UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

RECONNAISSANCE DES MATÉRIAUX PAR TEXTURE ET RÉFLECTIVITÉ MULTISPECTRALE DE SURFACE DANS LES ENVIRONNEMENTS NON CONTRÔLÉS

MÉMOIRE PRÉSENTÉ

COMME EXIGENCE PARTIELLE GÉNIE MÉCANIQUE

PAR

AHMED DHAHRI

SEPTEMBRE 2024

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

MAITRISE EN INGENIERIE-CONCENTRATION GENIE MECANIQUE (M. SC. A)

Direction de recherche :

Pr. Sousso Kelouwani	Directeur de recherche
Pr. Usef Faghihi	Codirecteur de recherche

Jury d'évaluation

Pr. Jean-Sébastien Dessureault	Evaluateur externe
Pr. Mohamed Habibi	Evaluateur externe
Pr. Sousso Kelouwani	Directeur de recherche

Résumé

La reconnaissance des matériaux en environnements non contrôlés est cruciale pour des domaines variés tels que l'infographie, la robotique, la réalité augmentée et le recyclage. Dans le domaine de la robotique, identifier les matériaux d'un objet permet aux bras mécaniques d'ajuster leurs méthodes de manipulation, améliorant ainsi la sécurité. De même, pour les véhicules autonomes, la compréhension des matériaux des surfaces peut optimiser la navigation en adaptant le comportement du véhicule aux conditions du terrain. Bien que nombreuses techniques de reconnaissance des matériaux existent, qui combinent la vision par ordinateur avec divers capteurs, il reste un défi crucial de trouver un équilibre optimal entre la généralité des techniques de vision, qui permettent une reconnaissance plus universelle des matériaux, et la précision des capteurs, qui peut être indispensable pour certaines applications spécifiques comme la robotique industrielle ou la navigation autonome.

Dans le contexte de ce projet, une nouvelle méthode basée sur un système d'imagerie multimodale est proposée. Le modèle utilisé combine le mécanisme d'attention et la convolution pour la reconnaissance des matériaux avec une plus grande précision. Cette approche se distingue des méthodes traditionnelles, basées sur l'image visuelle et multispectrale, en exploitant à la fois la texture et les caractéristiques de réflectivité propres aux surfaces des matériaux. Notre système utilise une combinaison de caméras — incluant une caméra de profondeur, une caméra RVB conventionnelle, une caméra proche-infrarouge, ainsi qu'un projecteur laser infrarouge. Ce système conçu spécifiquement pour capturer les textures visuelles et la distribution des réflexions à travers le spectre visible-infrarouge. Les données recueillies sont ensuite analysées par un modèle de fusion pour la reconnaissance des matériaux.

Les tests et évaluations montrent que notre modèle surpasse nettement les techniques existantes en termes de précision de reconnaissance des matériaux lorsque toutes les modalités sont utilisées ensemble (image de profondeur, image proche-infrarouge et projections laser infrarouge). Même lorsque limité aux images RVB seules, le modèle restant compétitif avec les meilleurs systèmes basés uniquement sur ces données, illustrant la robustesse et la flexibilité de notre approche. Cela démontre l'efficacité de l'intégration des différentes modalités avec les informations visuelles.

Remerciements

Je tiens à exprimer ma profonde gratitude au Professeur Sousso Kelouwani, du département de génie mécanique à l'UQTR, pour ses conseils avisés et son soutien constant tout au long de ce projet.

Un merci particulier à M. Ali Amamou, chercheur postdoctoral à l'Institut de Recherche sur l'Hydrogène, dont la disponibilité et les insights précieux lors de nos discussions ont grandement enrichi ce travail.

Je remercie également tous les collègues du laboratoire pour leurs échanges stimulants qui ont, de près ou de loin, façonné ce projet.

Enfin, je suis infiniment reconnaissant envers ma famille et mes amis pour leur soutien sans faille et leurs encouragements continus durant mes études.

Table des matières

Liste des tableaux

Liste des figures

Liste des acronymes

ACRV Apprentissage Contrastif de la Représentation Visuelle. 72

ACSOR Capteur capacitif adaptatif pour la télémétrie d'obstacles. x, 5

CCD Dispositif à couplage de charge. 6

CMOS Métal-oxyde-semiconducteur complémentaire. 7

CNN Réseaux de neurones convolutionnels. 19

CRF Champ Aléatoire Conditionnel. 77

CUReT Base de données sur la réflectance et la texture de Columbia-Utrecht. 38

FDRB Fonction de distribution de la réflectance bidirectionnelle. 37

FDRBVE Fonction de Distribution de Réflectance Bidirectionnelle Variable dans l'Espace. 45

FDRDSB Fonction de Distribution de Réflectance de Dispersion de Surface Bidirectionnelle. 44

FMD Base de données des matériaux Flickr. 34

FTB Fonction de Texture Bidirectionnelle. 44

IR Infrarouge. 8, 9

KTH-TIPS Textures sous différents éclairages, poses et échelles. 39

LWIR Infrarouge à ondes longues. 8

MINC Jeu de données des matériaux en contexte. 34

MWIR Infrarouge à ondes moyennes. 8

NWIR Infrarouge à ondes proches. 8

SWIR Infrarouge à ondes courtes. 8

ToF Temps de vol. x, 7, 9

ViT Transformateur de vision. 22

Chapitre 1 - Introduction

1.1 Contexte général

La reconnaissance des matériaux est essentielle dans de nombreux domaines et trouve son application dans une vaste gamme de contextes, allant de la navigation robotique [17], qui permet aux robots de se déplacer de manière adaptative dans leur environnement, aux processus industriels tels que le recyclage [18], où elle aide à identifier et à trier différents matériaux, ou au contrôle qualité [19]. De plus, elle enrichit les expériences de réalité augmentée en permettant une reconstruction plus précise et réaliste des éléments de l'environnement réel [20].

L'un des défis majeurs dans la reconnaissance des matériaux réside dans le compromis entre la versatilité offerte par la vision par ordinateur et l'efficacité des capteurs spécifiques. La vision par ordinateur permet une grande flexibilité, car elle ne nécessite pas un environnement contrôlé ni une configuration matérielle complexe pour être déployée. Cette capacité à fonctionner dans des conditions non préparées ou non contrôlées ouvre les portes à une large gamme d'applications pratiques. Cependant, cette versatilité peut parfois se faire au détriment de la précision et de l'efficacité que peuvent offrir des capteurs spécialisés conçus pour des tâches spécifiques. En effet, les algorithmes de vision par ordinateur, bien qu'adaptables à de nombreuses situations, s'appuient généralement sur des caractéristiques visuelles généralisées, telles que la couleur, la texture ou la forme [21]. Ces informations peuvent ne pas être suffisamment spécifiques pour différencier certains matériaux dans des conditions environnementales complexes (lumière, poussière, angles de vue). Par contraste, les capteurs spécialisés, comme les spectromètres ou les capteurs tactiles, sont optimisés pour détecter des propriétés physiques ou chimiques précises, mais au prix d'une limitation dans leur capacité à s'adapter à des environnements variés. Ainsi,

la vision par ordinateur offre une portée plus large en termes d'applications, mais peut manquer de la précision requise pour des scénarios où une détection fine et spécifique est nécessaire, comme dans le contrôle qualité de matériaux sensibles ou la manipulation robotique de haute précision [19].

La détection à distance s'avère particulièrement bénéfique dans le secteur industriel, offrant une réduction significative du temps d'opération et limitant la nécessité de tests destructifs. Cette capacité à acquérir des données détaillées sans interaction physique directe avec les matériaux cibles constitue un atout majeur, surtout dans des domaines tels que la gestion des déchets, le contrôle de qualité dans l'industrie pharmaceutique et la manipulation robotique.

En somme, la reconnaissance des matériaux dans des environnements non contrôlés est une technologie en pleine expansion, offrant des perspectives prometteuses pour l'amélioration des capacités robotiques, l'optimisation des processus industriels et l'enrichissement des expériences de reconstruction graphique de l'environnement. Le développement continu dans ce domaine est essentiel pour surmonter les défis liés au compromis entre versatilité et efficacité, ainsi que pour maximiser les avantages de la télédétection dans les applications industrielles [19].

1.2 Enjeux scientifiques de la reconnaissance des matériaux

La reconnaissance des matériaux peut être optimisée en utilisant des capteurs et des systèmes de vision, chacun offrant des avantages uniques adaptés à diverses applications. Les capteurs se distinguent par leur fiabilité et leur précision dans la détection des caractéristiques physiques des matières. Cependant, leur utilisation dans un environnement non contrôlé présente des défis majeurs en raison de diverses interférences

environnementales telles que les conditions limoneuses, l'orientation et la position de la cible, la complexité de l'environnement, la poussière, etc. D'autre part, la vision par ordinateur excelle en termes de versatilité, permettant l'identification d'une large gamme de caractéristiques visuelles comme la forme, la couleur et la texture, sans contact direct avec l'objet analysé.

1.2.1 Critiques des méthodes conventionnelles

Les capteurs de proximité sont conçus pour identifier la présence d'un élément, soit par contact direct, soit à une faible distance, offrant ainsi une flexibilité d'application remarquable. Leur utilisation, qu'elle soit isolée ou en combinaison avec d'autres capteurs, contribue à accroître la précision et la fiabilité dans l'identification des matériaux. Le choix spécifique du capteur dépend fortement de l'application visée et des caractéristiques uniques des matériaux à détecter. De nombreux phénomènes physiques peuvent être exploités pour l'identification des matériaux, chacun offrant une méthode adaptée à l'application ciblée:

- **Phénomène électromagnétique :** Divers types de capteurs exploitent les phénomènes électromagnétiques pour fonctionner. Parmi eux, les capteurs à effet Hall, illustrés dans la figure 1.1, les capteurs magnéto-résistifs et les magnétomètres à induction magnétique sont notables, comme souligné dans Lenz (1990) [22]. Ces dispositifs sont conçus pour détecter la présence de champs magnétiques autour des matériaux, ce qui les rend particulièrement efficaces pour l'identification des métaux ferreux. En revanche, leur efficacité diminue considérablement pour les matériaux non ferreux.
- **Phénomène diélectrique :** L'identification des matériaux via des capteurs capacitifs se base sur l'analyse des variations de la capacité électrique induites par

Figure 1.1 Capteur magnétique à effet Hall [1].

la présence de matériaux au voisinage des électrodes du capteur. Ces dispositifs exploitent le fait que la permittivité électrique — une propriété intrinsèque des matériaux qui influence leur comportement dans un champ électrique — modifie la capacité mesurée par le capteur. Lorsqu'un matériau, typiquement un isolant, s'intercale entre les électrodes, il modifie le champ électrique établi, entraînant un changement dans la capacité détectée. Cette propriété permet aux capteurs capacitifs de fournir des données précieuses, capables de distinguer une large gamme de matériaux [2] Cependant, ces capteurs, comme présentés dans la figure 1.2, sont sensibles à l'hétérogénéité des matériaux et aux interférences des conditions non contrôlées, telles que la poussière et l'humidité.

- **Spectroscopie :** Ces capteurs utilisent généralement l'infrarouge pour identifier les matériaux en fonction de leur absorption ou réflexion de la lumière IR, une technique fréquemment employée dans la spectroscopie infrarouge. Les spectromètres effectuent des tests élémentaires et des analyses spectrochimiques de divers matériaux sous conditions très variables. Le spectromètre, illustré dans la figure (1.3), analyse les longueurs d'onde et l'intensité de la lumière réfléchie ou émise par une décharge d'étincelle ou une source de rayonnement sur l'échantillon

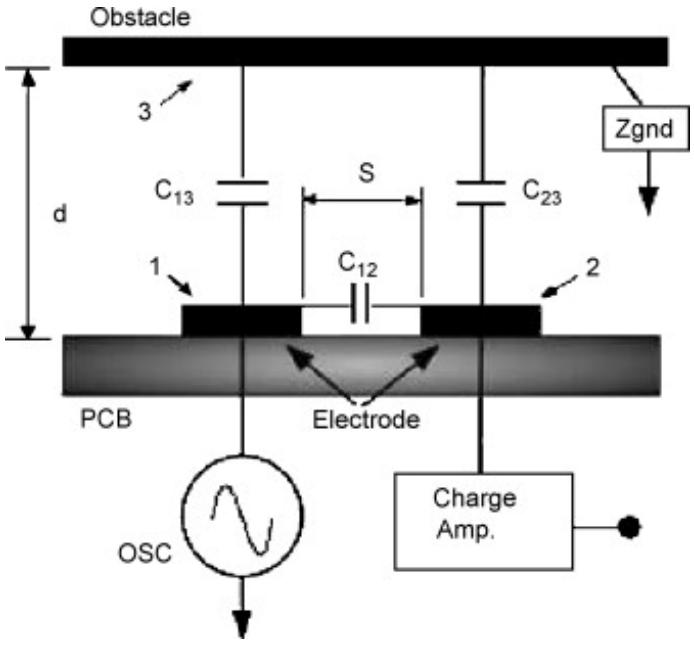


Figure 1.2 Capteurs capacitifs ACSOR [2].

pour identifier et quantifier les éléments présents. La spectroscopie explore les vibrations moléculaires; les groupes fonctionnels des atomes sont associés à des bandes d'absorption infrarouge caractéristiques, qui correspondent aux vibrations fondamentales des groupes [23]. Avec l'accès à une vaste base de données de matériaux, une unité de calcul compare le spectre reçu aux modèles enregistrés [24]. C'est une méthode scientifique très précise, mais elle nécessite des conditions strictement contrôlées.

- **Phénomène acoustique :** Ces capteurs (figure 1.4) utilisent les ondes sonores pour étudier les propriétés des matériaux. La manière dont les ondes sonores sont réfléchies, transmises ou absorbées par un matériau peut révéler des informations sur sa structure, sa densité et sa composition. Le capteur à ultrasons, en particulier, exploite la différence d'énergie du signal d'écho ultrasonore renvoyé par les surfaces pour identifier les matériaux [4].

Figure 1.3 Spectromètre xSort [3].

Chaque type de capteur de proximité possède des fonctions distinctes qui le rendent particulièrement adapté à certaines applications. Le choix du capteur idéal est influencé par divers facteurs, notamment les propriétés physiques des objets à détecter et les exigences spécifiques de l'application, comme la similarité aux autres matériaux et la portée de la cible à détecter. Pour assurer le bon fonctionnement de ces capteurs, il est nécessaire de mettre en place un environnement adapté qui réponde à leurs besoins spécifiques, tel qu'une distance de détection optimale. Cette nécessité d'adapter l'environnement aux spécificités des capteurs peut effectivement rendre leur déploiement coûteux, notamment en termes d'installation. De plus, cette contrainte peut compliquer l'intégration des capteurs dans des environnements complexes, où les conditions idéales pour leur fonctionnement ne sont pas toujours réalisables.

1.2.2 Apport des méthodes d'imagerie

Les caméras sont des dispositifs opto-électroniques conçus pour la capture des images adaptées à une large gamme d'applications. Elles transforment le rayonnement capté par des systèmes optiques en signaux électriques grâce à des capteurs tels que les CCD

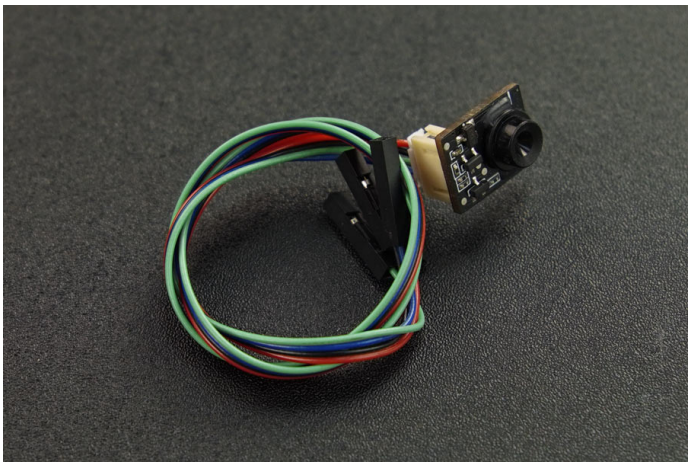


Figure 1.4 Capteur ultra-sonore ToF [4].

ou CMOS, permettant l'observation de phénomènes visibles ou invisibles à l'œil nu. Cette capacité dépend de la spécificité de la caméra, de la gamme spectrale qu'elle cible (infrarouge, ultraviolet, infrarouge thermique) et de la technique d'imagerie employée (imagerie de profondeur, imagerie ToF, imagerie de champ lumineux). Parmi les différents types de caméras, on peut notamment citer :

- **Imagerie visuelle :** Pour créer l'image d'une scène, la caméra visuelle utilise une série de lentilles qui concentrent les rayons lumineux sur un capteur comme montré dans la figure (1.5). Ce capteur échantillonne la lumière et enregistre les informations sous forme électronique, qui sont ensuite converties en données numériques. L'imagerie offre des compromis entre performances, complexité et coût. Il n'est donc pas surprenant que l'imagerie par caméra numérique soit considérée comme l'un des domaines de recherche et d'application les plus dynamiques, et que de nombreux produits commerciaux capitalisant sur ses principes aient déjà émergé dans diverses applications sur le marché. La popularité extrême et toujours croissante des appareils photo numériques grand public à capteur unique stimule les activités de recherche dans les domaines de l'acquisition,

du traitement et du stockage d'images numériques couleur [25]. Dans le contexte de la reconnaissance d'images, il est notable que les bases de données d'images visuelles constituent les ressources les plus vastes, diversifiées et accessibles au public, alimentant ainsi les avancées dans ce domaine [26].

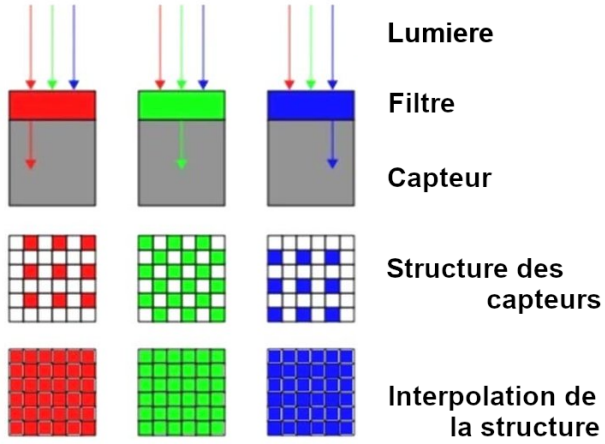


Figure 1.5 Structure des capteurs de la caméra visuelle [5].

- **Imagerie infrarouge et thermique :** La bande IR est généralement divisée en bande réflective et thermique. Les bandes réflectives et thermiques sont classées en infrarouge proche (NWIR; 0,75 à 2,4 μm), infrarouge à ondes courtes (SWIR; 0,9 à 2,4 μm), infrarouge à ondes moyennes (MWIR; 3 à 5 μm) et à ondes longues (LWIR; 8–14 μm). La lumière NWIR et SWIR est produite par des objets chauds tels que des ampoules ou des lasers IR, et n'est pas visible par les humains. Les caméras d'imagerie thermique-infrarouge ne nécessitent aucun éclairage artificiel. Pour cette raison, souvent qualifiées de passives (figure 1.6) pour une illustration d'une caméra thermique). La plupart des objets à température ambiante, y compris le corps humain, émettent de l'énergie thermique IR, ce qui constitue la principale différence entre l'imagerie dans les bandes IR réflectives et thermiques [27].
- **Imagerie de profondeur :** Deux principes de fonctionnement pour ces systèmes

Figure 1.6 Caméra d'imagerie thermique FLIR A38 [6].

d'imagerie de profondeur à faible coût ont été déployés avec succès ces dernières années. D'une part, les capteurs de lumière structurée, qui sont des systèmes stéréo actifs, utilisent un projecteur infrarouge (IR) pour projeter un motif lumineux connu sur la scène. Une caméra infrarouge détecte le motif déformé et estime la disparité en utilisant des algorithmes stéréo classiques. Une gamme de matériels peu coûteux et légers basés sur ce principe a été lancée, incluant des dispositifs tels que Microsoft Kinect, Asus Xtion Pro Live, Intel RealSense (Figure 1.7), Primesense Carmine, et Occipital Structure Sensor. D'autre part, les capteurs de temps de vol (ToF), moins courants tels que le Kinect V2, émettent un signal lumineux et mesurent le temps nécessaire pour voyager jusqu'à l'objet et revenir. Les cartes de profondeur obtenues de ces capteurs contiennent souvent du bruit significatif, ce qui rend généralement nécessaire un prétraitement de la profondeur [28].

Figure 1.7 Caméra de profondeur Intel RealSense D435 [7].

- **Imagerie multispectrale et hyper-spectrale :** L'imagerie hyper-spectrale est une méthode utilisée en télédétection, capturant le spectre électromagnétique des longueurs d'onde visibles aux infrarouges. Cette technique utilise des capteurs pour générer de multiples bandes spectrales distinctes d'une scène, transformant chaque pixel d'une image hyper-spectrale en un vecteur multidimensionnel qui reflète les intensités spectrales variées à des longueurs d'onde spécifiques. Cette capacité à détecter de fines variations spectrales permet des applications étendues dans divers secteurs. Les recherches récentes montrent que la classification de ces images, qui catégorise chaque pixel selon sa signature spectrale unique, est un domaine de recherche dynamique, suscitant un intérêt significatif dans le domaine de la télédétection. Le processus de classification, cependant, rencontre deux obstacles principaux : la variabilité des signatures spectrales due à des facteurs tels que l'éclairage et la limite de disponibilité des bases de données publiques face à la haute dimensionnalité des données hyper-spectrales. Ce dernier problème conduit souvent à des difficultés dans l'apprentissage des modèles et affecte la performance des algorithmes de classification [29].
- **Imagerie acoustique et ultra-sonore :** Les caméras acoustiques, illustrées dans la figure 1.8, peuvent aider à localiser les fuites sous pression dans les systèmes à air comprimé ou à détecter les décharges partielles dans les systèmes électriques

haute tension. Construites avec plusieurs microphones, elles produisent une image acoustique précise qui affiche visuellement des informations ultrasoniques, même dans des environnements industriels bruyants. L'image acoustique est souvent superposée en temps réel sur l'image d'un appareil photo numérique, facilitant la localisation précise de la source sonore [30].

Figure 1.8 Caméra d'imagerie acoustique FLIR Si124 [6].

