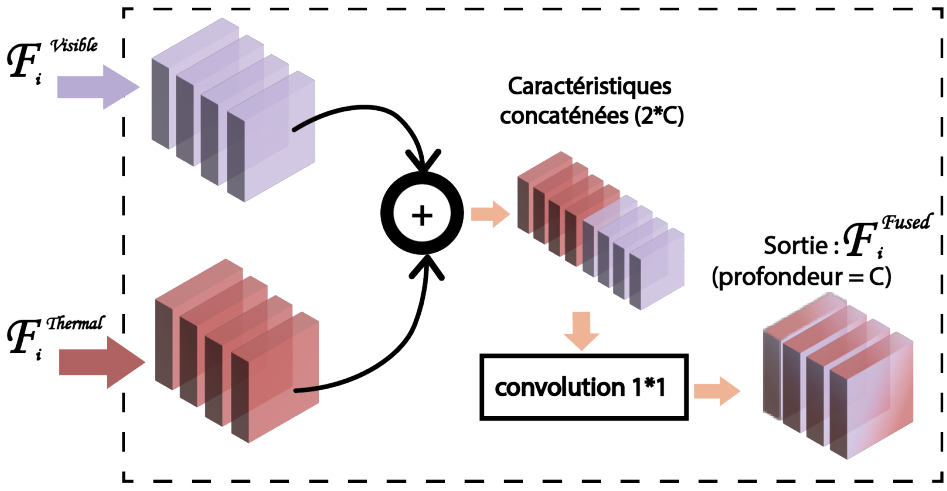


FIGURE 5.11 – Illustration de la fusion de caractéristiques par concaténation et convolution apprise.

CAFF est comparé ici à l'**opération de similarité cosinus (cos-sim)**, effectuée entre les cartes de caractéristiques de chaque spectre [134,135]. Cette opération de mesure de la similarité entre deux cartes de caractéristiques est présentée dans la section 5.1.2. Dans l'équation (5.12), F_i^{Thermal} et F_i^{Visible} représentent les cartes de caractéristiques mono-spectrales. Une convolution 1*1 est ajoutée, remodelant la matrice de similarité extraite pour l'ingestion de l'information dans la tête du *Transformer*.

$$\mathrm{F}_i^{\text{Correlations}} = \frac{F_i^{\text{Thermal}} \cdot F_i^{\text{Visible}}}{\|F_i^{\text{Thermal}}\| \times \|F_i^{\text{Visible}}\|}, \quad \forall i \in \{1, \ldots, n\} \tag{5.12}$$

Base	Tête	Extracteur	mAP75	mAP
LLVIP	DINO	Resnet-50-CAFF*	**78.2**	**67.0**
		Resnet-50-concat	77.9	66.9
		Resnet-50-cos-sim	77.8	66.5
FLIR-aligned	DINO	Resnet-50-CAFF*	**40.6**	37.9
		Resnet-50-concat	35.1	**38.9**
		Resnet-50-cos-sim	38.3	36.6
FLIR-aligned	DINO	SwinL-CAFF	**51.6**	**50.5**
		SwinL-concat	48.6	48.4
		SwinL-cos-sim	42.6	43.7

TABLE 5.9 – Comparaison des différentes méthodes de fusion étudiées : CAFF/CAFF*, concaténation des caractéristiques et similarité cosinus sur les bases LLVIP et FLIR-aligned.

Le Tableau 5.9 montre les scores de détection d'objets obtenus par les fusions alternatives sur LLVIP et FLIR. Il montre une diminution de performance jusqu'à 6,8 % lors de l'utilisation de la similarité cosinus sur le jeu de données FLIR-aligned, SwinL-cos-sim versus SwinL-CAFF dans le tableau, au lieu de notre proposition d'attention croisée. Si CAFF fournit généralement de meilleures performances de détection multi-spectrales, l'écart entre les différents modules est réduit sur LLVIP. La tâche de détection y est plus facile, avec principalement des données de vision nocturne ou en faible luminosité plus favorables à la modalité IR. La concaténation de caractéristiques uniques ,-concat dans le tableau, offre une haute performance, proche de CAFF et CAFF*. Celle-ci dépasse même la mAP avec l'extraction de caractéristiques Resnet-50 sur FLIR-aligned, soulignant la capacité des détecteurs utilisant les *Transformers* à extraire directement des corrélations significatives sur des cartes de caractéristiques empilées puis fusionnées de deux modalités. La fusion par concaténation seule serait une alternative en contexte de puissance de calcul contrainte car elle obtient des performances à peine inférieures à notre approche. La Figure 5.12 donne des exemples de de détection des différents modules de fusion, à seuil de 50 %, sur LLVIP : si les approches CAFF (a) et par concaténation (b) présentent des détections assez proches, il y a une baisse de la capacité de détection pour la fusion par similarité cosinus (c), induisant ici 2 détections manquées.

5.5.5 Conclusion sur les expériences conduites

Nous avons donc pu montrer sur différents problèmes de détection multi-spectrale visible et infrarouge que notre modèle fournissait des gains de performances par rapport à la littérature disponible (Section 5.5.2). Nous avons montré sa meilleure robustesse à un désalignement systématique de la modalité infrarouge par rapport à une architecture disponible dans la littérature (Section 5.5.3). Enfin, l'étude d'ablations nous a permis de

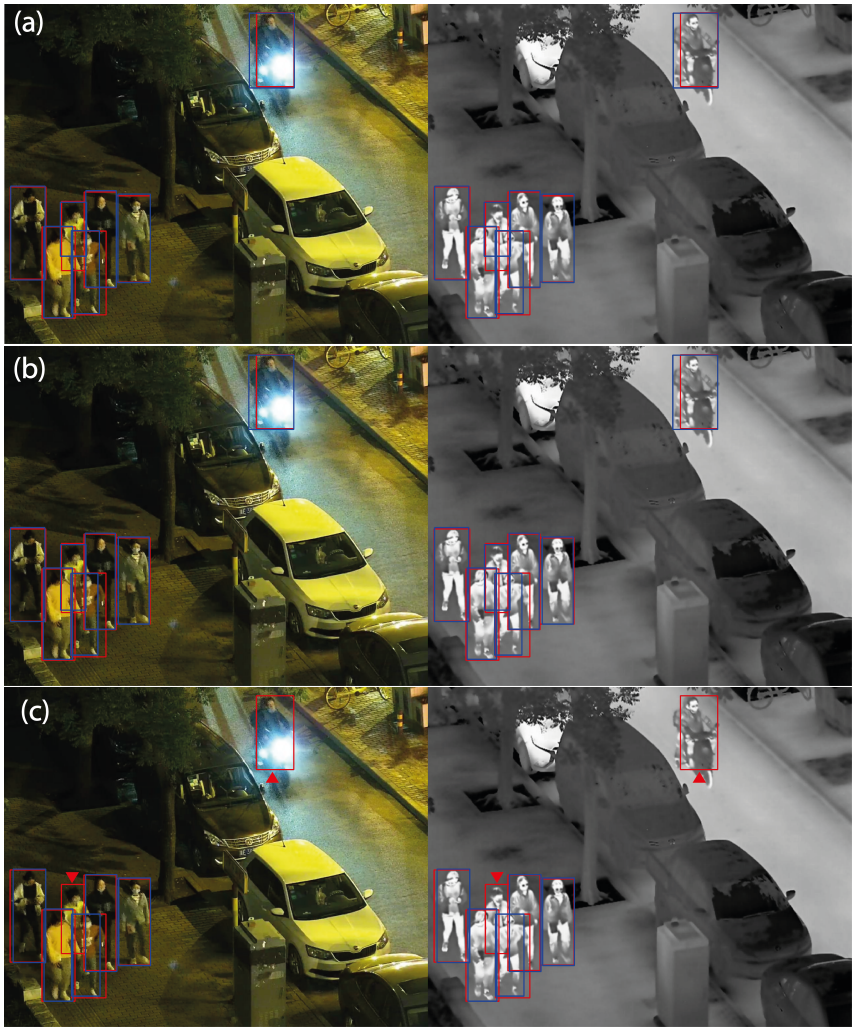


FIGURE 5.12 – Paires d'images visibles et infrarouges issues du jeu de test LLVIP. (a) est la détection de CAFF, (b) est la concaténation et (c) est la similarité cosinus. La vérité terrain est encadrée en rouge tandis que la détection de notre modèle est encadrée en bleu. Le seuil de confiance est de 50 %. Un triangle rouge indique la détection manquée.

montrer le degré de modularité de l'approche proposée ainsi que le bénéfice de l'attention croisée comparée à d'autres opérations de fusion (Section 5.5.4).

5.6 Conclusion

Dans ce chapitre, nous avons développé une nouvelle architecture de fusion construite à partir des *Transformers* pour la détection, et basée sur la fusion des caractéristiques extraites *via* une opération d'attention croisée. Nous avons vu que le modèle proposé, CAFF-DINO, surpassait les performances des architectures proposées actuellement dans la littérature.

Notre architecture est générique sur le plan architectural et peut être facilement ré-adaptée pour les futurs *Transformers* de détection qui émergeront dans la littérature de la détection d'objets mono-spectrale. Cela a été illustré dans les études d'ablations conduites

où notre module de fusion a été couplé à une assez grande variété de détecteurs basés sur les *Transformers*, permettant une convergence du modèle et des performances de détection allant jusqu'à approcher l'architecture basée sur DINO. Le modèle proposé est aussi générique du point de vue des modalités d'imagerie d'entrée : des fusions d'informations radar et visible, ou tout autre type de données d'image, sont alors envisageables.

Des perspectives d'amélioration du module de fusion peuvent être relevées, néanmoins quelques-unes semblent être plus évidentes pour permettre des gains de performance de détection supplémentaires. Symétriser l'opération d'attention croisée en est une : nous avions tenté l'ajout d'une opération symétrique d'attention croisée corrélant cette fois la modalité visible par rapport à l'infrarouge, puis la concaténation des deux caractéristiques d'attention croisée formées. Néanmoins, notre tentative entraînait un doublement du coût calculatoire du module de fusion et l'ensemble n'arrivait pas à convergence durant l'apprentissage. Ce problème est sans doute explicable par une trop grande augmentation du nombre de paramètres, qui reste un point critique de notre modèle. Une simplification opératoire de l'attention croisée suivant la dimensionalité des caractéristiques à traiter est aussi envisageable pour réduire la complexité globale de l'opération. Nous pourrions aussi nous inspirer des *Transformers* linéaires et des *Transformers* longs développés pour le langage naturel [136, 137].

La mise en place de méthodes de compression dimensionnelle et spatiale des caractéristiques mono-modales avant l'opération de fusion pourrait aussi régler le problème de l'explosion du nombre de paramètres, au moins partiellement. Cela pourrait potentiellement être réalisé au travers d'interpolations ou de convolutions des espaces latents mono-spectraux. Ces opérations économiseraient le nombre de paramètres, et donc ouvriraient à plus de marge calculatoire, en jouant sur la dimensionalité des vecteurs traités. Mais cela se ferait au prix d'une perte d'information, suivant le degré de compression : pour ces perspectives, une problématique de dosage complexe à équilibrer apparaît. Des expériences préliminaires montraient qu'une division par deux des cartes de caractéristiques traitées avant l'opération d'attention croisée, par interpolation non apprise, entraînait une perte de mAP létale de quelques pourcentages de l'architecture sur la base de données LLVIP. Néanmoins, ces tests restaient préliminaires : une interpolation apprise par convolution ainsi qu'une expérimentation claire permettant de choisir la réduction de dimension appliquée à chaque niveau d'abstraction permettrait sans doute de stabiliser les performances, en réduisant le coût en mémoire et en calcul.

L'architecture est entraînée avec les extracteurs gelés permettant notamment de réduire le nombre total de paramètres à entraîner, et donc la dépendance au volume de données. Cependant, les caractéristiques extraites par les extracteurs pour chaque spectre ne sont donc sans doute pas optimales. Il y a des solutions potentielles à ce problème dans la littérature de l'apprentissage profond pour la détection, par exemple par pré-entraînement des extracteurs de caractéristiques sur les bases infrarouge et visible seules, qui permettraient de les spécialiser. Un pré-entraînement auto/non-supervisé pourrait se révéler envisageable à terme, en vue d'optimiser les caractéristiques formées par les extracteurs pour chaque modalité avant leur fusion.

L'implémentation d'une fusion inspirée de la méthode CLIP [54] associée à de l'apprentissage contrastif [56] en supplément de l'attention croisée, pourrait amener à un gain de performance en enrichissant les caractéristiques de corrélation inter-spectre formées. Néanmoins, il faudrait alors procéder à l'entraînement sur des machines plus performantes

que celles disponibles directement à l'ONERA, telles que les calculateurs nationaux comme Jean-Zay, limitant alors la disponibilité et la versatilité de l'architecture proposée. Nous pourrions enfin chercher à rendre volontairement l'apprentissage plus robuste aux désalignements par l'introduction d'un décalage spatial aléatoire, ce qui n'a pas été encore étudié dans le cadre de cette recherche.

Nous pouvons maintenant mettre en œuvre notre méthode de fusion pour le contrôle non destructif, ici la fusion Flying Spot/visible. C'est l'objet du prochain chapitre.

Chapitre 6

Fusion infrarouge-visible pour la localisation de fissures en thermographie Flying-Spot

Le plus grand mystère n'est pas que nous soyons jetés au hasard entre la profusion de la vie et celle des astres ; c'est que, dans ce que Pascal appelle notre prison, nous tirions de nous-même des images assez puissantes pour nier notre néant...

Lazare, A. Malraux.

Sommaire

Nous avons élaboré une nouvelle architecture de fusion au chapitre précédent et ses capacités de détection et localisation ont été démontrées sur des données de fusion visible-infrarouge passives de référence. Nous l'exposons donc maintenant aux données infrarouges actives de thermographie Flying Spot couplées au spectre visible et à leurs spécificités, puis la comparons avec un ou plusieurs modèles élaborés dans la littérature. Ce dernier chapitre se concentre sur la constitution et l'utilisation par l'apprentissage de bases de données appairées infrarouge-visible, traitant de la problématique du recalage face à des différences de champs importantes entre les deux modalités. Par simplification de l'appariement et de la constitution des bases de données expérimentales, la fusion est réalisée sur les images thermiques individuelles en cours de balayage laser, couplées à l'enregistrement image-par-image d'une caméra visible.

Dans ce chapitre, nous présentons dans un premier temps la phase de collecte de données multi-spectrales FST-visible dans la section 6.1. Dans la section 6.2, l'adaptation de la synthèse d'images par diffusion pour la production de paires visible-infrarouge supplémentaires est développée. La section 6.3 présente ensuite les différents résultats

expérimentaux, comparant une architecture de fusion de la littérature avec le modèle développé dans le chapitre 5. Enfin, la section 6.4 développe les bénéfices de la fusion par rapport à la détection mono-spectrale des fissures pour la FST sur le jeu de données disponible.

La partie de ce travail consacrée à la synthèse de paires pour l'augmentation de données a fait l'objet d'un acte de conférence présenté à *SPIE Thermosense, Defense+Commercial Sensing 2024*, primé du prix du meilleur papier des doctorants.

Les notions d'entraînement de modèles pour la vision par ordinateur sont abordées plus en détail dans les annexes A.1 et A.2. L'idée de transfert d'apprentissages est synthétisée dans l'annexe A.5. Le lecteur peut aussi se référer à l'annexe A.6 sur les *Transformers* et le mécanisme d'attention. L'IA générative pour l'augmentation de données est présentée au Chapitre 2 et l'architecture de fusion CAFF-DINO fait l'objet du Chapitre 5.

6.1 Données multi-spectrales

Cette section se concentre sur la collecte de données multi-spectrales appariées. Pour chaque type d'échantillon employé, les paramètres thermographiques suivent ceux décrits dans le chapitre 1. Le protocole d'annotation est d'ailleurs sensiblement similaire à celui développé pour les données FST seules : c'est la modalité infrarouge qui sert de spectre de référence et d'annotation, le capteur thermique étant placé à la normale de l'échantillon. Pour chaque base, nous avons collecté une version partiellement recalée avec seulement un centrage-découpage paramétré manuellement des deux images autour de la source d'échauffement. Cela nous permet d'éviter des déformations dues aux différences de champs observés entre la modalité visible et l'infrarouge, tout en évitant une perte d'information spatiale visible issue d'une interpolation. Une base de données recalée a néanmoins été constituée par appariement multi-spectral pour fournir un point de comparaison avec les échantillons métalliques complexes revêtus.

Il est prévu à terme d'intégrer les données multi-spectrales, constituant la base FLYD II, ainsi qu'un supplément de données recalées encore en constitution sur ces échantillons, dans le dépôt github https://github.com/kevinhelvig/FLYD-ii après une vague d'acquisitions supplémentaires.

6.1.1 Banc d'essai expérimental

Le banc de thermographie laser de l'ONERA DMAS a été modifié par l'adjonction d'une caméra visible, illustré dans la Figure 6.1, afin de permettre des acquisitions par FST et en spectre visible partiellement synchrones. La synchronisation a été réalisée par logiciel ici. L'axe optique de la caméra visible forme un angle approximatif de 30 degrés avec l'axe normal caméra infrarouge - échantillon. Cet angle est le minimum pour pouvoir enregistrer avec la caméra visible sans avoir la lame dichroïque dans le champ. Un filtre orange a été de plus ajouté sur le capteur visible permettant d'atténuer la saturation induite par le faisceau laser. La fréquence d'acquisition globale de l'ensemble est fixée à 5 Hz, qui est la valeur maximale de la caméra visible et permettant de maximiser le nombre d'images produites modulo une certaine perte de qualité d'enregistrement. Les balayages sont effectués à une distance allant jusqu'à un centimètre avec le défaut cible, augmentant la diversité des réponses collectées. La vitesse de balayage a été fixée à 0.5

mm/s. La puissance a varié de 0.5 W pour les échantillons complexes, à 1.5 W dans le cas des échantillons métalliques d'essais mécaniques. Les données en spectre visible ont de plus été débruitées par une approche de type *filtre bilatéral* et rehaussement de contraste CLAHE [138]. La caméra visible utilisée pour ce chapitre de démonstration expérimentale est assez ancienne et l'acquisition a une qualité d'image dégradée.

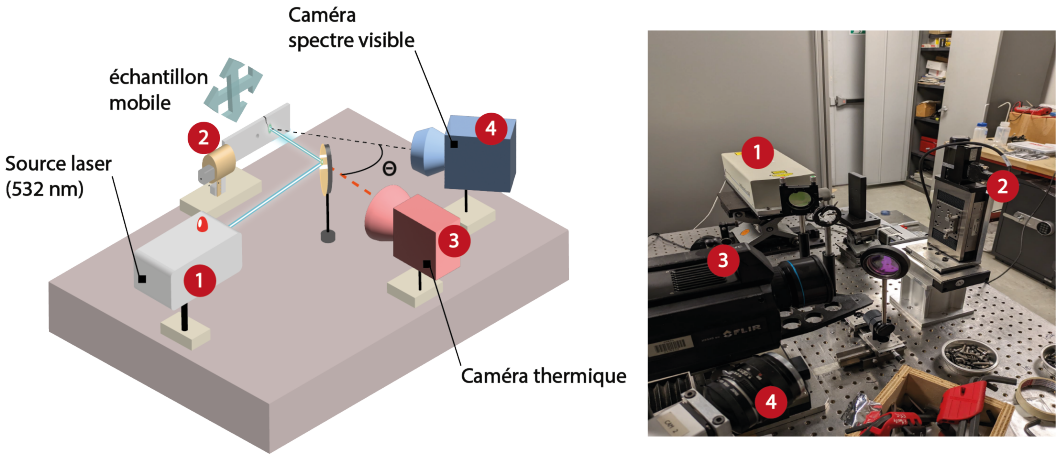


FIGURE 6.1 – Banc d'essai FST+visible pour acquisitions synchrones.

6.1.2 Base de données sur échantillons simples

Les trois échantillons constituant cette base sont des échantillons métalliques simples obtenus en deuxième moitié de thèse, et les enregistrements de deux d'entre eux sont utilisés pour l'entraînement tandis que les balayages d'un autre échantillon sont conservés pour l'évaluation des performances. Les images enregistrées dans le spectre IR et visible ont d'abord été appariées mais non parfaitement recalées : nous avons seulement procédé à un fenêtrage et centrage des deux images autour du maximum d'intensité, c'est-à-dire sur la tâche du laser. Les images visibles sont découpées au format (640, 480), tandis que celles en IR sont découpées au format (180, 180). Environ 400 paires d'images sont utilisées pour l'entraînement des réseaux neuronaux, tandis que 400 autres sont conservées pour l'évaluation des performances. Les boîtes englobantes de vérité-terrain localisant le défaut sont annotées en infrarouge, considéré comme la modalité de référence, tandis que le spectre visible est considéré comme une source d'information complémentaire pour les modèles neuronaux multi-spectraux. Cette modalité peut, par exemple, nous informer sur l'état de surface et sur les contours géométriques de la pièce examinée. Les boîtes englobantes sont définies en utilisant un protocole similaire au processus de marquage semi-automatisé mono-spectral avec correction humaine.

La Figure 6.2 illustre une paire d'images, les deux images étant centrées et découpées autour de la source laser : en visible et en IR respectivement. Cette paire d'images est enregistrée au début d'un balayage thermique et met en évidence l'information complémentaire entre FST et visible. Le défaut y est largement discernable en modalité visible sur ces enregistrements, donnant une première estimation préliminaire de sa longueur, tandis que le flux de chaleur n'a pas encore atteint la fissure cible : c'est un régime de

détection passif (pré-détection de la fissuration et de son amorce) qui peut être plus favorable au visible, comme évoqué dans la section 1.3.2. La Figure 6.3 donne un exemple similaire de paire, mais sans défaut. La largeur de fissuration est assez importante sur cette base de données et la détection du défaut en modalité visible n'est que peu affectée par le bruit de la caméra.

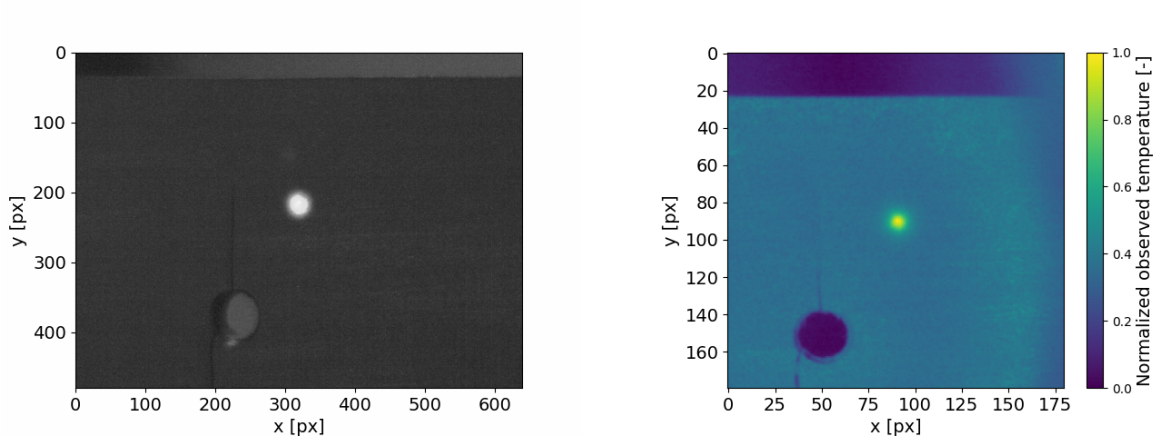


FIGURE 6.2 – Exemple de paire visible (gauche) et MW-IR (droite), sur une acquisition présentant un défaut. Le balayage est effectué relativement loin de la fissure et l'image est acquise en début d'échauffement.