Contrairement aux capteurs de proximité, qui sont limités dans leur capacité à gérer des environnements complexes, les caméras et les systèmes d'imagerie offrent une flexibilité bien plus grande dans l'analyse des environnements non contrôlés. Ces systèmes surpassent les capteurs traditionnels en exploitant des algorithmes d'intelligence artificielle et d'apprentissage automatique pour interpréter des informations visuelles complexes. Bien qu'ils ne soient pas idéaux pour mesurer avec précision les propriétés physiques spécifiques des matériaux, ces dispositifs apportent une compréhension

approfondie des aspects visuels et du contexte, fournissant une vue globale de l'environnement.

1.3 Problématique de recherche

La précision dans l'identification des matériaux est cruciale pour améliorer l'efficacité dans des applications variées telles que le recyclage, la réalité augmentée et divers processus industriels. Une identification précise est indispensable, particulièrement pour le tri des déchets destinés au recyclage. Face aux défis posés par les environnements non contrôlés, caractérisés par une variabilité de l'éclairage et des orientations d'objets imprévisibles, une plus grande précision dans la reconnaissance des matériaux est exigée pour plusieurs applications. Les systèmes de vision, qui s'appuient sur des indices visuels tels que la couleur, la texture et des informations contextuelles comme la reconnaissance d'objets et de scènes, offrent une adaptabilité accrue dans ces environnements complexes. Cependant, leur efficacité peut être compromise par des matériaux visuellement similaires et des conditions d'éclairage variables. Ainsi, améliorer la robustesse et la précision de la reconnaissance des matériaux en enrichissant les systèmes de vision avec des modalités complémentaires fournissant des informations supplémentaires non redondantes reste un sujet de recherche ouvert.

1.4 Objectifs de recherche

Dans le domaine de la reconnaissance et de l'identification des matériaux, l'absence d'une méthodologie assez robuste pour identifier précisément les matériaux représente une lacune significative dans les contextes environnementaux non contrôlés. Ce projet vise à élaborer une nouvelle architecture de système basée sur l'imagerie multispectrale et multimodale pour la reconnaissance des matériaux dans ces environnements. Pour réaliser

l'objectif principal mentionné, les sous-objectifs suivants doivent être abordés :

- Compréhension du problème de reconnaissance des images.
- Exploration des modalités utilisées.
- Proposition de modalités complémentaires pour la reconnaissance des matériaux.
- Proposition et développement de bases de données pour les différentes phases d'apprentissage.
- Évaluation des performances et validation expérimentale.

1.5 Méthodologie

Pour atteindre ces objectifs, une séquence chronologique est mise en place tout au long de ce projet. La démarche méthodologique adoptée pour ce travail se divise essentiellement en quatre chapitres principaux :

- Chapitre 1 - Introduction : Présenter le cadre applicatif de la recherche en matière de reconnaissance des matériaux, ainsi que les outils utilisés dans ce secteur, tant dans l'industrie que dans la recherche scientifique.
- Chapitre 2 - État de l'art: Citer les principaux travaux de recherche sur la reconnaissance des matériaux avec les caméras et les technologies de l'imagerie, en se concentrant sur les environnements non-contrôlés.
- Chapitre 3 - Conception de système et choix de modalités: Comparer les modalités les plus performantes dans le contexte de la reconnaissance des matériaux, puis, choisir et adapter ces méthodes pour la reconnaissance dans les environnements non-contrôlés.
- Chapitre 4 - Validation expérimentale: Présenter les aspects techniques, ainsi que les résultats expérimentaux qui valident le choix des modalités et la conception.

- Conclusion et perspectives: Conclure la totalité des étapes faites dans cette recherche. Puis, donner quelques perspectives pour les travaux qui peuvent être faits dans le futur.

Chapitre 2 - État de l'art

La reconnaissance des matériaux est considérée comme un objectif spécifique de la reconnaissance des images. Il est donc logique de porter notre attention sur les concepts initiés dans ce domaine. Ce chapitre traitera essentiellement de la problématique de la reconnaissance, des bases de données et des modèles de vision impliqués dans cet objectif.

2.1 Reconnaissance des images

Définition de la reconnaissance dans la vision : Les architectures de base, souvent appelées backbones, fournissent initialement une fonction de reconnaissance globale, qui est ensuite personnalisée pour des tâches spécifiques telles que la classification, la détection ou la segmentation. Cette personnalisation est réalisée en intégrant un module approprié à la tâche. La construction de systèmes de vision implique généralement une phase d'extraction des caractéristiques (feature extraction), qui alimente un module spécifique à une tâche. L'architecture de base peut générer un volume dense de données pour la détection et la localisation d'objets, comme illustré dans la figure (2.1). Chaque pixel représente une probabilité prédictive calculée en fonction de la région d'image correspondante, fournissant ainsi une carte de probabilité de la présence des objets dans l'image.. Bien que les architectures de base puissent être développés à partir de zéro sur des données spécifiques à une tâche, de nombreux architectures de base disponibles sont pré-entraînés sur de grands ensembles de données, puis affinés (fine-tuning) pour la tâche spécifique une fois intégrés dans la forme finale du réseau. Cette méthode, appelée apprentissage par transfert (transfer learning), offre plusieurs avantages. Premièrement, elle réduit considérablement les besoins en données spécifiques requises pour les applications de l'apprentissage profond a permis d'améliorer les performances sur un large éventail d'applications. Deuxièmement, elle peut accélérer l'apprentissage et réduire les

coûts de calcul, même lorsque de grandes quantités de données spécifiques à une tâche sont disponibles [8].

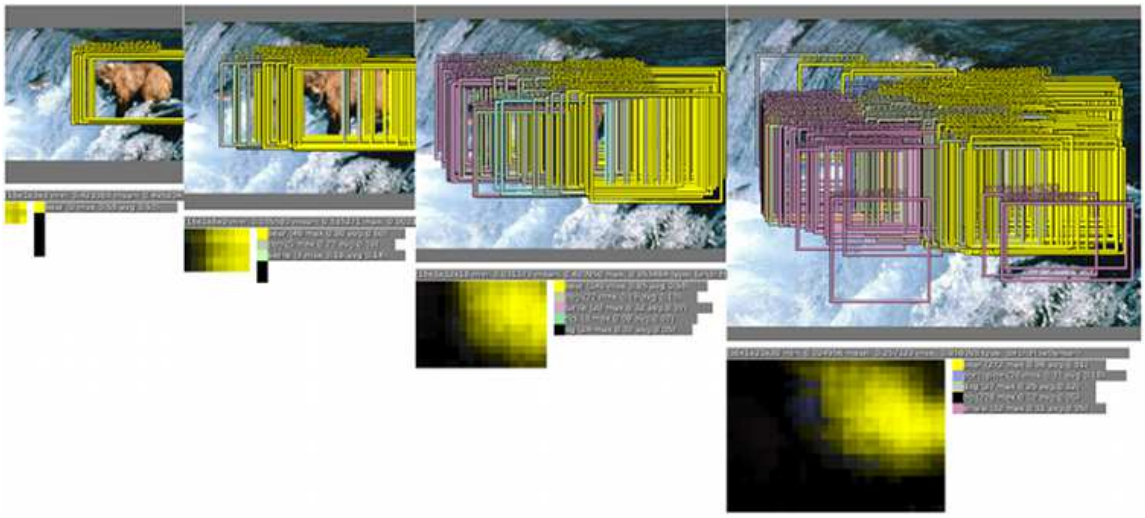


Figure 2.1 Des exemple de l'exécution d'une prédiction dense (dense prediction) d'une architecture de base pour une tâche de détection en variant la résolution de l'image [8]. Les rectangles sur l'image décrit les région de l'image ou il y a une forte probabilité d'existence d'un objet. La carte au dessous de chaque image est la carte de probabilité des objets correspondante.

Les sous-tâches de la reconnaissance dans la vision : L'apprentissage profond en vision par ordinateur représente une évolution moderne des réseaux neuronaux artificiels. Il comprend plusieurs tâches potentielles que peuvent accomplir les modèles de vision. Voici quelques-unes des tâches les plus courantes, présentées dans la figure (2.2) :

- La classification des images: La tâche de classification d'images consiste à étiqueter des images entrées avec une probabilité de présence d'une classe d'information visuelles particulière (chien, voiture, chat, texture, ligne de traçage dans la rue, ...). Plus précisément, étant donné un ensemble d'image $I = I1, ..., In$ et un ensemble de vecteurs d'étiquettes $Y = y1, ..., yK, yi \in 0, 1n1$ (où K représente le nombre de classes), la tâche est de produire un ensemble de prédictions $\hat{Y} = \hat{y}1, ..., \hat{y}K$, qui

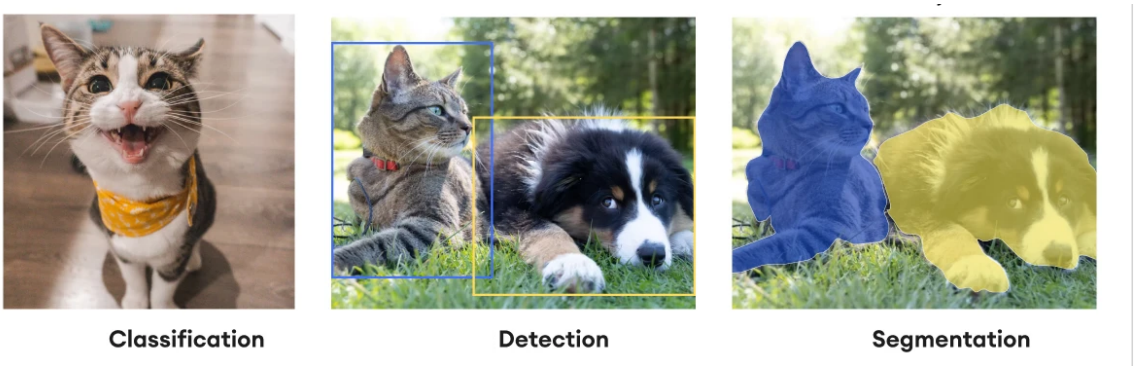


Figure 2.2 Types des sous-tâches de la reconnaissance d'objets [9].

correspondent autant que possible à Y. Comme ce type de tâches est typiquement résolu en utilisant des techniques d'apprentissage automatique, les ensembles Y et I sont divisés en sous-ensembles de test et d'entraînement, tandis que la performance d'une méthode de classification donnée est évaluée sur la partie de test [31].

- La détection: La tâche de détection d'objets est étroitement liée à celle de la classification d'images. La principale différence réside dans le fait qu'en plus de fournir des informations sur la présence ou l'absence d'une classe donnée dans une image particulière, il faut également extraire la position d'une instance (ou de plusieurs instances) de la classe (généralement sous forme de boîte englobante). Une détection est considérée comme un vrai positif lorsque la boîte englobante produite a un chevauchement suffisamment important avec un objet réel [31].
- La segmentation: La segmentation sémantique d'images implique de faire des prédictions au niveau des pixels, produisant des masques de sortie 2-dimensionnel. Ce processus divise les images en fonction des informations distinctes qu'elles contiennent [32].
- La récupération d'images: La récupération d'images implique la collecte d'images ayant le même objet. De nombreux modèles de vision sont utilisés et offrent de bonnes performances par rapport aux méthodes traditionnelles telles que VLAD

et le vecteur de Fisher. En utilisant des modèles de vision, un large ensemble de données peut être utilisé dans l'apprentissage d'un modèle de vision. Puis, ce modèle est utilisé pour l'extraction des descripteurs et peut également être appliqué à la recherche d'images pour obtenir une meilleure précision [32].

Apprentissage des modèles : L'apprentissage supervisé en vision par ordinateur implique la formation de modèles sur un ensemble de données étiquetées, où chaque exemple est associé à une annotation spécifiant le résultat attendu, tel que l'étiquette d'un objet dans une image. Cette méthode est largement utilisée pour des tâches telles que la classification d'images, la détection d'objets, et la segmentation sémantique. L'apprentissage supervisé est particulièrement efficace lorsque les données abondent. En l'absence de données suffisantes, des techniques d'adaptation au domaine, telles que l'apprentissage par transfert ou le réglage fin, sont employées. Ces méthodes permettent d'améliorer les performances dans un domaine cible en transférant des connaissances acquises dans un domaine source [33]. La connaissance du domaine source peut être obtenue par un apprentissage supervisé sur un ensemble de données pertinentes ou par un apprentissage semi-supervisé. Ce dernier est fréquemment utilisé dans la formation des grands modèles de Transformateur, exploitant de vastes ensembles de données collectées via le web scraping [34, 35]. Les méthodes d'apprentissage semi-supervisé reposent sur plusieurs hypothèses fondamentales [33] :

- Hypothèse de stabilité (Smoothness Assumption): Si deux points x_1, x_2 résidant dans une région à forte densité sont proches, leurs résultats correspondants y_1, y_2 devraient l'être également.
- Hypothèse de groupement (Cluster Assumption): Si des points se trouvent dans le même groupe, il est probable qu'ils appartiennent à la même classe.

- Hypothèse de manifold (Manifold Assumption): Les données (à haute dimension) se situent (grossièrement) sur un manifold de basse dimension.

En résumé, la reconnaissance constitue une fonction essentielle de la vision par ordinateur, servi par une architecture de base dans le domaine de l'apprentissage profond. Cette architecture de base a pour rôle de créer des caractéristiques pertinentes, lesquelles seront ensuite employées dans diverses tâches spécifiques en aval.

2.2 Modèles de reconnaissance d'images

Ils existent plusieurs types de modèles de reconnaissance de vision, qui exploitent différents types d'éléments et d'architectures. En premier lieu, les réseaux de neurones convolutifs sont très similaires aux réseaux de neurones traditionnels. Ils sont constitués de neurones qui ont des poids et des biais apprenables. Chaque neurone reçoit des entrées, effectue un produit scalaire et finit avec une non-linéarité. Les architectures du réseau neuronal convolutif (CNN) supposent explicitement que les entrées sont des images, ce qui nous permet d'intégrer certaines propriétés dans l'architecture. Ceux-ci rendent alors la fonction de l'action directe du réseau (forward function) plus efficace à mettre en œuvre et réduisent considérablement le nombre de paramètres dans le réseau [36]. Nous citons ici certains modèles CNN performants:

- *ResNet:* Ce réseau adopte l'apprentissage résiduel toutes les quelques couches empilées (voir figure 2.3) Un élément de base est défini comme dans l'équation (2.1):

$$y = F(x, W_i) + x \tag{2.1}$$

Ici x et y sont les vecteurs d'entrée et de sortie des couches. La fonction $F(x, W_i)$

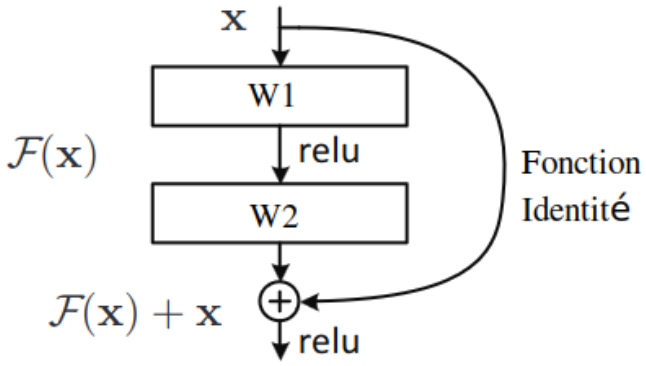


Figure 2.3 Bloc résiduel de ResNet [10].

représente la cartographie résiduelle à apprendre et les biais sont omis pour simplifier les notations. W_i représente l'ensemble des paramètres du réseau.

Dans la figure (2.3), l'exemple comporte deux couches, $F = W_2 \cdot \sigma(W_1 \cdot x)$ dans lequel σ désigne ReLU. L'opération $F + x$ est effectuée par une connexion raccourcie et une addition élément par élément. Finalement, la deuxième non-linéarité est ajoutée après l'addition [10].

Resnet a obtenu d'excellents résultats grâce à des réseaux résiduels extrêmement profonds sur l'ensemble de données de classification ImageNet-1k. Le réseau résiduel à 152 couches ResNet152 était le réseau le plus profond présenté sur ImageNet, lors de sa proposition en 2015. Bien que le nombre de couches élevé, ce réseau a toujours une complexité inférieure à celle des VGGnets précédents. Il affiche 3,57% d'erreur top-5 sur l'ensemble de tests ImageNet et a remporté la première place au concours de classification ILSVRC 2015. Les représentations extrêmement profondes ont également d'excellentes performances de généralisation sur d'autres tâches de reconnaissance [10].

- *EfficientNet:* Il existe de nombreuses façons de dimensionner un CNN en fonction des contraintes de ressources. La profondeur du réseau peut être augmentée ou réduite en ajustant le nombre de couches, comme nous le voyons dans des variantes telles que ResNet-18 et ResNet-200 de la famille ResNet. En ajustant le nombre de canaux, WideResNet et MobileNets peuvent être dimensionnés en fonction de la largeur du réseau. Il est également bien connu qu'une plus grande résolution de l'image d'entrée améliore la précision en raison de l'augmentation du nombre d'opérations en virgule flottante. Bien que des études antérieures aient montré que la profondeur et la largeur du réseau sont toutes les deux importantes pour le pouvoir d'expression des ConvNets, EfficientNet fournit une fonction éprouvée efficace pour la mise à l'échelle des ConvNet pour les trois dimensions de largeur, de profondeur et de résolution du réseau. La méthode de dimensionnement composée proposée utilise un coefficient composé ϕ pour dimensionner uniformément la largeur, la profondeur et la résolution du réseau, comme indiqué dans les équations (2.2)

$$\begin{aligned} depth : d &= \alpha^{\phi} \\ width : w &= \beta^{\phi} \\ resolution : r &= \gamma^{\phi} \\ \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2\alpha, \beta, \gamma \geq 1 \end{aligned} \tag{2.2}$$

où α, β, γ sont des constantes qui peuvent être déterminées par une recherche par grille. ϕ est un coefficient défini par l'utilisateur qui contrôle le nombre de ressources supplémentaires disponibles pour le dimensionnement du modèle,

tandis que α, β, γ spécifient comment affecter ces ressources supplémentaires à la largeur, à la profondeur et à la résolution du réseau, de manière respective [11].

En plus des CNNs, ils existent les Transformateurs de vision. Les réseaux neuronaux Transformateurs sont capables d'attendre des performances égales ou supérieures à celles des CNN pour les tâches de classification d'images dans des bases de données à grande échelle. Ces transformateurs de vision (ViT) fonctionnent presque de la même manière que les transformateurs utilisés dans les tâches de traitement de Langage naturel, en utilisant l'auto-attention, plutôt que la convolution, pour agréger les informations à travers toute la zone de l'image. Certains modèles de Transformateurs de vision sont présentés ci-dessous.

- *ViT:* inspiré par le succès de la dimensionnement des Transformateurs dans les tâches linguistiques, ViT a appliqué un Transformateur standard directement aux images, avec le moins de modifications possibles. Pour ce faire, l'image est divisée en morceaux et la séquence linéaire des embeddings de ces morceaux est fournie en entrée à un Transformateur. Les morceaux d'image sont traités de la même manière que les tokens (mots) dans une application de traitement du langage. Nous entraînons le modèle à la classification d'images de manière supervisée lorsqu'ils sont entraînés sur des ensembles de données de taille moyenne tels qu'ImageNet sans régularisation forte, ces modèles produisent des précisions modestes de quelques points de pourcentage en dessous des ResNets de taille comparable. Ce résultat apparemment décourageant est prévisible : Les transformateurs n'ont pas certains des biais inductifs inhérents aux CNN, tels que l'équivariance de la translation et la localité, et ne se généralisent donc pas bien lorsqu'ils sont entraînés sur des quantités insuffisantes de données. Cependant, la situation change si les modèles sont entraînés sur des ensembles de données plus importants (14 à 300

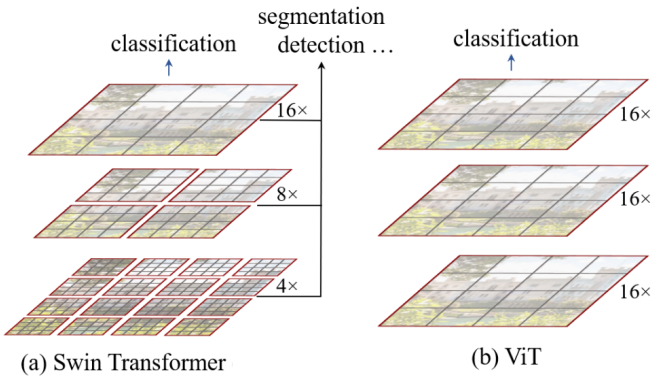


Figure 2.4 Différence entre l'architecture de Swin et ViT [11].

millions d'images). À ce niveau, l'entraînement à grande échelle l'emporte sur le biais inductif. ViT obtient d'excellents résultats lorsqu'il est pré-entraîné à une échelle suffisante et transféré à des tâches comportant moins de points de données.

- *Transformeur-Swin :* Ce modèle fonctionne en créant des cartes de données hiérarchiques et maintient une complexité de calcul linéaire par rapport à la taille de l'image. Il commence par traiter des morceaux de petite taille et fusionne progressivement les morceaux voisins dans des couches plus profondes du Transformateur. Cette approche hiérarchique permet au Transformeur-Swin d'intégrer des techniques avancées de prédiction dense, similaires à celles utilisées dans les réseaux de pyramides de données ou U-Net, tout en conservant une complexité de calcul linéaire. Ceci est réalisé en calculant l'auto-attention localement au sein de fenêtres non chevauchantes qui subdivisent l'image (voir figure 2.4). La conception hiérarchique du Transformeur-Swin, associée à l'approche innovante de fenêtres en décalage, en fait un outil polyvalent pour une variété de tâches de vision. Il offre une amélioration significative des performances dans divers domaines tels que la classification d'images, la détection d'objets et la segmentation sémantique [11].