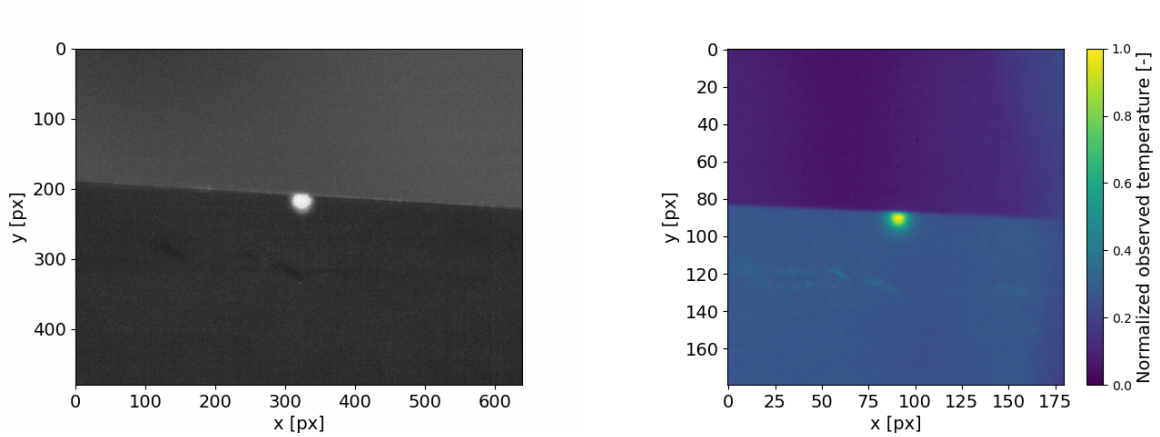


FIGURE 6.3 – Exemple de paire visible (gauche) et MW-IR (droite), sur une acquisition saine.

6.1.3 Base de données sur échantillons complexes

La base DAWN-MS a été constituée à partir des échantillons métalliques complexes revêtus présentés au chapitre 1 ainsi que par des pièces supplémentaires obtenues en fin de thèse afin d'augmenter la diversité des fissures détectées. Nous pouvons notamment relever la présence d'un endommagement de fissuration avec délaminage partiel de la barrière thermique. La puissance de balayage fut limitée à 0.5 W afin de réduire la saturation des images produites sur certains des nouveaux échantillons présentant un revêtement

fortement dégradé. Les balayages laser de 2,5 mm, parallèles au défaut, sont réalisés à une vitesse de 0.5 mm/s, avec une taille de spot large de 1.5 mm. La région des balayages est d'ordre centimétrique, de part et d'autre de la fissure. Nous avons ainsi une base présentant les deux régimes de données comme discuté dans la section 1.3.2 : un mode de pré-détection à longue distance de la fissure où les informations passives liées à la fissure sont prédominantes pour la détection (repérage de l'amorçage de la fissuration), et le mode actif où la discontinuité thermique liée à la fissure est importante (mode de caractérisation en longueur de fissuration). Le défaut est beaucoup moins ouvert sur cette base de données et le bruit caméra affecte de manière importante la détection en modalité visible.

À partir des enregistrements réalisés, nous avons constitué une base de données contenant environ 1000 paires pour le jeu d'entraînement et 200 pour l'évaluation des métriques de localisation. Nous avons supprimé quelques paires mal appariées ou mal acquises et effectué une sélection aléatoire des paires restantes, afin d'augmenter la difficulté de l'apprentissage.

6.1.4 Estimation d'homographie et recalage FST-visible

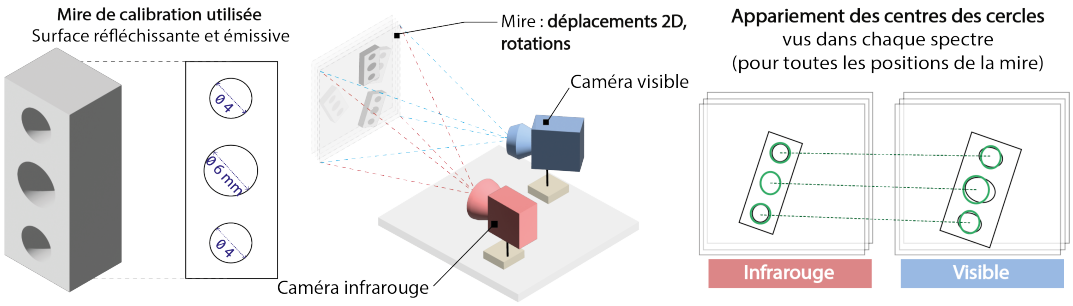


FIGURE 6.4 – Concept d'appariement proposé, utilisant la mire visible dans les deux modalités et la détection de cercles.

Un recalage d'images a été réalisé à partir d'une mire consistant en une pièce percée assez émissive, bien visible à la fois en spectre infrarouge et en spectre visible lors de l'acquisition passive. Cela nous permet de ramener les images visibles dans le repère de la caméra infrarouge. La mire sélectionnée présente des structures caractéristiques facilement identifiables sur chaque spectre (des cercles, au nombre de trois) : la position de chaque cercle est identifiée sur chaque modalité, pour chaque paire, donnant les points pour le recalage. Nous multiplions le nombre de paires jusqu'à approximativement une vingtaine de couples, afin d'avoir suffisamment de points et de redondances en cas de mauvaise identification et/ou de non détection par l'algorithme de détection de cercle. La Figure 6.4 fournit une illustration conceptuelle de notre proposition d'appariement des deux modalités.

La Figure 6.5 récapitule l'algorithme de recalage proposé pour la mire, avec les images visibles et infrarouges, le calcul de l'homographie permettant de passer du repère visible vers l'infrarouge, ainsi que les images de la mire après recalage. Nous distinguons déjà des phénomènes marginaux d'erreur de recalage sur les images visibles redimensionnées, qui sont sans doute dus aux imperfections de l'estimation du recalage ainsi qu'à la non-planéité

de la mire utilisée. Le recalage utilisé ici est obtenu à partir de 20 paires infrarouge-visibles avec une variation aléatoire de l'orientation et de la position de la mire.

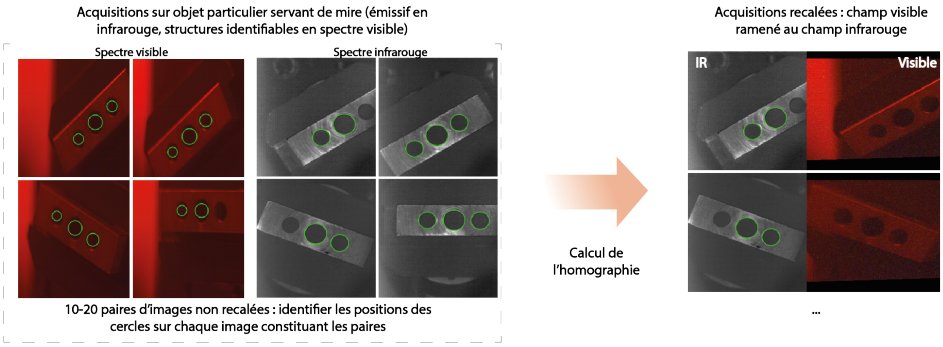


FIGURE 6.5 – Protocole de recalage multi-spectre infrarouge visible. Les acquisitions sont réalisées de manière passive avec une mire visible dans les deux spectres.

Ce procédé de recalage a été appliqué à la base DAWN-MS et nous permet de comparer la performance obtenue avec et sans recalage dans le contexte du CND. Les Figures 6.6 et 6.7 montrent des exemples de triplets d'image FST puis d'images visibles sans et avec recalage sur les échantillons métalliques complexes (cas avec et sans défaut). Si les déformations sont finalement assez modérées autour de la fissure, correspondant à une région assez plane, quelques-unes peuvent être distinguées à l'œil, en particulier sur les régions non planes. Nous pouvons aussi noter sur les exemples fournis plusieurs points : d'abord, le cas avec défaut est un cas où l'objet est à peine perceptible en modalité visible, favorable à l'infrarouge. De plus, le visible apporte des informations de contours externes et de géométries non observables dans la modalité infrarouge.

Cette base de données nous sert donc à comparer les performances de localisation de modèles de fusion avec recalage ou avec recalage imparfait. De plus, le jeu d'images recalées est utilisé afin de comparer les performances architecturales en mono et en multi-spectres, par facilité en termes de travail d'annotation de la position de la fissure.

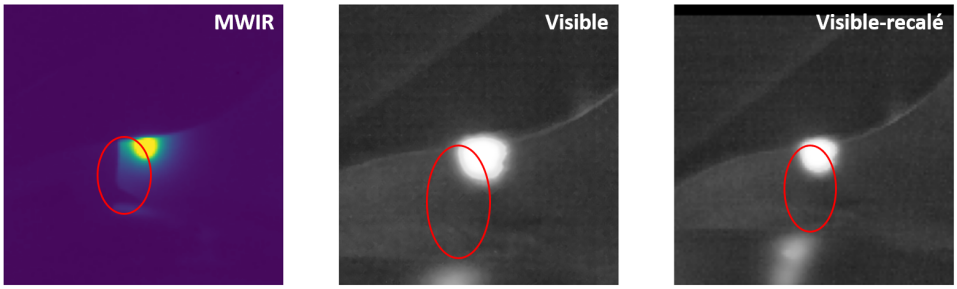


FIGURE 6.6 – Exemple d'acquisition avec fissure : image thermique individuelle - visible imparfaitement recalé (centrage+découpe) et visible recalé. Le défaut entouré en rouge est ici à peine perceptible en spectre visible.

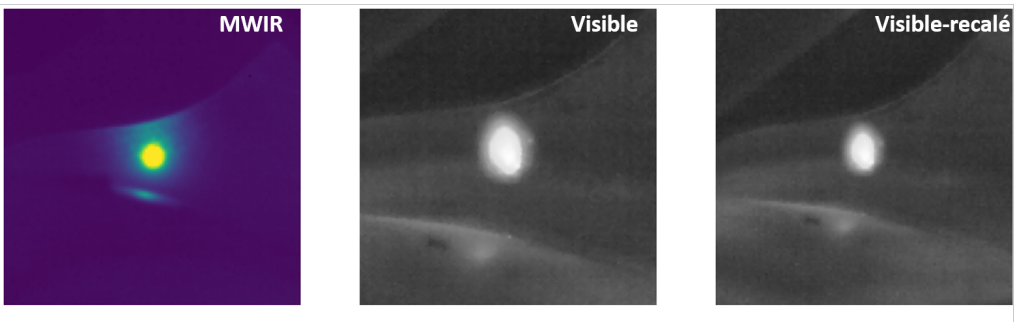


FIGURE 6.7 – Exemple d'acquisition sans fissure : image thermique individuelle - visible imparfaitement recalé (centrage+découpe) et visible recalé.

6.2 Synthèse de paires pour l'augmentation de données multi-spectrales

Cette section présente notre protocole de synthèse de paires d'images FST-visible par diffusion pour l'augmentation de données vu au Chapitre 2. Nous allons à nouveau recruter des modèles type Stable Diffusion couplés aux méthodes d'ajustement de modèle *DreamBooth* [139] et guidant ensuite la synthèse d'images grâce à l'approche Control-Net [140]. Les détails concernant les modèles de diffusion guidés par le texte comme Stable Diffusion sont fournis au Chapitre 2.

La section 6.2.1 fournit une base bibliographique sur la synthèse bi-spectrale, notamment sous l'angle du transfert d'une modalité à l'autre. La section 6.2.2 présente la méthode que nous proposons, basée sur les modèles de diffusion et un guidage par carte de contours. Enfin, les sections 6.2.3 fournissent les résultats de synthèse de paires pour les acquisitions sur les échantillons simples de FLYD-II.

6.2.1 Éléments de bibliographie sur le transfert de domaine

La synthèse d'images mono-spectrales a déjà été abordée au chapitre 2. Néanmoins, la synthèse de paires d'images multi-spectrales est un problème d'un autre ordre, nécessitant que le modèle de synthèse ait assimilé les propriétés de chaque modalité mais aussi le passage d'une modalité à l'autre. Pour le passage du spectre visible à l'infrarouge, nous pouvons notamment penser à un piéton et à la façon dont ses caractéristiques visibles vont être traduites en infrarouge : des phénomènes comme la défocalisation autour de cibles observées comme des piétons peuvent par exemple induire un effet d'ébavurage léger autour des objet.

Le transfert de domaine en apprentissage machine vise à adapter un modèle d'un domaine source à un domaine cible ayant des distributions de données différentes. Parmi les architectures notables, CycleGAN [33] a démontré la capacité de passer des images d'un domaine artistique, par exemple, à un autre sans correspondances directes, utilisant des générateurs et des discriminateurs pour apprendre les caractéristiques et correspondances entre chaque modalité de manière non supervisée. De plus, UNIT, pour *Unified Image-to-Image Translation*, [141] a proposé une approche basée sur des autoencodeurs variationnels pour le transfert de domaine unifié. D'autres modèles tels que MUNIT (pour

Multimodal Unsupervised Image-to-Image Translation) [142] et Star-GAN [143] ont étendu ces concepts pour gérer des conversions multi-domaines visuelles et multimodales sonores, respectivement. Toutes ces approches atteignent un certain degré d'efficacité dans la bonne traduction des caractéristiques d'un spectre à l'autre, néanmoins elles restent généralement assez imparfaites sur le plan du photoréalisme : le diable erre dans des détails comme les contours d'objets ou l'éclairage des scènes, imparfaitement reproduits.

Plusieurs travaux récents ont été proposés dans le domaine de la vision par ordinateur, utilisant des modèles de diffusion pour le transfert de domaine avec un modèle spécifique à deux branches bruit-vers-image [144], ou bien employant directement des modèles texte-vers-image pour traduire une image dans une autre modalité sans guidage spécifique [145]. Les approches réemployant les grands modèles de diffusion texte-image ont l'avantage de limiter le travail de composition architecturale et d'entraînement tout en atteignant des performances état de l'art sur des métriques comme le FID.

6.2.2 Méthode Control-Net

Control-Net [140] est une architecture neuronale permettant de conditionner à plusieurs niveaux d'abstraction la génération d'images. Elle peut être vue comme un modèle se greffant sur le modèle de diffusion et guidant la génération par l'injection d'une information résiduelle tirée d'informations a priori d'entrée sous la forme d'une carte 2D d'entrée telle qu'une carte de profondeur, de contours ou bien une carte de segmentation. Le modèle Control-Net utilisé est tiré du dépôt suivant : https://github.com/lllyasviel/ControlNet.

L'approche Control-Net permet des générations très cohérentes et assez bien généralisables à de nouveaux concepts sans ré-entraînement, dans de nombreux cas de figure, tant côté photoréalisme que génération artistique [140]. La Figure 6.8 fournit un schéma de principe permettant d'illustrer ce couplage entre Control-Net et un modèle type Stable Diffusion.

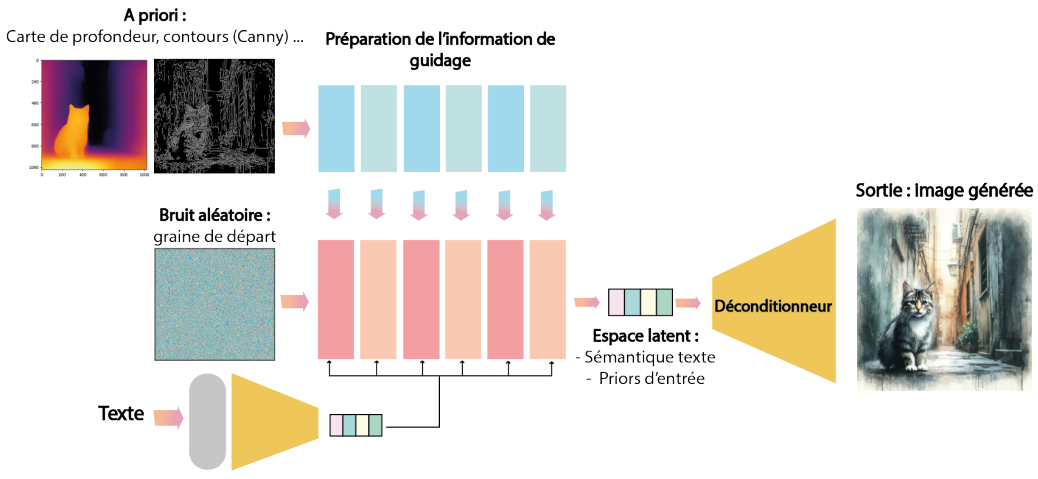


FIGURE 6.8 – Schéma de principe de l'approche Control-Net [140].

Cette approche peut nous permettre de moduler la synthèse d'images dans un spectre B après avoir synthétisé une première image dans une modalité de départ A, à partir des propriétés considérées comme clés dans les deux images, comme les contours de la pièce mécanique observée.

6.2.3 Génération de paires à partir d'échantillons-éprouvette

Nous proposons un protocole de génération de paires infrarouge-visible cohérentes pour des images thermiques individuelles synthétiques. Cette approche repose sur l'utilisation combinée de deux modèles de diffusion entraînés séparément. Un premier modèle de diffusion génère la base d'images visibles. On extrait ensuite par filtre de Canny une carte de contours [146]. Cette carte est utilisée pour guider le modèle de diffusion générant la base d'images infrarouges cohérentes avec les images visibles synthétiques appariées. Le Control-Net utilisé pour le guidage sémantique est dans cette étude un modèle déjà pré-entraîné et disponible en ligne via la librairie *Hugging Face - Diffusers* [57], utilisé tel quel sans ré-apprentissage. Un tri manuel permet dans un deuxième temps de nettoyer la base de données synthétiques en élaguant les images mal appariées ou les artefacts de génération, qui restent marginaux. Ce processus de génération est illustré avec la Figure 6.9.

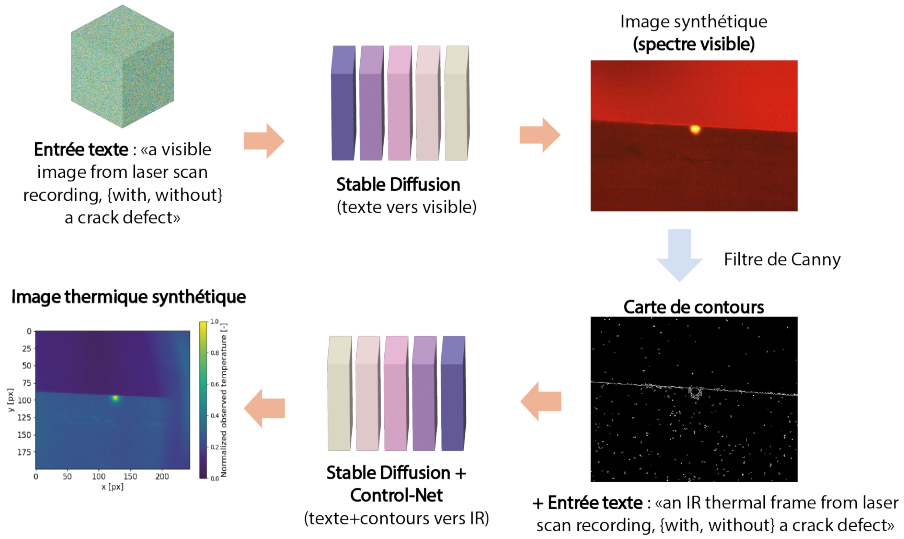


FIGURE 6.9 – Schéma de principe sur la méthode Control-Net.

À partir de 400 paires d'images FST-visibles partiellement recalées sur éprouvettes constituant le sous-échantillon d'entraînement de FLYD-II, cette approche nous a permis de produire approximativement 2000 paires infrarouge-visible cohérentes après nettoyage des paires incohérentes, correspondant à approximativement moins de 5 % du total produit. Quelques exemples de génération sont fournis dans la Figure 6.10.

Ces bases de données sont utilisées pour pré-entraîner nos modèles de localisation multi-spectrale sur les bases de données expérimentales que nous avons constituées, en vue d'augmenter les performances de détection de fissures.

6.3 Validations expérimentales

Dans cette section, nous évaluons les performances de détection FST-visible obtenues sur les différentes bases de données constituées. Notre approche CAFF-DINO est ainsi comparée à l'architecture CFT-YOLO-v5 [117], qui présente quelques similitudes de

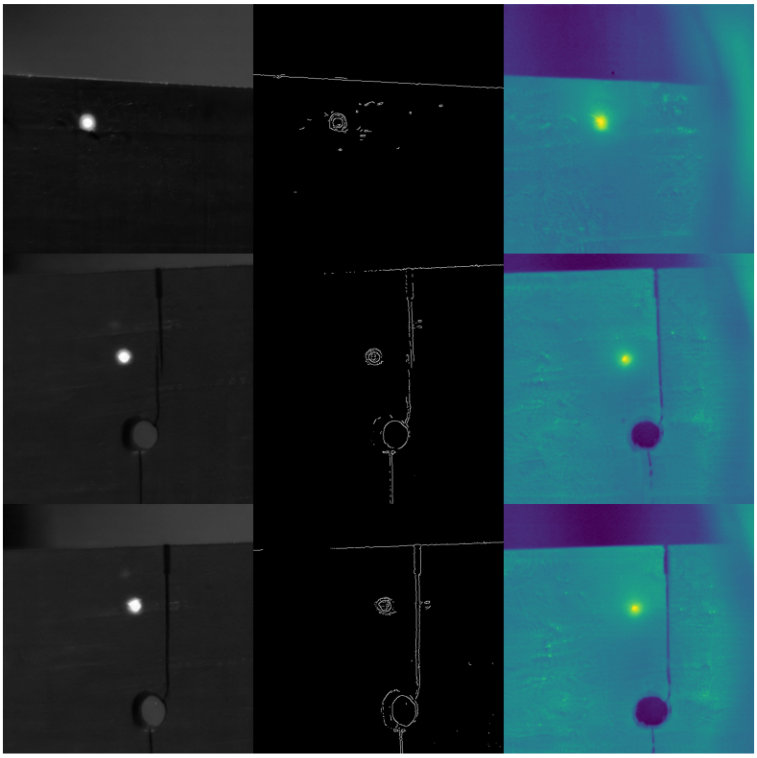


FIGURE 6.10 – Exemples de paires d'images synthétiques générées par notre protocole pour la base FLYD-II. A gauche, on trouve l'image visible (niveaux de gris), au centre la carte de contours (filtre de Canny), puis enfin l'image FST synthétique tout à droite.

principe architecturaux, notamment dans l'utilisation du mécanisme d'attention. C'est la *mean Average Precision* (abréviée en mAP) et ses variations qui sont ici mesurées pour quantifier la localisation de la fissure. Cette métrique est présentée au Chapitre 4, section 4.3.1. Sur chaque base, nous nous intéressons à l'influence de l'apport en données synthétiques sur la performance de localisation.

L'entraînement de l'architecture CAFF-DINO est réalisé durant 12 époques, tandis que l'entraînement de CFT-YOLOv5 [117] dure 100 époques. Tous les autres hyper-paramètres sont réglés selon les valeurs par défaut des implémentations d'origine de ces modèles [117].

6.3.1 Localisation sur les données partiellement recalées

Sur la base FLYD-II, la Table 6.1 compare les performances dans la tâche de localisation des fissures obtenues avec les deux modèles évalués ici. Une augmentation importante de la mAP allant jusqu'à 11 % (8 % pour mAP50) est obtenue pour l'architecture de la littérature, indiquant une capacité améliorée à détecter et localiser les défauts grâce à l'apport en données synthétiques. Il faut noter qu'un sur-apprentissage a été observé

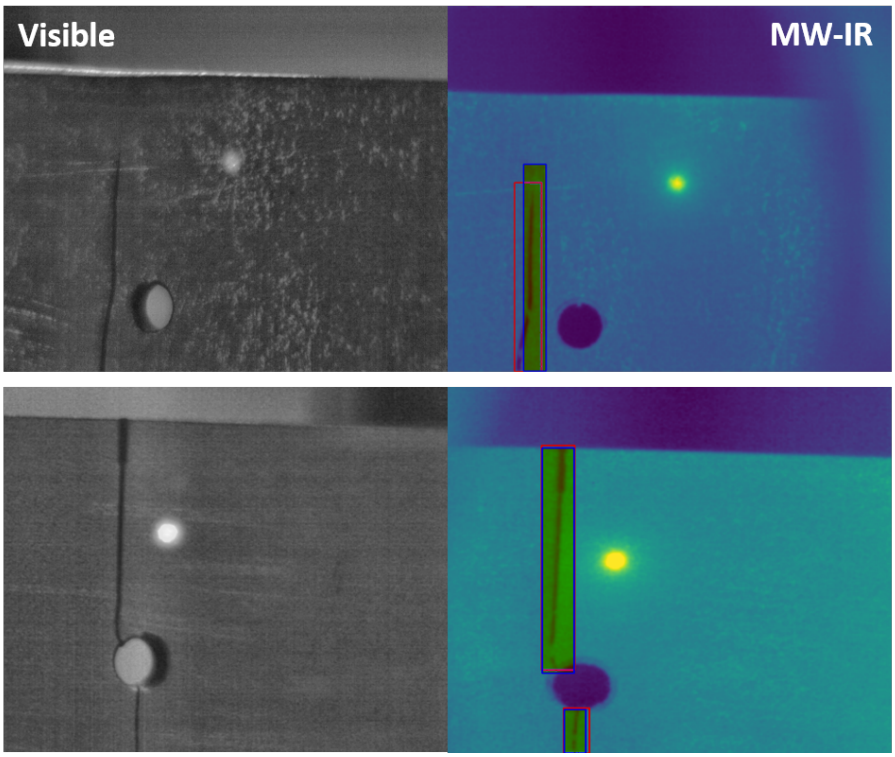


FIGURE 6.11 – Exemple de détection de fissures réalisée avec CAFF-DINO sur deux paires de la base FLYD-II. La vérité terrain correspond à la boîte rouge, tandis que la détection fournie par le modèle est en bleu.

avec un entraînement direct, impliquant un effondrement de la mAP pendant la phase d'apprentissage. Ce phénomène semble évité avec un pré-entraînement sur des paires de données synthétiques : la mAP converge vers les meilleures performances au cours de la phase d'apprentissage, indiquant comment la procédure proposée atténue le manque de données expérimentales. Ce bénéfice est également observé avec notre architecture, qui présente les gains de performances les plus importants. Notre modèle montre de meilleures performances que l'architecture de la littérature, avec un différentiel de mAP de 15 % en faveur de notre modèle, après un pré-apprentissage sur des données synthétiques. La Figure 6.11 fournit un exemple qualitatif de détection avec notre architecture. La Figure 6.12 montre un exemple de détection par CAFF-DINO sur une paire synthétique générée à partir de notre procédure de génération d'images.