- *Eva-02:* L'approche d'EVA-02 en matière d'apprentissage des représentations visuelles consiste à produire une représentation hiérarchique et robuste des données qui soit efficace en termes de ressources de calcul, ce qui la rend accessible à une recherche et à une application plus larges, sans nécessiter d'infrastructure importante. Elle démontre des performances supérieures dans diverses tâches de vision représentatives tout en utilisant beaucoup moins de paramètres et de coûts de calcul par rapport aux approches antérieures de pointe. Les innovations architecturales comprennent l'utilisation de RoPE 2D rotary position embedding, sub-LN comme couche de normalisation, swich activation comme réseau feedforward. Les variantes d'EVA-02, avec des tailles de modèle allant de 22M à 304M de paramètres, ont montré des performances impressionnantes dans des tâches telles que la classification d'images, la détection d'objets, la segmentation d'instances et la segmentation sémantique, surpassant souvent les modèles les plus grands. Par exemple, une variante de petite taille d'EVA-02 avec seulement 22 millions de paramètres atteint une précision de 85,8% sur la base des données de validation ImageNet-1K, tandis qu'un modèle plus grand avec 304 millions de paramètres atteint une précision exceptionnelle de 90,0% sur la base des données de validation [37].
- *DaViT:* Ce modèle introduit un mécanisme d'attention double, contenant l'attention de la fenêtre spatiale et l'attention du groupe de canaux, pour capturer le contexte global tout en maintenant l'efficacité de calcul. L'auto-attention à fenêtres spatiales multiples divise la dimension spatiale en fenêtres locales, où chaque fenêtre contient plusieurs jetons spatiaux. Chaque jeton est également divisé en plusieurs têtes. L'auto-attention à tête unique par groupe de canaux regroupe les jetons de canaux en plusieurs groupes. L'attention est portée sur chaque groupe de canaux

avec un canal entier au niveau de l'image comme jeton. Un jeton de canal qui capture des informations globales est également ajouté. En utilisant alternativement les deux types d'attention, ce modèle présente l'avantage de capturer à la fois les interactions locales à petite échelle et les interactions globales au niveau de l'image.

Depuis la percée des Vision Transformateurs ViT, ils ont été plus performants que les modèles de vision convolutifs CNN dans les tâches de classification. Cela s'explique par leur capacité à agréger le contexte global dès les premières phases du réseau. Toutefois, ces modèles sont exempts du biais inductif local propre aux CNN. Certains réseaux ont donc tenté de fusionner ces deux méthodes tout en préservant leurs avantages.

- *CoAtNet* il combine les forces des réseaux convolutifs et des transformateurs pour créer un modèle hybride qui offre des performances de pointe dans diverses bases de données sous différentes contraintes de ressources. Les CoAtNets reposent sur deux idées clés : premièrement, la convolution en profondeur et l'auto-attention peuvent être unifiées grâce à une simple attention relative et, deuxièmement, l'empilement vertical des couches de convolution et des couches d'attention de manière ciblée peut améliorer de manière significative la généralisation, la capacité et l'efficacité [38].

  Les CoAtNets y parviennent en intégrant la convolution en profondeur couramment utilisée dans les couches d'attention avec une attention relative simple et en empilant les couches de convolution et d'attention d'une manière qui combine efficacement leurs forces, comme illustré dans la figure 2.5 (Les blocs en rouge sont des bloc de mécanisme d'attention. Les blocs en jaune sont des blocs de convolution en profondeur). Cette approche permet aux CoAtNets d'hériter des capacités remarquables de généralisation des CNN grâce à leurs biais inductifs

favorables et de bénéficier de la scalabilité supérieure des modèles Transformateurs avec des données abondantes, ce qui conduit à une convergence plus rapide et à une efficacité améliorée. CoAtNet atteint une efficacité supérieure à celle de ViT en poussant le record de la précision top-1 d'ImageNet-1K à 90,88% [38].

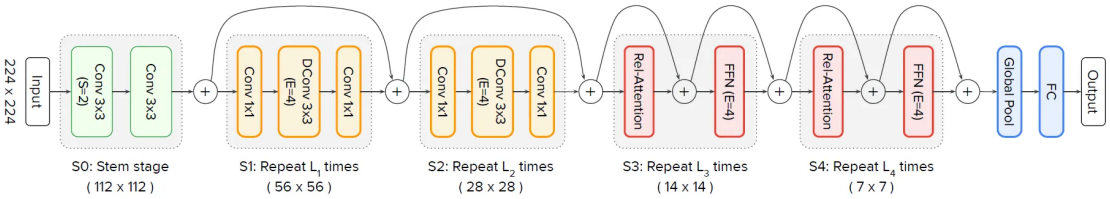


Figure 2.5 Architecture CoAtNet [12].

- *MOAT:* Comme CoAtNet, MOAT est une architecture hybride CNN-Transformateur qui combine à la fois l'auto-attention et les convolutions. Cependant, MOAT combine des blocs transformateurs d'auto-attention et des convolutions de profondeur en un seul bloc MOAT. Ce bloc MOAT est ensuite utilisé comme élément de base pour l'architecture MOAT. MOAT est encore plus efficace que ViT et atteint une précision de 89,1% dans le top 1 d'ImageNet-1K lorsqu'il est entraîné uniquement sur la base de données ImagNet [39].

2.3 Modèles de fusion d'images

La fusion de capteurs est un aspect essentiel de la plupart des systèmes autonomes. Elle intègre les données acquises à partir de plusieurs modalités de détection pour réduire le nombre d'incertitudes et surmonter les lacunes des capteurs individuels fonctionnant indépendamment. De plus, la fusion de capteurs aide à développer un modèle cohérent capable de percevoir les environs avec précision dans diverses conditions environnementales [40].

Il existe trois approches principales pour fusionner les données sensorielles provenant de différentes modalités de détection [40]:

- Fusion de haut niveau (fusion tardive) : Chaque capteur exécute l'algorithme de traitement de manière indépendante et procède ensuite à la fusion. Les approches de fusion de haut niveau sont souvent adoptées en raison de leur moindre complexité relative par rapport aux approches de fusion de bas niveau et de niveau intermédiaire. Cependant, la fusion de haut niveau peut fournir des informations inadéquates lorsque les classifications ayant une valeur de confiance plus faible sont rejetées, par exemple en cas de chevauchement de plusieurs obstacles.
- Fusion de niveau intermédiaire (fusion de descripteurs) : Il s'agit d'un niveau d'abstraction entre la fusion de bas niveau et la fusion de haut niveau. Elle fusionne les informations relatives à plusieurs objectifs extraites des données des capteurs correspondants (mesures brutes), telles que les informations de couleur des images ou les caractéristiques de localisation des radars et des LiDAR, et effectue ensuite la reconnaissance et la classification sur la base des informations multi-capteurs fusionnées. Cependant, sa faiblesse réside dans la perception limitée de l'environnement et la perte d'informations contextuelles.

  Dans la vision par ordinateur, la fusion de niveau intermédiaire est faite par des modèles qui sont conçus pour maximiser le choix des informations complémentaires. Par exemple, il y a Dense Fuse, un réseau de fusion basé sur la convolution. DenseFuse traite les deux modalités dans deux branches séparées pour extraire des descripteurs afin qu'il les additionne [41]. Par ailleurs, il existe Swin Fuse, qui est similairement un réseau de fusion intermédiaire. Pourtant, il est basé sur les blocs Transformeur-Swin [42].
- Fusion de bas niveau (fusion précoce) : Les données de chaque capteur sont fusionnées au niveau d'abstraction le plus bas (données brutes). Par conséquent, toutes les informations sont conservées et peuvent potentiellement améliorer la

précision de la détection des obstacles. Par exemple, l'intégration d'une image de caméra dans un nuage de points 3D de la même vue. Dans la pratique, l'approche de fusion à bas niveau s'accompagne d'une multitude de défis, notamment au niveau de sa mise en œuvre. Elle nécessite un étalonnage extrinsèque précis des capteurs pour fusionner avec précision leurs perceptions de l'environnement.

2.4 Reconnaissance des matériaux

La reconnaissance des matériaux est une tâche spécialisée dans le domaine de la reconnaissance d'objets, qui vise à identifier et à classifier différents matériaux, tels que le métal, le plastique, le bois, etc., à partir d'images ou d'autres formes de données sensorielles. Pour accomplir cette tâche, des modèles d'apprentissage automatique généralistes, tels que les réseaux de neurones convolutifs et les modèles basés sur les Transformateurs, sont fréquemment utilisés. Ces techniques permettent une classification précise des matériaux en exploitant leurs caractéristiques visuelles et texturales distinctes capturées dans les données collectées.

2.4.1 Descripteurs des matériaux dans l'image visuelle

La distribution des images réelles présente des caractéristiques non gaussiennes, capturant des auto-corrélations complexes à un niveau supérieur qui sont définies par le contexte de l'image, comme les objets, la scène et les conditions d'éclairage [43]. Ces caractéristiques sont essentielles pour reconnaître avec précision les matériaux dans des environnements non contrôlés. Les caractéristiques de bas niveau de l'image comprennent des éléments tels que la couleur, les gradients et la texture. Ces attributs sont identifiés par des modèles qui prédisent localement les caractéristiques en fonction des pixels environnants [44]. Pour déduire ces caractéristiques, des informations sont compilées

à partir de diverses plages de pixels non voisins. Pour une classification efficace des matériaux dans les images naturelles, il est essentiel d'extraire des caractéristiques de haut niveau et de bas niveau, chacune par le biais de processus différents.

Concernant les caractéristiques de bas niveau, plusieurs méthodes s'appuient sur ce type de caractéristique, en utilisant des filtres de bas niveau et des algorithmes de regroupement. Ces techniques permettent d'analyser les textures, les couleurs et les gradients en examinant les variations locales au sein de l'image, facilitant ainsi la distinction entre différents matériaux et éléments dans une scène.

- **GLCM:** Il consiste à réaliser des expériences statistiques sur la matrice (ou les matrices) contenant les cooccurrences des intensités de pixels à des angles et distances donnés. Ces expériences statistiques fournissent intuitivement des mesures de propriétés telles que la douceur, la rugosité et la régularité, par exemple, sur la distribution des pixels sur la texture [45].
- **LBP:** Un histogramme est calculé avec la distribution des configurations binaires des pixels de l'image, sur la base du seuillage de la fenêtre environnante de chaque pixel lorsque l'intensité du pixel voisin est inférieure ou supérieure à la valeur centrale. La figure 2.6 illustre ce processus et présente la mise en œuvre la plus simple de la méthode LBP. Une fenêtre de 3×3 peut être considérée et un histogramme de 256 cases peut être généré étant donné les 2^8 combinaisons possibles des fenêtres binaires [46].
- **Banques des filtres:** Cette approche consiste à transformer l'image d'entrée en tenant compte d'un ensemble complet de banques de filtres. Supposons que le nombre total de filtres NF, il en résultera également un ensemble d'images transformées NF, après application du filtre F_k, où $1 \leq k \leq NF$, sur l'image d'entrée I. Un vecteur caractéristique peut alors être calculé, par exemple, en concaténant la

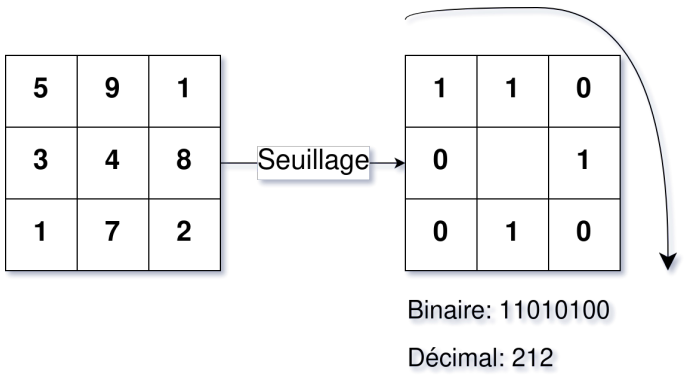


Figure 2.6 Execution de la méthode LPB.

moyenne et l'écart-type de la magnitude de chacune des I_k images transformées [47].

- **FV-CNN:** Le descripteur FV-CNN est un descripteur de texture basé sur le vecteur de Fisher (FV). Le FV regroupe les caractéristiques locales de manière dense dans les régions décrites en supprimant les informations spatiales globales, ce qui les rend plus aptes à décrire les textures que les objets. Le vecteur de Fisher est calculé sur la sortie d'une seule (dernière) couche convolutive du CNN. En évitant le calcul des couches entièrement connectées, l'image d'entrée n'a pas besoin d'être redimensionnée à une taille spécifique. En revanche, les caractéristiques convolutionnelles denses sont extraites à plusieurs échelles et regroupées en un seul FV (Voir figure 2.7), tout comme pour SIFT. Les caractéristiques convolutives regroupées sont extraites immédiatement après le dernier opérateur de filtrage linéaire et ne sont pas normalisées d'une autre manière [13, 48].
- **Wavelet-CNN:** Le modèle est similaire à celui des CNN utilisant des skip-connections. Il intègre directement les approches spectrales dans les CNN, en particulier celles basées sur l'analyse multi-résolution à l'aide de la transformée en ondelettes. La différence essentielle réside dans le fait que certaines couches des CNN à ondelettes n'ont pas de paramètres entraînables. Au lieu de cela, ces couches

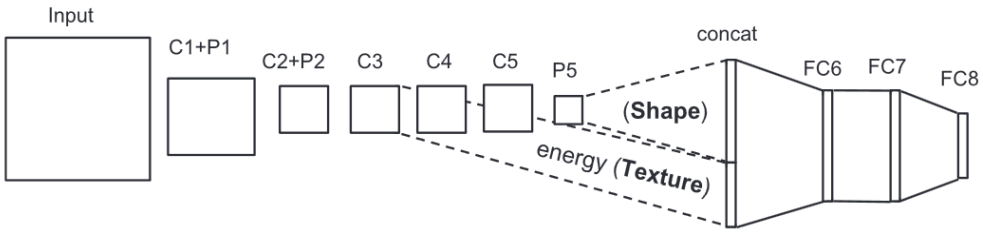


Figure 2.7 Architecture T-CNN [13].

effectuent une analyse multi-résolution à l'aide de paramètres fixes définis par la transformée en ondelettes. L'extraction de caractéristiques dans le domaine des fréquences présente deux avantages. Tout d'abord, un filtre spatial peut facilement être rendu sélectif en améliorant certaines fréquences et en en supprimant d'autres. Cette sélection explicite de certaines fréquences est difficile à contrôler dans les CNN. En outre, la structure périodique d'une texture peut être représentée par une certaine fréquence spatiale dans le domaine spectral. [49].

Pour les caractéristiques de haut niveau, diverses méthodes se concentrent sur ce type de caractéristiques, en employant des modèles avancés tels que les CNNs ou les Transformateurs de vision. Ces modèles sont capables de traiter des informations complexes, grâce à leur capacité à apprendre des représentations profondes et abstraites des données. Ils utilisent des bases de données volumineuses qui conservent les informations sur les objets et le contexte global de l'image.

- **CNNs:** Certaines techniques ont utilisé le réglage fin et l'adaptation de domaine sur des modèles pré-entraînés tels que VGG, AlexNet et GoogleNet sur des ensembles de données qui conservent le contexte des matériaux dans l'image, comme MINC. Ces modèles ont atteint 85,6% en termes de précision de test sur leur ensemble de données [50]. De manière similaire, [51] a employé l'apprentissage par transfert sur un modèle ResNet-50 en utilisant l'ensemble de données d'images de déchets,

atteignant une précision de test de 87% sur le même ensemble de données.

- **Transformateurs:** Les Transformaters de vision sont une approche innovante dans le domaine de la vision par ordinateur qui se base sur l'architecture des Transformateurs, initialement développée pour les tâches de traitement du langage naturel, pour gérer les données visuelles. Ces modèles traitent les images en les divisant en morceaux et en appliquant des mécanismes d'attention pour capturer les dépendances locales et à longue portée au sein de l'image, leur permettant d'exceller dans une variété de tâches de vision par ordinateur. Contrairement aux CNNs traditionnels, les ViTs ne dépendent pas des biais inductifs des opérations de convolution, leur permettant d'apprendre des représentations visuelles plus flexibles. Les ViTs ont démontré des améliorations significatives par rapport aux approches conventionnelles, en particulier dans les contextes où la disponibilité de grands ensembles de données leur permet de profiter de leur capacité à apprendre à partir des données [52].

Dans le contexte de la reconnaissance des matériaux, quelques travaux ont utilisé les Transformateurs de vision pour classifier les images des matériaux. Dans [53], il est démontré que les Vision Transformateurs détiennent une promesse exceptionnelle dans le domaine de la reconnaissance des matériaux, en montrant une capacité à différencier les différences et les transitions entre les états de la même matière. Cette capacité découle de la compréhension profonde de la structure des données visuelles par l'architecture, lui permettant de saisir les motifs complexes qui définissent et différencient les matériaux. Cette capacité les permet à exceller dans l'apprentissage avec peu d'exemples (few-shot learning), où leurs vastes représentations apprises permettent au modèle de s'adapter rapidement à de nouveaux concepts avec un minimum de formation.

Par ailleurs, dans le domaine de la segmentation d'image, l'intégration d'une tête de segmentation basée sur l'attention améliore la précision de la segmentation en se concentrant sur les caractéristiques les plus pertinentes, tout en préservant les détails spatiaux cruciaux pour une bonne précision. Cette approche surmonte les limites rencontrées avec les modèles profonds traditionnels, en évitant la dégradation des informations à la sortie et la performance et en optimisant l'intersection sur Union (IoU), un indicateur clé pour l'évaluation de la segmentation d'image. Les têtes de segmentation fondées sur l'attention démontrent ainsi leur supériorité dans le traitement des tâches de segmentation complexes, marquant une avancée significative dans ce domaine [54].

2.4.2 Modalités de reconnaissance des matériaux

L'intégration d'autres modalités à la vision par ordinateur peut effectivement réduire la distribution a priori des données d'entrée du système et augmenter la précision des résultats. Ces modalités supplémentaires enrichissent le système de caractéristiques discriminatives supplémentaires, facilitant la distinction entre des matériaux visuellement similaires. Parmi ces modalités, on trouve:

- **Imagerie visuelle:** Au lieu d'améliorer la perception visuelle par l'ajout de capteurs physiques [55–57], un sous-ensemble de chercheurs a proposé des solutions qui dépendent uniquement des caméras visuelles, en mettant l'accent sur l'amélioration de la compréhension des images. Cette discussion scientifique se concentre principalement sur deux modalités critiques : les caractéristiques locales (texture) et les caractéristiques non locales globales (contexte : objets, scène, …). Les caractéristiques locales examinent les attributs de texture et de couleur de l'image de surface, tandis que les caractéristiques non locales englobent

les caractéristiques plus larges des objets et des scènes. Une distinction cruciale entre les deux approches réside dans la nature des ensembles de données qu'elles utilisent. Par exemple, la base de données des matériaux Flickr (FMD) [58], principalement utilisée pour les caractéristiques locales, comprend principalement des textures de surface. En revanche, le jeu de données des matériaux en contexte (MINC) [50], utilisé pour les caractéristiques globales, offre une vue plus holistique en incluant à la fois la surface cible et son environnement dans des contextes non controlés. Les caractéristiques locales peuvent être catégorisées en caractéristiques artisanales—telles que les caractéristiques de texture invariantes à l'illumination de Markov [59]; les banques de filtres; et les filtres ondelettes [13]—ainsi que les caractéristiques convolutionelles extraites automatiquement [60] et les modèles convolutionels profonds appliqués aux bases de données de texture [18]. D'autre part, les méthodes basées sur les caractéristiques globales utilisent principalement des modèles convolutionels ou transformateurs, entrainés sur des formats de bases de données qui incluent le contexte d'objets [50, 51, 53, 61, 62] et de scène [54]. Ces méthodes de caractéristiques globales ont montré qu'elles surpassaient leurs homologues locales, atteignant des précisions de test de 85,6 % sur la base de données MINC [50] et 87 % sur l'ensemble de données d'images de dechets [63], les rendant plus adaptées aux applications réelles. Cependant, il est important de noter que la texture et l'apparence visuelle ne sont pas uniques à une matière, ce qui peut entraîner des prédictions inexactes pour des matériaux visuellement similaires et en présence d'entrées adverses.

- **Imagerie proche infrarouge:** La différence d'intensité des images dans le proche infrarouge n'est pas seulement due à la couleur particulière du matériau (Voir figure 2.8), mais aussi aux caractéristiques d'absorption et de réflectivité du colorant.

Cette indépendance relative des informations relatives au proche infrarouge et à la couleur fait des images NIR un candidat de choix pour la classification des matériaux [14].

Figure 2.8 Exemple de la différence entre l'imagerie visuelle (a) et l'imagerie proche-infrarouge (b) [14].

- **Imagerie multispectrale:** Les caméras multispectrales surpassent les caméras visuelles en termes de classification et les capacités descriminatives. Par exemple, Namin et Peterson [64] ont introduit une méthode basée sur des caméras multispectrales avec six bandes visuelles et une bande proche-infrarouge. Leur méthode a extrait les caractéristiques locales des sept bandes pour chaque pixel. Des caractéristiques de texture telles que le GLCM et les caractéristiques du spectre de Fourier sont exploitées pour rendre le système plus robuste à différentes conditions d'éclairage. Ensuite, les pixels sont classifiés en utilisant les classificateurs SVM et Adaboost. En conséquence, les données de test ont été classées dans dix classes pré-définies en utilisant SVM et Adaboost avec des précisions moyennes de validation croisée de 91,9 % et 89,1 %, respectivement, sur l'ensemble de données de la caméra FluxData, surpassant les méthodes basées sur les caméras visuelles.
- **Imagerie hyperspectrale:** L'imagerie hyperspectrale est une technique importante en télédétection, qui collecte le spectre électromagnétique du visible au proche

infrarouge. Les capteurs d'imagerie hyperspectrale fournissent souvent des centaines de bandes spectrales étroites provenant de la même zone de la surface de la terre. Dans les images hyperspectrales, chaque pixel peut être considéré comme un vecteur à haute dimension dont les entrées correspondent à la réflectivité spectrale dans une longueur d'onde spécifique [29].