Entraînement	Modèle	mAP50 (↑)	mAP75 (↑)	mAP (↑)
Sans pré-entraînement sur données synthétiques	CFT YOLO-v5	0.76	0.38	0.34
	CAFF-DINO	0.51	0.01	0.15
Avec pré-entraînement sur données synthétiques	CFT YOLO-v5	0.84	0.36	0.45
	CAFF-DINO	**0.98**	**0.69**	**0.60**

TABLE 6.1 – Comparaison des performances de localisation sur la base de données FLYD-II (mAP50, mAP75, mAP) sans pré-entraînement sur des données synthétiques et avec entraînement sur des données synthétiques.

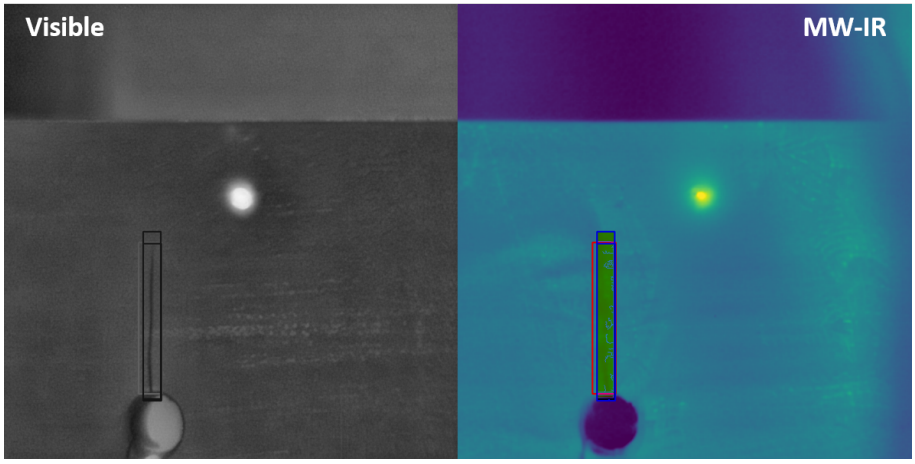


FIGURE 6.12 – Exemple de détection de fissures réalisée avec CAFF-DINO sur une paire synthétique générée à partir de la base FLYD-II. La vérité-terrain correspond à la boîte rouge tandis que la détection fournie par le modèle est en bleu.

Sur la base DAWN-MS, nous constatons les mêmes gains de performance que pour la base de données précédente, comme montré dans la Table 6.2 regroupant les scores de localisation avant et après ingestion de données synthétiques. Ici, le volume de données artificielles joue encore un rôle de palliatif au manque de données expérimentales, néanmoins les gains obtenus sont bien plus réduits, avec une augmentation de 3 % de la mAP75 pour l'architecture CAFF-DINO. Notre architecture a des performances supérieures ici aussi face au modèle alternatif, avec un différentiel de mAP50 de 34 % en faveur de notre modèle. Le modèle de la littérature fournit des performances faibles comparées à notre modèle, indiquant que l'architecture CFT-YOLO v5 aurait plus de mal à s'adapter aux écarts et distorsions présents entre les deux spectres.

Entraînement	Modèle	mAP50 (↑)	mAP75 (↑)	mAP (↑)
Sans pré-entraînement sur données synthétiques	CFT YOLO-v5	0.55	0.01	0.13
	CAFF-DINO	0.93	0.41	0.46
Avec pré-entraînement sur données synthétiques	CFT YOLO-v5	0.58	0.01	0.17
	CAFF-DINO	**0.92**	**0.44**	**0.46**

TABLE 6.2 – Comparaison des performances de localisation sur la base de données DAWN-MS (mAP50, mAP75, mAP) sans pré-entraînement sur des données synthétiques et avec entraînement sur des données synthétiques.

Pour conclure, les deux expériences mettent en évidence l'avantage de l'utilisation d'un pré-entraînement sur des données synthétiques pour les performances de détection multispectrale de fissures.

6.3.2 Performances sur les données recalées

La Table 6.3 regroupe les évaluations de performance de localisation de fissures sur la version recalée de la base DAWN-MS. Nous constatons un gain de performance marginal comparativement à l'image non recalée fournie en Table 6.2, pour CAFF-DINO.

Entraînement	Modèle	mAP50 (↑)	mAP75 (↑)	mAP (↑)
Sans pré-entraînement sur données synthétiques	CFT YOLO-v5	0.72	0.22	0.32
	CAFF-DINO	0.91	0.30	0.42
Avec pré-entraînement sur données synthétiques	CFT YOLO-v5	0.73	0.37	0.38
	CAFF-DINO	**0.94**	**0.46**	**0.49**

TABLE 6.3 – Comparaison des performances de localisation sur la base de données DAWN-MS recalée (mAP50, mAP75, mAP) sans pré-entraînement sur des données synthétiques et avec entraînement sur des données synthétiques.

Néanmoins, l'architecture CFT-YOLO v5 présente un net différentiel de performance en faveur de la base recalée, indiquant une moindre capacité à s'ajuster à l'absence de recalage propre.

Nous constatons aussi, comme précédemment, un bénéfice net dans les scores de localisation de tous les modèles après un pré-entraînement sur données synthétiques. La Figure 6.13 fournit un exemple de détection sur une paire recalée de la base DAWN-MS. Nous pouvons voir que le réseau arrive sur cet exemple à utiliser l'information infrarouge seule pour détecter la fissure, qui est pratiquement invisible en spectre visible du fait du bruit important dans ce spectre.

6.4 Étude des bénéfices propres à la fusion d'informations

La Table 6.4 fournit des scores de détection et de localisation de fissures mono-spectrales pour les architectures primaires mono-modales, avec nos scores après apprentissage sur synthétiques pour comparaison. Sur la base de données recalée, la fusion des informations infrarouge et visible permet un gain en qualité de détection comparativement au mono-spectral IR ou visible pour l'architecture CAFF-DINO. Nous mesurons ainsi un gain de mAP de 5 % entre notre modèle de fusion et l'architecture DETR-DINO native en infrarouge seul. Les données de la base DAWN-MS sont assez bruitées en spectre visible. Cela pourrait favoriser dans ce cas la détection en spectre infrarouge de manière artificielle, bien que notre architecture montre tout de même de bonnes capacités à détecter la fissure sur cette modalité.

Les gains de qualité de détection précise (mAP) par rapport à la détection en visible seule sont très clairs pour notre architecture. Un gain dans la capacité de détection relativement modéré par rapport à l'infrarouge seul est aussi relevé. Il y a peu de cas dans la base actuelle qui soient défavorables à cette modalité : une explication pour l'augmentation de performance serait la réduction des hallucinations des modèles dues aux éléments géométriques comme les contours, qui sont mieux compris lors de l'apprentissage après fusion des deux spectres. Nous notons néanmoins que l'architecture CFT-YOLO v5 montre une baisse de performance par rapport à l'architecture YOLOv5 native. Le modèle ne semble pas parvenir à fusionner efficacement les informations de chaque modalité.

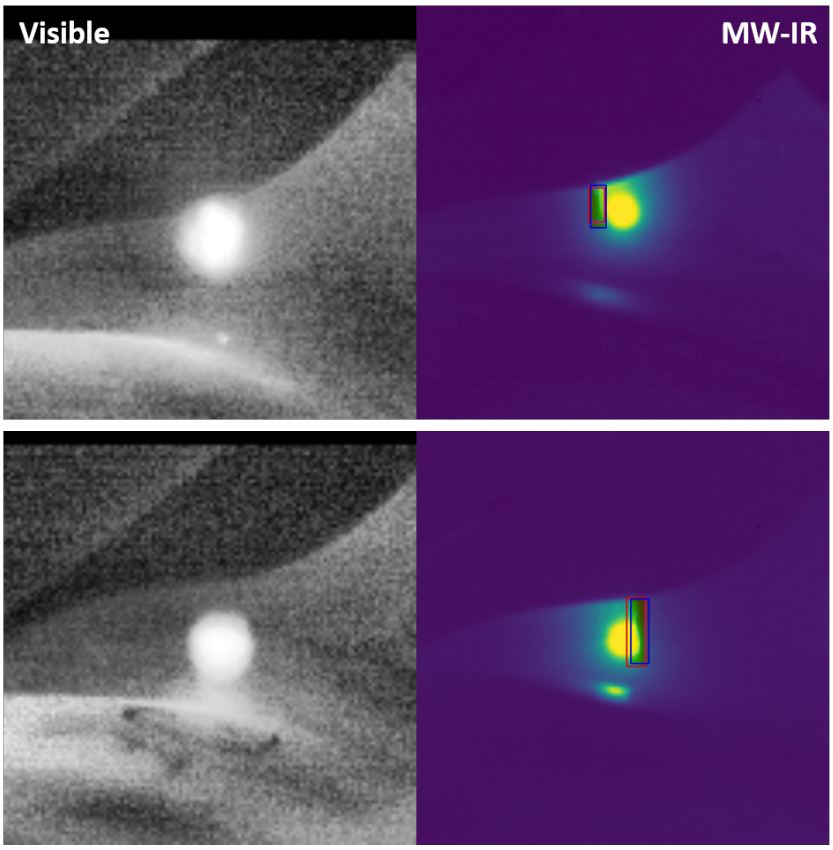


FIGURE 6.13 – Exemple de détection de fissures réalisée avec CAFF-DINO sur deux paires d'images recalées de la base DAWN-MS. La vérité terrain correspond à la boîte rouge, tandis que la détection fournie par le modèle est en bleu.

Spectre	Modèle	mAP50 (↑)	mAP75 (↑)	mAP (↑)
Visible	YOLO-v5 [87]	pas de convergence		
	DETR-DINO [128]	0.87	0.30	0.40
Infrarouge	YOLO-v5 [87]	0.88	-	0.42
	DETR-DINO [128]	0.88	0.39	0.44
Visible + Infrarouge	CFT-YOLO-v5 [117]	0.73	0.37	0.38
	CAFF-DINO	**0.94**	**0.46**	**0.49**

TABLE 6.4 – Performances mono-modales sur les données recalées de DAWN-MS. Les architectures mono-spectrales d'origine sont entraînées suivant le même protocole de pré-entraînement sur synthétiques que les modèles scores multi-spectraux.

6.5 Conclusion

Ce chapitre nous a donc permis de réaliser la fusion IR-visible avec la thermographie laser Flying Spot par apprentissage pour la détection de fissures sur matériaux métalliques. Nous avons pu comparer notre proposition d'architecture à des modèles de la littérature, validant les gains de performances comparées pour la localisation de fissures. Nous avons

ensuite vu que, si les données recalées fournissaient généralement de meilleures performances de détection, ce gain est assez marginal, montrant que le simple recalage partiel par centrage sur la source laser peut suffire dans ce contexte grâce à notre architecture de fusion, alors que le recalage peut entraîner des déformations importantes de l'image. Enfin, nous avons fait une comparaison entre les performances de localisation pour chaque modalité seule, et avons illustré le bénéfice de la fusion des deux modalités.

Les perspectives de poursuite de travail liées à ce chapitre sont d'abord la collecte de données supplémentaires afin d'étendre la base de fusion et d'augmenter la capacité de généralisation des modèles entraînés sur celle-ci. Si ce chapitre fournit une validation méthodologique du principe et des avantages de la fusion pour la localisation de fissures avec la FST couplée spectre visible, un changement du moyen d'imagerie optique est sans doute nécessaire vers une caméra plus récente, moins bruitée et pouvant travailler à des fréquences d'acquisition plus élevées. Le passage à une caméra visible neuve est ainsi prévu pour l'intégration sur le banc d'essai. Les modèles entraînés ici pourront être utilisés comme pré-annotateurs automatiques, comme nous l'avons fait au chapitre 4, section 4.4.1. C'est dans ce contexte qu'une base FLYD-II présentant à la fois des données partiellement recalées (centrage-découpage) et recalées à l'aide du protocole présenté dans ce chapitre devrait être mise à disposition de la communauté. Le travail présenté ici, une fois étendu, fera de plus l'objet d'un article de revue à terme.

Sur le plan de la synthèse de paires infrarouge-visibles cohérentes, l'approche proposée fonctionne assez bien quand le type de données d'entrée reste spécifique : soit les échantillons simples, soit les échantillons complexes à imiter. Nos essais préliminaires pour généraliser l'approche en réajustant nos modèles de diffusion en même temps sur les données des échantillons simples et complexes ne se sont pas révélés fructueux pour l'instant. Une solution envisageable pourrait être le réajustement du modèle Control-Net utilisé, ce qui affinerait le guidage de la génération dans chaque modalité. Le transfert de domaine par diffusion est un domaine encore en cours d'investigation dans la littérature, et la référence [147] fournit une autre approche envisageable pour diversifier la synthèse de paires bispectrales et augmenter la cohérence de la synthèse par rapport aux données expérimentales. La solution proposée est neuro-inspirée et utilise des modules ajoutés pour adapter un générateur spécialisé dans un spectre donné et gelé à de nouvelles modalités ou styles avec peu d'exemples d'apprentissage.

Nous allons maintenant dresser une conclusion de l'ensemble du travail de thèse mené, construisant le bilan de cette recherche et identifiant quelques perspectives de poursuite particulièrement du côté du traitement appliqué à l'imagerie thermique Flying Spot.

Conclusion et perspectives

Dans la longue histoire du tout, j'aurai toujours un faible pour le temps assez bref où les hommes auront vécu, dans l'angoisse et dans l'orgueil, sur cette planète reculée, perdue au fond de l'univers et qu'ils appelaient la Terre.

Presque rien sur presque tout,
J. D'Ormesson.

Dans les chapitres précédents, nous avons présenté le travail de thèse conduit au cours de ces trois années. Nous dressons d'abord une conclusion des travaux menés dans cette thèse et des contributions scientifiques produites. Puis, nous discutons des diverses perspectives possibles ouvertes par cette recherche pour d'autres travaux.

Bilan des recherches conduites

Nous avons élaboré des approches d'apprentissage profond pour l'automatisation de la détection de défauts par thermographie laser FST. Le travail mené a participé à l'exploitation de ces méthodes encore peu développées en thermographie laser.

Le premier enjeu de ce travail était de disposer de données suffisantes pour entraîner des réseaux de neurones de détection et de localisation de défauts. Dans le chapitre 1, nous avons ainsi pu collecter divers volumes de données de thermographie laser Flying Spot à partir d'un échantillon de minerai assez réduit en étendant la plage des paramètres de travail en dehors du point de fonctionnement théorique optimal. D'abord, des bases de données expérimentales et simulées ont été constituées sur des éprouvettes d'essais mécaniques de forme simple et homogène. Puis, d'autres bases ont été bâties après expérimentations sur des échantillons complexes revêtus plus difficiles à traiter. Afin d'augmenter artificiellement la quantité de données disponibles, nous avons déployé au Chapitre 2 des approches d'augmentation de données utilisant les modèles de diffusion par débruitage pour générer des images thermiques de synthèse. Ces méthodes ont été employées pour synthétiser des cartographies thermiques reconstruites ainsi que des images thermiques individuelles. Le mélange des propriétés des images telles que l'orientation ou la longueur de fissuration nous permet d'augmenter artificiellement à la fois la quantité et la diversité des bases de données disponibles.

Pour exploiter les différents types de données et la progressivité dans la complexité

entre images thermiques simulées, synthétiques et réelles, nous avons proposé au Chapitre 3 une approche d'apprentissage progressif pour apprendre à détecter des fissures de plus en plus complexes. Cette méthode part de scènes simples simulées par méthodes éléments-finis puis utilisant la synthèse d'images plus complexes et riches par modèles de diffusion, et enfin les données expérimentales réelles. Nous avons utilisé dans le Chapitre 4 une approche similaire par transfert des représentations depuis des surfaces génériques réelles vers des surfaces plus complexes, ce qui permettait encore une fois des gains de performances pour la tâche de localisation de l'endommagement. Ce chapitre montre aussi que l'utilisation de surfaces génériques synthétisées par diffusion pour pré-apprendre réhausse d'autant plus la capacité du modèle à localiser la fissure en FST. Nous avons ainsi élaboré des méthodes applicables à la détection et à la localisation de défauts en FST, permettant de supporter justement des régimes relativement pauvres en données, ré-exploitant des caractéristiques issues de données plus simples, collectées sur des échantillons disponibles ou bien des données synthétiques. Nous y avons aussi montré que nos méthodes étaient performantes pour localiser les défauts de type fissure tant pour des cartographies thermiques reconstruites de l'ensemble du balayage laser que pour des images thermiques individuelles prises au cours d'un balayage. La Figure 7.1 montre des exemples de détection de fissures avec une localisation par boîte englobante et une première segmentation du défaut par filtre de contours sur des images thermiques individuelles prises sur des éprouvettes et sur des échantillons complexes.

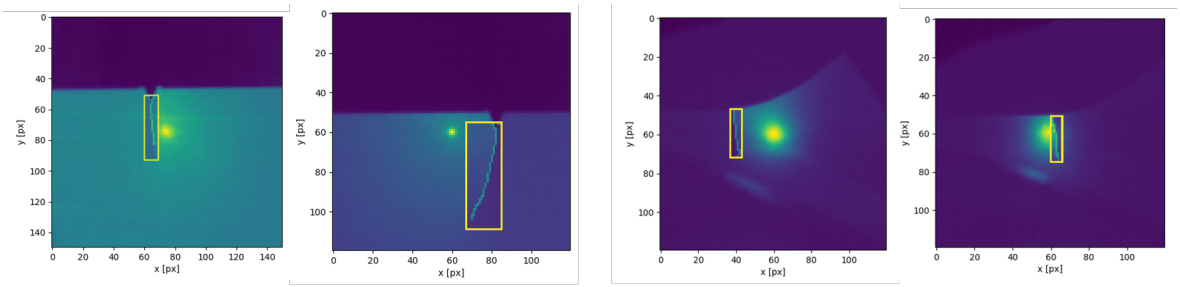


FIGURE 7.1 – Détection de fissures sur images thermiques individuelles, avec localisation et première segmentation du défaut par filtre de contours, sur éprouvettes à gauche et sur échantillons complexes à droite.

La troisième partie du travail de thèse s'est focalisée sur la fusion de données par apprentissage entre les spectres infrarouge et visible. Dans le Chapitre 5, nous avons présenté notre architecture de fusion multi-spectrale infrarouge-visible basée sur les *Transformers* de détection et le mécanisme d'attention croisée. Nous avons validé la pertinence de cette contribution méthodologique sur des bases de données publiques de fusion infrarouge passif et visible de référence en vision par ordinateur pour la détection multi-spectrale de piétons. Un exemple de détection multi-spectrale réalisée par notre modèle CAFF-DINO sur la base publique LLVIP est montré sur la Figure 7.2. Nous avons de plus montré la robustesse de cette approche à des décalages systématiques entre la modalité visible et la modalité infrarouge. Le Chapitre 6 a ensuite présenté la première collecte de données FST-visibles appariées avec deux méthodes de recalage différentes, ainsi qu'une méthode de génération d'images appariées suffisamment cohérente par modèles de diffusion. Nous avons ainsi pu entraîner notre réseau de fusion multi-spectrale pour la détection de fissures et démontrer le bénéfice de la fusion FST-visible par rapport aux spectres pris individuel-

lement sur les performances de localisation. Des exemples de détection multi-spectrale de fissures sont fournis par les Figures 7.3 et 7.4, respectivement sur échantillon simple et sur échantillon complexe.

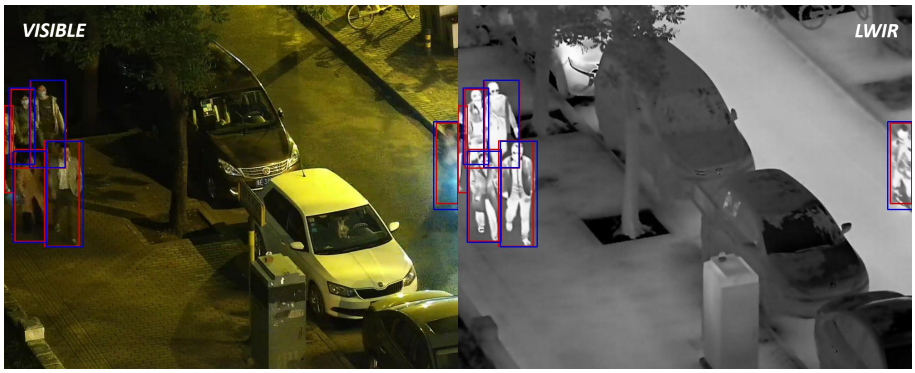


FIGURE 7.2 – Exemple de paire visible-infrarouge issue de la base LLVIP. La vérité-terrain est encadrée en rouge, les détections réalisées par CAFF-DINO sont en bleu.

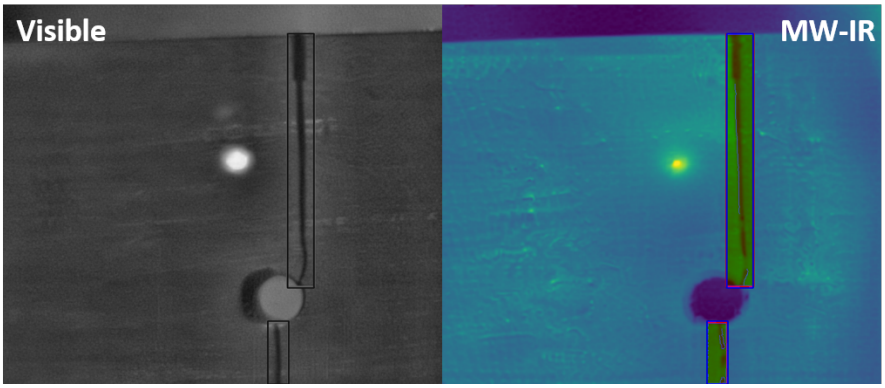


FIGURE 7.3 – Exemple de paire visible-infrarouge issue de la base FLYD-II. La vérité-terrain est encadrée en rouge, les détections réalisées par CAFF-DINO sont en bleu.

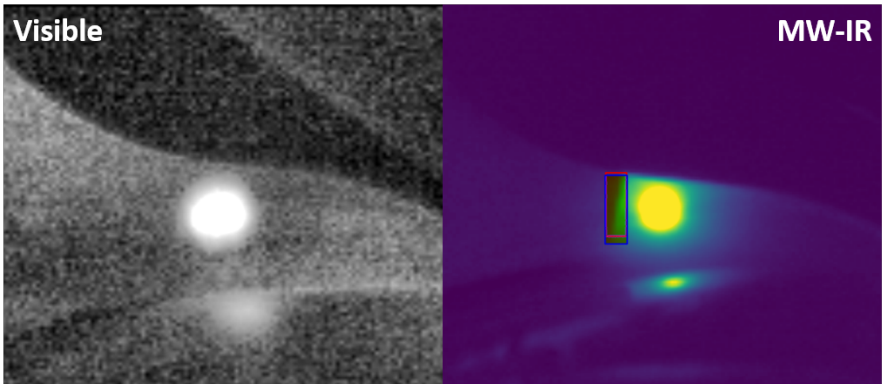


FIGURE 7.4 – Exemple de paire visible-infrarouge issue de la base DAWN-MS recalée. La vérité-terrain est encadrée en rouge, les détections réalisées par CAFF-DINO sont en bleu.