- **Variance de réflectivité:** Peut être réalisée par une caméra à champ lumineux [65] ou une caméra stéréo [66]. Les caméras à champ lumineux génèrent une image 4D en prenant plusieurs vues d'une scène tout en changeant d'ouverture. Wang et al. ont utilisé des caméras à champ lumineux pour générer un ensemble de données de classification des matériaux en utilisant la caméra Lytro Illum. En adaptant des réseaux convolutionnels pré-entraînés aux images 4D, ils ont résolu le problème élevé de calcul et de coût de mémoire lié au traitement des images 4D. Leur méthode proposée a résulté en une augmentation de précision de 7 % par rapport aux caméras visuelles, prouvant ainsi que les informations dépendantes de la réflectance contribuent aux informations visuelles dans la reconnaissance [65]. En revanche, les caméras stéréos ont été utilisés par Xue et al, [66]. Les caméras stéréo prennent l'image d'une scène avec une petite variations angulaire, ce qui enrichit l'image visuelle avec des indices radiométriques.
- **Imagerie à temps de vol:** Le phénomène de rugosité peut être capturé à l'aide d'une caméra à temps de vol [67] ou d'un LiDAR à temps de vol [68]. Ces types de caméras fonctionnent en illuminant la scène avec une source de lumière modulée et en observant la lumière réfléchie. Le déphasage entre l'illumination et la réflexion est mesuré et traduit en distance, Ces dispositifs sont capables à générer une reconstruction 3D de la surface cible, ainsi que avoir une information sur la rugosité de la surface. Kim et al. ont utilisé une caméra à temps de vol pour

identifier les propriétés réflectives des matériaux, en combinant ces données avec des images visuelles traditionnelles, ce qui a permis d'améliorer la précision de 10% par rapport aux techniques utilisant uniquement des images visuelles [67].

- **Imagerie acoustique:** Le phénomène de rugosité peut être capturé à l'aide d'une caméra à ultrasons [69]. La technique à ultrason permet de capturer les signaux rétro-diffusés par le biais d'un C-scan ultrasonique, ce qui permet d'obtenir un cube de données tridimensionnel. Ce cube est composé d'une série d'images bi-dimensionnelles C-scan, chacune capturée à des profondeurs variables, révélant la variance statistique des signaux de diffusion de la micro-structure. Les signaux ultrasoniques sont recueillis à l'aide d'un transducteur piézoélectrique à large bande dont la fréquence centrale est de 5 MHz [69].
- **Fonction de distribution de la réflectance bidirectionnelle (FDRB):** La fonction de distribution de réflexion bidirectionnelle a également été un domaine de concentration, offrant des descripteurs pour les propriétés de surface telles que la rugosité, la transparence et l'absorption de la lumière. Cette fonction est précise pour caractériser les matériaux. Des recherches approfondies [66,70–73], ont validé l'efficacité de la prédiction des types de matériaux grâce à l'acquisition précise et robuste des données FDRB. Diverses configurations d'équipement, telles que les dômes d'illumination [70, 72] et les réflecteurs semi-hémisphériques associés à des spectro-radiomètres Visible-IR [71], ont été utilisées à cet effet. Néanmoins, ces équipements nécessitent des conditions spécifiques, limitant leur applicabilité dans des environnements non contrôlés.
- **Imagerie thermique:** La méthode se compose de deux types de caractéristiques : la perméation de l'eau et un cycle de chauffage/refroidissement. Pour la perméation de l'eau, un modèle 3D décrivant le taux et la taille de la perméation de l'eau est

extrait. Pour le chauffage et le refroidissement, nous résolvons une variation de l'équation de la chaleur pour les paramètres constants [15] (Voir figure 2.9).

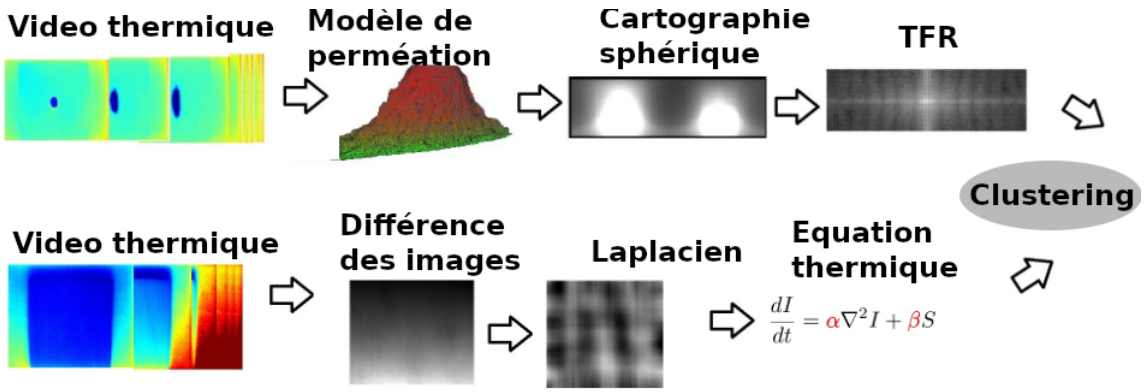


Figure 2.9 Classification des matériaux avec caméra thermique [15].

Les méthodes utilisant des caméras ToF, des caméras acoustiques, des mesures de réflectivité, des caméras hyperspectrales, ainsi que les caractéristiques locales des images visuelles, requièrent généralement des conditions spécifiques pour fonctionner efficacement. Ces exigences les rendent souvent inadaptées pour aborder les complexités des environnements naturels, où de tels contrôles ne peuvent pas toujours être garantis.

2.4.3 Base de données de la reconnaissance des matériaux

Il existe plusieurs bases de données publiques dédiées à la reconnaissance des matériaux, qui jouent un rôle crucial dans le développement et l'évaluation des modèles d'apprentissage automatique. La plupart de ces bases de données se concentrent sur l'apprentissage des caractéristiques locales, notamment la texture, et incluent des ensembles de données tels que les exemples ci-dessous. Ces bases de données sont principalement axés sur la texture des matériaux.

- Base de données sur la réflectance et la texture de Columbia-Utrecht (CUReT) [74] : Une collection des images de textures sous différents angles d'éclairage et de vue.

- FMD [58] : Un ensemble de données qui contient des images de matériaux collectées à partir de Flickr, visant à représenter une variété de textures dans des conditions d'éclairage naturelles.
- Textures sous différents éclairages, poses et échelles (KTH-TIPS) [75]: Une base de données qui fournit des images de textures sous différents angles, éclairages et échelles.
- Base de données de texture Kylberg [76] : Contient des images de textures avec une variation contrôlée, conçue pour l'étude de la classification des textures.

MINC, en revanche, se distingue comme l'unique base de données à grande échelle qui intègre le contexte de l'objet. Conçue pour permettre aux modèles d'aller au-delà de la simple texture, MINC vise à fournir une compréhension plus riche en reconnaissant l'objet ou l'environnement où ces textures apparaissent. Cela permet aux modèles d'apprentissage profond, tels que les CNN et les Transformateurs, de mieux interpréter les matériaux en tenant compte de leur contexte d'utilisation ou de leur emplacement, ce qui est essentiel pour des applications pratiques telles que la reconnaissance d'objets dans des environnements non-contrôlés [50].

L'intégration du contexte dans la reconnaissance des matériaux ouvre la voie à des systèmes de vision par ordinateur plus sophistiqués et précis, capables de comprendre non seulement ce qu'ils voient mais aussi dans quelles circonstances les objets sont observés. Cela améliore significativement la capacité des systèmes à interagir de manière intelligente avec leur environnement, en offrant des perspectives plus nuancées et des réponses plus adaptées à la complexité du monde réel.

2.5 Analyse de l'état de l'art

Les caméras visibles, particulièrement efficaces pour capturer une riche variété d'informations complexes sur les matériaux, bénéficient considérablement de la disponibilité à grande échelle de bases de données d'images publiques ainsi que de modèles profonds pré-entraînés sur ces immenses ensembles de données. Cependant, elles présentent des limites telles que la similarité visuelle entre différents matériaux, la variabilité de l'apparence au sein d'un même type de matière, et la complexité des informations capturées. Pour surmonter ces défis, il est souvent recommandé de combiner plusieurs modalités de vision complémentaires à l'imagerie visible, dans le but de réduire la variabilité de la distribution probabiliste a priori.

Selon l'état de l'art, les méthodes les plus efficaces de reconnaissance des matériaux intègrent l'analyse de la texture [13, 18, 59, 60] et du contexte dans les images [50, 51, 53, 54, 61–63]. Dans certains cas, ces caractéristiques sont combinées avec d'autres modalités, chacune présentant ses avantages et ses limites. Par exemple, la fusion d'images visuelles avec d'autres informations spectrales, telles que celles issues de caméras multispectrales [14,64] ou hyperspectrales [29], améliore les capacités discriminatoires du système, mais ne résout pas totalement le problème de la similarité visuelle des matériaux. De plus, l'ajout de modalités comme l'imagerie à temps de vol [67, 68], acoustique [69], thermique [15] - ainsi que la FDRB [66, 70–73] apporte des informations supplémentaires et complémentaires, telles que la rugosité de la surface, les propriété thermique et les propriétés de réflexion, bien qu'elles nécessitent des conditions spécifiques pour fonctionner efficacement.

L'exploitation de la variance de réflectivité pour enrichir les images visuelles augmente la capacité discriminatoire du système [65, 66]. Néanmoins, il est crucial de reconnaître

que les méthodes se focalisant uniquement sur cette variance rencontrent des difficultés, notamment dans la détection des phénomènes d'auto-occlusion et d'auto-ombrage, et ne parviennent pas à fournir une caractérisation complète des matériaux [77]. Ces approches omettent des facteurs essentiels tels que la géométrie de l'objet et l'éclairage externe, indispensables pour une analyse exhaustive de la distribution de la réflectivité.

2.6 Conclusion

La reconnaissance des matériaux dans le domaine de la vision par ordinateur est une tâche spécifique qui s'appuie sur des concepts avancés de reconnaissance d'images. Elle implique l'utilisation de modèles de vision, tels que les réseaux de CNNs et les Transformateurs, qui sont entraînés et affinés sur des bases de données spécialisées pour identifier efficacement les caractéristiques des matériaux à partir d'images. Ces bases de données, telles que CUReT, FMD, KTH-TIPS, et MINC, fournissent une variété de textures et de contextes pour améliorer la précision de la reconnaissance. La reconnaissance des matériaux bénéficie également de l'intégration de différentes modalités sensorielles, enrichissant ainsi les caractéristiques discriminatives du système pour distinguer entre des matériaux visuellement similaires.

Chapitre 3 - Conception de système et choix de modalités

Ce chapitre présente la méthodologie suivie pour intégrer la texture multispectrale aux caractéristiques de réflectivité de la surface dans les spectres visuel et infrarouge. La première phase est la modélisation de l'installation utilisée. La seconde phase est la création d'un modèle d'apprentissage profond qui fusionne les données acquises par le système installé.

3.1 Caractéristiques physiques des surfaces des matériaux

L'information introduite dans le système se présente sous forme d'images réelles capturées dans le spectre infrarouge-visible. Ces images possèdent des caractéristiques spatiales et spectrales distinctives. En outre, elles intègrent le phénomène de réflexion, capturant ainsi la manière dont les surfaces interagissent avec la lumière.

3.1.1 Caractéristiques spectrales de l'information

Une image RGB capture la lumière dans le spectre visible, qui s'étend approximativement de 380 à 740 nanomètres (nm). Ce spectre englobe toutes les couleurs perceptibles par l'œil humain. Le modèle de couleurs RVB divise spécifiquement ce spectre visible en trois grandes bandes; canal rouge : capte la lumière dans les grandes longueurs d'onde, approximativement entre 620 et 740 nm, ce qui correspond aux couleurs rouges. canal vert : capte la lumière dans les longueurs d'onde moyennes, approximativement de 495 à 570 nm, correspondant aux couleurs vertes. canal bleu capte la lumière dans les longueurs d'onde les plus courtes, environ de 450 à 495 nm, ce qui correspond aux couleurs bleues. En outre, les images dans le proche infrarouge capturent la lumière dans le spectre du proche infrarouge, qui s'étend approximativement de 750 à 1400 manomètres (nm). Ce spectre se situe juste au-delà du spectre de la lumière visible,

qui se termine à environ 740 nm, et n'est pas visible pour l'œil humain.

La similitude observée entre la vision dans le visible et le proche infrarouge est principalement attribuée aux caractéristiques de réflexion et d'absorption étroitement liées des matériaux dans ces deux gammes spectrales. Notamment, les capteurs d'imagerie courants pour la capture de la lumière visible et proche infrarouge sont calibrés pour des gammes de longueurs d'ondes spécifiques de [400nm à 700nm] et [650nm à 1000nm]. Cette gamme de longueurs d'onde adjacentes explique en partie pourquoi les réponses visuelles et dans le proche infrarouge des matériaux tendent à être localement similaires, influençant ainsi la manière dont les matériaux sont perçus et analysés dans ces différentes techniques d'imagerie. Cette relation est prouvée en calculant les corrélations de la réponse réfléchissante de 7000 matériaux bruts trouvés dans la base de données Splib07 [78]. Les corrélations sont présentées dans le tableau 3-1.

Tableau 3-1 Corrélation entre la réponse réflective dans le spectre proche infrarouge (NIR) - visible. Le spectre proche infrarouge est compris entre 700 et 900 nm et le spectre visible est compris entre 400 et 700 nm.

	700nm	800nm	900nm
400nm	0.42	0.42	0.42
500nm	0.77	0.77	0.77
600nm	0.55	0.51	0.51
700nm	1.0	1.0	1.0

3.1.2 Caractéristiques spatiales de l'information

Pour une classification efficace des matériaux dans les images naturelles, il est essentiel d'extraire à la fois les caractéristiques de haut niveau et de bas niveau, chacune par une méthodologie distincte. Les caractéristiques de bas niveau, incluant la couleur, les gradients et la texture, sont identifiées à l'aide de modèles qui analysent le voisinage local des pixels pour prédire ces attributs avec précision. Ces modèles se penchent sur

l'entourage de chaque pixel pour comprendre les fonctions caractéristiques visuels et les variations présentes. D'autre part, les distributions réelles des images se caractérisent par leur nature non gaussienne [43]. Ces corrélations, influencées par le contexte de l'image - tels que les objets qu'elle contient, la scène représentée et les conditions d'éclairage - jouent un rôle crucial dans la reconnaissance des matériaux dans des environnements non contrôlés ou naturels. Pour capturer ces caractéristiques de haut niveau, les informations sont agrégées à partir de diverses plages de pixels non adjacentes, permettant une compréhension plus globale des propriétés des matériaux en considérant le contexte plus large dans lequel ils apparaissent. Cette approche double, analysant à la fois les informations locales et globales des pixels, permet une prédiction précise des classes de matériaux dans le monde réel.

3.1.3 Modèle de réflectivité des surfaces

La perception humaine des matériaux dépend de la lumière réfléchie, transmise et/ou absorbée par les objets qui atteignent l'observateur. L'apparence des matériaux peut varier de manière significative en fonction d'un large éventail de propriétés telles que la couleur, la douceur, la géométrie, la rugosité, la réflectivité, l'angle sous lequel le matériau est vu et les directions d'éclairage. L'un des principaux défis de l'infographie consiste à mesurer simplement et précisément l'apparence des caractéristiques des matériaux à partir d'objets du monde réel [16]. Il existe plusieurs types de modélisation des réflectivités [16]:

1. FDRDSB - Fonction de Distribution de Réflectance de Dispersion de Surface Bidirectionnelle. C'est la modélisation la plus générale. Elle inclut tous les phénomènes physiques qui décrivent le comportement des rayons lumineux.
2. FTB - Fonction de Texture Bidirectionnelle. Elle décrit les mêmes phénomènes que le BSSRDF, mais pour les matériaux texturés.

3. FDRBVE - Fonction de Distribution de Réflectance Bidirectionnelle Variable dans l'Espace. Elle est similaire au BTF, mais la mesure de SVBRDF ne prend pas en compte les phénomènes d'auto-ombrage, d'auto-occlusion et d'auto-réflexion.
4. FDRB - Fonction de Distribution de Réflectance Bidirectionnelle. C'est la représentation la moins complexe. Elle constitue l'élément de base de la SVBRDF. La FDRB est une fonction radiométrique, actuellement utilisée avec plus ou moins de précision dans les systèmes de rendu photoréaliste. Elle décrit, de manière générale, comment l'énergie incidente est redirigée dans toutes les directions à travers un hémisphère au-dessus de la surface. Historiquement, la FDRB a été définie comme une représentation simplifiée de la réflectance pour les surfaces opaques : la FDRB suppose que la lumière entrant dans un matériau le quitte à la même position.

L'estimation de la fonction de distribution de la réflectivité bidirectionnelle d'un matériau dans des conditions naturelles est une tâche difficile. Généralement, l'obtention d'une mesure précise de la FDRB nécessite l'utilisation d'équipements de laboratoire spécialisés, tels que des génio-réflecteurs ou des dômes hémisphériques, associés à des conditions d'éclairage contrôlées, comme illustré dans la figure 3.1. Ces outils et paramètres sont essentiels car ils permettent une analyse détaillée de la façon dont la lumière interagit avec les surfaces des matériaux sous différents angles et dans différentes conditions. Toutefois, en l'absence d'un tel équipement spécialisé et d'environnements d'éclairage spécifiques, il est plus difficile d'obtenir une estimation fiable de la FDRB, ce qui souligne la complexité de la saisie et de l'interprétation des caractéristiques optiques des matériaux. Par conséquent, pour pouvoir estimer une mesure grossière de la FDRB, d'autres simplifications ont été appliquées à l'équation du modèle. **La FDRB est généralement mesuré par des équipements de laboratoire ce qui n'est pas utile**

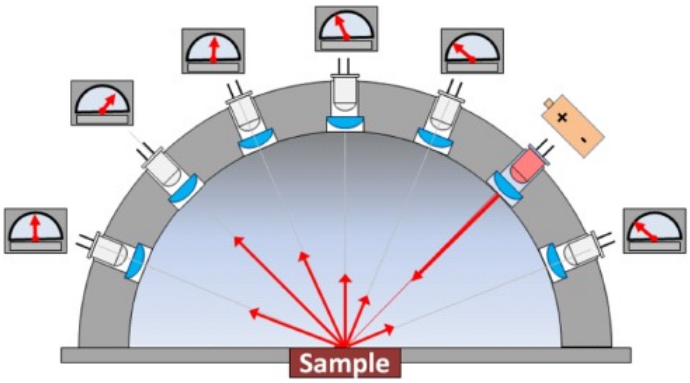


Figure 3.1 Installation pour mesure de FDRB [16].

pour les environnements non-contrôlés. le but est d'estimer la FDRB d'une manière uni-variée à partir d'un seul point.

La formule générale du modèle de FDRB est présenté par l'équation (3.1) [16].

$$brdf(\theta_r, \phi_r, \theta_i, \phi_i) = \frac{dL(\theta_r, \phi_r)}{dE(\theta_i, \phi_i)} = \frac{dL(\theta_r, \phi_r)}{E(\theta_i, \phi_i)\cos\theta_i d\phi_i} \tag{3.1}$$

L est l'illumination réfléchie provenant d'un point de mesure, θ_r et ϕ_r sont, respectivement, les angles zénithaux et azimutaux de la direction de vue. E est la radiance allant vers un point de mesure et θ_i et ϕ_i sont, respectivement, les angles zénithaux et azimutaux de la direction de la lumière. Pour réduire davantage la complexité du modèle, les hypothèses suivantes sont considérées:

- Toutes les surfaces des matériaux sont considérées isotropes (ce qui est le cas pour la plupart des surfaces de matériaux réels). La fonction FDRB serait symétrique autour de l'axe normal de la surface. Donc, on peut ignorer ϕ_i et ϕ_r.
- La direction de la lumière et la direction de la caméra sont jointes, donc $\theta_i = \theta_r$ et $\phi_i = \phi_r$.

Par la suite, équation (3.1) est réduite à équation (3.2):

$$brdf(\theta_r) = \frac{dL(\theta_r)}{dE(\theta_r)} = \frac{dL(\theta_r)}{E(\theta_r)\cos\theta_r d\theta_r} \tag{3.2}$$

L'équation (3.2) rend trois inconnus que l'on va chercher à estimer dans des conditions naturelles.

Pour mieux expliquer la relation entre la FDRB et sa capacité à prédire les matériaux, nous avons choisi le modèle de Phong [79] pour extraire les constantes des caractéristiques. Le modèle de réflexion de Phong décrit l'aspect visuel de la FDRB. Ce modèle divise les composantes de l'illumination en trois parties : les réflexions ambiantes, diffuses et spéculaires. Ces phénomènes physiques peuvent être observés individuellement par l'œil humain. Ce modèle est décrit par l'équation (3.3) et l'équation (3.4). En appliquant les mêmes simplifications que celles démontrées précédemment sur le modèle Phong actuel, on obtient l'équation (3.5).

$$dL(\theta_r, \phi_r, \theta_i, \phi_i) = dL(\vec{Lp}(\theta_i, \phi_i), \vec{V}(\theta_r, \phi_r)) \tag{3.3}$$

$$dL(\vec{Lp}, \vec{V}) = k_a I_a + k_d I_d (\vec{Lp}.\vec{N}) + k_s I_s (\vec{R}.\vec{V})^n \tag{3.4}$$

$$dL(\theta_r) = k_a I_a + k_d I_d \cos(\theta_r) + k_s I_s \cos^n(2.\theta_r) \tag{3.5}$$

Où $\vec{V}$ est la direction de vue ou l'angle de mesure et $\vec{L_p}$ est la direction de la lumière. k_a, k_d et k_s sont respectivement les constantes de réflexion ambiante, diffuse et spéculaire. Ces constantes sont spécifiques aux propriétés de surface du matériel. I_a, I_d et I_s sont

respectivement les illuminations de réflexions diffuse et spéculaire. Le n représente la constante de brillance décrivant la taille des reflets spéculaires. Et enfin, $\vec{N}$ représente la direction normale de la surface.

Pour chaque θ_r, les illuminations I_a, I_d et I_s dépendent de la direction et de l'intensité de la source lumineuse E, de leurs constantes respectives k et suivent la loi de conservation de l'énergie. Par conséquent, l'équation de la FDRB peut être réécrite en termes de variables mesurées et de constante de caractérisation dans l'équation (3.6).

$$brdf = brdf_{k_a,k_d,k_s,n}(\theta_r, E, I_x) = \frac{k_a I_a(E) + k_d I_d(E)\cos(\theta_r) + k_s I_s(E)\cos^n(2.\theta_r)}{E(\theta_r)\cos\theta_r d\omega_i} \quad (3.6)$$

Cette équation montre que k_a, k_d, k_s et n peuvent être des constantes caractéristiques pour les matériaux. En tant qu'exemple concret de l'estimation approximative de la FDRB, la largeur et la valeur moyenne de la distribution de la FDRB peuvent être estimées en utilisant son aspect visuel, comme le montrent respectivement les figures (3.2.A) et(3.2.B). Ces deux critères sont traduits visuellement par les réflexions spéculaires (étroitesse de la distribution FDRB) et l'albédo (intégrale de la distribution FDRB). Ces phénomènes détectés peuvent varier sur une surface matérielle homogène donnée, ce qui permet de caractériser davantage la distribution de la FDRB.