L'utilisation des modèles génératifs par diffusion pour la synthèse d'images thermiques et l'augmentation des volumes de données disponibles a été largement explorée et déployée durant ce travail de thèse, montrant l'efficacité de ces approches pour stabiliser et augmenter les performances d'apprentissage des méthodes basées sur la donnée en FST. L'apprentissage profond y a démontré sa capacité à discriminer efficacement la présence d'une fissuration en FST puis à localiser le défaut sur l'image d'examen, tant sur carte thermique reconstruite que sur image thermique individuelle isolée d'un film thermique.

Perspectives

Intégration au banc d'essai, caractérisation du défaut. Les perspectives à court-terme concernent l'intégration logicielle des modèles de détection développés, afin de faciliter leur utilisation du banc par des opérateurs non-formés. La caractérisation précise du défaut au-delà de la seule localisation puis segmentation par filtre de Canny en est une autre.

Augmenter la diversité des fissures et la quantité de données. Poursuivre l'extension des bases de données serait bénéfique pour l'application dans l'industrie : celle-ci permettrait de conforter les performances obtenues sur une diversité d'échantillons et de fissures en croissance. Malgré les efforts d'augmentation de performance menés, les méthodes par apprentissage restent sensibles à une confrontation avec des données d'entrée acquises dans des conditions trop différenciées par rapport à celles de la base d'entraînement. La solution la plus simple reste l'augmentation, au fil des expérimentations conduites à l'ONERA, de la diversité des échantillons d'entrée afin de consolider les systèmes d'apprentissage sur le long terme. L'utilisation de modèles de diffusion pour la synthèse d'images diversifiées s'est révélée très efficace ; néanmoins, la synthèse en elle-même présente un coût calculatoire non négligeable.

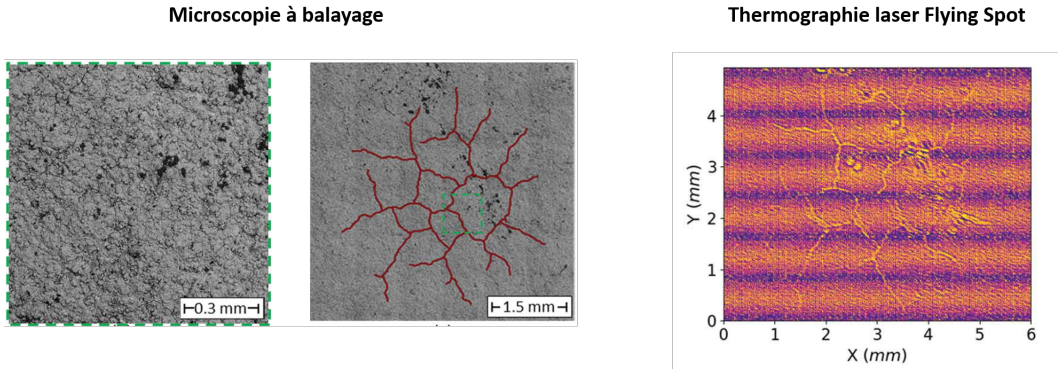


FIGURE 7.5 – Imagerie par microscope à balayage à gauche, et reconstruction cartographique Flying Spot à droite, d'un endommagement par micro-faïençage d'une matrice céramique composite. Tiré de [6].

Ensuite, il serait pertinent de considérer l'observation de défauts moins unitaires et compacts parfois étudiés en FST, tels que les réseaux de fissurations obtenus par faïençage ou par impact, typiques des surfaces céramiques fragiles. Un exemple de ce type d'endommagement est illustré sur la Figure 7.5 où l'on peut voir la capacité de la FST à mettre en

évidence ce type de défaut avec des moyens d'instrumentation moins lourds que la microscopie électronique et pouvant être employés *in situ*. L'image FST est ici une composition de balayages aller-retours horizontaux. Néanmoins, ces endommagements posent des défis du côté de l'acquisition expérimentale, avec l'utilisation d'objectifs microscopiques du fait de la finesse du réseau de fissuration, qui contraint également les paramètres d'expérimentation. Du côté des traitements, une problématique est la recherche d'un moyen de segmenter proprement ce réseau de fissures.

Se passer de synthèse d'images avec les modèles génératifs ? Utiliser des stratégies d'augmentation par diffusion s'est révélé très efficace durant ce travail de thèse. Néanmoins, leur coût calculatoire en entraînement et en génération d'images thermiques est une limitation que l'on ne peut pas nier, malgré l'approche par réajustement de modèles comme *Stable Diffusion* qui mitige partiellement ces points. Dans l'optique où les images synthétisées servent de *proxy* intermédiaire pour les caractéristiques formées par le modèle de diffusion, passer de la synthèse d'images au transfert des caractéristiques apprises par distillation de connaissances sans génération d'images [76,77] pourrait réduire le coût lors de l'entraînement de nos modèles. L'exploitation de l'espace latent formé par un processus de diffusion pour la détection d'objets sans entraîner un second modèle sur des données de synthèse ou bien distiller les représentations formées par le modèle génératif sont des approches complexes, bien que des pistes existent avec des analogies possibles dans la littérature de la détection d'anomalies [78–80].

Adjonction d'autres modalités. L'ajout d'autres bandes infrarouges que le MWIR pourrait aussi être pertinent suivant les matériaux des futurs échantillons. Une part des travaux antérieurs menés à l'ONERA sur les échantillons complexes était notamment focalisée sur la bande SWIR, nécessitant généralement des échauffements plus importants pour pouvoir observer la diffusion de chaleur [5]. Une perspective à plus long terme est la sélection autonome des longueurs de travail éliminant d'autant plus la part de travail opérateur sur l'ensemble du système d'inspection.

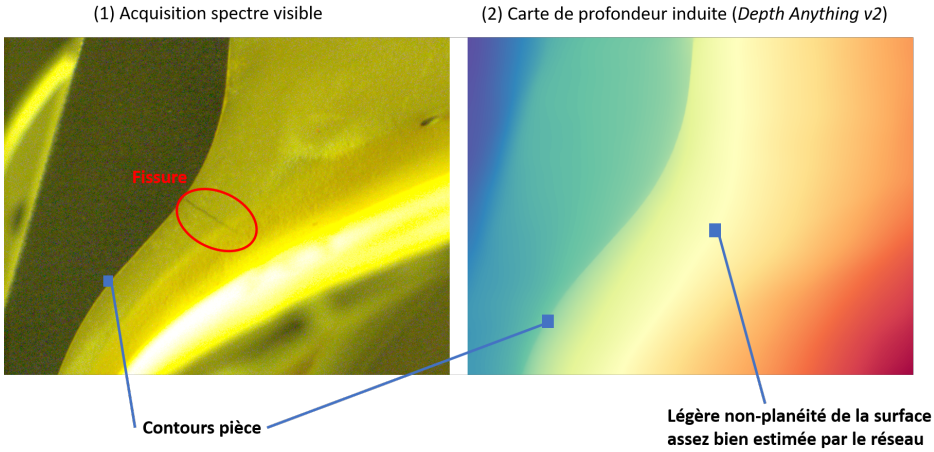


FIGURE 7.6 – À gauche, image visible haute définition d'un échantillon complexe, prise de vue centrée sur la fissuration. À droite, carte de profondeur déduite par le réseau *Depth Anything v2* [148,149] sans réapprentissage.

L'introduction d'informations complémentaires liées à la pièce étudiée pourrait être bénéfique pour limiter la quantité d'erreurs produites par le modèle et associées à des

incompréhensions sur la surface observée, comme les non-planéités ou les éléments géométriques de la surface. Nous pourrions ainsi ajouter au banc une caméra 3D ou bien tenter d'exploiter les méthodes récentes d'estimation 3D monoculaire. Cette dernière approche est illustrée sur la Figure 7.6 où le modèle de fondation *Depth Anything v2* [148, 149] arrive à assez bien déduire sans réapprentissage les aspects géométriques et 3D des pièces revêtues complexes à partir d'une seule image visible. Un réajustement de ce modèle pour la donnée FST à partir de cartographies de profondeur déduites en spectre visible monoculaire semble tout à fait à portée. *Marigold* est un modèle similaire qui pourrait servir lors d'une ablation des différentes 3D monoculaires apprises [150].

Modèles de fondation et thermographie. Nous pourrions envisager le développement d'un modèle de fondation pour la thermographie.

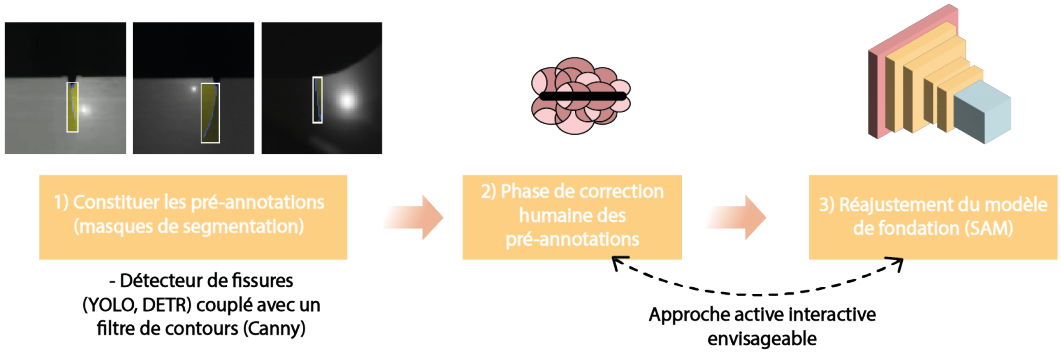


FIGURE 7.7 – Proposition de protocole de réajustement d'un modèle de fondation type SAM, à partir d'un pré-annotateur couplé à un filtre de contours.

Un modèle de fondation est un modèle massif entraîné pour former des caractéristiques versatiles afin d'être agnostique au problème final d'utilisation. Il peut alors être facilement réajusté sur de nouvelles tâches pauvres en données avec des performances convenables. Nous pourrions par exemple nous baser sur l'ajustement de modèles déjà existants comme le *Segment Anything Model* (SAM) [151]. Des exemples de réadaptation de ce modèle existent en détection d'anomalies industrielles optiques [152] ou pour l'examen médical [153]. La Figure 7.7 fournit une ébauche de protocole d'adaptation pour le modèle de fondation SAM avec un de nos détecteurs neuronaux combiné à un filtre de contours pour extraire des masques de segmentation de la fissure : ces masques serviraient dans une optique de réajustement du modèle de fondation. Une alternative au réajustement de SAM qui a été évoquée au Chapitre 4 est la construction d'une Mixture d'Experts [102], c'est-à-dire l'utilisation d'un mécanisme de sélection appelé *switch* afin de combiner plusieurs modèles spécialisés suivant le type de données d'entrée. Ces systèmes de mixtures sont bien plus complexes que les méthodes ensemblistes qui moyennent la réponse de plusieurs modèles spécialisés pour augmenter la qualité de la réponse de sortie, mais s'en rapprochent conceptuellement [98, 99]. Cela présenterait l'avantage de permettre un entraînement modulaire où chaque réseau-expert serait entraîné sur un domaine de données précis, maîtrisant alors en partie le coût en temps machine global en phase d'entraînement. Cette idée de Mixture est présentée dans une première ébauche de protocole Figure 7.8 et permettrait de rehausser encore les capacités de généralisation face à de nouveaux échantillons inconnus notamment. Elle permettrait aussi d'ajouter de nouveaux experts suivant les besoins, spécialisés dans de nouvelles familles d'échantillons

par exemple. Ici encore, le recours aux données synthétiques ou bien aux caractéristiques apprises associées permettrait d'entraîner des modèles plus massifs et plus généralisants.

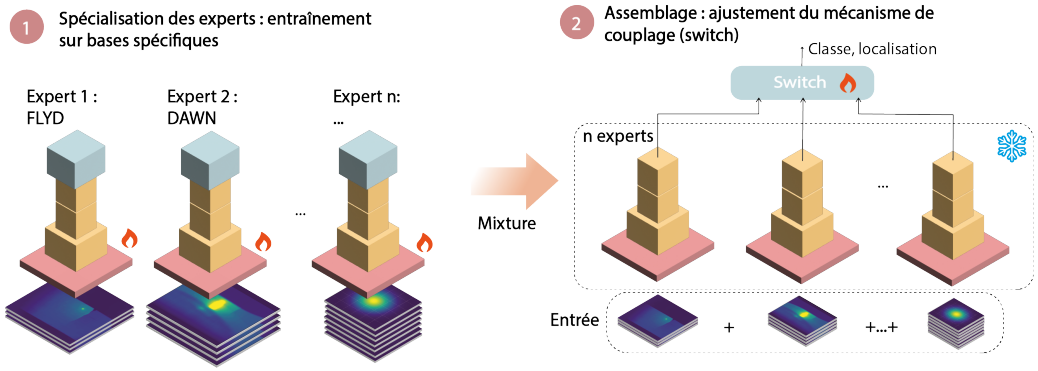


FIGURE 7.8 – Illustration simplifiée d'une première proposition de composition d'une Mixture d'Experts à partir des bases de données FST collectées au cours de la thèse. La flamme indique une structure neuronale entraînée et le flocon une structure à poids gelés. Le modèle est inspiré de l'architecture DAMEX [106].

Modèles informés par la thermique. L'essor dans la littérature des modèles informés par la physique ouvre des perspectives de réduction supplémentaire de la dépendance aux données, c'est-à-dire l'introduction d'à priori liés aux lois physiques de la diffusion thermique, par *design* ou par la fonction de coût d'apprentissage par exemple. Néanmoins, la question de la performance de détection de défauts brute se pose, ces approches étant encore récentes, ouvertes, et en cours d'investigation. Nous pourrions nous inspirer de ce qui se fait en mécanique des fluides, comme par exemple l'exploration de solutions architecturales élaborées sur des modèles de diffusion pour la modélisation des cartographies synthétiques d'écoulements turbulents [154].

D'un autre côté, l'émergence des modèles d'états neuronaux (en anglais *State-Space Models*) [155] comme alternative moins coûteuse aux *Transformers*, basés sur un dérivé des architectures récurrentes afin de calculer un état T dans une séquence de données, est un nouveau type d'architecture prometteur pour le traitement séquentiel comme pour le traitement d'images [156–158]. S'ils pourraient se révéler pertinents dans l'élaboration d'un traitement appris de la séquence vidéo thermique à coût calculatoire plus réduit par rapport aux *Transformers*, ils pourraient aussi être plus propices que d'autres modèles d'apprentissage machine pour l'incorporation par *design* des lois dérivées de la physique. Nous pourrions procéder de manière analogue aux équations différentielles neuronales (*Neural ODE* en anglais) qui consistent en la résolution d'équations différentielles par réseaux de neurones [159]. Mentionnons ici un travail récent combinant ces modèles avec les lois de diffusion pour le transport de chaleur et de matière [160].

Nous avons vu que les données synthétiques produites par diffusion permettaient d'augmenter les performances de modèles spécialisés, en introduisant des variations dans les propriétés des images réelles. Néanmoins, comme dit plus haut, il peut être plus économique et plus performant de se passer de la donnée synthétique comme *proxy* des apprentissages du modèle de diffusion et de tenter d'exploiter directement le modèle générateur pour la détection. Ainsi, nous pourrions transférer des caractéristiques apprises par le modèle génératif pour d'autres tâches comme la localisation par exemple en intégrant

une approche progressive : la loi physique peut être introduite lors d'un apprentissage sur données simulées par éléments finis, dans la fonction de coût, ce qui va contraindre cette première formation de caractéristiques latentes. Ensuite, un ajustement statistique pur sur des données réelles permettrait d'ajuster les caractéristiques apprises à des phénomènes plus spécifiques et moins accessibles par la modélisation physique comme les contours de pièce ou les artefacts. Cela fait l'objet d'une proposition de thèse à l'ONERA dans la continuité des travaux présentés ici. La Figure 7.9 fournit une première idée de construction d'architectures PINN inspirée de la détection d'anomalies, où un modèle génératif apprend par la synthèse des caractéristiques thermiques plus cohérentes avec la physique qui serait intégrée sous la forme d'une fonction de coût mixte. Ces caractéristiques formées seraient ensuite transférées pour construire un détecteur de défauts.

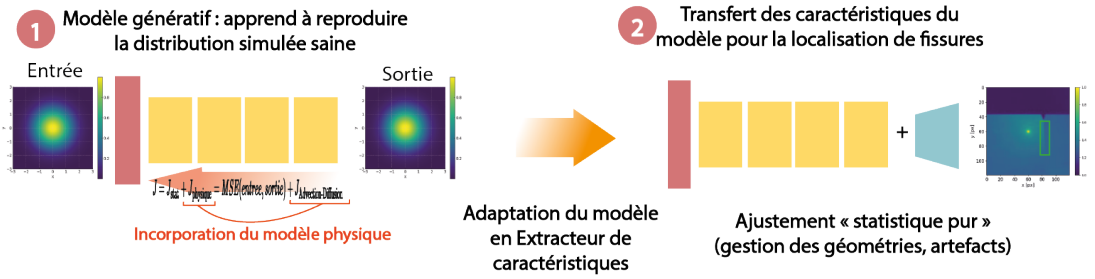


FIGURE 7.9 – Proposition de protocole d'apprentissage informé par la physique inspiré de la détection d'anomalies.

Explicabilité, modèles du langage et thermographie. Une autre perspective de recherche à moyen terme concerne l'utilisation de modèles vision+langage. Dans le cadre de la thermographie laser, cela pourrait reposer sur la construction d'un modèle argumentant ses choix qui s'inscrirait dans le contexte de l'IA eXplicable (XAI). Ce sujet fait l'objet d'une thèse à l'ONERA : ces arguments peuvent ensuite subir des objections de la part des opérateurs humains, ou permettre d'identifier des biais dans l'apprentissage en cas d'erreur. Ce type de modèle peut aussi être particulièrement utile pour rédiger des rapports de manière autonome, après examen et détection du défaut, permettant d'ouvrir la voie à une automatisation poussée des examens non destructifs dans leur ensemble. La question fondamentale de la performance de détection de ces approches, tant de manière généraliste que sur les données thermographiques, est stimulante, devant le vieux paradoxe de l'apprentissage machine, où performances brutes et interprétabilité sont souvent antinomiques.

Une limitation importante pour cette perspective de recherche est néanmoins le manque de rapports écrits riches dans le domaine du CND, qui présente très peu compte-rendu d'inspection rédigés, comparativement au milieu médical par exemple, où des descriptions d'examens textuels très riches sont disponibles en relativement bonne quantité, avec des bases de données texte+image compilant examens multi-senseurs et rapports d'examen pour la détection de cancers ou les stades de développement de la maladie d'Alzheimer [161, 162]. L'utilisation du langage pourrait étonnamment se révéler plus pertinente dans le cadre de la robotisation du moyen d'essai pour le pilotage autonome en temps réel ou pour la prise de décision des zones d'intérêt cibles. La robotique cognitive, qui consiste en un déploiement d'un modèle de langage pour la prise de décision du système robotique, semble devenir une alternative à l'apprentissage par renforcement qui s'adapterait

au moins aussi bien à un contexte avec peu/pas de données. Elle pourrait à terme présenter une déployabilité à partir, d'une part, d'un simple réajustement de modèle de langage et, d'autre part, de la mise en place de systèmes mémoire *Retrieved Augmented Memory* (RAG) [163]. Cela permettrait une adaptation autonome en ligne du système d'examen, avec une potentielle adaptation au cours du cycle de vie du moyen d'inspection. Nous pouvons penser ici à un changement de pièce examinée ou d'outillage.

L'utilisation de tels LLM est très récente et encore largement en perspective, ainsi les possibilités de développement et les performances de telles approches restent ouvertes malgré quelques travaux notables fournissant une première base d'investigation [164–166].

Prospectives fondamentales à long terme. Des perspectives de plus long terme sont aussi dégagées depuis ce travail de thèse et de la nature du signal thermique mesuré en FST, bien que plutôt concentrées vers la recherche fondamentale en apprentissage. D'une part, le champ thermique au cours d'un enregistrement de balayage pourrait être assez bien adapté au codage événementiel pour des réseaux de neurones à impulsion (SNN) [167], *modulo* une conversion appropriée du signal fourni par la caméra au cours d'un balayage. Cela ouvrirait à de la recherche architecturale centrée sur ces SNN, qui sont une famille d'architectures encore grandement en développement mais dont le volume de recherches produit connaît une forte croissance car orientées vers la frugalité en données d'apprentissage. Ces modèles d'apprentissage, bien plus proches du modèle du neurone issu des neurosciences computationnelles [113], peuvent être intéressants en particulier dans le cadre d'une implémentation embarquée, sur carte neuromorphique et frugale en données. Cette famille de modèles pourrait tout à fait servir au développement d'un modèle PINN, comme montré dans ce travail [168]. Il n'y a pour l'heure aucune recherche notable sur le développement de senseurs thermiques événementiels.

Une autre grande approche frugale ouverte est bien entendu la construction de modèles spécialisés par Recherche neuronale d'architectures (*Neural Architecture Search*, NAS) [169]. Cette famille d'approches consiste en la construction autonome de nouvelles architectures d'apprentissage pour résoudre un problème donné, au travers de méthodes de sélection génétiques ou basées sur des méta-modèles : l'architecture ainsi produite réduisant largement le nombre de paramètres du réseau de neurones ainsi généré à performance équivalente et donc leur besoin en données. Cette approche, longtemps restée confidentielle, a connu quelques succès tels que l'élaboration d'Efficient-Net [170, 171] qui est un modèle très robuste aux hallucinations, ou YOLO-NAS en localisation d'objets [172]. Si elles pourraient avoir un intérêt en soi pour la donnée FST mono-spectrale seule, c'est bien du côté de la construction de détecteurs multi-spectraux frugaux en données qu'une opportunité de recherche méthodologique existerait. Ainsi, pratiquement aucun travail de recherche n'explore l'application de méthodes NAS pour de l'information multi-modale à l'heure actuelle. Par exemple, les méthodes dérivées de *Differentiable Architecture Search* (DARTS) [173–176] ont démontré leur capacité à construire des architectures par sélection qui sont au niveau de l'état de l'art ou supérieures avec un coût calculatoire contrôlé. Un travail d'adaptation à la donnée multi-modale via les *Transformers*, qui sont absents de ces méthodes de sélection, pourrait se révéler pertinent.

Ce travail de thèse ouvre donc la voie à de multiples perspectives de poursuite d'études, tant sur le plan de l'application de la thermographie couplée à l'apprentissage machine que du développement méthodologique en apprentissage multi-modal.

Quatrième partie

Annexes

Annexe A

Apprentissage profond

A.1 Bases de l'apprentissage profond

Cette section fournit des notions fondamentales en apprentissage profond. Des ressources complémentaires en ligne peuvent être trouvées sur le github suivant [177]. De plus, les références [178, 179] fournissent des livres accessibles tout en étant richement détaillés sur les méthodes de construction d'architectures d'apprentissage profond.

Introduction et modèle du Perceptron. L'apprentissage profond est une sous-catégorie de l'apprentissage automatique qui utilise des réseaux de neurones artificiels avec plusieurs couches cachées afin de répondre à un problème. L'idée de profondeur vient du nombre de couches empilées parfois très grand, dans un réseau de neurones moderne des années 2020. L'objectif est de permettre aux machines d'apprendre à réaliser une tâche à partir de données d'entrée en extrayant de manière autonome les caractéristiques pertinentes face au problème à résoudre [180].

Les réseaux de neurones utilisés dans le cadre de l'apprentissage profond moderne suivent dans leurs principes le modèle du *Perceptron* de Rosenblatt [181, 182]. Ils sont constitués de neurones artificiels organisés en couches : une couche d'entrée, plusieurs couches cachées et une couche de sortie. Chaque neurone applique une fonction d'activation à une combinaison linéaire de ses entrées (un nœud ou synapse), appelés **paramètres**, et les couches cachées permettent de modéliser des relations non linéaires complexes entre une entrée et une réponse de sortie. Ce modèle du neurone est illustré par la Figure A.1.