3.1.4 Système d'acquisition

Dans cette section, nous explorons l'interaction entre les propriétés optiques des matériaux et l'imagerie capturée par les caméras stéréo IR-vision RGB-D disponibles sur le marché, qui fournissent les modalités que nous avons suggérées. La caméra que nous avons choisie est l'Intel RealSense D435 (voir figure 3.3). Cette caméra fournit une image visuelle, une image proche-infrarouge, une image de profondeur et une projection

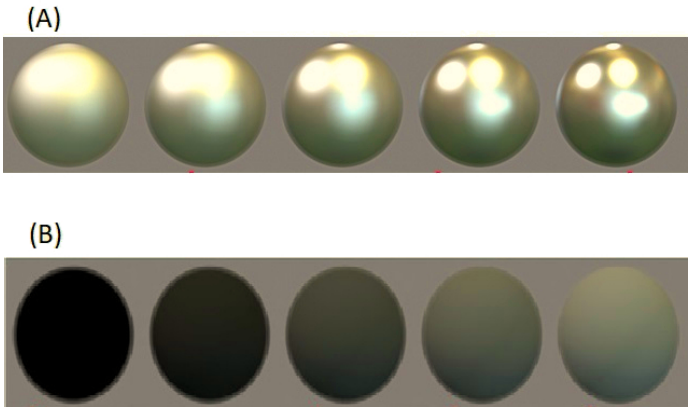


Figure 3.2 Représentation des types de fonctions FDRB apparaissant dans les objets générés par Blender : A-Type de réflexion : diffuse à spéculaire. B-Luminosité de la surface.

des faisceaux laser dans le spectre proche-infrarouge. En utilisant ce type de caméra, ces conditions peuvent être réalisées :

Figure 3.3 Caméra de profondeur de type Intel RealSense d435 [7].

- Les directions de la caméra et de la source lumineuse sont jointes.
- À l'aide de la lumière des projecteurs laser, de multiples faisceaux laser sont émis dans de nombreuses directions vers la surface du matériau, dans le spectre infrarouge proche.
- L'illumination de réflexion est estimée en comparant l'image proche-infrarouge contenant la réflexion des faisceaux laser créés et l'image visible contenant uniquement l'illumination naturelle de la scène. Il est démontré dans la section

(3.1.1) que les matériaux ont pratiquement les mêmes propriétés de réflexion et d'absorption entre le spectre visible et le spectre proche infrarouge. Une représentation de cette opération est donnée dans la figure (3.4).

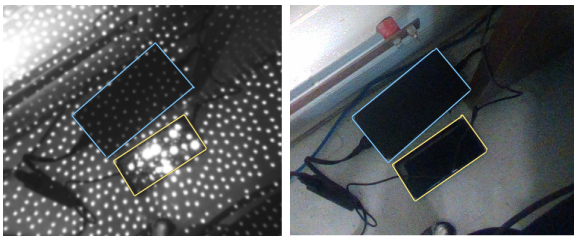


Figure 3.4 Images visuelle et proche-infrarouge de deux matériaux visuellement similaires dans le spectre visible. Les surfaces à l'intérieur des rectangles bleu et jaune sont, respectivement, du plastique et du verre. Des caractéristiques de réflexion discriminantes ont pu être observées dans l'image proche-infrarouge.

- Les normales de la surface sont déduites de l'image de la caméra de profondeur.
- Les limites d'une seule surface homogène peuvent être apprises automatiquement par le réseau de reconnaissance. Ainsi, la distribution de la FDRB peut être estimée dans chaque région d'une surface, si les conditions adéquates sont réunies. Pour recueillir le maximum d'informations à partir d'une image de réflexion, la zone d'image doit présenter une résolution spatiale suffisante et ne pas contenir d'apparences non homogènes excessives.
- La caméra de profondeur, les caméras NIR et RGB, les projecteurs laser sont alors capables de détecter plusieurs points de données en fonction de l'éclairage incident et de la normale de la surface, comme le montre la figure (3.5).

D'après l'équation (3.2), l'estimation de la FDRB est possible si les composants E, dL et θ_r peuvent être mesurés à partir d'un seul point par notre caméra de profondeur, ce qui est validé par les conditions réalisées. E: L'intensité de lumière incidente que l'on peut contrôler. Il s'agit de l'intensité des faisceaux laser. dL: L'intensité de la lumière réfléchie

que l'on va mesurer par la différence entre les images visuelle et proche-infrarouge. θ_r: l'angle entre le rayon réfléchi et la normale de la surface. Ce terme peut être estimé à partir de l'information sur la géométrie, acquise par l'image de profondeur. Il est évident que l'échantillonnage de la FDRB dépend de l'angle zénithal θ_r. Ce modèle est appelé échantillonnage spatial uni-varié, qui consiste à estimer une distribution en capturant plusieurs échantillons d'une surface homogène, en fonction d'une variable dL, à partir d'une image (figure 3.5).

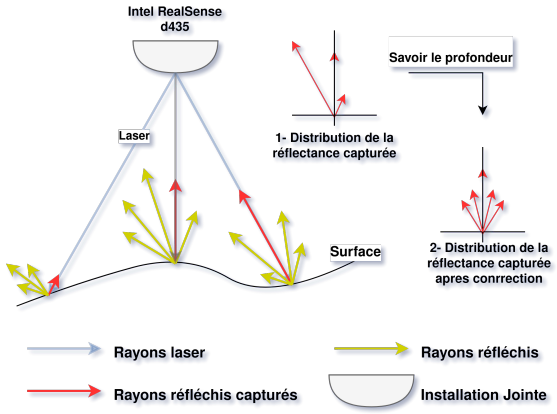


Figure 3.5 Représentation de l'échantillonnage spatial uni-variant à partir d'une image 2D pour obtenir une estimation de la FDRB.

Revenons à l'équation (3.6), qui est une fonction avec une variable θ_r et cinq constantes: n, k_a, k_d, E et k_s. Il serait possible d'estimer les constantes de caractérisation en estimant la distribution de la FDRB par un échantillonnage spatial univarié, comme le montre la figure (3.5), et en contrôlant les variables à l'aide de notre configuration. Bien que la méthode simplifiée ne permette pas d'obtenir directement une mesure précise de la FDRB, le modèle a permis d'estimer une mesure approximative de la distribution. Par conséquent, ces données offriront des informations plus discriminantes à notre modèle en complément d'autres modalités (texture et contexte).

3.2 Système d'apprentissage

Les modèles d'apprentissage profond, en particulier les Transformateurs de vision et les réseaux de neurones convolutifs, qui intègrent à la fois des caractéristiques de bas niveau et de haut niveau pour améliorer la précision des prédictions sont examinés dans cette section. Nous explorons également le concept de fusion d'images, un aspect critique lorsqu'il s'agit de traiter des modalités multiples d'images provenant de différentes caméras. Pour relever les défis posés par les ensembles de données de taille limitée, nous avons mis en œuvre des stratégies telles que les techniques d'apprentissage semi-supervisé. De plus, nous avons utilisé l'exploitation d'ensembles de données publiques RGB à grande échelle comme moyen de régularisation, ce qui a permis d'atténuer davantage les problèmes associés aux ensembles de données plus petits.

3.2.1 Modèle de classification

Différences entre les Transformateurs de vision et les réseaux neuronaux convolutifs: Les modèles d'apprentissage profond sont essentiels pour extraire des caractéristiques de haut niveau qui encapsulent efficacement le contexte des images. Parmi ces modèles, les transformateurs excellent particulièrement grâce à leurs mécanismes d'attention. Ces mécanismes leur permettent de capturer efficacement de nombreuses informations contextuelles en une seule opération, ce qui les rend aptes à comprendre les relations et les structures complexes des images. En revanche, les réseaux neuronaux convolutifs ont tendance à extraire des caractéristiques de haut niveau dans leurs couches profondes, où chaque pixel est associé à un champ réceptif spécifique. Ce processus est plus progressif, les réseaux neuronaux convolutifs passant de caractéristiques simples à des caractéristiques complexes, couche par couche. Cependant, les CNN possèdent un avantage certain dans l'extraction de caractéristiques de bas niveau telles que les

textures et les bords. Leur architecture convolutive est intrinsèquement adaptée à cette tâche, permettant l'identification efficace de motifs et de détails de base dans les couches initiales. Cette différence fondamentale souligne les forces et les applications uniques des transformateurs et des CNN dans le domaine du traitement et de l'analyse d'images.

Caractéristiques de haut niveau et de bas niveau et leurs chemins dans les modèles de vision : Les caractéristiques de bas niveau jouent un rôle essentiel dans la prédiction des attributs de texture, qui sont cruciaux pour la reconnaissance des matériaux. Il est donc avantageux que les modèles déployés facilitent un parcours rationalisé pour ces caractéristiques en termes de transformations (noyaux de convolution ou blocs d'attention) vers la sortie. Les connexions résiduelles, qui sont intégrées dans presque toutes les architectures de réseau modernes, y contribuent efficacement. En revanche, les caractéristiques de haut niveau, telles que le contexte de l'objet et de la scène dans le manifeste des caractéristiques des images naturelles, ne peuvent pas être déduites par les premières transformations dans le CNN. Un modèle plus sophistiqué est donc nécessaire pour apprendre la fonction qui met en correspondance toutes ces modalités. Il est donc nécessaire d'utiliser des modèles de Transformateur et de réseaux neuronaux convolutionnels profonds, qui sont aptes à traiter des tâches de mise en correspondance aussi complexes.

Critères d'évaluation des modèles : La référence standard pour l'évaluation des modèles de classification d'objets est généralement la première précision sur l'ensemble de données ImageNet-1k. Cependant, il est essentiel de prendre en compte d'autres mesures lors de l'évaluation de ces modèles. Par exemple, la précision de zéro-shot donne un aperçu de la capacité d'un modèle à établir des connexions abstraites et à s'adapter à la reconnaissance de nouveaux sujets. À l'inverse, la précision du top-5,

Tableau 3-2 Métriques des modèles dans différentes familles de modèles, sur l'ensemble de données Imagenet-1k.

Modèle	Précision Top-1 test	Précision Zéro-coup	Précision Top-5 test	Famille
CoatNet-6	90.45%	84.2%	99.3%	Hybride
Moat-4	89.1%	–	–	Hybride
Eva-02-L	90.0%	80.4%	99.0%	Attention
VIT-G/14	91.1%	88.3%	99.1%	Attention
Swin-v2-G	90.2%	84.0%	97.2%	Attention
DaVit-G	90.4%	–	–	Attention
ResNet-152	85.6%	24.1%	97.2%	Convolution
EffNet-L2	88.6%	–	98.1%	Convolution

ou du top-k en général, bien que moins cruciale, démontre la capacité d'un modèle à différencier les classes, en particulier lorsque les différences sont subtiles. Pour valider le choix du modèle utilisé dans cette recherche, nous avons dressé une liste des modèles récents les plus performants sur l'ensemble de données ImageNet-1k, présentés dans le tableau 3-2. Nous avons délibérément inclus dans cette sélection des modèles basés sur les CNNs et les Transformateurs pour des raisons spécifiques : Les Transformateurs de vision ont fait preuve d'une compétence remarquable en matière d'apprentissage à partir de zéro, soulignant leur capacité à s'adapter rapidement à de nouvelles tâches et à reconnaître des objets non vus. Inversement, les CNN ont été délibérément choisis pour leur tendance inhérente à capturer les caractéristiques locales, une caractéristique cruciale pour l'extraction d'informations liées à la texture.

3.2.2 Modèle de fusion de données

Étant donné que la méthode que nous proposons implique trois modalités d'entrée, image visuelle, proche-infrarouge et profondeur, nous estimons qu'il est nécessaire d'incorporer des techniques de fusion d'images pour combiner efficacement ces modalités:

- Pour les modèles basés sur la convolution, trois méthodes de fusion seront expérimentées. Dans le cas d'une fusion précoce, seule la dimension de la couche

d'entrée sera modifiée. Pour la fusion de niveau intermédiaire, DenseFuse sera intégré dans l'architecture. En ce qui concerne la fusion tardive, trois réseaux distincts seront empilés, avec une tête de classification ajoutée à la fin.

- Pour les Transformateurs, seule l'option de fusion tardive est pratiquement possible, car toute modification de l'architecture nécessite la répétition d'un pré-entraînement approfondi.

3.2.3 Pre-entraînement et apprentissage des représentations visuelles

Pré-entraînement sur l'ensemble des données RGB grand-échelle - Phase 1 : Étant donné la vaste disponibilité d'ensembles de données étiquetées RGB, nous pouvons entraîner efficacement le segment du réseau dédié au RGB. Au cours de cette phase initiale de pré-entraînement, il est essentiel que les poids relatifs aux autres modalités restent statiques, afin de garantir que l'apprentissage du réseau se concentre sur les informations RGB (au cas de fusion bas niveau ou niveau intermédiaire). Ce régime de pré-entraînement se déroule selon un processus séquentiel en deux étapes. Dans un premier temps, la voie RGB du réseau est entraînée exclusivement avec des données RGB afin d'établir une solide compréhension fondamentale de cette modalité. Par la suite, le pré-entraînement s'étend à toutes les modalités, en utilisant toute la gamme des données disponibles. Cette approche inclusive facilite une expérience d'apprentissage holistique, permettant au réseau d'intégrer et d'interpréter toute la gamme des entrées sensorielles.

La phase initiale de l'opération s'appuie sur un ensemble de données à grande échelle pour affiner le biais du modèle, en le rapprochant de la véritable distribution des données tout en maintenant la variance à un niveau minimal. Pour ce faire, d'autres voies du réseau sont désactivées ou gelées. Dans le cas d'une fusion de niveau intermédiaire, des tenseurs zéro sont attribués aux branches qui ne sont pas entraînées, ce qui réduit au silence tout

signal d'activation pour ces branches masquées. L'utilisation de la fonction d'activation ReLU, dépourvue de biais, garantit que les neurones correspondants restent inactifs, préservant ainsi leurs poids au cours du processus de rétropropagation et empêchant toute modification involontaire des prédictions du modèle.

Apprentissage des représentations visuelles multimodales - Phase 2 : Dans la phase suivante de pré-entraînement, l'ensemble des données de toutes les modalités d'entrée est utilisé afin de réduire davantage l'écart entre le biais appris par le modèle et le véritable biais sous-jacent des données, situé dans un espace de caractéristiques abstraites. Cette étape est cruciale car elle permet non seulement de réduire la variance, mais aussi d'augmenter considérablement la précision prédictive du modèle. En intégrant toutes les informations sensorielles disponibles, le réseau affine ses paramètres afin de refléter plus précisément la complexité et les nuances des données d'entrée.

Les algorithmes d'apprentissage profond, y compris les réseaux neuronaux convolutifs et les Transformateurs s de vision, nécessitent généralement des ensembles de données à grande échelle pour atteindre une convergence optimale et des capacités de généralisation robustes. Cette exigence pose souvent un défi dans notre scénario, où la disponibilité des données étiquetées est limitée. Dans ce cas, le pré-entraînement semi-supervisé s'est avéré être une stratégie efficace. Cette approche tire parti de la combinaison d'une petite quantité de données étiquetées et d'un plus grand ensemble de données non étiquetées. Elle permet ainsi au modèle d'apprendre des représentations significatives à partir du vaste ensemble de données non étiquetées, tandis que les données étiquetées guident l'apprentissage vers des résultats plus précis et plus spécifiques. Cette méthode permet non seulement d'améliorer les performances du modèle dans les environnements où les données sont rares, mais elle contribue également à rendre les applications d'apprentissage

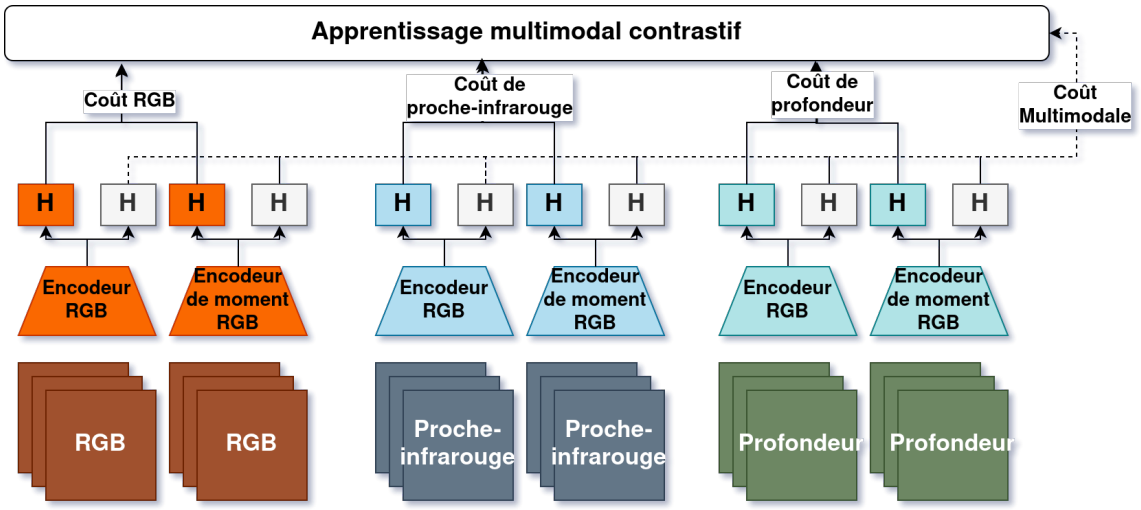


Figure 3.6 Cadre d'apprentissage de représentation visuelle contrastive multimodale. Tous les coûts utilisées visent à maximiser l'accord entre les logits prédits en utilisant des points de données similaires et le désaccord entre les points de données négativement similaires. Pour la perte multimodale, la présence de toutes les modalités n'est pas nécessaire.

automatique plus efficaces et plus polyvalentes. La méthodologie d'apprentissage des représentations visuelles dans les réseaux de données multimodales, telle que démontrée dans Yuan et al. (2021) [80], est sous-tendue par un double objectif (figure 3.6) : la préservation des similarités à la fois intra-modales et inter-modales. Cette technique exploite les caractéristiques uniques intrinsèques à chaque modalité, tout en capturant simultanément les riches informations sémantiques qui émergent des corrélations entre les différentes modalités. Une telle stratégie améliore considérablement la sophistication et la précision des représentations visuelles que le réseau est capable d'apprendre. L'exploitation simultanée des caractéristiques intra-modales et des relations intermodales permet d'obtenir une représentation plus robuste et sémantiquement plus riche des données.

3.2.4 Apprentissage semi-supervisé

Apprentissage semi-supervisé - Phase 3 : Afin d'optimiser l'utilisation des données et d'atténuer le risque de overfitting, des techniques d'apprentissage semi-supervisé avec régularisation de la cohérence (Consistency regulation) sont appliquées pendant l'apprentissage pour toutes les modalités. La régularisation de la cohérence est une technique qui améliore la stabilité du modèle et augmente la résistance au bruit dans les données d'apprentissage. Plus précisément, nous utilisons la méthode d'assemblage temporel (Temporal ensembling) [81], qui repose sur une collection de modèles avec des architectures et des hyper-paramètres identiques, mais des initialisations et des augmentations d'entrée différentes. Pendant la formation, chaque modèle de l'ensemble est exposé à des entrées variées, et leurs prédictions sont agrégées pour former une sortie collective, comme illustré dans la figure (3.7.C). La régularisation de la cohérence impose une pénalité sur les divergences entre les sorties des modèles individuels et la sortie de l'ensemble agrégé. En outre, la moyenne temporelle est employée à chaque point de données pour renforcer la stabilité des prédictions au cours des itérations de formation successives. Le bloc de la moyenne mobile exponentielle (EMA), représenté dans la figure (3.7.C), sert à tempérer tout changement soudain de cohérence tout au long de la phase d'apprentissage. Son objectif est de garantir des prédictions stables et fiables à chaque point d'inférence, guidées par l'équation 3.7. La méthode de régularisation de cohérence se fait parallèlement avec l'apprentissage supervisé avec le même ensemble de données.

$$y_{ema} = \alpha y_{ema} + (1-\alpha) y_{pred} \tag{3.7}$$

Les blocs de la processus d'apprentissage, avec les 3 phases, est décrit dans la figure 3.7 sont expliqué comme :Bloc A : Pré-Entraînement de l'architecture de base RGB

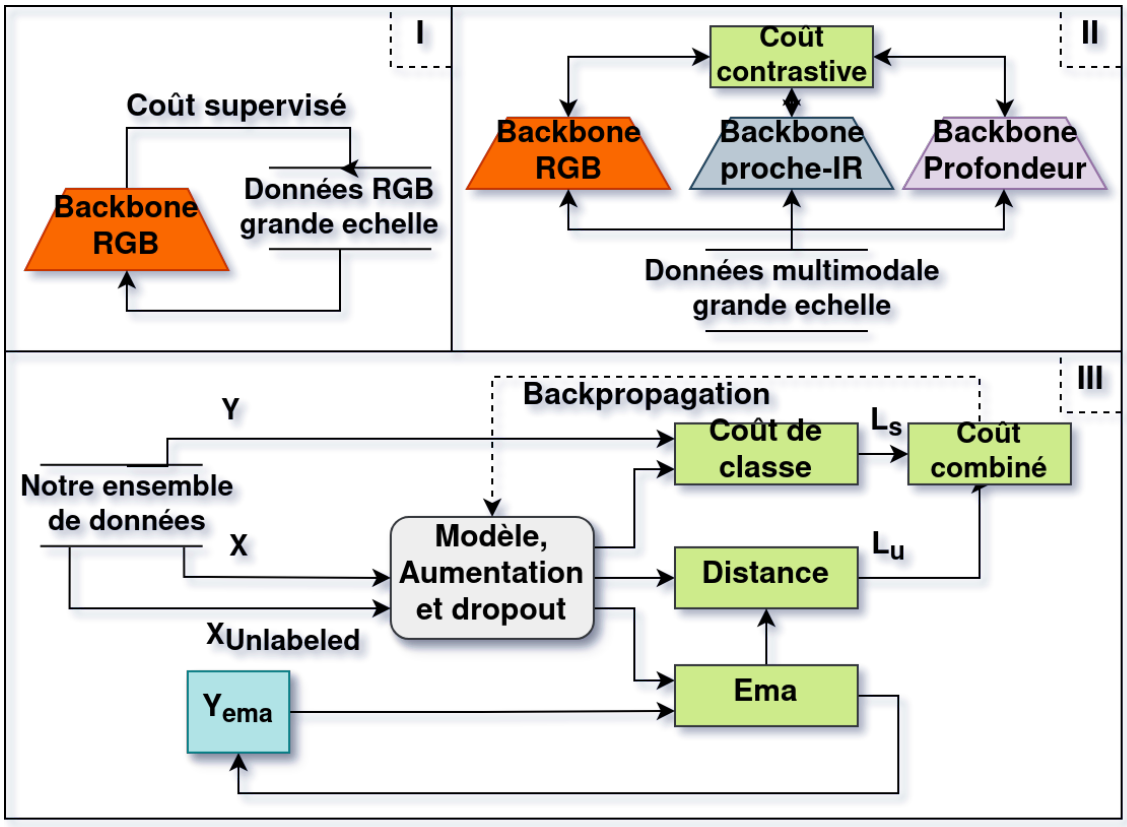


Figure 3.7 Principaux blocs de notre processus d'apprentissage.

avec un ensemble de données RGB à grande échelle, étiqueté, visant à l'apprentissage de la représentation visuelle. Bloc B : Pré-entraînement à l'apprentissage contrastif multimodale pour l'apprentissage de la représentation visuelle inter-modale. Bloc C : Processus d'apprentissage semi-supervisé de la représentation : Perte de classe : perte d'entropie croisée de reconnaissance. Distance : Il s'agit de la perte non supervisée, qui est littéralement la distance MSE entre 2 prédictions stochastiques du même point de données mais avec des augmentations différentes. *EMA* : la moyenne mobile exponentielle de l'étiquette prédite. X_u : données non étiquetées. X_l, Y : données étiquetées. Les blocs sont annotés en fonction de l'ordre séquentiel de leur exécution.