Nous pouvons formaliser ce modèle du perceptron, aussi appelé couches denses ou complètement connecté, comme suit :

$$\mathbf{y} = f_{\text{activation}}(\mathbf{\Theta x} + \mathbf{b})$$

où $\mathbf{x}$ est le vecteur d'entrée de dimension $(n, 1)$. $\mathbf{\Theta}$ est la matrice des poids de dimension (m, n). $\mathbf{b}$ est le vecteur des biais de dimension $(m, 1)$. $\mathbf{y}$ est le vecteur de sortie de dimension $(m, 1)$. $f_{\text{activation}}$ est la fonction d'activation appliquée élément par élément, permettant de non-linéariser la réponse.

Cette expression peut être détaillée, ici dans le cas de vecteurs-colonnes x et y :

$$y_i = f_{\text{activation}}\left(\sum_{j=1}^{n} \Theta_{ij} x_j + b_i\right) \quad \text{pour } i \in \{1, 2, \ldots, m\}, \quad m \in \mathbb{N}^{+*}$$

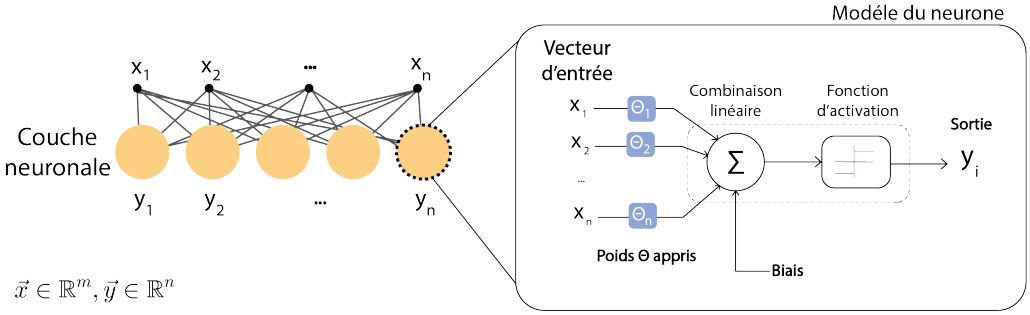


FIGURE A.1 – Illustration du modèle du neurone artificiel en apprentissage profond moderne basé sur le modèle du Perceptron, à partir de la couche neuronale : celle-ci est appelée couche dense ou complètement connectée.

où y_i est le i-ième élément du vecteur de sortie $\mathbf{y}$. Θ_{ij} est l'élément de la i-ième ligne et j-ième colonne de la matrice des poids $\mathbf{\Theta}$. x_j est le j-ième élément du vecteur d'entrée $\mathbf{x}$. b_i est le i-ième élément du vecteur des biais $\mathbf{b}$.

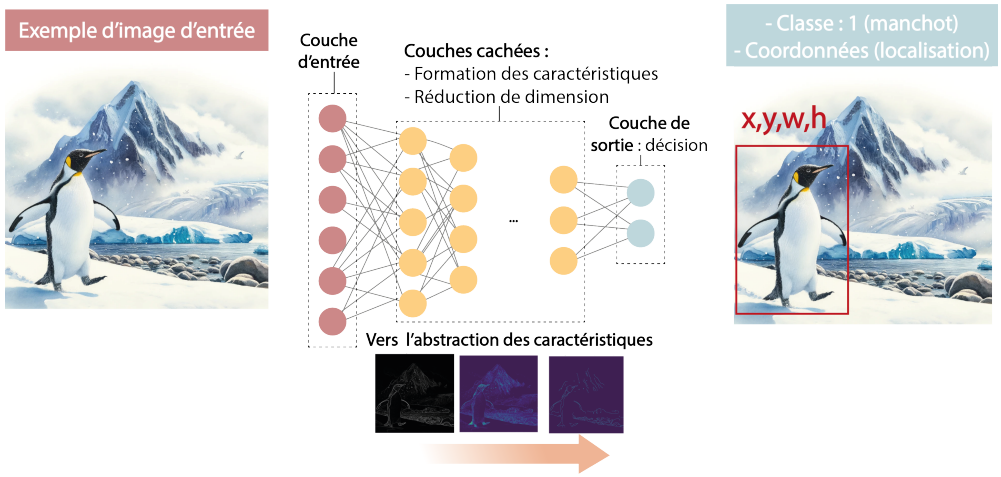


FIGURE A.2 – Illustration du fonctionnement d'un réseau de neurones profond moderne en vue de réaliser une tâche de vision.

Les fonctions d'activation jouent différents rôles suivant leur position dans le réseau de neurones, comme illustré dans la Figure A.3. Les fonction d'activation situées dans le réseau (couche d'entrée et couches cachées) servent généralement au réseau de neurones à former des opérations non-linéaires afin de mieux modéliser des phénomènes complexes associés aux données d'entrée [183]. La principale fonction d'activation interne au réseau est souvent la fonction RELU (*REctified Linear Unit*). Les fonctions d'activation placées sur les dernières couches neurales servent généralement à reprojeter ou à normaliser le vecteur de neurones de sortie (entre 0 et 1), par exemple afin de redistribuer cette sortie en probabilités de classification, parfois appelés confiance. Les fonctions courantes pour cette tâche sont généralement Softmax et Sigmoid [183].

Couches neuronales spécifiques. Certaines couches d'un réseau de neurones peuvent en plus présenter des propriétés spécifiques, qui ont souvent été développées empiriquement, afin de mitiger des problèmes survenant durant l'entraînement d'un modèle neuronal. Typiquement, nous pouvons trouver les couches *Dropout* [184] où l'activité d'une

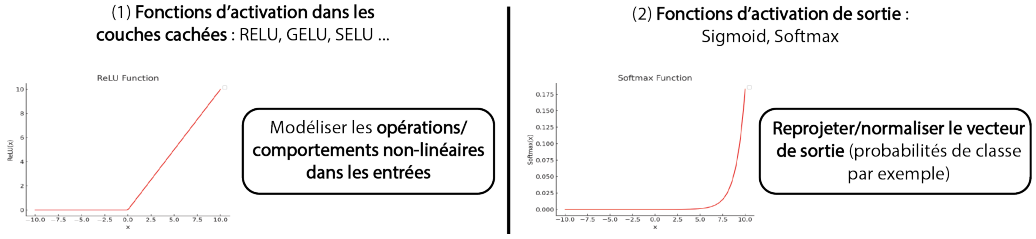


FIGURE A.3 – Synthèse des différentes fonctions d'activation neuronales et de leur rôle.

proportion donnée de neurones est aléatoirement coupée durant l'entraînement : cela réduit artificiellement le nombre de paramètres à entraîner et donc le risque de surentraînement. Cette couche est illustrée par la Figure A.4. Notons aussi les couches *LayerNorm* et *BatchNorm* qui viennent normaliser les valeurs des neurones à l'échelle de la couche, stabilisant l'entraînement et mitigant le risque de divergence catastrophique [185].

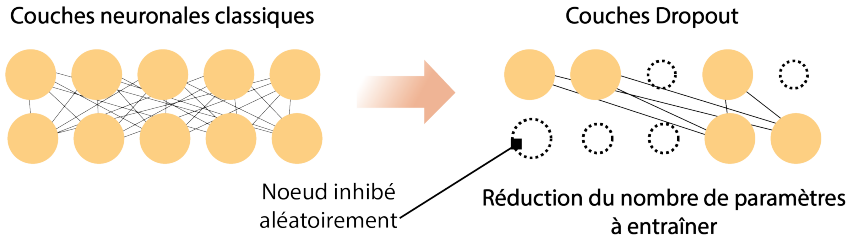


FIGURE A.4 – Illustration du principe d'une couche de Dropout.

Expressivité, représentations, caractéristiques. Il existe deux approches principales pour comprendre le fonctionnement des réseaux de neurones. La première est liée à l'analyse fonctionnelle et a été développée notamment par les mathématiques de Kolmogorov [186]. Elle considère un réseau de neurones comme un **approximateur universel de fonction** : plus le réseau est profond, plus il peut modéliser facilement des fonctions complexes pour réaliser une tâche. Nous retrouvons aussi le terme d'**expressivité** pour qualifier cette propriété d'un modèle d'apprentissage [125, 187]. La deuxième approche, appelée **théorie de la représentation**, a été largement initiée par Bengio [188]. Elle considère que les modèles forment des représentations statistiques matricielles ou vectorielles des données à différents degrés d'abstraction. La profondeur du réseau permet alors de former des représentations de plus en plus abstraites, d'extraire des structures de données de plus en plus fines, afin de mieux répondre à la tâche demandée. La Figure A.2 illustre un prototype de base de réseau de neurones pour la vision, avec les différentes couches de l'entrée vers la réalisation de la tâche. L'idée de caractéristiques peut être assimilée à la formation de filtres par apprentissage par un modèle neuronal, qui est un angle d'analyse assez courant des modèles d'apprentissage en vision. Ainsi en lieu et place de filtres extrayant des éléments précis de la scène observé comme des contours, un modèle neuronal va pouvoir former des filtres abstraits et capturant des éléments beaucoup plus fins de l'image. Ce concept est illustré dans son principe dans la Figure A.5 : nous pouvons y voir la formation de caractéristiques de plus en plus fines et abstraites, depuis l'extraction d'un contours général jusqu'au filtrage de caractéristique précises.

Lorsque les annotations associées à une tâche ne sont pas forcément disponibles (ou en quantité très limitée), une **approche auto/non-supervisée** est alors envisageable.

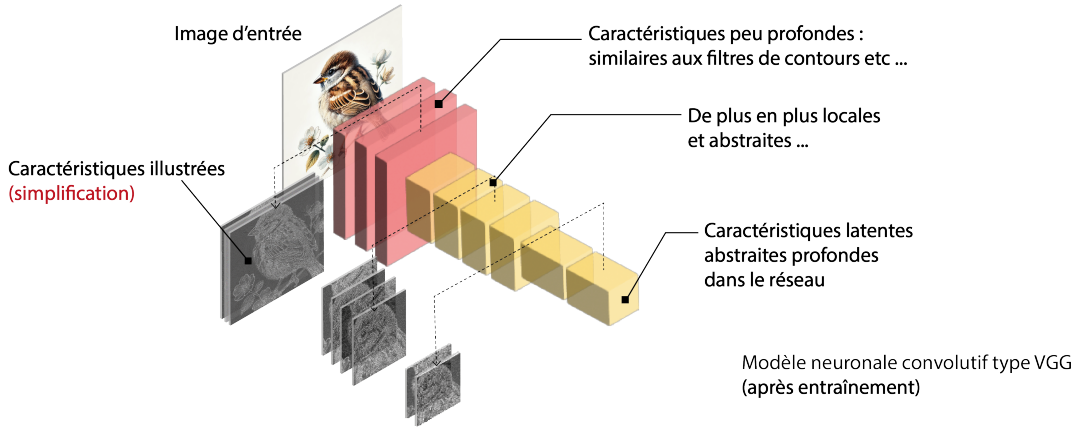


FIGURE A.5 – Illustration du concept de caractéristiques formées dans un réseau de neurones de manière simplifiée, dans le cadre de la vision par ordinateur.

Sans trop développer ces approches non-explorées dans la thèse, nous décrivons ici deux méthodes possibles. D'abord, nous avons l'approche de pré-apprentissage par reconstruction où le label est remplacé par l'image d'entrée elle-même. Le modèle, le plus souvent un encodeur-décodeur, doit alors apprendre à reconstruire son image d'entrée. Ce type de modèle est décris dans l'annexe A.4. L'encodeur forme alors des représentations liées aux éléments de la scène observée comme les contours ou textures, qui sont ré-employables dans un contexte supervisé contraint en données. Cette approche est décrite plus en détails dans la section d'annexe A.5 dédiée au transfert d'apprentissages. Une autre approche auto/non supervisée est l'utilisation d'un modèle d'apprentissage machine non-supervisé (comme un modèle k-plus-proches-voisins) qui va former des pseudo-labels pour pré-apprendre à un modèle. Cette dernière technique est plus adaptée lorsque le volume de données est très important mais que l'annotation est trop coûteuse à obtenir (plusieurs millions d'images et peu/pas d'experts humains disponibles).

Fonction de coût et type de tâche. Les réseaux de neurones sont entraînés à l'aide de techniques d'optimisation telles que la **descente de gradient** [189], qui ajuste les poids des connexions entre les neurones, couche après couche, pour minimiser une fonction de coût décrivant l'écart entre la réponse souhaitée et la réponse obtenue. La fonction de coût est définie suivant la tâche à réaliser. Pour des tâches de classification ou de régression supervisées, c'est-à-dire où l'on peut associer un label de sortie à une donnée d'entrée, il existe des fonctions de coût purement statistiques nées des recherches **en apprentissage machine supervisé** comme **l'entropie croisée** [190], qui est la plus courante et mesure directement l'écart entre l'annotation souhaitée et la réponse du modèle d'apprentissage. Nous trouvons d'autres fonctions suivant la tâche réalisée et la distribution des données, comme la fonction Hinge pour les tâches de classification multi-classes [191, 192]. Les fonctions de coût peuvent être composées, comme dans le cas de l'entraînement d'un modèle de localisation d'objets où l'on va avoir une fonction de perte créditée sur la classe de l'objet-cible et une autre perte créditée sur les points localisant l'objet (régression). Notons que l'introduction de fonctions de coût basées sur une loi de comportement comme une loi de diffusion flux/matière ... est une des voies possibles pour concevoir un réseau de neurones informé par la physique.

Lors d'un **apprentissage supervisé standard**, les données sont vues plusieurs fois, après une redistribution aléatoire. C'est la notion d'**époque**, correspondant au nombre de cycle d'entraînement où les données sont vues par le réseau. Pour accélérer l'entraînement, on calcule le gradient moyen sur un nombre donnée d'images d'entrée correspondant à **la taille de lot (ou *batch*)**. Cela permet aussi de normaliser le gradient sur plusieurs images d'entrées ce qui peut lisser des variations brusques de l'erreur lors de l'apprentissage. Les données d'entraînement sont séparées en deux ensembles : un jeu d'entraînement (*train-set*) servant à entraîner le modèle, et un jeu de test (*test-set*) servant strictement pour évaluer les performances du modèle et n'étant pas vu lors de l'entraînement.

Les concepts de distribution des données et de tâche à effectuer sont synthétisés dans la Figure A.6.

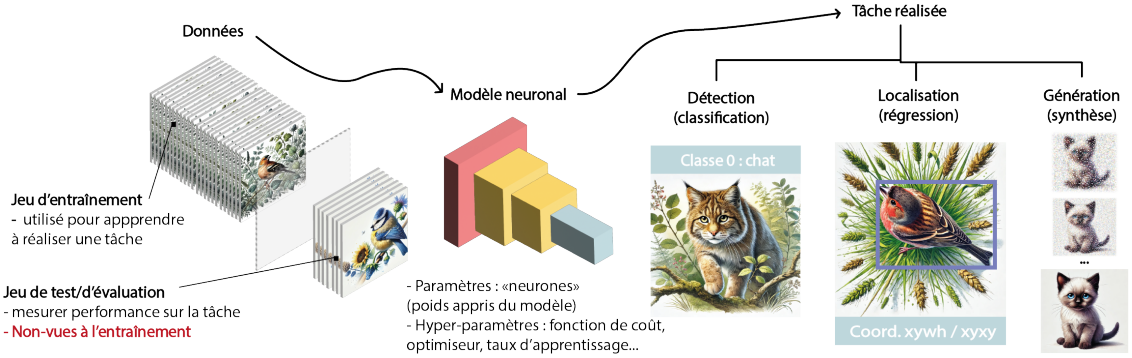


FIGURE A.6 – Synthèse sur la distribution des données et sur les différentes tâches de vision réalisables par un système d'apprentissage machine.

Optimisation des nœuds neuronaux. L'expression mère de l'optimisation des nœuds neuronaux est décrite ci-dessous, et a été popularisée par Hinton [189]. C'est la rétro-propagation, permettant d'informer chaque couche du modèle de l'erreur réalisée lors de la réalisation de la tâche :

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

où θ représente les poids du modèle, η est **le taux d'apprentissage**, et J est la fonction de coût. Pour un réseau de neurones, la mise à jour des poids via la règle de la chaîne dans la rétro-propagation est donnée par :

$$\frac{\partial J}{\partial \theta^{(l)}} = \delta^{(l+1)} \cdot \frac{\partial z^{(l)}}{\partial \theta^{(l)}}$$

où $\delta^{(l+1)}$ est l'erreur rétro-propagée de la couche, liée à la fonction de coût, et $(l+1)$ et $z^{(l)}$ est l'entrée de la couche l.

Si cette première équation décrit le fondement de l'adaptation des poids du modèle à la perte associée à la tâche à remplir, elle n'assure pas forcément la convergence stable du modèle dans un puits de potentiel local de la fonction de coût, qui n'est pas globalement convexe a priori (elle ne l'est presque jamais). Ainsi, historiquement, d'autres optimiseurs ont été développés, ajoutant un aspect stochastique sur le taux d'apprentissage forçant à explorer toute la granularité de la fonction de coût (optimiseur SGD [193]). D'autres optimisations plus récentes introduisent en plus un mécanisme de *momentum* pour augmenter

les chances de convergence du modèle d'entrée dans un minimum local de la fonction de coût, comme pour l'optimiseur ADAM [194]. C'est un domaine de recherche encore en changement constant afin d'accélérer au maximum la recherche des minima locaux dans les espaces décrits par la fonction de coût. Le pseudo-code 3 fournit un squelette de base de la rétro-propagation.

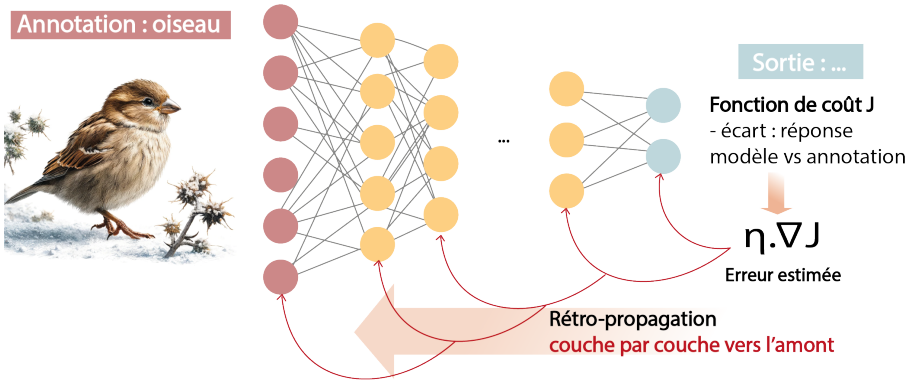


FIGURE A.7 – Illustration du principe de base de la rétro-propagation. Nous estimons l'erreur faites par le modèle grâce à une fonction de coût puis cette erreur est propagée aux noeuds du modèle.

Algorithm 3 Apprentissage par rétro-propagation dans un réseau

1: Initialiser les poids θ du réseau noté f_θ {Paramètres}
2: Initialiser η le taux d'apprentissage, le nombre d'époque {Hyper-Paramètres}
3: Définir J la fonction de coût et l'optimiseur
4: **repeat**
5: Sélectionner un lot de données d'entraînement
6: **for** chaque exemple (x_i, y_i) dans le lot **do**
7: $\hat{y}_i \leftarrow f_\theta(x_i)$ {Propagation avant : estimation de la sortie modèle}
8: **for** chaque couche l de sortie à entrée **do**
9: {Rétro-propagation}
10: $\frac{\partial J(\hat{y}_i, y_i)}{\partial \theta^{(l)}} \leftarrow \delta^{(l+1)} \cdot \frac{\partial z^{(l)}(\hat{y}_i, y_i)}{\partial \theta^{(l)}}$ {Calculer gradients liés au coût J rapport aux poids}
11: Mettre à jour les poids : $\theta^{(l)} \leftarrow \theta^{(l)} - \eta \cdot \frac{\partial J(\hat{y}_i, y_i)}{\partial \theta^{(l)}}$ {par l'optimiseur}
12: **end for**
13: **end for**
14: **until** critère d'arrêt atteint (nombre d'époques)

Ouverture : d'autres modèles du neurone. En plus des réseaux de neurones artificiels traditionnels, il existe tout un ensemble de modèles alternatifs, bien que ceux-ci soient généralement confinés à des communautés restreintes, illustrés de manière simplifiée dans la Figure A.8. Sans trop détailler, parmi eux nous pouvons trouver les réseaux de neurones à impulsion [167, 195, 196] basés sur le modèle biologique de l'axone du calmar de Hodgkins-Huxley [197], qui est à la racine des neurosciences computationnelles [113, 198, 199]. Nous y trouvons aussi les réseaux dits "à réservoirs" reposant sur un neurone modélisé par des poids cachés non-appris en plus du noeud d'entrée qui reste appris (*reservoir*

computing) [200, 201]. Il y a enfin les nouveaux réseaux de neurones dits de "Kolmogorov-Arnold" proposant d'apprendre la fonction d'activation plutôt que les poids [202, 203]. Ces approches seulement évoquées ici offrent des perspectives intéressantes dans un contexte de recherche fondamentale, par exemple pour le développement de nouvelles approches frugales en données. Néanmoins elles restent bien moins couramment utilisées dans les applications les plus matures de l'apprentissage machine pour l'industrie.

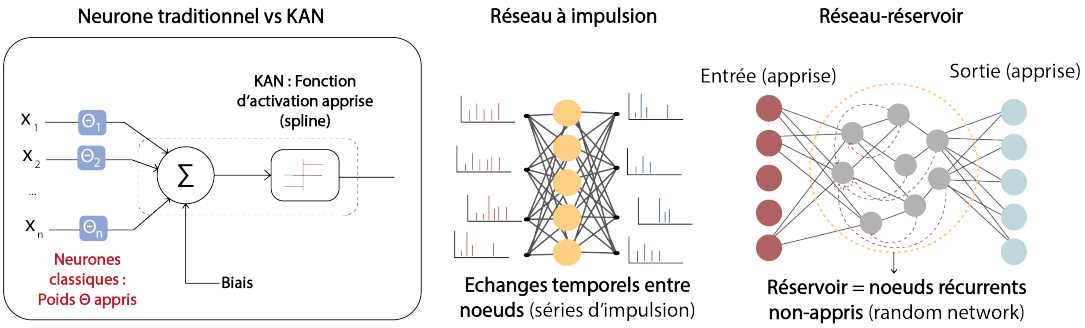


FIGURE A.8 – Illustration des modèles alternatifs au neurone artificiel. Nous y distinguons les KAN, les SNN et les réservoirs.

A.2 Réseaux de neurones pour la vision

Les réseaux de neurones convolutifs (CNN) sont les structures type pour les tâches de vision par ordinateur. Un CNN se compose de couches d'extraction de caractéristiques et de couches de prise de décision [204].

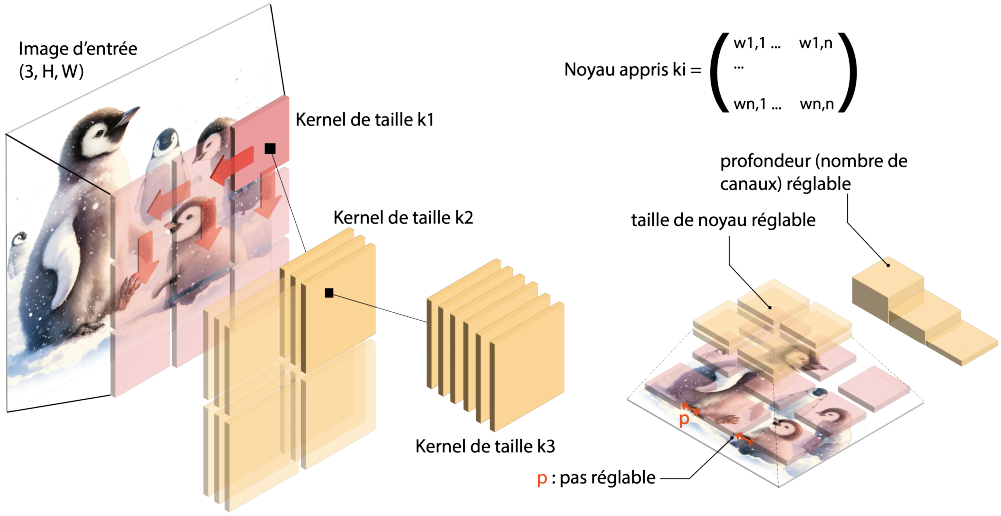


FIGURE A.9 – Illustration du fonctionnement des couches de convolution. Un noyau appris balaye l'image pour former des caractéristiques.