3.3 Conclusion

Dans ce chapitre, on présente une étude approfondie visant à améliorer la reconnaissance des matériaux en utilisant des caractéristiques de texture et de réflectivité multispectrale. L'étude se concentre sur l'intégration de modalités d'images multi-spectrales, dans les spectres visible et infrarouge proche, pour surmonter les limitations des approches basées uniquement sur l'aspect visuel. En exploitant la réflectivité et la texture des surfaces, le travail propose un système d'apprentissage profond capable de distinguer les matériaux dans des conditions non contrôlées, en tenant compte des phénomènes physiques comme l'auto-occlusion et l'auto-ombrage. Ce système utilise des techniques avancées de modélisation de la réflectivité et de l'apprentissage machine, y compris l'apprentissage semi-supervisé et la fusion de données, pour offrir une classification précise des matériaux à partir d'images capturées par une caméra de profondeur, démontrant ainsi l'importance de combiner différentes caractéristiques et modalités pour une reconnaissance efficace des matériaux.

Chapitre 4 - Validation expérimentale

Ce chapitre présente une explication détaillée des procédures proposées. Ces procédures englobent plusieurs étapes clés, à savoir la création du jeu de données, la sélection et l'apprentissage des modèles de fusion et de classification. De plus, on explique comment l'intégration de la modalité de réflexion est réalisée à travers de notre approche.

4.1 Jeu de données

La méthodologie suivie comprend des critères spécifiques à respecter : la taille, la diversité, le bon échantillonnage et le nombre de catégories. Comme indiqué dans [50] avec quelques changements inclus :

- Taille: Pour résoudre le problème du manque de données et de la difficulté d'acquisition des données, des méthodes semi-supervisées sont employées. Ainsi, les données recueillies doivent être suffisantes pour l'apprentissage semi-supervisé, qui nécessite une taille inférieure à celle des données nécessaires à l'apprentissage supervisé, de l'ordre d'une ou deux magnitudes.
- Nombre de catégories: Les mêmes catégories qu'ils [50] ont été utilisé, tout en supprimant les catégories qui ne sont pas présentes dans les environnements intérieurs et en fusionnant les catégories plus similaires afin de minimiser la taille des données requises.

4.1.1 Acquisition de données brutes

Étant donné qu'aucun jeu de données publique ne fournit les modalités nécessaires pour la méthode proposée, il devient indispensable de créer notre propre jeu de données. Cette dernière est élaborée à travers une démarche spécifique d'acquisition de données et d'annotation manuelle. Cette approche dépendante de la configuration matérielle utilisée,

un changement dans cette configuration entraînerait une modification conséquente du profil des données collectées. Ces données sont capturées à l'aide de la caméra de profondeur Intel RealSense d435i, qui présente un champ de vision horizontal et vertical de 87° x 58° pour les capteurs de profondeur et d'infrarouge proche, et de 69° x 42° pour le capteur RVB. La caméra a été réglée pour synchroniser l'image de profondeur avec l'image RVB et pour produire des images de profondeur en mode haute densité. Il est aussi important de souligner que le capteur infrarouge proche et les lasers projetés fonctionnent à une longueur d'onde de 850 nm. La figure 4.1 représente le flux de données pour extraire automatiquement des images structurellement différentes, en minimisant le flou de mouvement, à partir d'un flux vidéo multi-modale. Une sélection des images repose sur trois critères principaux :

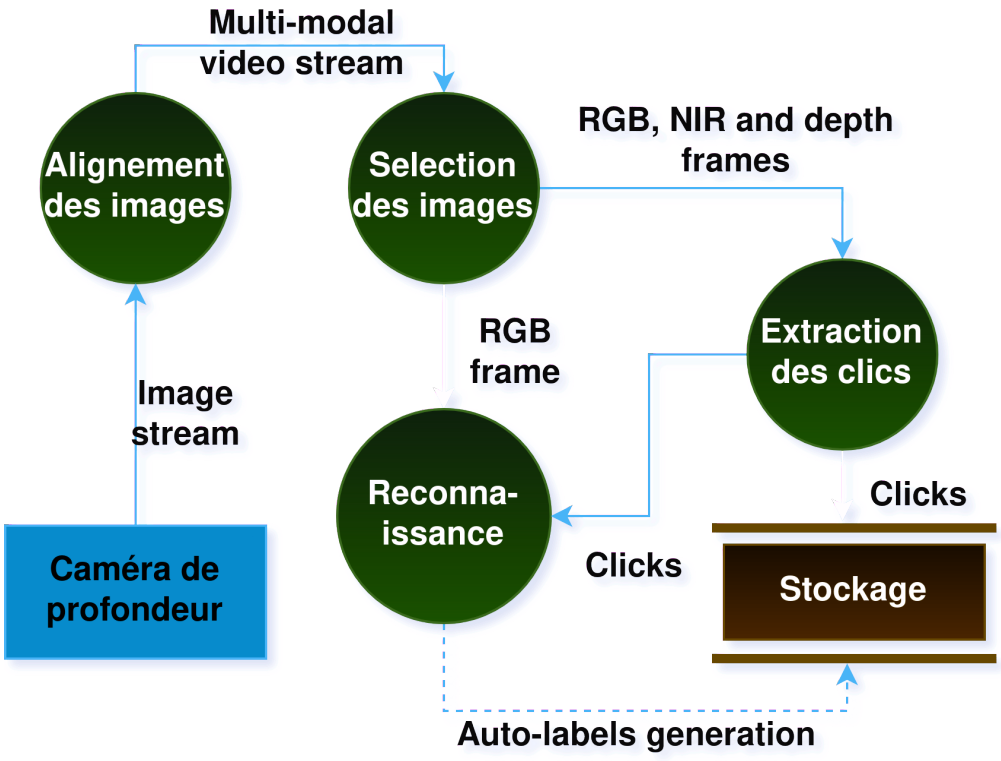


Figure 4.1 Flux de données de la procédure d'acquisition de données.

- L'écart de la distance des histogrammes de couleur entre l'image instantanée I_t et la précédente image sélectionnée I_{i-1} doit surpasser un seuil défini, comme le montre l'équation 4.1.
- La distance temporelle entre I_t et I_{i-1} doit excéder un seuil spécifique, tel que décrit par l'équation 4.2.
- La variance du laplacien de l'image I_t doit être supérieure à un seuil donné pour éliminer les images affectées par le flou de mouvement, comme indiqué dans l'équation 4.3.

Lorsque ces trois conditions sont remplies, l'image instantanée I_t est intégrée à la collection des images sélectionnées en tant que I_i. Après l'extraction des images, les *morceaux* de l'image finale seraient extraits à l'aide des méthodes de clic décrites dans [50].

$$|H_c(I_{i-1}) - H_c(I_t)| > H_{th} \tag{4.1}$$

$$T(I_t) - T(I_{i-1}) > T_{th} \tag{4.2}$$

$$Var(\triangle I_t) > v_{th} \tag{4.3}$$

H_c représente le vecteur de l'histogramme de couleurs. I_t est l'image instantanée, tandis que I_i et I_{i-1} correspondent aux images sélectionnées. $Var()$ désigne la variance statistique et $\triangle$ l'opérateur laplacien. Après l'extraction des images, les échantillons d'images finaux seraient extraits en utilisant les méthodes de sélection décrites dans [50].

4.1.2 Échantillonnage et annotation

Après la procédure d'acquisition du jeu de données, 7 500 caractéristiques d'entrée provenant de 1 300 images de scènes ont été produites. Plutôt qu'un découpage aléatoire, les données ont été divisées en ensembles d'apprentissage et de test en fonction des différentes scènes capturées, dans le but d'obtenir des mesures de test plus précises. En outre, 7 500 morceaux ont été extraits au hasard des images brutes et pseudo-étiquetés à l'aide d'un ensemble de modèles (CoatNet-6, Eva-02 et EffNet-L2) entraînés sur des données RVB à grande échelle, comme illustré dans la Figure 4.1. Ces modèles ont été pré-entraînés à la fois sur MINC et sur les morceaux initiaux pour le pseudo-étiquetage. En outre, 10 000 morceaux supplémentaires ont été récoltés au hasard à partir des données brutes pour former un jeu de données non étiquetées, facilitant ainsi le processus d'apprentissage semi-supervisé.

Pour l'apprentissage des branches RVB de nos réseaux proposés, nous avons d'abord utilisé l'ensemble complet des 23 catégories proposées par MINC. Ensuite, nous avons réduit le nombre de catégories de sortie à 15, comme expliqué ci-dessous :

- La céramique et la pierre polie sont combinées car elles sont rares et difficiles à collecter, et elles ont des caractéristiques réfléchissantes et visuelles similaires.
- Tapis est fusionné avec Tissu.
- Feuillage, Nourriture, Cheveux, Peau, Ciel et Eau sont fusionnés avec d'autres, car ils ne peuvent pas être considérés comme des matériaux.
- Papier peint est également fusionné avec le plastique, le papier ou la peinture en fonction du matériau cohérent. Lors de l'apprentissage sur MINC, cette catégorie n'est pas prise en compte.

Les situations où les données sont manquantes ou incomplètes sont fréquentes. Une

approche simple consiste à exclure simplement les échantillons incomplets de l'analyse, mais cela peut entraîner une perte d'informations précieuses, dans les cas où les données sont rares. Une autre solution consiste à considérer les données manquantes comme une catégorie distincte afin d'éviter les fausses prédictions dans les situations où la profondeur ou les points laser IR sont manquants.

4.1.3 Pré-traitement des données

Des caméras différentes sont utilisées pour capturer les images RVB et proche-infrarouge, résultant en un désalignement entre les deux images. Pour aborder ce problème, il est nécessaire de réduire le désalignement entre les deux images avant qu'elles ne soient introduites dans l'étape de fusion. Comme mentionné précédemment, l'image de profondeur est alignée avec l'image RVB, donc, l'alignement doit être appliqué sur l'image proche-infrarouge. Ainsi, pour atteindre l'alignement, nous avons sélectionné les recadrages redimensionnés les mieux adaptés de l'image proche-infrarouge pour chaque image multimodale en fonction de la distance de Chanfreiner et mesuré la profondeur moyenne du même cadre. Par la suite, nous avons conduit une régression linéaire sur le centre et la taille de la culture en fonction de la profondeur moyenne. En appliquant ces fonctions sur les cadres proche-infrarouge, nous avons obtenu un désalignement négligeable, comme montré dans la 4.2.

4.1.4 Base de données grande échelle

Plusieurs ensembles de données publiques sont accessibles en ligne, notamment FMD, CUReT [82], KTH-TIPS [83] et MINC. Parmi ces ensembles de données, celui de MINC se distingue par sa taille impressionnante, avec 1,2 million de points de données. Ce jeu de données à grande échelle correspond parfaitement aux exigences des modèles

Figure 4.2 Images RVB et proche-infrarouge superposés avec alignement à gauche et sans alignement à droite.

de classification modernes de pointe et est particulièrement bien adapté aux tâches de reconnaissance difficiles dans les scénarios du monde réel. En conséquence, nous avons délibérément choisi d'utiliser le jeu de données MINC à des fins de pré-entraînement. Cette décision a pour double objectif d'améliorer la robustesse des prédictions de notre modèle dans la branche RVB pendant l'entraînement et d'atténuer les risques potentiels de surapprentissage.

4.2 Sélection des modèles d'apprentissage

Pour chaque choix d'un modèle, des critères scientifiques doivent être sélectionnés. Les modèles seront évalués selon ces critères pour valider notre choix.

4.2.1 Modèle de classification

Pour valider les stratégies de sélection de modèles décrites dans la Section 3.2.1, notamment après leur performance post-évaluation sur ImageNet-1k, nous avons effectué des évaluations supplémentaires sur le jeu de données MINC. Cette évaluation s'est concentrée sur les capacités des modèles à reconnaître les matériaux dans un contexte de tâche "in the wild", qui implique d'intégrer à la fois le contexte de texture et d'objet dans un environnement non contrôlé. Nos critères de sélection pour les modèles étaient basés sur leur performance au sein de leurs familles respectives. De la famille CNN, nous avons choisi ResNet-152 et EfficientNet-L2, tandis que CoatNet-6 a été sélectionné de la famille hybride. Pour la famille des Transformateurs , nous avons opté pour ViT-G, EVA-02-L, et Swin-V2-G, en prenant en compte leur précision zéro-coup élevée comme facteur décisif. De plus, GoogLeNet, un modèle à l'état de l'art (SOTA), a également été inclus dans notre ligne de test.

Nous avons entraîné tous les réseaux pendant 10 itérations d'apprentissage sur le jeu de données MINC en utilisant l'optimiseur Adam avec une taille de lot de 8 et un taux d'apprentissage de base de 10^{-4}, diminuant de 25% tous les 100000 itérations avec une taille de morceau d'entrée de 256, ainsi qu'une décroissance du taux d'apprentissage selon une fonction cosinus. Les résultats des tests de précision sur le jeu de données MINC, résumés dans le tableau 4-1, montrent une variation notable des performances entre les différents modèles, en particulier en termes de précision Top-1. CoatNet-6

se distingue comme le modèle le plus performant, atteignant une précision Top-1 de 90,5% et une précision Top-5 de 99,4%, surpassant ainsi les autres architectures. Cependant, ces résultats doivent être interprétés dans le cadre de l'application spécifique de reconnaissance des matériaux dans des environnements non contrôlés ("in the wild").

Les variations observées entre les modèles CNN, hybrides et Transformateurs révèlent plusieurs aspects importants. Tout d'abord, CoatNet-6, en tant que modèle hybride combinant les avantages des réseaux convolutifs et des Transformateurs , semble mieux s'adapter aux défis des environnements non contrôlés. Sa capacité à capturer de meilleurs informations locales (texture) et globales (contexte d'objet) permet d'améliorer la robustesse de la classification dans des conditions où la variabilité des textures et des objets est élevée. Cela justifie sa supériorité par rapport aux autres modèles.

En revanche, les performances plus faibles du ViT-G/14, avec une précision Top-1 de 71,1%, mettent en évidence les limitations des Transformateurs purs dans le cas où la taille du jeu de données est limitée. Cela suggère que, malgré leur succès sur des tâches de classification sur des jeux de données large-échelle, ces modèles pourraient ne pas être suffisamment adaptés aux scénarios où les informations contextuelles doivent être apprises à partir du cas particulier.

De plus, les variations observées dans les précisions Top-5 (où tous les modèles obtiennent des résultats supérieurs à 98%) démontrent que, même si certains modèles sont moins performants en termes de choix de la classe correcte en premier lieu, ils parviennent tout de même à inclure la bonne classe parmi leurs cinq premières prédictions. Cela souligne une capacité potentielle à être utilisée dans des applications où une certaine tolérance à l'erreur est acceptable, comme la pré-classification suivie d'une analyse humaine.

Tableau 4-1 Test de la précision sur le jeu de données MINC.

Modèle	Précision top-1 test	Précision top-5 test
CoatNet-6	**90.5%**	**99.4%**
Eva-02-L	89.2%	99.0%
ResNet101	87.5%	98.9%
EfftNet-L2	85.5%	98.7%
Swin-v2-L	88.2%	98.9%
ViT-G/14	71.1%	93.9%
GoogLeNet	85.5%	98.1%

4.2.2 Modèle de fusion de données

Dans notre étude, nous présentons une analyse comparative de notre approche avec la méthode de l'état de l'art proposée par Bell et al. [50]. Pour cette comparaison, nous avons sélectionné trois modèles - CoatNet-6, EVA-02-L et EfficientNet-L2 - et les avons entraînés sur notre jeu de données. Ces modèles ont été choisis pour servir de base à notre modèle de fusion. Nous avons exploré différentes méthodes de fusion, allant de la fusion tardive à la fusion précoce, comme le montre la figure 4.3, dans tous les architecture de base. En outre, nous avons intégré une fusion de niveau intermédiaire avec l'extracteur de caractéristiques DenseFuse pour les modèles EfficientNet-L2 et CoatNet-6. Cela nous a permis d'évaluer l'impact des différentes stratégies de fusion et l'efficacité de l'extraction de caractéristiques pour améliorer les performances du cadre d'apprentissage multi-modale. Les résultats du tableau 4-2 montrent une variation des performances selon la méthode de fusion utilisée. La fusion tardive se démarque clairement par rapport aux autres, notamment avec des gains de précision top-1 substantiels, particulièrement pour les modèles CoatNet-6 et EVA-02-L. Par exemple, pour le modèle CoatNet-6, la précision top-1 passe de 82.7% pour la version entraînée uniquement avec les données RVB à 88.9% avec la fusion tardive. Cette tendance est similaire pour EVA-02-L, où la précision augmente de 80.2% à 86.4% avec la fusion tardive.

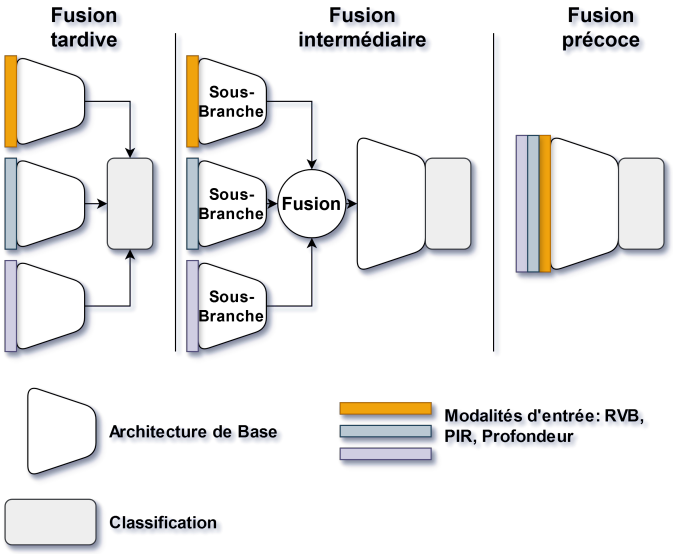


Figure 4.3 Représentation des méthodes de fusion à travers différentes architectures de bases.

Ces résultats suggèrent que la fusion tardive, en conservant la structure de l'architecture de base des modèles pré-entraînés, permet de mieux exploiter les informations complémentaires issues des différentes modalités sans altérer la représentation spatiale et sémantique apprise par le modèle. En revanche, la fusion précoce, bien qu'elle intègre les données multi-modales dès les premières couches du réseau, semble entraîner une dilution de l'information pertinente. Cela se traduit par des performances moins bonnes, comme en témoigne la baisse de précision observée avec EVA02, où la fusion précoce donne une précision de seulement 65.5%, contre 80.2% pour le modèle basé uniquement sur les données RVB et 86.4% pour la fusion tardive.

La fusion intermédiaire, bien qu'elle montre des performances légèrement supérieures à la fusion précoce dans certains cas, notamment avec CoatNet-6, n'atteint pas les niveaux de la fusion tardive. Cela peut s'expliquer par le fait que, bien qu'elle préserve une partie des informations structurales, elle introduit une complexité supplémentaire qui n'est pas pleinement exploitée par les modèles utilisés.

Tableau 4-2 Comparaison entre les méthodes de fusion.

Modèle	notre jeu de données Précision top-1
$EffNetL2_{RVB}$	79.9%
$EffNetL2_{tardive}$	**83.3%**
$EffNetL2_{prcoce}$	75.3%
$EffNetL2_{Intermdiaire}$	76.8%
$EVA02_{RVB}$	80.2%
$EVA02_{tardive}$	**86.4%**
$EVA02_{prcoce}$	65.5%
$CoatNet6_{RVB}$	82.7%
$CoatNet6_{tardive}$	**88.9%**
$CoatNet6_{prcoce}$	80.2%
$CoatNet6_{Intermdiaire}$	81.0%

4.3 Évaluation et processus d'apprentissage

Notre méthodologie d'apprentissage est divisée en trois phases distinctes : apprentissage supervisé à l'aide de vastes ensembles de données RVB, pré-entraînement avec des données multimodales et apprentissage semi-supervisée incorporant des données provenant de toutes les modalités. Des augmentations spécifiques sont adaptées à chaque phase. Pour l'entraînement avec les données RVB, nous mettons en œuvre les mêmes augmentations que celles détaillées dans [50]. Dans le cadre de l'assemblage temporel, nous utilisons des augmentations préservant la texture, tout en introduisant diverses augmentations de bruit telles que l'application de flous et de bruit blanc aux images RVB. En outre, les augmentations de position, y compris le retournement et le recadrage, sont appliquées uniformément à toutes les images. L'utilisation d'augmentations au cours de la phase supervisée a été stratégiquement choisie pour réduire le surapprentissage excessif. Ces augmentations reflètent celles utilisées dans l'assemblage temporel afin de garantir la cohérence du régime d'apprentissage. L'ensemble du processus d'apprentissage est illustré en détail dans la figure 3.7, et la fonction de coût adoptée est décrite dans l'équation 4.4.

$$L_{M,Aug}(x,y) = C_E(M_{Aug}(X_l), Y_l) + MSE(M_{Aug}(X_u), Y_{ema}(X_u)) \tag{4.4}$$

C_E, MSE et M_{Aug} sont respectivement l'entropie croisée, l'erreur quadratique moyenne des pertes et le modèle avec augmentation et abandon stochastiques.