Les couches d'extractions de caractéristiques sont généralement des couches de convolution pour les modèles conçus pour la vision par ordinateur. Ceux-ci appliquent des filtres à l'image d'entrée pour détecter des motifs locaux tels que les bords, les textures et les

formes, qui vont être formés lors de l'apprentissage. Ces couches peuvent être suivies de couches dites de *pooling* moyennant ou maximisant les valeurs observées de manière non apprise, mais réduisant la résolution des caractéristiques extraites tout en préservant les informations essentielles. Une illustration générale du principe des convolutions apprises est fourni Figure A.9. L'opération de convolution peut être formalisée comme suit :

$$\text{Conv}^{K}_{C_0 \to C_1}(\text{Img}) = \sum_{c_m=0}^{C_1-1} \sum_{c_l=0}^{C_0-1} K_{c_m,c_l} * \text{Img}_{c_l}$$

où Img représente une entrée de type matricielle (image ou carte de caractéristiques) de dimension (C_0, H_0, W_0). K est le noyau de convolution défini par des poids appris, supposé carré de taille (k, k). C_1 est le nombre de canaux de sortie. Img_{c_l} représente le c_l-ième canal de l'entrée Img, de dimensions (H_0, W_0). K_{c_m,c_l} est le noyau de convolution pour le canal de sortie c_m et le canal d'entrée c_l, de dimensions (K, K). $*$ désigne l'opération de convolution.

L'opération de convolution pour un point de l'image de sortie à une profondeur de canal c_l peut quant à elle être formalisée comme suit :

$$\text{Conv}^{K}_{c_l}(\text{Img})_{h,w} = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} K_{i,j} \cdot \text{Img}_{h+i,w+j}$$

où h et w correspondent aux coordonnées spatiales de la matrice de sortie de dimension (C_1, H_1, W_1).

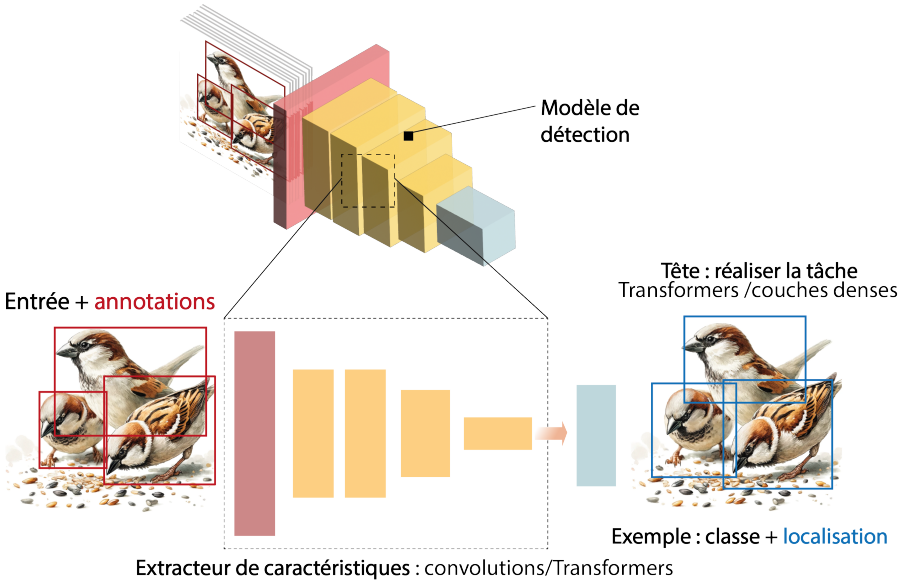


FIGURE A.10 – Illustration de la structure de base d'un modèle de vision, avec l'extracteur de caractéristique et la tête.

Les couches de prise de décision sont généralement des couches entièrement connectées (ou dites densément connectées), qui utilisent les caractéristiques extraites pour classer l'image ou effectuer une autre tâche de vision. Nous en arrivons ainsi au prototype de modèle d'apprentissage profond montré en Figure A.10 avec un bloc neuronal d'extraction

de caractéristiques (*backbone* en anglais). Ce bloc forme des représentations ou caractéristiques statistiques lors de l'apprentissage afin de résoudre une tâche remplie par le modèle. Ces caractéristiques sont utilisées par la tête (*head* en anglais) suivant la tâche : séparation des représentations suivant leurs classes, détection par boîte englobante ou polygones.

Liens avec l'apprentissage machine traditionnel. Ce découpage entre une étape de filtrage et une étape de prise de décision (classification ou régression) est un héritage des approches d'apprentissage machine traditionnel illustrée en Figure A.11, où des filtres choisis étaient appliqués afin d'extraire des caractéristiques-clés comme les filtres HOG [71], Canny [146] ou SIFT [205]. Après une mise en forme de ces caractéristiques sous forme d'un vecteur, celles-ci étaient traitées par un système d'apprentissage machine classique afin de réaliser la tâche, tels qu'une machine à vecteur de support (SVM) [206, 207], forêt aléatoire [208], ou plus récemment les approches dites de *gradient boosting* [209].

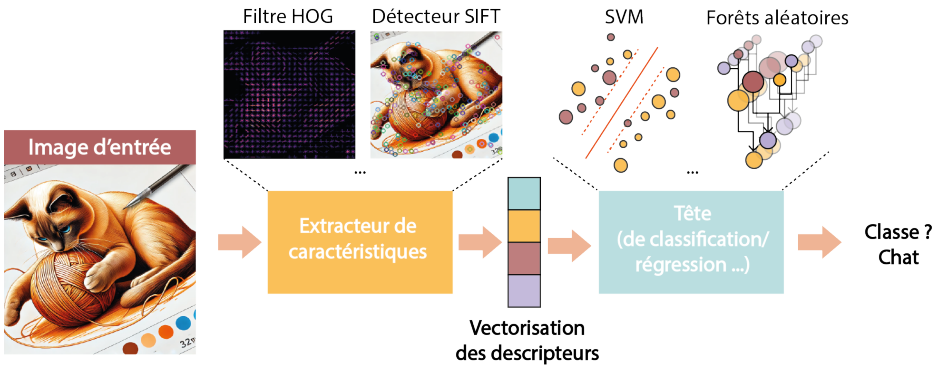


FIGURE A.11 – Protocole de traitement d'images type par apprentissage machine traditionnel.

A.3 Augmentation de données

L'augmentation est une stratégie basique afin d'augmenter artificiellement le minerai en données disponibles pour procéder à un apprentissage. Des stratégies d'augmentation simples sont aujourd'hui implémentées de manière quasi-automatique lors d'un apprentissage machine, et bien qu'elle montre tout son intérêt surtout dans les cas où le volume de données pour apprentissage est contraint (sous les 1000 images pour donner une valeur empirique). Les augmentations vont jouer sur l'orientation et la forme des images permettant alors un rebrassage limité des caractéristiques images : ce "brassage" permet souvent de faire apparaître des propriétés extrêmement utiles dans les réseaux de neurones, comme l'invariance par rotation ou la robustesse de ces modèles au bruit image. Ces propriétés sont absentes ou bien très difficiles à atteindre pour d'autres systèmes d'apprentissage machine comme les machines à vecteurs de support (SVM) [210].

La Figure A.12 fournit une illustration de la plupart des augmentations de données majeures. Nous y retrouvons les miroirs horizontaux et verticaux (*flips*), la rotation, le fenêtrage (*crop*), le bruitage Gaussien ... Un tirage aléatoire permet généralement de choisir au cours de chaque époque si une augmentation est appliquée et cela à chaque époque, avec une probabilité associée fixée généralement à 0.5 par augmentation : cela fait qu'une image est rarement vue avec le même jeu d'augmentations durant l'ensemble

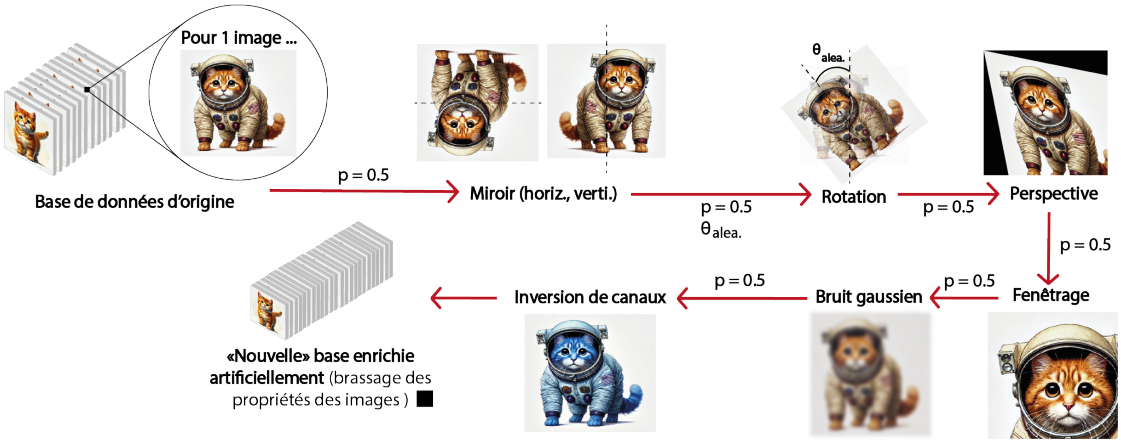


FIGURE A.12 – Protocole d'augmentations de données typiques sur une image afin d'augmenter artificiellement la quantité de données disponible. p désigne la probabilité que l'augmentation soit appliquée sur l'image.

des époques d'un apprentissage complet. La documentation de la librairie Pytorch fournit un éclairage d'implémentation plus détaillé des augmentations de données pour les images [211].

D'autres stratégies d'augmentation de données existent aujourd'hui, comme la synthèse d'images en utilisant un réseau de neurones dédié et qui est étudiée dans le Chapitre 2. Notons aussi la possibilité d'utiliser la simulation numérique, avec des outils comme la simulation éléments-finis. La génération numérique d'environnements complexes pour l'augmentation de données est aussi possible via des logiciels comme Blender [212], comme illustré dans la Figure A.13.

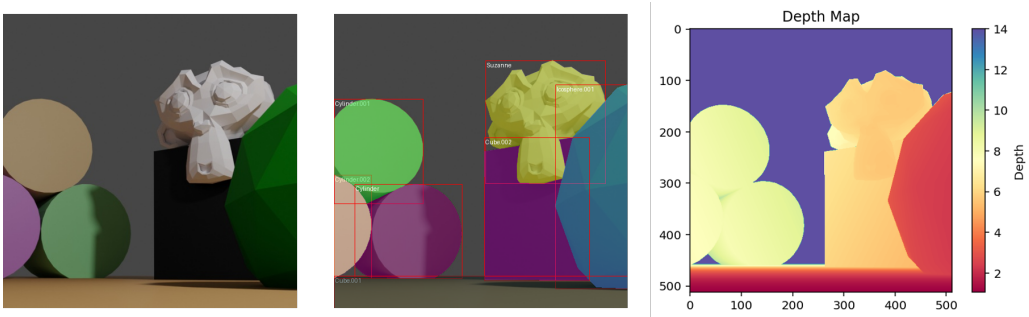


FIGURE A.13 – Exemple de scène générée avec Blender [212]. L'image synthétique est à gauche, la localisation d'objets est au centre. La carte de profondeur est à droite.

A.4 Structures de base : auto-encodeurs, notion de résidu

L'auto-encodeur constitue une classe de réseaux de neurones utilisés pour apprendre une représentation compacte d'un ensemble de données, souvent appelée l'espace latent. Ils se composent de deux parties principales : l'encodeur et le décodeur. L'encodeur va

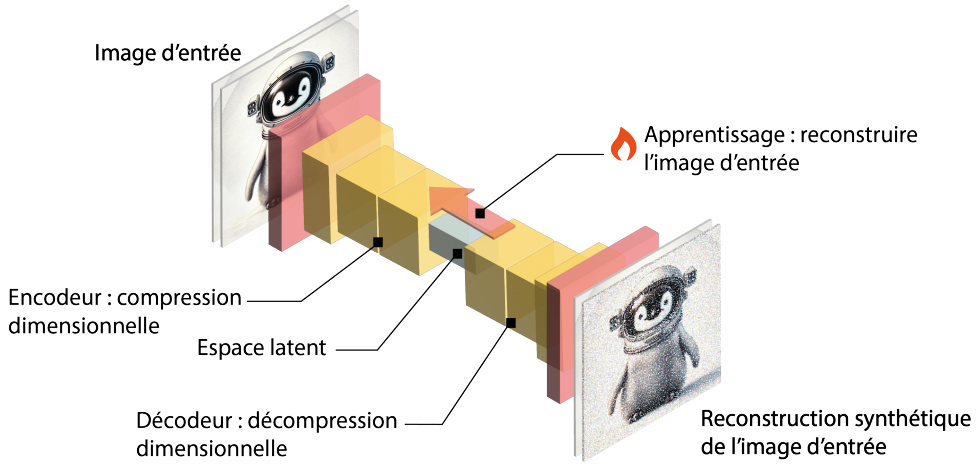


FIGURE A.14 – Illustration de la structure d'un auto-encodeur.

apprendre à comprimer dimensionnellement les données d'entrée. Tandis que le décodeur reconstruit ensuite les données d'origine à partir de cette représentation compacte. On entraîne souvent un auto-encodeur avec une fonction de perte de reconstruction, notamment l'erreur moyenne des moindres carrés. Les auto-encodeurs trouvent des applications dans la réduction de dimensionnalité, la détection d'anomalies [49], mais notamment la génération de nouvelles données similaires aux données d'entrée, comme expliqué au Chapitre 2. La Figure A.14 fournit une illustration de la structure typique d'un auto-encodeur.

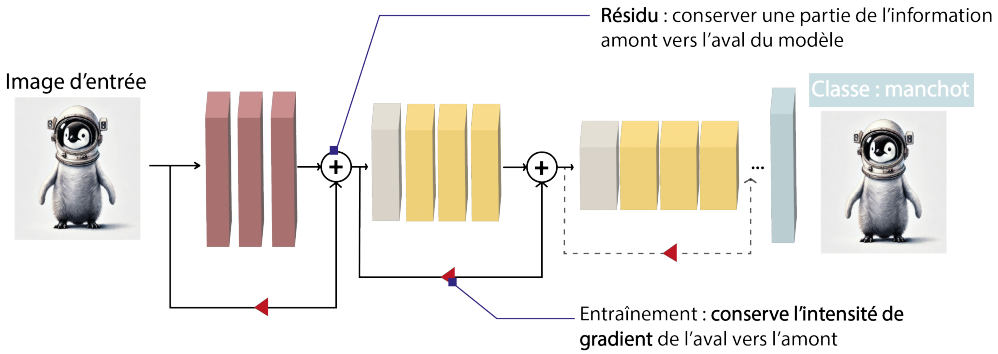


FIGURE A.15 – Illustration du principe des architectures résiduelles.

Les réseaux à résidu (ResNet) sont des architectures de réseaux de neurones profonds introduite pour résoudre le problème de la dégradation des performances dans les réseaux très profonds [91]. En effet, un empilement très massif de couches neurales finit par provoquer un "épuisement de gradient" (*gradient vanishing* en anglais) pour les couches les plus en amont du réseau, ralentissant ainsi fortement le processus d'entraînement. La principale innovation des modèles résiduels comme ResNet est l'utilisation de "blocs résiduels". Chaque bloc résiduel correspond à une connexion directe (ou connexion de saut) qui contourne une ou plusieurs couches puis se concatène au bloc de couche choisi. Cela permet de conserver des informations extraites en début de modèle pour des couches neu-

rales beaucoup plus proches de la décision et de faciliter la propagation du gradient lors de l'entraînement en augmentant le gradient transmis aux couches en début de réseau. Cette approche a permis la construction de réseaux beaucoup plus profonds, tels que ResNet-50 ou ResNet-101, tout en maintenant de hautes performances en classification d'images et en d'autres tâches de vision par ordinateur, souvent utilisés comme extracteurs de caractéristiques pour des modèles de localisation. La Figure A.15 fournit une illustration de la structure typique d'un réseau à résidu.

A.5 Transfert d'apprentissage, réajustement de modèle

Le transfert d'apprentissage est une technique qui consiste à utiliser un modèle pré-entraîné sur une grande base de données pour une nouvelle tâche. Cette approche est particulièrement utile lorsque les données disponibles pour la nouvelle tâche sont considérées comme limitées (de 100 à 1000 images pour donner un ordre de grandeur empirique).

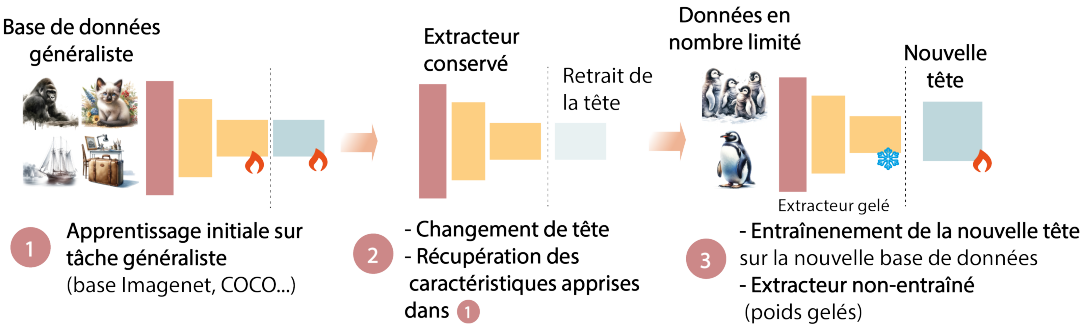


FIGURE A.16 – Illustration d'un protocole de transfert et réajustement de modèle typique.

Le réajustement d'un modèle (ou *fine-tuning* en anglais) est un processus de transfert d'apprentissage où les poids du modèle pré-entraîné sont ré-ajustés (ou affinés) sur la nouvelle tâche. Cela permet au modèle de conserver les connaissances générales acquises lors de l'entraînement initial tout en s'adaptant aux spécificités de la nouvelle tâche. Cette nouvelle tâche peut être un nouveau jeu de données plus restreint, mais aussi une tâche généralement plus gourmande en données comme de la localisation d'objets, pour du pré-entraînement d'Extracteurs de caractéristiques. Un processus typique de transfert et ajustement est illustré par la Figure A.16, montrant le passage sur une base de données généraliste pour former des caractéristiques puis les transférer sur la tâche spécifique.

Une approche de transfert d'apprentissage assez répandue dans des contextes de manque de données ou de données non-annotées est une approche non-supervisée de pré-apprentissage par reconstruction. Le modèle apprend d'abord à former des caractéristiques en reconstruisant simplement la distribution d'entrée, greffé à un décodeur pour former un auto-encodeur pour la synthèse d'images comme décrit dans l'annexe A.4. Le modèle est ensuite réajusté sur la tâche finale, comme pour de la détection d'objets. Ce principe est illustré dans la Figure A.17.

Typiquement, on remplace les dernières couches du modèle qui correspondent à la prise de décision du modèle et pré-entraîné sur une tâche généraliste, par de nouvelles couches

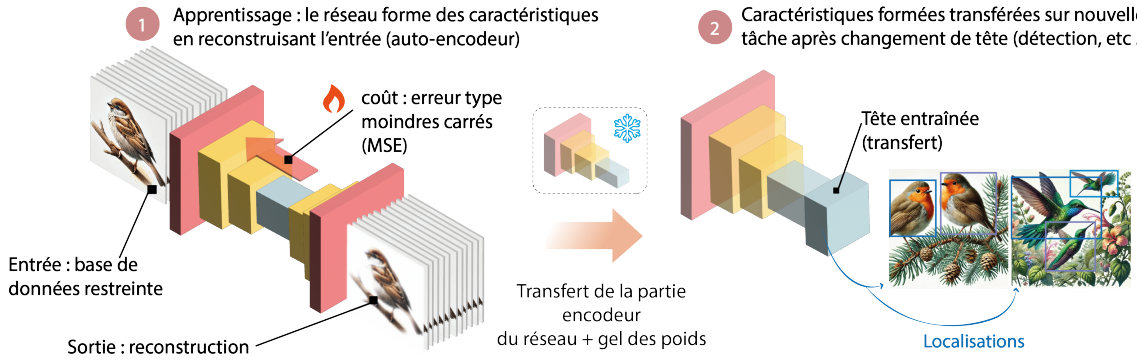


FIGURE A.17 – Protocole de pré-apprentissage non-supervisé par reconstruction des images d'entrée.

adaptées à la nouvelle tâche grâce à un ré-entraînement. On peut entraîner le modèle en utilisant un taux d'apprentissage nul (poids gelés) pour les couches pré-entraînées et un taux d'apprentissage plus élevé pour les nouvelles couches. Suivant le besoin nous pouvons aussi ré-entraîner partiellement les couches gelées avec un taux d'apprentissage très faible. D'autres approches de réajustement de modèles sont apparues dans les années récentes suite à l'explosion dans la littérature des modèles de langage, afin d'en minimiser le coût en ré-entraînement. Elles sont mentionnées ici pour le lecteur plus curieux et avancé : la *Low Rank Adaptation* (LORA) où un modèle intermédiaire de faible dimension est ajusté pour remplir la tâche-cible, et quelques uns de ses variants les plus récents [213–216].

A.6 Transformers et mécanisme d'attention

Le *Transformers* est une architecture de réseau de neurones introduite par Vaswani et al. en 2017 [67] et d'abord été utilisée dans le traitement du langage naturel [217]. Leur innovation clé est le mécanisme d'attention qui permet au modèle de se concentrer sur différentes parties de l'entrée lorsqu'il génère une sortie. Le mécanisme d'attention est illustré dans la Figure A.18 : à gauche le processus de calcul de l'attention dans un *Transformer* est fourni, puis à droite une intuition du fonctionnement de l'attention comme inter-association des caractéristiques de l'image, à plusieurs échelles et à différentes régions de celle-ci.

Le mécanisme d'attention calcule une série de poids qui déterminent l'importance relative de différentes parties de l'entrée. Ces poids sont utilisés pour combiner les représentations des différentes parties de l'entrée, permettant au modèle de créer une représentation contextuelle enrichie. Dans un contexte de vision par ordinateur, cette architecture a été adaptée pour former les *Transformers* de vision (ViT) [68]. L'entrée matricielle de type image est convertie en une séquence de fenêtres dont nous pouvons calculer l'équation d'attention. La séquence est ordonnée grâce à un mécanisme appris ou non permettant d'attribuer une position dans la structure de données à chaque fenêtre (encodage sinusoïdal par exemple).

Nous pouvons exprimer l'auto-attention comme suit :

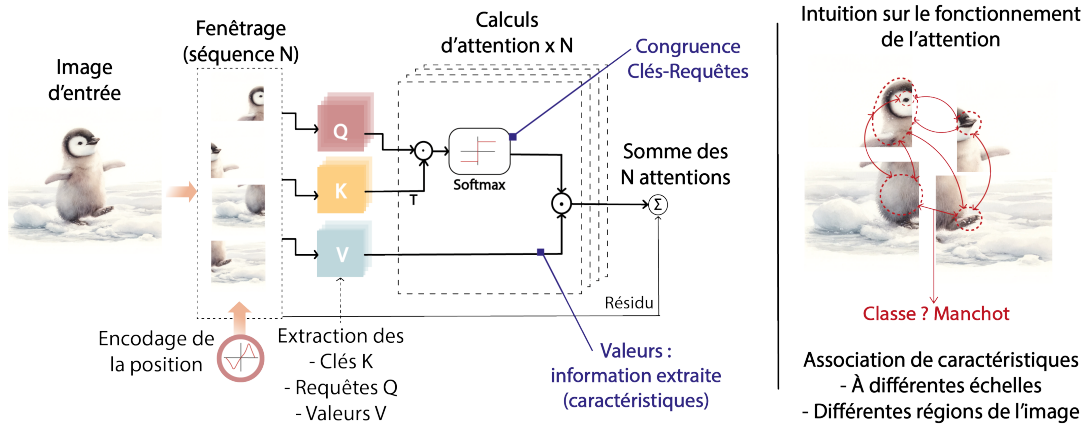


FIGURE A.18 – A gauche illustration du calcul d'attention dans un *Transformer* à partir d'une image exemple. A droite, intuition sur le fonctionnement de l'attention. Le modèle extrait et associe des caractéristiques de l'image à différentes échelles pour remplir sa tâche.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

où Q (*Queries*), K (*Keys*) et V (*Values*) sont des matrices obtenues en multipliant les entrées par des matrices de poids apprises, extrayant alors de l'entrée une information spécifique. Plus précisément :

- Q (*Queries*) : Les requêtes sont des vecteurs représentant l'élément dimensionnel interrogé. Intuitivement, elles correspondent aux questions que nous posons à chaque partie de la séquence d'entrée pour déterminer leur pertinence.
- K (*Keys*) : Les clés sont des vecteurs représentant des attributs importants des éléments d'entrée. Intuitivement, elles peuvent représenter l'indice de mise en correspondance par rapport aux requêtes Q i.e. la compatibilité entre les requêtes et les éléments de la séquence.
- V (*Values*) : Les valeurs sont les représentations des éléments d'entrée après transformation. Intuitivement, elles contiennent les caractéristiques, ou informations dans la structure de données que nous voulons récupérer et combiner, en fonction des poids d'attention qui sont appris pour réaliser une tâche donnée.