4.3.1 Évaluation de pré-entraînement

Phase1: Nous avons effectué des phases de pré-entraînement en utilisant trois réseaux - CoatNet-6, EVA-02-L et EfficientNet-L2 - sur le jeu de données MINC, chacun pour une durée de 5 itérations d'apprentissage. Le pré-entraînement a utilisé l'optimiseur Adam avec une taille de lot de 8 et un taux d'apprentissage initial de 10^{-4}. Un planificateur de taux d'apprentissage en cosinus a été mis en œuvre pour diminuer progressivement le taux au cours de la période d'apprentissage. Chaque réseau a traité des données d'entrée comprenant des morceaux de résolution de 256 pixels. Les résultats, présentés dans la figure 4-1, ont révélé que tous les réseaux présentaient des niveaux de performance comparables lorsqu'ils étaient intégrés à des techniques de fusion.

Phase 2: Notamment, Apprentissage Contrastif de la Représentation Visuelle (ACRV) a été appliqué exclusivement à la fusion tardive. Pour faciliter l'apprentissage de modèles multi-modaux, nous avons sélectionné des ensembles de données spécifiques accessibles au public pour le pré-entraînement des réseaux. Pour la RVB-Profondeur, nous avons utilisé les ensembles de données ScaNet et SUN RVB-D, qui contiennent respectivement 12 000 et 10 335 points de données. Pour RVB-proche-infrarouge, le jeu de données de l'EPFL, comprenant 1 900 points de données, a été utilisé. Les résultats de la précision zéro-coup, présentés dans le tableau 4-3, montrent des variations intéressantes entre les modèles avec et sans ACRV. Il est essentiel d'interpréter ces variations pour comprendre les implications de l'apprentissage contrastif dans ce contexte.

Tableau 4-3 Précision moyenne zéro-coup avec et sans ACRV

Architecture de base	Précision zéro-coup sans ACRV	Précision zéro-coup avec ACRV
$EffNetL2_{tardive}$	24.2%	25.9%
$EVA02_{tardive}$	62.1%	69.5%
$CoatNet6_{tardive}$	61.0%	68.8%

Tout d'abord, nous constatons que l'application de l'ACRV n'a pas toujours conduit à une amélioration des performances. Par exemple, pour le modèle EfficientNet-L2 avec fusion tardive, la précision zéro-coup diminue légèrement avec l'intégration de l'ACRV (de 25,9% à 24,2%). Cette baisse pourrait s'expliquer par le fait que l'architecture EfficientNet-L2, dans ce contexte particulier, n'exploite pas pleinement les avantages de l'apprentissage contrastif des représentations visuelles. Cela indique que l'efficacité de l'ACRV peut dépendre fortement de l'architecture du réseau et des modalités utilisées.

En revanche, pour les modèles EVA-02 et CoatNet-6, nous observons des gains de performance notables avec ACRV. Par exemple, pour EVA-02, la précision zéro-coup passe de 69,5% à 62,1% sans ACRV, ce qui montre une augmentation significative d'environ 7,4%. Un comportement similaire est observé avec CoatNet-6, où la précision augmente de 68,8% à 61,0%, soit une amélioration de 7,8%. Ces résultats suggèrent que ces architectures, en particulier EVA-02 et CoatNet-6, sont plus aptes à capturer les représentations complexes des données multi-modales via l'apprentissage contrastif, ce qui améliore leur capacité à généraliser à des tâches de reconnaissance zéro-coup.

4.3.2 Évaluation entre modalités et état de l'art

Dans la première expérience, les trois modèles choisis - CoatNet-6, EVA-02-L et EfficientNet-L2 - sont évalués en utilisant à la fois des méthodes de fusion précoce et tardive. La deuxième expérience a adopté une approche plus structurée pour évaluer les

performances des modèles dans différentes conditions. Dans un premier temps, nous avons testé les modèles en utilisant uniquement des données RVB afin d'établir une base de référence. Nous avons ensuite effectué un test complet en utilisant toutes les modalités disponibles. Enfin, nous avons exclu les informations de profondeur pour nous concentrer sur les modalités RVB et proche infrarouge.

Les résultats, présentés dans le tableau 4-4, montrent que nos modèles surpassent les méthodes de pointe existantes sur le jeu de données MINC RVB, avec des améliorations significatives lorsqu'ils sont entraînés avec des modalités supplémentaires. Par exemple, pour le modèle CoatNet-6, la précision passe de 89,1% en utilisant uniquement les données RVB à 94,3% lorsque la fusion tardive des données RVB, PIR et profondeur est appliquée. Cela souligne l'importance des informations supplémentaires apportées par la modalité de profondeur, qui permet de mieux capturer la géométrie et la texture des matériaux. La fusion tardive, en particulier, a démontré son efficacité en intégrant ces informations complémentaires tout en préservant les détails capturés par chaque modalité.

Un autre exemple notable est celui d'EfficientNet-L2, où la précision progresse de 81,2% avec uniquement les données RVB à 87,3% avec la fusion tardive des modalités RVB, PIR et profondeur. Cette amélioration de plus de 6 points montre que la profondeur apporte des informations essentielles qui aident à la discrimination des matériaux, notamment pour des surfaces aux caractéristiques géométriques complexes. La légère baisse de performance observée lors de la fusion précoce (84,3% pour CoatNet-6) par rapport à la fusion tardive (94,3%) confirme que l'intégration des modalités doit se faire de manière optimisée afin de maximiser la complémentarité des données.

L'ajout progressif des modalités au réseau permet également de valider leur complémentarité. Par exemple, pour le modèle EVA-02-L, la précision passe de 87,2% en

utilisant uniquement les données RVB à 91,4% avec la fusion tardive des données RVB, PIR et profondeur. Toutefois, l'élimination de l'information de profondeur entraîne une baisse notable, avec une précision de seulement 65,5% lors de l'utilisation des modalités RVB et PIR avec fusion précoce. Cela démontre que l'information de profondeur est indispensable pour extraire des indices géométriques essentiels à la classification de certains matériaux.

Tableau 4-4 Comparaison entre modalités et état de l'art

Architecture	Précision MINC top-1	Précision MINC top-5	Précision top-1 Notre jeu de données
$GoogLeNet_{RVB}$	81.1%	97.2%	75.6%
$EffNetL2_{RVB}$	84.5%	98.7%	81.2%
$EffNetL2_{RVB-PIR-tardive}$	—	—	82.0%
$EffNetL2_{RVB-PIR-Depth-prcoce}$	—	—	85.3%
$EffNetL2_{RVB-PIR-Depth-tardive}$	—	—	87.3%
$EVA02_{RVB}$	86.5%	99.3%	87.2%
$EVA02_{RVB-PIR-tardive}$	—	—	88.4%
$EVA02_{RVB-PIR-Depth-prcoce}$	—	—	65.5%
$EVA02_{RVB-PIR-Depth-tardive}$	—	—	91.4%
$CoatNet6_{RVB}$	89.7	99.4%	89.1%
$CoatNet6_{RVB-PIR-tardive}$	—	—	91.7%
$CoatNet6_{RVB-PIR-Depth-prcoce}$	—	—	84.3%
$CoatNet6_{RVB-PIR-Depth-tardive}$	—	—	**94.3%**

Après avoir effectué une analyse comparative de l'algorithme que nous proposons, qui combine CoatNet6 et la fusion tardive, par rapport aux méthodes basées sur les RVB pour chaque catégorie, comme indiqué dans le tableau 4-5, nous avons observé des améliorations notables en termes de précision dans presque toutes les catégories. Ces résultats suggèrent que l'approche multimodale permet de mieux capturer certaines caractéristiques des matériaux, en particulier celles qui sont difficiles à distinguer en se basant uniquement sur les informations RVB. Une analyse plus fine des catégories révèle des écarts significatifs dans certaines d'entre elles, notamment le métal, le miroir, le carrelage et le plastique. Par exemple, pour la catégorie du métal, la méthode RVB

appliquée à notre jeu de données montre une forte baisse de précision (58,2%), tandis que l'approche multimodale atteint une précision de 77,8%. Cette augmentation de près de 20% peut être attribuée à la complexité du métal en tant que matériau : les reflets et la texture variable rendent difficile son identification uniquement via les caractéristiques de couleur. La fusion tardive permet de mieux exploiter ces informations contextuelles, compensant ainsi les limitations des méthodes RVB.

Le **miroir** présente un cas encore plus frappant, où la précision passe de 23,5% avec l'approche RVB à 73,0% avec notre méthode multimodale. L'échec des méthodes RVB à reconnaître correctement le miroir peut être expliqué par l'aspect hautement réfléchissant de ce matériau, qui confond les modèles traditionnels d'apprentissage basés sur la couleur. Notre méthode, en intégrant d'autres modalités, parvient à surmonter cette limitation en capturant des indices structurels et spatiaux supplémentaires, réduisant ainsi la confusion entre le miroir et d'autres surfaces brillantes.

Pour le **carrelage**, une augmentation notable de la précision est également observée (de 72,4% à 89,3%), ce qui suggère que l'approche multimodale aide à distinguer ce matériau dans des conditions d'éclairage variées. Le carrelage, souvent soumis à des reflets, présente des motifs répétitifs qui peuvent perturber les méthodes basées sur le RVB, mais sont mieux capturés par la fusion de modalités.

Le **plastique** est un autre exemple intéressant : bien que la méthode RVB atteigne déjà une performance relativement correcte (71,7%), l'utilisation de notre approche multimodale augmente la précision à 80,8%. Cette amélioration peut être expliquée par la diversité des plastiques en termes de texture et de finition (lisse, mat, brillant), des propriétés que notre méthode parvient à mieux distinguer en intégrant des informations complémentaires provenant de modalités différentes.

Dans d'autres catégories, telles que le **bois** et la **pierre**, où les méthodes basées sur RVB sont déjà performantes, les gains obtenus par notre approche sont moins prononcés mais tout de même significatifs. Par exemple, pour le bois, la précision passe de 96,5% à 97,7%, ce qui démontre que, même pour des matériaux plus facilement reconnaissables, notre approche parvient à apporter des améliorations subtiles, notamment en renforçant la robustesse face à des variations de texture ou de contexte environnemental.

Tableau 4-5 Comparaison entre notre méthode et l'état de l'art en termes de précision par catégorie.

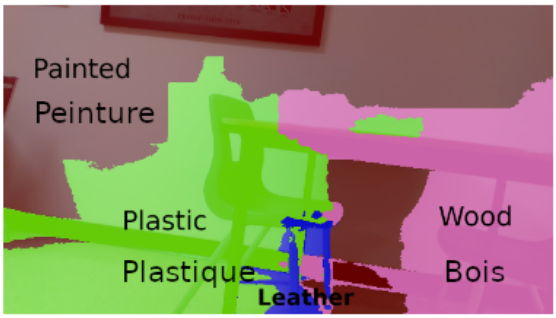
Catégorie	Précision RVB MINC	Précision RVB notre jeu de données	Précision multimodale notre jeu de données
Brique	89,4%	88,1%	92,3%
Pierre polie	89,7%	89,4%	92,2%
Tissu	95,1%	79,8%	84,3%
Verre	90,5%	83,4%	84,9%
Cuir	77,3%	88,6%	90,1%
Métal	87,9%	**58,2%**	**77,8%**
Miroir	73,8%	**23,5%**	**73,0%**
Autre	89,4%	83,2%	87,2%
Peint	90,1%	92,7%	83,8%
Papier	68,6%	84,3%	89,3%
Plastique	66,2%	**71,7%**	**80,8%**
Pierre	74,8%	91,4%	96,0%
Carrelage	84,2%	**72,4%**	**89,3%**
Bois	94,5%	96,5%	97,7%

4.3.3 Expérience de segmentation

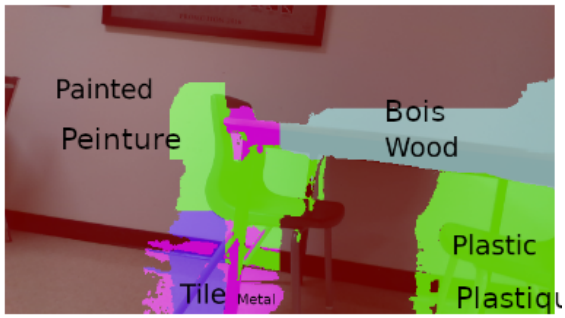
En plus de la classification basée sur les morceaux , le modèle de reconnaissance pourrait être utilisé pour une expérience de segmentation en l'utilisant comme une fenêtre glissante sur l'ensemble de l'image pour effectuer une prédiction dense. Cependant, les méthodes de classification basées sur les morceaux peuvent ne pas avoir de cohérence locale dans leurs résultats, de sorte qu'un Champ Aléatoire Conditionnel (CRF) a été attaché à l'arrière-plan pour générer le résultat final avec une plus grande précision,

comme démontré dans [50]. La figure 4.4 présente les résultats. La figure montre que le modèle proposé, entraîné sur toutes les modalités, donne des résultats plus cohérents sur le plan visuel. Il est plus précis dans la détection du plastique et du métal, ce que d'autres méthodes ne parviennent pas à faire.

(a) Image d'entrée.



(b) CoatNet6 entraîné sur RGB.



(c) CoatNet6 entraîné sur toutes les modalités.

Figure 4.4 Expérience de segmentation à l'aide de la CRF : comparaison entre les résultats de CoatNet6 entraîné sur des données RVB et des données de modalité complète avec fusion tardive.

4.4 Conclusion

En conclusion, ce chapitre a exposé les résultats obtenues dans le cadre de la reconnaissance des matériaux basée sur les modalités des profondeur, réflectivité et images multispectrale. Ces résultats prouve la complémentarité des modalités sélectionnées entre eux, ce qui a contribuer à l'amélioration des performances de la méthode proposée en termes de précision. Le cadre expérimental de la recherche présenté dans ce chapitre est basé sur cinq volets principales :

- La création d'un jeu de données multimodales spécifiquement dédiée à l'apprentissage du système développé, en réponse à l'absence de jeux de données publiques couvrant les modalités requises par notre méthode.
- La sélection des métriques pour évaluer les modèles de fusion et de classification d'images, afin de sélectionner les plus convenables à la méthode envisagée.
- Le développement du système final, qui incorpore les modèles préalablement sélectionnés pour optimiser les processus de reconnaissance.
- Le pré-entraînement des modèles avec des jeux de données publique à grande échelle pour améliorer la capacité de généralisation.
- L'entraînement et l'évaluation inter-modale du système final, afin de valider l'intégration efficace des différentes modalités.

L'intégration graduelle des modalités au système proposé (commençant par l'image visuelle seule, suivie de l'ajout de l'imagerie proche-infrarouge, puis de l'ensemble des modalités disponibles) a entraîné une amélioration progressive de la précision de la reconnaissance, démontrant ainsi la complémentarité des modalités sélectionnées. Cette augmentation de la précision moyenne est principalement due à la capacité améliorée du système à identifier avec plus de précision certaines catégories spécifiques, telles que

les miroirs, les métaux, le plastique et le carrelage. Ces catégories se caractérisent par des surfaces lisses et réfléchissantes qui présentent des aspects visuels variés selon les conditions d'éclairage. Par conséquent, ces résultats mettent en évidence l'importance cruciale de la réflectivité pour renforcer la reconnaissance de ces classes de matériaux.

Chapitre 5 - Conclusion et perspectives

Nous proposons une méthode de reconnaissance des matériaux qui intègre de manière unique la texture multispectrale et les caractéristiques de réflectivité de la surface dans le spectre visuel-infrarouge. Notre approche est structurée en deux phases : la première implique la fusion de caméras multimodales pour consolider les caractéristiques et traiter le désalignement des caméras, et la seconde est un modèle de classification spécialement conçu pour la reconnaissance des matériaux. Cette méthode se distingue par sa capacité à identifier avec précision les classes de surfaces matérielles dans des environnements non contrôlés, en combinant les caractéristiques de réflectivité proposées et la texture du matériau dans le spectre PIR-visible. Nous utilisons des projecteurs laser PIR pour générer des réflexions dans le cadre PIR, qui sont ensuite modélisées à l'aide des modèles analytique et Phong. Nos résultats indiquent qu'en estimant la lumière réfléchie à partir de la disparité entre les images RVB et PIR, et en déduisant la géométrie de la surface à partir des données de profondeur, nous pouvons obtenir une approximation fiable des caractéristiques de réflexion.

L'efficacité de notre méthode est encore renforcée par la mise en œuvre de solutions visant à résoudre les problèmes de limitation des données couramment rencontrés dans les approches multi-spectrales et hyper-spectrales précédentes. Ces solutions comprennent l'adoption d'un cadre d'apprentissage semi-supervisé et l'utilisation stratégique d'ensembles de données RVB à grande échelle pour la régularisation. Les évaluations comparatives de notre modèle par rapport aux méthodes de pointe existantes sur l'ensemble de données MINC et sur notre propre ensemble de données ont révélé que notre approche permet d'obtenir une précision de prédiction comparable en utilisant uniquement des données RVB. Plus particulièrement, il y a une augmentation significative

de 6 % de la précision des tests lorsque toutes les modalités d'entrée sont utilisées sur notre ensemble de données.

5.1 Perspectives

Pour améliorer ce travail, les recommandations suivantes pourraient être envisagées :

- **Extension aux capteurs multispectraux supplémentaires** : L'intégration de capteurs à infrarouge à ondes courtes ou moyennes permettrait d'enrichir la texture multispectrale utilisée, augmentant ainsi la précision et les capacités discriminantes du modèle, particulièrement dans des environnements où les conditions lumineuses varient fortement.
- **Projections laser à plusieurs longueurs d'onde** : L'ajout de projections laser couvrant diverses longueurs d'onde pourrait réduire la sensibilité aux radiations infrarouges présentes dans la lumière ambiante. Cette approche renforcerait encore plus la robustesse du modèle, notamment dans des environnements non contrôlés.
- **Évaluation du vieillissement des structures mécaniques** : Une application importante de notre méthode pourrait être le suivi et l'évaluation du vieillissement des structures mécaniques. En identifiant les modifications subtiles des caractéristiques de surface au fil du temps, notre modèle pourrait être utilisé pour détecter les signes précoces d'usure ou de dégradation, améliorant ainsi la maintenance prédictive des infrastructures.
- **Analyse sémantique dans les environnements virtuels** : L'adoption de cette approche dans le cadre de la réalité augmentée pourrait permettre une meilleure reconnaissance des matériaux. Cette avancée faciliterait l'interaction utilisateur-environnement dans des simulations, avec des applications potentielles dans des domaines tels que la formation industrielle et la conception virtuelle.

- **Optimisation des réseaux neuronaux profonds pour la reconnaissance des matériaux dans des cas spécifiques** : Le développement de jeux de données spécifiques aux matériaux et aux textures pourrait permettre une optimisation ciblée des modèles de réseaux neuronaux utilisés. Cette spécialisation améliorerait la précision du modèle, en particulier dans les tâches de classification de matériaux complexes dans des environnements non structurés.
- **Applications robotiques et industrielles** : Enfin, notre méthode pourrait être intégrée dans des systèmes industriels automatisés ou des robots intelligents, pour des tâches allant de la reconnaissance en temps réel des matériaux à la gestion automatisée des processus de recyclage. L'implémentation de cette technologie dans des systèmes embarqués offrirait des opportunités d'automatisation avancée dans divers secteurs.

Références

[1] Factoryfuture, "Capteur à effet-hall," https://www.factoryfuture.fr/capteur-effet-hall/, 2024, accessed: 2024-02-02.

[2] N. Kirchner, D. Hordern, D. Liu, and G. Dissanayake, "Capacitive sensor for object ranging and material type identification," *Sensors and Actuators A: Physical*, vol. 148, no. 1, pp. 96–104, 2008.

[3] "Spectro xsort," https://www.spectro.com/products/xrf-spectrometer/xsort-xrf-gun-handheld-analyzer#, 2024, accessed: 2024-02-02.

[4] "Ultrasonic tof material detection sensor," https://www.dfrobot.com/product-2688.html, 2024, accessed: 2024-02-02.

[5] M. Selek, "A new autofocusing method based on brightness and contrast for color cameras," *Advances in Electrical and Computer Engineering*, vol. 16, pp. 39–44, 01 2016.

[6] Flir, "Flir," https://www.flir.ca/, 2024, accessed: 2024-02-02.

[7] I. Realsense, "Intel realsense," https://www.intelrealsense.com/depth-camera-d435i/, 2024, accessed: 2024-02-02.

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[9] "Image segmentation detailed overview," https://www.superannotate.com/blog/image-segmentation-for-machine-learning, 2024, accessed: 2024-02-02.

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

[11] B. Koonce and B. Koonce, "Efficientnet," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 109–123, 2021.

[12] "Review — coatnet: Marrying convolution and attention for all data sizes," https://sh-tsang.medium.com/review-coatnet-marrying-convolution-and-attention-for-all-data-sizes-1462b9bc25ac, 2024, accessed: 2024-02-02.

[13] V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters*, vol. 84, pp. 63–69, 2016.

[14] N. Salamati, C. Fredembach, and S. Süsstrunk, "Material classification using color and nir images," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 2009, pp. 216–222.

[15] P. Saponaro, S. Sorensen, A. Kolagunda, and C. Kambhamettu, "Material classification with thermal imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4649–4656.

[16] D. Guarnera, G. C. Guarnera, A. Ghosh, C. Denk, and M. Glencross, "Brdf representation and acquisition," in *Computer Graphics Forum*, vol. 35. Wiley Online Library, 2016, pp. 625–650.

[17] Z. Erickson, N. Luskey, S. Chernova, and C. C. Kemp, "Classification of household materials via spectroscopy," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 700–707, 2019.