Le terme $\frac{QK^T}{\sqrt{d_k}}$ calcule la similarité entre les requêtes et les clés, normalisée par la racine carrée de la dimension des clés d_k. Une fonction exponentielle Softmax est ensuite appliquée pour obtenir des poids d'attention normalisés, qui sont ensuite utilisés pour pondérer les valeurs V.

Un *Transformer* est construit suivant les concepts de l'apprentissage profond par empilement relativement massif de modules d'attention. Il est ainsi composé de plusieurs blocs de couches d'attention et de couches denses de neurones positionnées en série. Les blocs de couches d'attention sont responsables de l'alignement et de la mise en correspondance des différentes parties de l'entrée, tandis que les couches denses traitent ces représentations enrichies pour produire les sorties.

Dans chaque bloc de transformer, le processus suit trois étapes principales :

— 1. Calcul de l'attention : Les poids d'attention sont calculés pour chaque paire de requêtes et de clés.

— 2. Combinaison des valeurs : Les valeurs sont combinées en utilisant les poids d'attention calculés.

— 3. Transformation finale : Les représentations combinées sont passées à travers des couches denses pour des transformations supplémentaires.

Ces étapes permettent aux *Transformers* de capter les dépendances à longue portée dans les données, rendant ce modèle particulièrement puissant pour le traitement de séquences et les tâches nécessitant une compréhension contextuelle globale. A contrario les réseaux construits seulement par convolutions extraient des informations locales mais ne peuvent que difficilement former des inter-associations entre les régions-fenêtres de l'entrée, dans leurs couches denses de prises de décision. Les *Transformers* semblent aussi bénéficier d'une meilleure expressivité, permettant de former plus facilement des fonctions liées au problème à résoudre par rapport aux méthodes par convolution [218–221].

Pour des informations plus avancées le lecteur peut se référer aux sources suivantes [222, 223]. Notons aussi des ressources vidéo disponibles en ligne, permettant d'obtenir une première compréhension plus intuitive de cette famille d'architectures, notamment dans le cadre du langage naturel [224–227] ainsi que le cours disponible en libre-accès suivant [228].

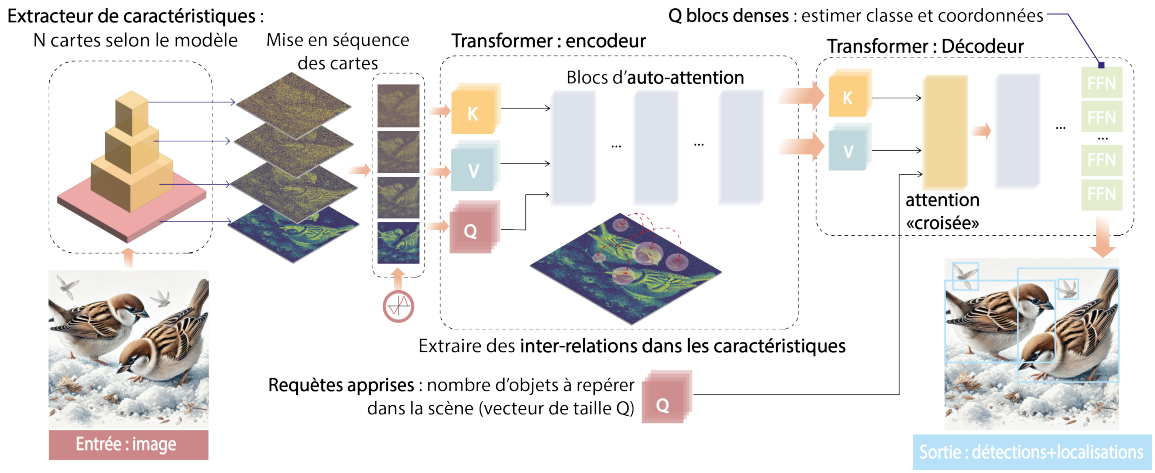


FIGURE A.19 – Illustration du principe général des *Transformers* de détection.

Transformers de détection. Les *Transformers* ont été adaptés pour la détection et la localisation d'objets sur des images, permettant le développement de modèles compétitifs avec les architectures précédentes basés sur les convolutions. Le premier représentant de cette approche est l'architecture DETR (*DEtection TRansformer*) [90], qui permet de détecter et de localiser des objets dans une image donnée. L'archétype de cette famille de modèles est illustré par la Figure A.19. On y trouve un extracteur de caractéristiques qui filtre l'image selon différentes profondeurs d'abstraction fixées suivant l'architecture utilisé : il appartient le plus souvent soit à la famille Resnet [91], soit à la famille Swin [69], reconnues pour leur efficacité dans ce type de tâche. Un bloc encodeur, basé sur l'auto-attention, extrait ensuite les inter-corrélations significatives entre ces caractéristiques extraites. Enfin, le décodeur utilise une attention croisée avec un vecteur de requêtes appris

pour interroger le modèle sur des localisations d'objet-candidats présents sur l'image. Ce vecteur de requête remplace de manière simplifiée un potentiel sous-réseau de proposition de région ou de candidats tels que présents dans les architectures Fast(er)-RCNN [123]. Une sélection des Q propositions de détection fournies par le réseau est généralement nécessaire afin d'éviter des sur-détections, grâce à une sélection à l'intensité du logit associé à la classe (confiance du modèle dans sa réponse), ou par une suppression des non-maxima (NMS) [229].

Ouverture : de la plausibilité biologique des *Transformers* ? Les rapprochements entre les modèles convolutifs, conçus de manière neuro-inspirée [204], et l'observation du fonctionnement neurologique de la perception visuel chez l'humain ont vite été faits [230]. En effet, les réseaux de neurones convolutifs reproduisent efficacement le traitement hiérarchique et spatial de l'information visuelle dans le cortex visuel humain [231–234]. Des travaux ont ainsi souligné une forte corrélation entre les caractéristiques formées par les modèles de vision par ordinateur modernes, comme les ResNets, et celles mesurées dans différentes régions neuronales du système visuel, comme la référence [235]. Un autre pointe une sensibilité comparable des deux types de circuit neuronaux -artificiels et biologiques- face aux hallucinations et attaques adverses en vue d'en provoquer, bien que les systèmes vivants semblent disposer de mécanismes de surcouches supplémentaires de correction d'erreurs encore en investigation [236]. De même certaines similarités plus limitées ont été mises en évidence entre les vivants et les machines dans le traitement du langage [237]

En revanche, les *Transformers* ont été à l'origine strictement conçus au travers des statistiques du traitement du langage naturel. Contrairement aux CNN, les *Transformers* n'ont pas été conçus avec l'inspiration de mécanismes neurologiques observés. Cependant, les recherches récentes [238–240] suggèrent que des mécanismes similaires à l'attention et aux *Transformers* pourraient être présents dans le cerveau humain, notamment dans des régions impliquées dans la gestion émotionnelle et l'interaction décision-action, telles que l'hippocampe et l'amygdale [112, 113]. Cela suggère que bien que les *Transformers* ne soient pas inspirés par la biologie, ils pourraient s'identifier à certains aspects du fonctionnement cérébral, en particulier ceux liés à l'intégration contextuelle et à l'attention sélective.

A.7 Explicabilité par méthodes de gradients

Interpréter et expliquer la décision d'un modèle neuronal est un défi majeur de l'apprentissage profond qui n'est pas encore résolu par la recherche actuelle. Il est en effet difficile d'établir une causalité claire entre l'image d'entrée, les couches cachées du modèle et les neurones rendant la décision. Néanmoins la littérature propose des méthodes basées sur la mesure du gradient, que nous récapitulons ici sous le terme de méthodes GRAD-CAM. Ces méthodes reposent sur la mesure de l'intensité des gradients propagées à travers une couche choisie dans le réseau, souvent la plus abstraite et profonde. Cette intensité est ensuite propagée à l'image d'entrée, permettant de récupérer une carte de température, ou carte d'activation neuronale. Cette carte de gradient permet d'identifier les régions de l'image impliquées dans la prise de décision du modèle. Nous pouvons alors associer cette décision à des caractéristiques de cette image, tels que des textures ou des contours. Néanmoins il faut rester prudent sur l'interprétabilité de ces méthodes : en effet

celles-ci permettent de corréler des décisions à des informations d'entrées, mais s'arrêtent là. Elles ne permettent pas d'identifier un rapport de causalité clair entre l'activation neuronale et la décision. La Figure A.20 fournit une illustration de synthèse sur les approches d'explicabilité par gradients.

Différentes variantes pour la mesure de l'activation neuronale existent, depuis GRAD-CAM [74] qui est l'approche originale jusqu'à Eigen-CAM [241] qui repose sur une décomposition en composantes ou en vecteurs propres de l'activation de la couche-cible afin d'affiner la corrélation entre l'intensité de gradients estimée et l'image. Notons l'existence de la librairie *Pytorch-GRAD-CAM* donne une implémentation modulaire d'une bonne partie de ces approches [73].

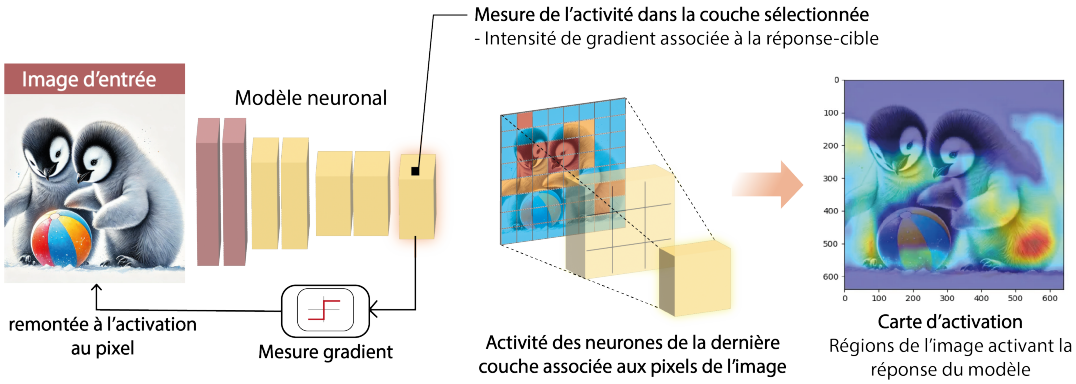


FIGURE A.20 – Illustration de principe des méthodes d'explicabilité par mesure du gradient.

Ouverture : explicabilité et dynamique non-linéaire (Chaos) ? L'étude des comportements des systèmes dynamiques complexes pourrait à terme fournir des approches transférables vers l'apprentissage profond. En effet, les approches chaotiques permettent souvent d'identifier et expliquer les différents modes propres de comportements convergents ou divergents de systèmes complexes à grande dimension, comme les fluides et la thermodynamique de l'atmosphère [242], la mécanique céleste avec un grand ensemble d'objets [243] ou l'évolution des populations animales [244]. De telles approches ont également montré certains succès en neurosciences, en identifiant des points de fonctionnement stables ou instables dans l'activité neuronale et leur lien dans la réussite de certaines tâches comme le raisonnement [245, 246]. Bien que peu explorées en intelligence artificielle, ces pistes pourraient enrichir la recherche sur l'explicabilité des modèles neuronaux, en tenant par exemple d'identifier des modes de rupture entre convergence et divergence lors de l'entraînement des modèles par rapport à des propriétés comme le nombre de paramètres ou la structure de l'architecture neuronale. La dynamique non-linéaire permettrait d'expliquer certains phénomènes anormaux parfois constatés dans l'apprentissage profond comme la disparition du sur-apprentissage lors d'entraînements très longs, analogue à un changement de mode propre pour les couches neuronales (*Grokking* en anglais) [247].

A.8 Convolutions et attention déformables

Cette section explique le concept de convolution déformable [248] et d'attention déformable [249] à la fondation d'architectures de détection d'objets comme Deformable-DETR [93]. La déformabilité, illustrée dans son principe avec la Figure A.21, permet de mieux capturer des objets de taille variable sur une image, en ajoutant des degrés de liberté supplémentaires aux noyaux de convolution.

Convolutions déformables Les convolutions déformables permettent aux réseaux de neurones d'apprendre des formes plus complexes et flexibles en ajustant par l'apprentissage les positions des noyaux. Cela est particulièrement utile pour les tâches où les objets peuvent apparaître à des échelles ou des orientations variées. Cela peut être particulièrement utile dans le contexte de la détection d'objets fins de taille variable comme les fissures, là où des architectures basées sur des *Transformers* conventionnels peuvent être mises en difficulté.

La convolution déformable peut être formalisée mathématiquement comme suit :

$$\text{Conv-Déformable}(Img, K, \Delta p) = \sum_{i=1}^{k} \sum_{j=1}^{k} Img(i + \Delta p_i, j + \Delta p_j) \cdot K(i, j)$$

où Img représente l'image d'entrée, K le noyau de convolution, et Δp les décalages appris pour chaque position (i, j), permettant au modèle de s'ajuster dynamiquement aux variations spatiales.

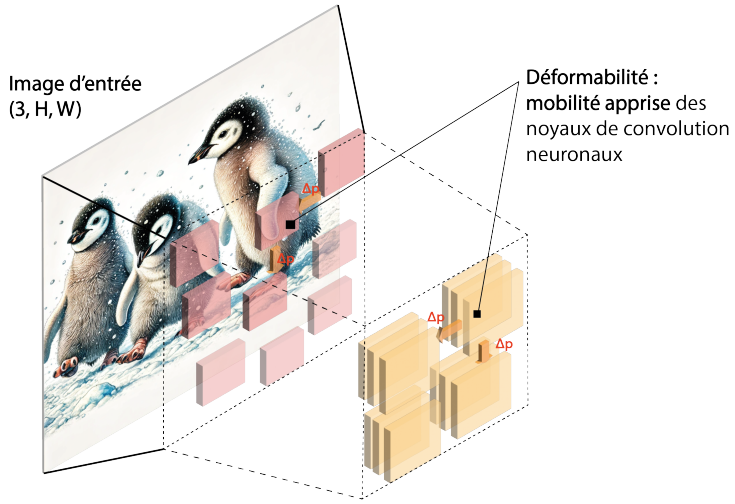


FIGURE A.21 – Illustration du mécanisme de déformabilité avec des noyaux de convolution.

Attention déformable. L'attention déformable est une extension du mécanisme d'attention standard, introduite pour améliorer la capacité des modèles à se concentrer sur des régions pertinentes de l'image, même si elles sont déformées ou déplacées. Cela est particulièrement utile dans des applications telles que la détection d'objets et la segmentation d'image.

L'attention déformable améliore le mécanisme d'attention en permettant des déformations spatiales des régions d'attention. Au lieu de calculer les poids d'attention pour des positions fixes, l'attention déformable apprend à ajuster dynamiquement ces positions,

ce qui permet de mieux capturer les relations spatiales complexes au-delà du mécanisme d'attention standard.

L'attention déformable peut être formulée comme suit :

$$\text{Attn-Déformable}(Q, K, V, \Delta p) = \sum_{k=1}^{K} Attn(Q, K, V + \Delta p_k)$$

où Δp_k représente les décalages appris pour chaque position k, permettant au modèle de s'ajuster dynamiquement aux variations spatiales.

Annexe B

Liste des publications

Actes de conférence

— Helvig, K., Abeloos, B., & Trouvé-Peloux, P. (2024). CAFF-DINO : Multi-spectral Object Detection Transformers with Cross-attention Features Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 3037-3046).

— Helvig, K., Trouvé-Peloux, P., Gavérina, L., Roche, J. M., & Abeloos, B. (2024, June). Synthetic visible-IR images pairs generation for multi-spectral NDT using flying spot thermography and deep learning. In Thermosense : Thermal Infrared Applications XLVI (Vol. 13047, pp. 14-26). SPIE.

— Helvig, K., Trouvé-Peloux, P., Gavérina, L., Roche, J. M., Abeloos, B., & Pradère, C. (2023, July). Laser flying-spot thermography : an open-access dataset for machine learning and deep learning. In Sixteenth International Conference on Quality Control by Artificial Vision (Vol. 12749, pp. 334-340). SPIE.

— Helvig, K., Trouvé-Peloux, P., Abeloos, B., Gavérina, L., & Roche, J. M. (2023, August). Détection de fissure sur matériaux métalliques par apprentissage profond et thermographie laser flying-spot : approche par apprentissage progressif. In GRETSI 2023.

— Helvig, K., Gaverina, L., Trouvé-Peloux, P., Roche, J. M., Abeloos, B., Pradere, C., & Le Besnerais, G. (2022, July). Towards deep learning fusion of flying spot thermography and visible inspection for surface cracks detection on metallic materials. In The biannual Quantitative InfraRed Thermography (QIRT) 2022.

Articles de revue

— Helvig, K., Trouvé-Peloux, P., Gaverina, L., Abeloos, B., & Roche, J. M. (2023). Automated crack detection on metallic materials with flying-spot thermography using deep learning and progressive training. Quantitative InfraRed Thermography Journal, 1-20.

— Helvig, K., Trouvé-Peloux, P., Gavérina, L., Roche, J. M., & Abeloos, B. (2024). Database for transfer learning in crack detection and localization on metallic materials

using flying spot thermography and deep learning. Journal of Electronic Imaging, 33(3), 031202-031202.

Bibliographie

[1] Abdoulahad Thiam, Jean-Christophe Kneip, Eugen Cicala, Yannick Caulier, Jean-Marie Jouvard, and Simone Mattei. Modeling and optimization of open crack detection by flying spot thermography. *NDT & E International*, 89 :67–73, July 2017.

[2] Ludovic Gavérina, Mohamed Bensalem, Adrian Bedoya, José González, Alain Sommier, Jean-Luc Battaglia, Agustin Salazar, Arantza Mendioroz, Alberto Oleaga, Jean-Christophe Batsale, and Christophe Pradere. Constant Velocity Flying Spot for the estimation of in-plane thermal diffusivity on anisotropic materials. *International Journal of Thermal Sciences*, 145 :106000, November 2019.

[3] Jean-Claude Krapez. Résolution spatiale de la caméra thermique à source volante. *International Journal of Thermal Sciences*, 38(9) :769–779, October 1999.

[4] Edward J. Kubiak. Infrared Detection of Fatigue Cracks and Other Near-Surface Defects. *Applied Optics*, 7(9) :1743–1747, September 1968. Publisher : Optical Society of America.

[5] Thierry Maffren. *Détection et caractérisation de fissures dans des aubes de turbine monocristallines pour l'évaluation de leurs durées de vie résiduelles*. Theses, Conservatoire national des arts et metiers - CNAM, April 2013.

[6] Bruno Passilly Thibaut Archer, Pierre Beauchêne and Jean-Michel Roche. Use of laser spot thermography for the non-destructive imaging of thermal fatigue microcracking of a coated ceramic matrix composite. *Quantitative InfraRed Thermography Journal*, 18(3) :141–158, 2021.

[7] L. Gaverina, M. Bensalem, A. Bedoya, J. González, A. Sommier, J. L. Battaglia, A. Salazar, A. Mendioroz, A. Oleaga, J. C. Batsale, and C. Pradere. Constant Velocity Flying Spot for the estimation of in-plane thermal diffusivity on anisotropic materials. *Int. J. Therm. Sci.*, 145 :106000, November 2019.

[8] Teng Li, Darryl P Almond, D Andrew S Rees, and B Weekes. Crack imaging by pulsed laser spot thermography. *Journal of Physics : Conference Series*, 214 :012072, March 2010.

[9] Teng Li, Darryl P Almond, and D Andrew S Rees. Crack imaging by scanning laser-line thermography and laser-spot thermography. *Measurement Science and Technology*, 22(3) :035701, March 2011.

[10] Yacine Mokhtari, Ludovic Gavérina, Clemente Ibarra-Castanedo, Matthieu Klein, P Servais, Jean Dumoulin, and Xavier Maldague. Comparative study of Line Scan

and Flying Line Active IR Thermography operated with a 6-axis robot. In *Proceedings of the 2018 International Conference on Quantitative InfraRed Thermography*. QIRT Council, 2018.

[11] Agustín Salazar, Arantza Mendioroz, and Alberto Oleaga. Flying spot thermography : Quantitative assessment of thermal diffusivity and crack width. *Journal of Applied Physics*, 127(13) :131101, April 2020. Publisher : American Institute of Physics.

[12] Stefano Sfarra, Ludovic Gavérina, Christophe Pradère, Alain Sommier, and Jean-Christophe Batsale. Integration study among flying spot laser thermography and terahertz technique for the inspection of panel paintings. *Journal of Thermal Analysis and Calorimetry*, 2022.

[13] Nelson W Pech-May and Mathias Ziegler. Detection of Surface Breaking Cracks Using Flying Line Laser Thermography : A Canny-Based Algorithm. *Engineering Proceedings*, 8(1) :22, November 2021.

[14] Peter W. Tse and Gaochao Wang. Sub-surface defects detection of by using active thermography and advanced image edge detection. *J. Phys. Conf. Ser.*, 842(1) :012029, May 2017.

[15] Victor Klamert, Matthias Schmid-Kietreiber, and Mugdim Bublin. A deep learning approach for real time process monitoring and curling defect detection in Selective Laser Sintering by infrared thermography and convolutional neural networks. *Procedia CIRP*, 111 :317–320, January 2022.

[16] Tamas Aujeszky, Georgios Korres, and Mohamad Eid. Material classification with laser thermography and machine learning. *Quantitative InfraRed Thermography Journal*, 16(2) :181–202, April 2019.

[17] Wenxiong Shi, Zhangyu Ren, Wei He, Junsong Hou, Huimin Xie, and Sheng Liu. A technique combining laser spot thermography and neural network for surface crack detection in laser engineered net shaping. *Optics and Lasers in Engineering*, 138 :106431, March 2021.

[18] Robin M Schmidt. Recurrent Neural Networks (RNNs) : A gentle Introduction and Overview, November 2019. arXiv :1912.05911.

[19] MM Groz, E Abisset-Chavanne, JC Batsale, and A Sommier. Active thermal super-resolution based on laser flying spot technique coupled with ir thermography and wavelet transform. *Quantitative InfraRed Thermography Journal*, pages 1–24, 2024.

[20] COMSOL Multiphysics. Introduction to comsol multiphysics®. *COMSOL Multiphysics, Burlington, MA, accessed Feb*, 9 :2018, 1998.

[21] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio : Data labeling software, 2020-2022. Open source software available from https ://github.com/heartexlabs/label-studio.

[22] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging : A review. *Journal of Imaging*, 9(4), 2023.

[23] Xiaohui Zhang, Ahana Gangopadhyay, Hsi-Ming Chang, and Ravi Soni. Diffusion model-based data augmentation for lung ultrasound classification with limited

data. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 664–676. PMLR, 10 Dec 2023.

[24] Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H Chi, and Sagar Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv :2002.03532*, 2020.

[25] Zheng Li, Yuxuan Li, Penghai Zhao, Renjie Song, Xiang Li, and Jian Yang. Is synthetic data from diffusion models ready for knowledge distillation ? *arXiv preprint arXiv :2305.12954*, 2023.

[26] Hatef Otroshi Shahreza, Anjith George, and Sébastien Marcel. Synthdistill : Face recognition with knowledge distillation from synthetic data. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[27] He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6008–6016, June 2024.

[28] Dana H. Ballard. Modular learning in neural networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1*, AAAI'87, page 279–284. AAAI Press, 1987.

[29] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. 1986.

[30] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook : data mining and knowledge discovery handbook*, pages 353–374, 2023.

[31] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE : Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[32] William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders. In *International Conference on Learning Representations*, 2022.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[34] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan : Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

[35] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[36] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International*

Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[37] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++ : Improving flow-based generative models with variational dequantization and architecture design. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2722–2730. PMLR, 09–15 Jun 2019.

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, June 2015.

[39] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net : Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, Cham, Switzerland, November 2015.

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[43] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen : An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Deep Floyd. Deep floyd if : Text-to-image model. https://github.com/deep-floyd/IF, 2023.

[45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl : Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[46] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https ://cdn. openai. com/papers/dall-e-3. pdf*, 2(3) :8, 2023.

[47] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[48] Jonathan Heek, Emiel Hoogeboom, and Tim Salimans. Multistep consistency models. *arXiv preprint arXiv :2403.06807*, 2024.

[49] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection : A survey. *Machine Intelligence Research*, 21(1) :104–135, January 2024.

[50] Haohao Xu, Shuchang Xu, and Wenzhen Yang. Unsupervised industrial anomaly detection with diffusion models. *J. Visual Commun. Image Represent.*, 97 :103983, December 2023.

[51] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[52] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Los Alamitos, CA, USA, jun 2015. IEEE Computer Society.

[53] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4) :600–612, April 2004.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca : Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[56] Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Homin Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning : A survey on image-based generative and discriminative training. *Transactions on Machine Learning Research*, 2023.

[57] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers : State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[58] Nathan Letheule, Flora Weissgerber, Sylvain Lobry, and Elise Colin. Automatic simulation of sar images : Comparing a deep-learning based method to a hybrid method. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4958–4961, 2023.

[59] P Ivan Pavlov (1927). Conditioned reflexes : An investigation of the physiological activity of the cerebral cortex. *Annals of Neurosciences*, 17(3) :136–141, July 2010.

[60] Y. Bengio et al. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, June 2009. Association for Computing Machinery.

[61] Ozsel Kilinc and Giovanni Montana. Follow the object : Curriculum learning for manipulation tasks with imagined goals. In *Deep RL Workshop NeurIPS 2021*.

[62] Jiwen Zhang, Jianqing Fan, Jiajie Peng, et al. Curriculum learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34 :13328–13339, 2021.

[63] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019.

[64] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[65] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, et al. Learn like a pathologist : curriculum learning by annotator agreement for histopathology image classification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2473–2483, 2021.

[66] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv :1409.1556 [cs]*, April 2015. arXiv : 1409.1556.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[68] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[69] Z. Liu et al. Swin Transformer : Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada, October 2021. IEEE.

[70] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021.

[71] N. Dalal et al. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893, San Diego, CA, USA, 2005. IEEE.

[72] David M. W. Powers. Evaluation : from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*, October 2020.

[73] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

[74] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vision*, 128(2) :336–359, February 2020.

[75] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[76] Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu. Teacher-student architecture for knowledge distillation : A survey. *CoRR*, 2023.

[77] Tianxun Zhou and Keng-Hwee Chiam. Synthetic data generation method for data-free knowledge distillation in regression neural networks. *Expert Systems with Applications*, 227 :120327, 2023.

[78] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion : Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[79] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv :2305.15956*, 2023.

[80] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8472–8480, 2024.

[81] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[82] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[83] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer : Hierarchical Vision Transformer using Shifted Windows, August 2021. arXiv :2103.14030 [cs].

[84] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet : Marrying convolution and attention for all data sizes. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[85] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit) : General visual representation learning. In *Computer Vision–ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[86] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet : A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[87] Glenn Jocher. Ultralytics yolov5, 2020.

[88] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision : From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4) :1680–1716, 2023.

[89] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9 : Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv :2402.13616*, 2024.

[90] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020 : 16th European Conference, Glasgow, UK, August*

23–28, 2020, Proceedings, Part I, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag.

[91] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[92] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[93] Zhu Xizhou, Su Weijie, Lu Lewei, Li Bin, Wang Xiaogang, and Dai Jifeng. Deformable DETR : deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[94] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO : Common Objects in Context, February 2015. arXiv :1405.0312 [cs].

[95] Xin Liu, Xudong Yang, Lianhe Shao, Xihan Wang, Quanli Gao, and Hongbo Shi. Gm-detr : Research on a defect detection method based on improved detr. *Sensors*, 24(11), 2024.

[96] Alaa Tharwat and Wolfram Schenck. A Survey on Active Learning : State-of-the-Art, Practical Challenges and Research Directions. *Mathematics*, 11(4) :820, February 2023.

[97] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning : a state of the art. *Artif. Intell. Rev.*, 56(4) :3005–3054, April 2023.

[98] M. A. Ganaie, Minghui Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. Ensemble deep learning : A review. *Eng. Appl. Artif. Intell.*, 115 :105151, October 2022.

[99] Ibomoiye Domor Mienye and Yanxia Sun. A Survey of Ensemble Learning : Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, 10 :99129–99149, September 2022.

[100] Haofei Yu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Mmoe : Mixture of multimodal interaction experts. *arXiv preprint arXiv :2311.09580*, 2023.

[101] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers : Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120) :1–39, 2022.

[102] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35 :7103–7114, 2022.

[103] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv :2209.01667*, 2022.

[104] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv :2208.02813*, 2022.

[105] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8583–8595. Curran Associates, Inc., 2021.

[106] Jain Yash, Behl Harkirat, Kira Zsolt, and Vineet Vibhav. DAMEX : Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[107] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection : Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015.

[108] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[109] Zuhui Hu, Yaguang Jing, and Guoqing Wu. Decision-level fusion detection method of visible and infrared images under low light conditions. *EURASIP Journal on Advances in Signal Processing*, 2023(1) :38, 2023.

[110] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[111] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE, 2020.

[112] John Mendoza and Anne Foundas. *Clinical neuroanatomy : a neurobehavioral approach*. Springer Science & Business Media, 2007.

[113] Dana H Ballard. *Brain computation as hierarchical abstraction*. MIT press, 2015.

[114] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2020.

[115] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 403–411, June 2023.

[116] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 72–80, 2021.

[117] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv :2111.00273*, 2021.

[118] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo : a visual language model for few-shot learning. *Advances in neural information processing systems*, 35 :23716–23736, 2022.

[119] Jaewoo Park, Chenghao Quan, Hyungon Moon, and Jongeun Lee. Hyperdimensional computing as a rescue for efficient privacy-preserving machine learning-as-a-service. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–8. IEEE, 2023.

[120] Arthur E Hoerl and Robert W Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.

[121] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 58(1) :267–288, 1996.

[122] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[123] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn : Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[124] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag.

[125] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5) :359–366, 1989.

[126] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu. Detrs with hybrid matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19702–19712, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

[127] Liu Shilong, Li Feng, Zhang Hao, Yang Xiao, Qi Xianbiao, Su Hang, Zhu Jun, and Zhang Lei. DAB-DETR : Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

[128] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO : DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

[129] C. Chen, Q. Fan, and R. Panda. Crossvit : Cross-attention multi-scale vision transformer for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[130] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *Computer Vision – ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 139–158, Berlin, Heidelberg, 2022. Springer-Verlag.

[131] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion : Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145 :109913, 2024.

[132] Mayur Dhanaraj, Manish Sharma, Tiyasa Sarkar, Srivallabha Karnam, Dimitris G. Chachlakis, Raymond Ptucha, Panos P. Markopoulos, and Eli Saber. Vehicle detection from multi-modal aerial imagery using YOLOv3 with mid-level fusion. In *Proceedings Volume 11395, Big Data II : Learning, Analytics, and Applications*, volume 11395, pages 22–32. SPIE, May 2020.

[133] Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M. Ni. Lite detr : An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18558–18567, June 2023.

[134] Stijn van Dongen and Anton J. Enright. Metric distances derived from cosine similarity and pearson and spearman correlations, 2012.

[135] Takumi Nakagawa, Yutaro Sanada, Hiroki Waida, Yuhui Zhang, Yuichiro Wada, Kōsaku Takanashi, Tomonori Yamada, and Takafumi Kanamori. Denoising cosine similarity : A theory-driven approach for efficient representation learning. *Neural Networks*, 169 :226–241, 2024.

[136] Rui Li, Jianlin Su, Chenxi Duan, and Shunyi Zheng. Linear attention mechanism : An efficient attention for semantic segmentation. *arXiv preprint arXiv :2007.14902*, 2020.

[137] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer : The long-document transformer. *arXiv preprint arXiv :2004.05150*, 2020.

[138] S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller. Contrast-limited adaptive histogram equalization : speed and effectiveness. In *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, pages 337–345, 1990.

[139] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth : Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, June 2023.

[140] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[141] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

[142] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[143] Ruobai Wang, Yu Ding, Lincheng Li, and Changjie Fan. One-shot voice conversion using star-gan. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7729–7733. IEEE, 2020.

[144] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm : Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv :2104.05358*, 2021.

[145] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. One-shot unsupervised domain adaptation with personalized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 698–708, 2023.

[146] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6) :679–698, 1986.

[147] Yi Wu, Ziqiang Li, Chaoyue Wang, Heliang Zheng, Shanshan Zhao, Bin Li, and Dacheng Tao. Domain re-modulation for few-shot generative domain adaptation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[148] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything : Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[149] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv :2406.09414*, 2024.

[150] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[151] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[152] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv :2305.10724*, 2023.

[153] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nat. Commun.*, 15(654) :1–9, January 2024.

[154] Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478 :111972, 2023.

[155] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[156] Badri Narayana Patro and Vijay Srinivas Agneeswaran. Mamba-360 : Survey of state space models as transformer alternative for long sequence modelling : Methods, applications, and challenges. *arXiv preprint arXiv :2404.16112*, 2024.

[157] Xiao Liu, Chenxu Zhang, and Lei Zhang. Vision mamba : A comprehensive survey and taxonomy, 2024.

[158] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba : Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv :2401.09417*, 2024.

[159] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6572–6583, Red Hook, NY, USA, 2018. Curran Associates Inc.

[160] Akshay J. Dave and Richard B. Vilim. Physics-informed state-space neural networks for transport phenomena. *Engineering Applications of Artificial Intelligence*, 133 :108245, 2024.

[161] Yuanyuan Qu, Jinwen Feng, Xiaohui Wu, Lin Bai, Wenhao Xu, Lingli Zhu, Yang Liu, Fujiang Xu, Xuan Zhang, Guojian Yang, Jiacheng Lv, Xiuping Chen, Guo-Hai Shi, Hong-Kai Wang, Da-Long Cao, Hang Xiang, Lingling Li, Subei Tan, Hua-Lei Gan, Meng-Hong Sun, Jiange Qiu, Hailiang Zhang, Jian-Yuan Zhao, Dingwei Ye, and Chen Ding. A proteogenomic analysis of clear cell renal cell carcinoma in a Chinese population. *Nat. Commun.*, 13, 2022.

[162] Shangran Qiu, Matthew I. Miller, Prajakta S. Joshi, Joyce C. Lee, Chonghua Xue, Yunruo Ni, Yuwei Wang, Ileana De Anda-Duran, Phillip H. Hwang, Justin A. Cramer, Brigid C. Dwyer, Honglin Hao, Michelle C. Kaku, Sachin Kedar, Peter H. Lee, Asim Z. Mian, Daniel L. Murman, Sarah O'Shea, Aaron B. Paul, Marie-Helene Saint-Hilaire, E. Alton Sartor, Aneeta R. Saxena, Ludy C. Shih, Juan E. Small, Maximilian J. Smith, Arun Swaminathan, Courtney E. Takahashi, Olga Taraschenko, Hui You, Jing Yuan, Yan Zhou, Shuhan Zhu, Michael L. Alosco, Jesse Mez, Thor D. Stein, Kathleen L. Poston, Rhoda Au, and Vijaya B. Kolachalama. Multimodal deep learning for Alzheimer's disease dementia assessment. *Nat. Commun.*, 13(3404) :1–17, June 2022.

[163] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up : Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36, 2024.

[164] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager : An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

[165] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *arXiv preprint arXiv :2307.01848*, 2023.

[166] Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. *arXiv preprint arXiv :2402.19469*, 2024.

[167] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models : Single neurons, populations, plasticity*. Cambridge university press, 2002.

[168] Siqi Wang, Pietro Maris Ferreira, and Aziz Benlarbi-Delai. Physics Informed Spiking Neural Networks : Application to Digital Predistortion for Power Amplifier Linearization. *IEEE Access*, 11 :48441–48453, May 2023.

[169] Alexandre Heuillet, Ahmad Nasser, Hichem Arioui, and Hedi Tabia. Efficient automation of neural network design : A survey on differentiable neural architecture search. *ACM Comput. Surv.*, may 2024. Just Accepted.

[170] Mingxing Tan and Quoc Le. Efficientnet : Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[171] Mingxing Tan and Quoc Le. Efficientnetv2 : Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[172] Shay Aharon, Louis-Dupont, Ofri Masad, Kate Yurkova, Lotem Fridman, Lkdci, Eugene Khvedchenya, Ran Rubin, Natan Bagrov, Borys Tymchenko, Tomer Keren, Alexander Zhilko, and Eran-Deci. Super-gradients, 2021.

[173] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS : Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

[174] Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking architecture selection in differentiable nas. In *International Conference on Learning Representations (ICLR)*, 2021.

[175] Xunyu Zhu, Jian Li, Yong Liu, and Weiping Wang. Improving differentiable architecture search via self-distillation. *Neural Networks*, 167 :656–667, 2023.

[176] Jiwoo Mun, Seokhyeon Ha, and Jungwoo Lee. De-darts : Neural architecture search with dynamic exploration. *ICT Express*, 9(3) :379–384, 2023.

[177] peggy1502. Amazing-Resources. https://github.com/peggy1502/Amazing-Resources. Github, hub de compilation de ressources didactiques sur l'apprentissage profond.

[178] François. *The Little Book of Deep Learning*. Writers Republic LLC, 2023.

[179] Simone Scardapane. Alice's adventures in a differentiable wonderland–volume i, a tour of the land. *arXiv preprint arXiv :2404.17625*, 2024.

[180] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

[181] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5 :115–133, 1943.

[182] F. Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6) :386–408, November 1958.

[183] Andrinandrasana David Rasamoelina, Fouzia Adjailia, and Peter Sinčák. A review of activation function for artificial neural network. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 281–286. IEEE, 2020.

[184] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1) :1929–1958, 2014.

[185] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

[186] Andrei Nikolaevich Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, pages 953–956. Russian Academy of Sciences, 1957.

[187] Johannes Schmidt-Hieber. The Kolmogorov–Arnold representation theorem revisited. *Neural Networks*, 137 :119–126, May 2021.

[188] Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning : A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 :1798–1828, 08 2013.

[189] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088) :533–536, 1986.

[190] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions : Theoretical analysis and applications. In *International Conference on Machine Learning*, pages 23803–23828. PMLR, 2023.

[191] Claudio Gentile and Manfred KK Warmuth. Linear hinge loss and average margin. *Advances in neural information processing systems*, 11, 1998.

[192] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same ? *Neural computation*, 16(5) :1063–1076, 2004.

[193] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010 : 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[194] DP Kingma. Adam : a method for stochastic optimization. In *Int Conf Learn Represent*, 2014.

[195] Kashu Yamazaki, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. Spiking Neural Networks and Their Applications : A Review. *Brain Sciences*, 12(7), July 2022.

[196] Kai Malcolm and Josue Casco-Rodriguez. A Comprehensive Review of Spiking Neural Networks : Interpretation, Optimization, Efficiency, and Best Practices. *arXiv*, March 2023.

[197] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4) :500, August 1952.

[198] Paul Miller. *An introductory course in computational neuroscience.* MIT Press, 2018.

[199] A. S. Maida. Cognitive Computing and Neural Networks : Reverse Engineering the Brain. In *Handbook of Statistics*, volume 35, pages 39–78. Elsevier, Walthm, MA, USA, January 2016.

[200] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3) :127–149, 2009.

[201] Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing : A review. *Neural Networks*, 115 :100–123, July 2019.

[202] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan : Kolmogorov-arnold networks. *arXiv preprint arXiv :2404.19756*, 2024.

[203] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau. Convolutional kolmogorov-arnold networks. *arXiv preprint arXiv :2406.13155*, 2024.

[204] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010.

[205] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[206] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

[207] Xiaowu Sun, Lizhen Liu, Hanshi Wang, Wei Song, and Jingli Lu. Image classification via support vector machine. In *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, volume 1, pages 485–489. IEEE, 2015.

[208] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[209] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[210] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5) :1055–1064, 1999.

[211] Transforming and augmenting images — Torchvision 0.18 documentation.

[212] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2 : A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82) :4901, 2023.

[213] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora : Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[214] Artyom O Levin and Yuri S Belov. A study on the application of using hypernetwork and low rank adaptation for text-to-image generation based on diffusion models. In *2024 6th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*, pages 1–5. IEEE, 2024.

[215] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+ : Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning*, 2024.

[216] Shubhankar Borse, Shreya Kadambi, Nilesh Prasad Pandey, Kartikeya Bhardwaj, Viswanath Ganapathy, Sweta Priyadarshi, Risheek Garrepalli, Rafael Esteves, Munawar Hayat, and Fatih Porikli. Foura : Fourier low rank adaptation. *arXiv preprint arXiv :2406.08798*, 2024.

[217] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[218] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions ? In *International Conference on Learning Representations*.

[219] Shanda Li, Xiangning Chen, Di He, and Cho-Jui Hsieh. Can vision transformers perform convolution ? *arXiv preprint arXiv :2111.01353*, 2021.

[220] Jorge Pérez, Javier Marinković, and Pablo Barceló. On the turing completeness of modern neural network architectures. In *International Conference on Learning Representations*, 2019.

[221] Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O(n) connections are expressive enough : Universal approximability of sparse transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13783–13794. Curran Associates, Inc., 2020.

[222] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[223] cmhungsteve. Awesome-Transformer-Attention. https://github.com/cmhungsteve/Awesome-Transformer-Attention. Github, hub de bibliographie sur les Transformers.

[224] 3Blue1Brown. Mais qu'est-ce qu'un réseau de neurones ? | Chapitre 1, Apprentissage profond. https://www.youtube.com/watch?v=aircAruvnKk, October 2017.

[225] 3Blue1Brown. Mais qu'est-ce qu'une convolution ? https://www.youtube.com/watch?v=KuXjwB4LzSA, November 2022.

[226] 3Blue1Brown. Visualiser l'attention, un cœur de transformateur | Chapitre 6, Apprentissage profond. https://www.youtube.com/watch?v=eMlx5fFNoYc, April 2024.

[227] 3Blue1Brown. But what is a GPT ? Visual intro to transformers | Chapter 5, Deep Learning. https://www.youtube.com/watch?v=wjZofJX0v4M, April 2024.

[228] CS25 : Tranformers United ! https ://web.stanford.edu/class/cs25/index.html. Cours en ligne sur les Transformers (Univ. Stanford).

[229] Meiling Gong, Dong Wang, Xiaoxia Zhao, Huimin Guo, Donghao Luo, and Min Song. A review of non-maximum suppression algorithms for deep learning target detection. In *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, volume 11763, pages 821–828. SPIE, 2021.

[230] Grace W Lindsay. Convolutional neural networks as a model of the visual system : Past, present, and future. *Journal of cognitive neuroscience*, 33(10) :2017–2031, 2021.

[231] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score : Which artificial neural network for object recognition is most brain-like ? *BioRxiv*, page 407007, 2018.

[232] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439) :eaav9436, 2019.

[233] Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43) :e2200800119, 2022.

[234] Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, 20(1) :e1011792, 2024.

[235] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific reports*, 8(1) :3752, 2018.

[236] Chong Guo, Michael J Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James J DiCarlo. Adversarially trained neural representations may already be as robust as corresponding biological neural representations. In *Proceedings of the 38th International Conference on Machine Learning*, 2022.

[237] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3) :369–380, 2022.

[238] James CR Whittington, Joseph Warren, and Tim EJ Behrens. Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*, 2022.

[239] Tom M George, Kimberly L Stachenfeld, Caswell Barry, Claudia Clopath, and Tomoki Fukai. A generative model of the hippocampal formation trained with theta driven local learning rules. *Advances in Neural Information Processing Systems*, 36, 2024.

[240] Dong Kyum Kim, Jea Kwon, Meeyoung Cha, and Chul Lee. Transformer as a hippocampal memory consolidation model based on nmdar-inspired nonlinearity. *Advances in Neural Information Processing Systems*, 36, 2024.

[241] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam : Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

[242] Sridhar Muddada and BSV Patnaik. Application of chaos control techniques to fluid turbulence. *Applications of Chaos and Nonlinear Dynamics in Engineering-Vol. 1*, pages 87–136, 2011.

[243] Edward Belbruno. *Capture dynamics and chaotic motions in celestial mechanics : With applications to the construction of low energy transfers*. Princeton University Press, 2004.

[244] Robert Mccredie May. Chaos and the dynamics of biological populations. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 413(1844) :27–44, 1987.

[245] Luca Cocchi, Leonardo L Gollo, Andrew Zalesky, and Michael Breakspear. Criticality in the brain : A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, 158 :132–152, 2017.

[246] John M Beggs. *The cortex and the critical point : understanding the power of emergence*. MIT Press, 2022.

[247] Ziming Liu, Ouail Kitouni, Niklas S Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking : An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35 :34651–34663, 2022.

[248] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[249] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Apprentissage multi-capteurs pour le contrôle non destructif de matériaux
Mots clés : apprentissage automatique, fusion de données, contrôle non-destructif, apprentissage profond

Résumé : La détection de défauts dans les structures aéronautiques et spatiales est cruciale, tant pour la fabrication que pour la maintenance. Les méthodes de contrôle non-destructif doivent être rapides, précises, fiables, économiques et de plus en plus automatisées. La complémentarité des différentes techniques d'inspection suggère leur utilisation simultanée pour renforcer la fiabilité des informations ou permettre une détection automatique difficile avec une seule technique.

Dans ce travail de thèse, nous explorons l'utilisation des méthodes d'apprentissage profond et de synthèse d'images pour la détection et la localisation de fissures par thermographie infrarouge laser Flying Spot sur des matériaux métalliques. Notre première contribution se concentre sur la collecte de données sur banc d'essai et la génération d'images synthétiques utilisant des modèles de diffusion pour augmenter la quantité et la diversité d'images thermographiques disponibles. La deuxième contribution concerne l'utilisation ces jeux de données au travers de protocoles d'entraînement progressif et de transfert de caractéristiques afin d'améliorer la capacité des réseaux de neurones à discriminer et localiser les endommagements sur les images thermiques Flying Spot. Enfin, notre troisième contribution porte sur la construction d'une nouvelle architecture d'apprentissage profond pour la fusion multi-spectrale infrarouge-visible et sa mise en œuvre pour la détection de défauts sur des données couplées thermographiques et visibles. Cette thèse illustre le bénéfice de la fusion multi-spectrale au travers des méthodes d'apprentissage profond, en particulier dans le cadre de la thermographie laser.

Title : Multi-sensor learning for material non destructive testing
Keywords : machine learning, data fusion, non-destructive testing, Deep Learning

Abstract : Defect detection in aeronautical and space structures is crucial for both manufacturing and maintenance. Non-destructive testing methods must be fast, precise, reliable, cost-effective, and increasingly automated. The complementarity of different inspection techniques suggests their simultaneous use to enhance information reliability or enable automatic detection that would be challenging with a single technique.

In this thesis, we explore the use of deep learning and image synthesis methods for the detection and localization of cracks using laser Flying Spot infrared thermography on metallic materials. Our first contribution focuses on data collection on test benches and the generation of synthetic images using diffusion models to increase the quantity and diversity of available thermographic images. The second contribution involves using these datasets with progressive training protocols and feature transfer to enhance the neural networks' ability to discriminate and localize damage in Flying Spot thermal images. Finally, our third contribution proposes the construction of a new deep learning architecture for infrared-visible multi-spectral fusion and its implementation for defect detection on coupled thermographic and visible data. This thesis illustrates the benefit of multi-spectral fusion through deep learning methods, particularly in the context of laser thermography.