[18] S. M. Zainab, K. Khan, A. Fazil, and M. Zakwan, "Foreign object debris (fod) classification through material recognition using deep convolutional neural network with focus on metal," *IEEE Access*, vol. 11, pp. 10 925–10 934, 2023.

[19] J. Villalba-Diez, D. Schmidt, R. Gevers, J. Ordieres-Meré, M. Buchwitz, and W. Wellbrock, "Deep learning for industrial computer vision quality control in the printing industry 4.0," *Sensors*, vol. 19, no. 18, p. 3987, 2019.

[20] K. Miyawaki and S. Okabe, "Material recognition for mixed reality scene including objects' physical characteristics." in *KMIS*, 2019, pp. 219–224.

[21] H. He, D.-W. Sun, Z. Wu, H. Pu, and Q. Wei, "On-off-on fluorescent nanosensing: Materials, detection strategies and recent food applications," *Trends in Food Science & Technology*, vol. 119, pp. 243–256, 2022.

[22] J. E. Lenz, "A review of magnetic sensors," *Proceedings of the IEEE*, vol. 78, no. 6, pp. 973–989, 1990.

[23] C. Berthomieu and R. Hienerwadel, "Fourier transform infrared (ftir) spectroscopy," *Photosynthesis research*, vol. 101, pp. 157–170, 2009.

[24] irisndt, "Spectrometer material analysis," https://www.irisndt.com/uk/laboratory-services/spectrometer-material-analysis/, 2024, accessed: 2024-02-02.

[25] R. Lukac, *Single-sensor imaging: methods and applications for digital cameras*. CRC Press, 2018.

[26] paperswithcode, "Image classification on imagenet," https://paperswithcode.com/sota/image-classification-on-imagenet, 2024, accessed: 2023-09-20.

[27] S. Y. Chekmenev, A. A. Farag, W. M. Miller, E. A. Essock, and A. Bhatnagar, "Multiresolution approach for noncontact measurements of arterial pulse using thermal imaging," *Augmented Vision Perception in Infrared: Algorithms and Applied Systems*, pp. 87–112, 2009.

[28] R. Maier and D. Cremers, *RGB-D Vision*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020, pp. 1–11. [Online]. Available: https://doi.org/10.1007/978-3-642-41610-1_109-1

[29] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[30] "Industrial acoustic imaging camera flir si124," https://www.flir.eu/products/si124/?vertical=condition%20monitoring&segment=solutions, 2024, accessed: 2024-02-02.

[31] S. Master, "Large scale object detection," *Czech Technical University*, 2014.

[32] R. K. Sinha, R. Pandey, and R. Pattnaik, "Deep learning for computer vision tasks: a review," *arXiv preprint arXiv:1804.03928*, 2018.

[33] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020.

[34] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[35] M. Ali and S. Khan, "Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4675–4679.

[36] "Cs231n convolutional neural networks for visual recognition," https://cs231n.github.io/convolutional-networks/, 2024, accessed: 2024-02-02.

[37] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva-02: A visual representation for neon genesis," *arXiv preprint arXiv:2303.11331*, 2023.

[38] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[39] C. Yang, S. Qiao, Q. Yu, X. Yuan, Y. Zhu, A. Yuille, H. Adam, and L.-C. Chen, "Moat: Alternating mobile convolution and attention brings strong vision models," *arXiv preprint arXiv:2210.01820*, 2022.

[40] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[41] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[42] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "Swinfuse: A residual swin transformer fusion network for infrared and visible images. arxiv 2022," *arXiv preprint arXiv:2204.11436*, 2022.

[43] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of mathematical imaging and vision*, vol. 18, no. 1, pp. 17–33, 2003.

[44] G. Winkler, *Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction*. Springer Science & Business Media, 2012, vol. 27.

[45] R. Harlick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[46] P. Cavalin and L. S. Oliveira, "A review of texture classification methods and databases," in *2017 30th SIBGRAPI Conference on graphics, patterns and images tutorials (SIBGRAPI-T)*. IEEE, 2017, pp. 1–8.

[47] F. Bianconi and A. Fernández, "Evaluation of the effects of gabor filter parameters on texture classification," *Pattern recognition*, vol. 40, no. 12, pp. 3325–3335, 2007.

[48] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3828–3836.

[49] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks for texture classification," *arXiv preprint arXiv:1707.07394*, 2017.

[50] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3479–3487.

[51] O. Adedeji and Z. Wang, "Intelligent waste classification system using deep learning convolutional neural network," *Procedia Manufacturing*, vol. 35, pp. 607–612, 2019.

[52] J. Bi, Z. Zhu, and Q. Meng, "Transformer in computer vision," in *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE, 2021, pp. 178–188.

[53] M. S. Drehwald, S. Eppel, J. Li, H. Hao, and A. Aspuru-Guzik, "One-shot recognition of any material anywhere using contrastive learning with physics-based rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 524–23 533.

[54] F. Xu, M. S. Wong, R. Zhu, J. Heo, and G. Shi, "Semantic segmentation of urban building surface materials using multi-scale contextual attention network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 158–168, 2023.

[55] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2407–2416, 2016.

[56] J. DeGol, M. Golparvar-Fard, and D. Hoiem, "Geometry-informed material recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1554–1562.

[57] Z. Erickson, E. Xing, B. Srirangam, S. Chernova, and C. C. Kemp, "Multimodal material classification for robots using spectroscopy and high resolution texture imaging," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 452–10 459.

[58] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 103–110.

[59] P. Vácha and M. Haindl, "Texture recognition under scale and illumination variations," *Journal of Information and Telecommunication*, pp. 1–19, 2023.

[60] H. Zhang, J. Xue, and K. Dana, "Deep ten: Texture encoding network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 708–717.

[61] W. Lu, J. Chen, and F. Xue, "Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach," *Resources, Conservation and Recycling*, vol. 178, p. 106022, 2022.

[62] X. Xie, L. Yang, and W.-S. Zheng, "Learning object-specific dags for multi-label material recognition," *Computer Vision and Image Understanding*, vol. 143, pp. 183–190, 2016.

[63] A. H. Vo, M. T. Vo, T. Le *et al.*, "A novel framework for trash classification using deep transfer learning," *IEEE Access*, vol. 7, pp. 178 631–178 639, 2019.

[64] S. T. Namin and L. Petersson, "Classification of materials in natural scenes using multi-spectral images," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 1393–1398.

[65] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4d light-field dataset and cnn architectures for material recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 121–138.

[66] J. Xue, H. Zhang, K. Dana, and K. Nishino, "Differential angular imaging for material recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 764–773.

[67] J. Kim, H. Lim, S. C. Ahn, and S. Lee, "Rgbd camera based material recognition via surface roughness estimation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1963–1971.

[68] D. Tafone, L. McEvoy, Y. M. Sua, P. Rehain, and Y. Huang, "Surface material recognition through machine learning using time of flight lidar," *Optics Continuum*, vol. 2, no. 8, pp. 1813–1824, 2023.

[69] X. Zhang and J. Saniie, "Material texture recognition using ultrasonic images with transformer neural networks," in *2021 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 2021, pp. 1–5.

[70] O. Wang, P. Gunawardane, S. Scher, and J. Davis, "Material classification using brdf slices," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2805–2811.

[71] C. Hahlweg and H. Rothe, "Classification of optical surface properties and material recognition using mutlispectral brdf data measured with a semihemispherical spectro-radiometer in vis and nir," in *Optical Fabrication, Testing, and Metrology II*, vol. 5965. SPIE, 2005, pp. 150–161.

[72] C. Liu and J. Gu, "Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral brdf," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 86–98, 2013.

[73] M. Weinmann, J. Gall, and R. Klein, "Material classification based on training data synthesized using a btf database," in *European Conference on Computer Vision*. Springer, 2014, pp. 156–171.

[74] S. K. N. J. J. K. Kristin J. Dana Bram Van Ginneken, "Curet: Columbia-utrecht reflectance and texture database," https://www.cs.columbia.edu/CAVE/software/curet/, 2024, accessed: 2024-02-02.

[75] M. Fritz, E. Hayman, B. Caputo, and J.-O. Eklundh, "The kth-tips database," 2004.

[76] G. Kylberg, *Kylberg texture dataset v. 1.0*. Centre for Image Analysis, Swedish University of Agricultural Sciences and . . ., 2011.

[77] G. A. Atkinson and E. R. Hancock, "Two-dimensional brdf estimation from polarisation," *Computer Vision and Image Understanding*, vol. 111, no. 2, pp. 126–141, 2008.

[78] USGS, "Spectral library version 7," https://crustal.usgs.gov/speclab/QueryAll07a.php, 2024, accessed: 2023-02-20.

[79] B. T. Phong, "Illumination for computer generated pictures," *Communications of the ACM*, vol. 18, no. 6, pp. 311–317, 1975.

[80] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6995–7004.

[81] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[82] CUReT, "Columbia-utrecht reflectance and texture database," https://www.cs.columbia.edu/CAVE/software/curet/, 2024, accessed: 2023-11-20.

[83] KTH-TIPS, "The kth-tips and kth-tips2 image databases," https://www.csc.kth.se/cvap/databases/kth-tips/index.html, 2024, accessed: 2023-11-20.

Annexe A - Base de données

Dans ce chapitre, nous présentons un aperçu des jeux de données utilisés dans la partie expérimentale de notre travail. Ces jeux de données incluent des ensembles publics largement utilisés, à savoir FMD et MINC, ainsi qu'un jeu de données que nous avons nous-mêmes créé spécifiquement pour cette étude. Ces jeux de données constituent la base sur laquelle reposent nos expériences de reconnaissance des matériaux et d'évaluation des modèles d'apprentissage automatique développés dans ce travail.

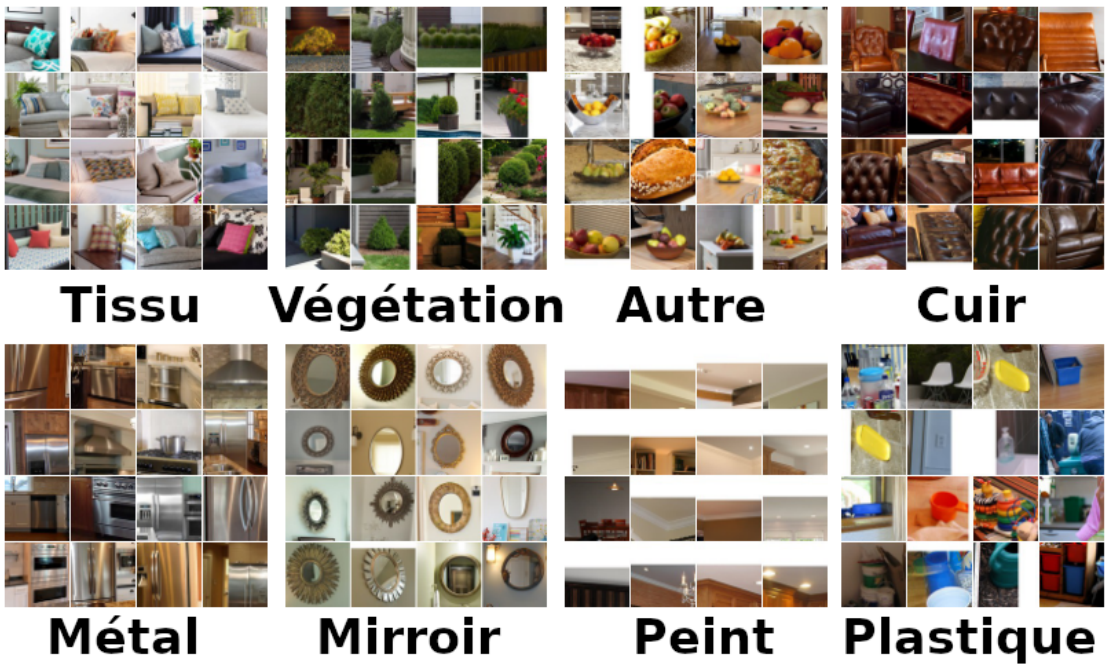


FIGURE A.1 – Aperçu de la base de données MINC.

FIGURE A.2 – Aperçu de la base de données FMD.

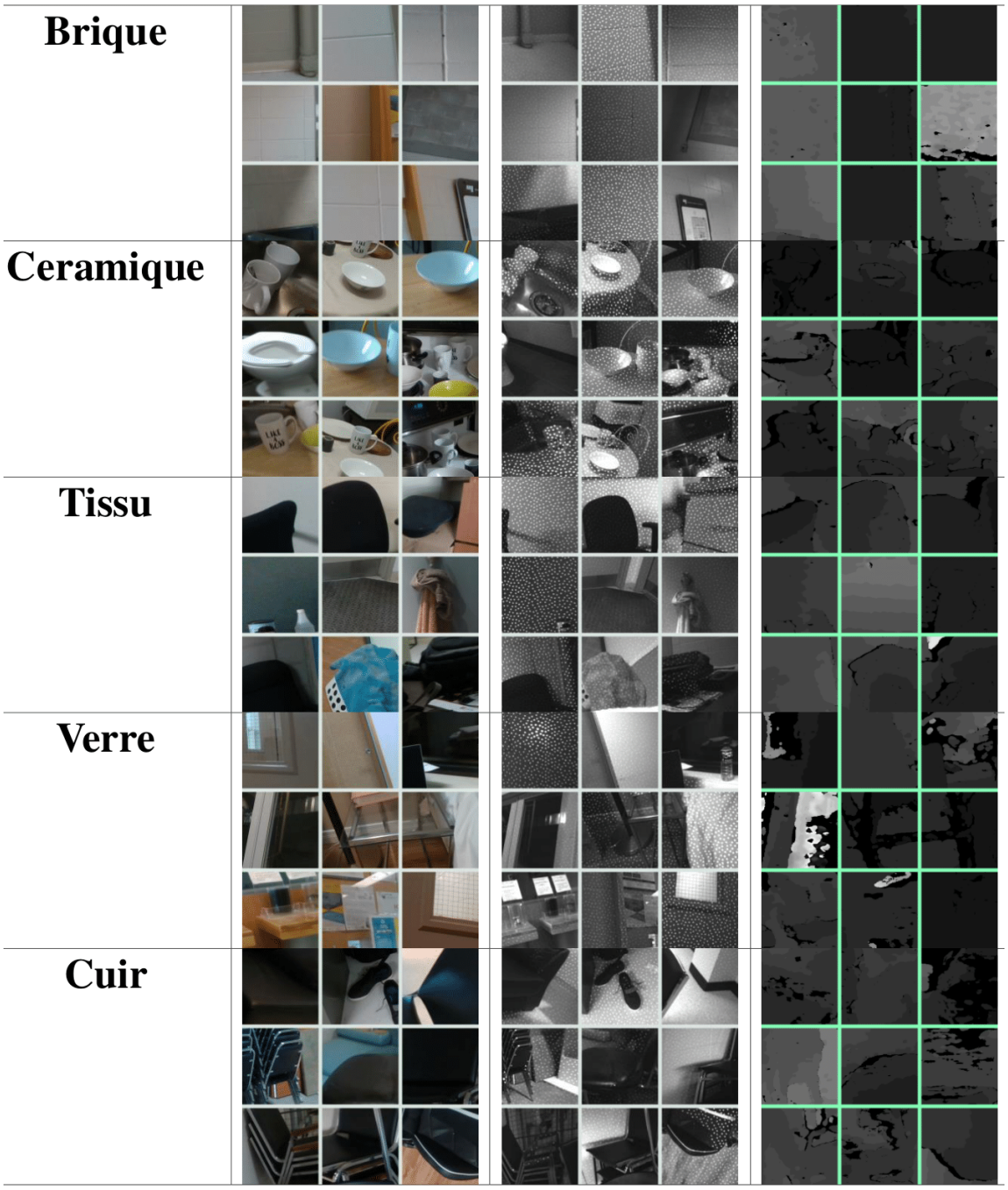


FIGURE A.3 – Aperçu de la base de données crée dans ce projet.

Annexe B - Article de journal proposé

Cet article a été soumis à la revue Machine Vision and Applications de Springer et est actuellement en révision. Le premier tour de révision est passé et nous attendons les réponses du journal.

Surface multi-spectral reflectivity and texture material recognition in non-constrained environments

Ahmed Dhahri[a,*], Ali Amamou[a], Usef Faghihi[b], Sousso Kelouwani[a], Jonathan Boisclair[a], Lotfi Zeghmi[a]

[a]*Department of mechanical engineering, Université du Québec à Trois-Rivières, Quebec, Canada*
[b]*Department of mathematics and computer science, Université du Québec à Trois-Rivières, Quebec, Canada*

Abstract

Material recognition is the process of distinguishing various materials based on their inherent physical properties. It plays a pivotal role in numerous applications, including manufacturing, recycling, and robotic handling. Conventional recognition methods predominantly employ sensor- and vision-based approaches. However, these methods often face challenges such as the similarity and variability in material appearance, environmental conditions, and geometric constraints. In this research, we introduce a multimodal, vision-based, attention-driven model for material recognition. Contrary to preceding texture-based and multi-spectral based methods, our approach harnesses both the texture and light reflection distribution characteristics intrinsic to material surfaces. The proposed method features a collocated system that combines depth, RGB, and near-infrared (Near-IR) cameras, along with infrared laser projector. This specific setup was selected to capture reflection distribution and texture across the visible-near-infrared spectrum. Subsequently, the data captured by this setup were processed by a recognition model within a fusion framework. Our results outperform previous methods in terms of accuracy when additional modalities (Depth, Near-IR, laser projectors) are available, while also exhibiting equivalent performance to top RGB-based models when solely reliant on RGB data. Thus, proving the complementarity of the added modalities with visible information.

Keywords: Material recognition, multi-spectral vision, BRDF, Vision-Transformers.

1. Introduction

Remote recognition of materials in non-constrained environments can contribute to computer vision, computer graphics, robotics [1], and many applications such as augmented reality and recycling [2]. In the context of robotics, knowledge of the object's material allows the robotic arm to customize its handling strategy to ensure greater safety. In addition, for self-driving vehicles, information about surface materials could improve their performance by adjusting their movements according to the surface of the surrounding terrain. Numerous visual-based and sensor-based methods for material recognition exist in literature and on the market [3]. However, the trade-off between the versatility of vision and the efficiency of sensors is a topic of extensive discussion. Consequently, devising solutions that efficiently harmonize these criteria could prove beneficial for these applications.

In the realm of material recognition, the integration of additional modalities has been a focal point of recent research, enhancing the perceptual capabilities of these systems. Hyper-spectral camera-based methods [4, 5, 6, 7, 8], capitalize on the acquisition of the spectrum for each pixel in an image. This approach adds significant discriminative information to the recognition system. However, the efficacy of these methods is contingent on lighting conditions, as the spectral value is influenced by the energy emitted at different wavelengths by light sources. This dependency renders hyper-spectral cameras less suitable for material recognition in uncontrolled environments. The Bidirectional Reflection Distribution Function (BRDF) has been another area of focus, serving as a potent descriptor for surface properties such as rugosity, transparency, and light absorption. This function is instrumental in characterizing the material of a surface. Extensive research [9, 10, 11, 12, 13], has validated the effectiveness of predicting material types through precise and robust acquisition of BRDF data. Various equipment setups, such as illumination domes [9, 12] and semi-hemispherical reflectors paired with Visible-IR spectro-radiometers [10], have been utilized for this purpose. Nonetheless, these used equipments necessitate controlled environments, limiting their applicability in non-constrained environments. In addition, Ultrasonic-based methods [14] operate by measuring the grain size of the target material. On the other hand, Time of Flight (ToF) sensor methods, including ToF [15] cameras and ToF LiDARs [16], identify materials by assessing surface roughness by the speed of the reflected wave. Despite their potential, these methods necessitate precise scanning aiming at the material of interest, rendering them unsuitable for non-constrained environments where such precision may be unattainable.

Instead of enhancing visual perception through the addition of physical sensors [17, 18, 1], a subset of researchers has proposed solutions that solely depend on visual cameras, with an emphasis on improving image understanding. This scientific

*Corresponding author
Email address: ahmed.dhahri@uqtr.ca (Ahmed Dhahri)

discussion primarily focuses on two critical modalities: local features (texture) and global non-local features (context). Local features scrutinize the texture and color attributes of the surface image, while non-local features encompass the broader characteristics of objects and scenes. A crucial distinction between the two approaches lies in the nature of the datasets they employ. For example, the Flickr Material Database (FMD)[19], predominantly used for local features, mainly comprises surface textures. In contrast, the Material In Context dataset (MINC)[20], utilized for global features, provides a more holistic view by including both the target surface and its surroundings within non-constrained environments. Local features can be categorized into hand-crafted features—such as Markovian, illumination-invariant texture features [21]; filter banks [22]; and wavelet filters [23]—as well as automatically extracted features, which include CNN feature vectors [24] and D-CNN models applied to texture datasets [2]. On the other hand, global feature-based methods predominantly utilize CNN or Vision Transformer models, trained on dataset formats that include the context of objects [20, 25, 26, 27, 28] and scenes [29]. These global feature methods have been shown to outperform their local counterparts, achieving test accuracies of 85.6% on the MINC dataset [20] and 87% on the Trash image dataset, making them more suitable for real-world applications. However, it is important to note that texture and visual appearance are not unique to a material, potentially leading to inaccurate predictions for visually similar materials and in the presence of adversarial inputs.

To overcome the challenges posed by the limitations of visual similarity, several techniques have been developed that expand the spectral range of imaging into the visual-infrared (RGB-IR) spectrum. This approach enriches the imaging process by integrating more spectral information, thereby introducing additional, distinctive features. Such enhancement not only aids in more accurately differentiating between materials or objects but also significantly refines the distribution of prior information, leading to more precise and reliable interpretations. The selection of infrared (IR) and near-infrared (NIR) spectra is particularly advantageous due to the ubiquitous presence of IR radiation in most lighting sources. This is in contrast to the specific requirements of hyper-spectral imaging, which makes them more dependent on lighting conditions. For instance, multiple types of setups were used, such as RGB-IR multi-channel cameras [30] and a collocated camera array comprising NIR (Near infrared) and RGB cameras [31]. This approach improved the accuracy of predictions by leveraging the differences in material appearance across multiple spectral ranges, which leads to further differentiation between visually similar materials. However, it's important to note that these methods, while advantageous in certain aspects, still share some of the inherent limitations of traditional RGB imaging techniques. Furthermore, the specificity of these setups leads to data that is highly dependent on the particular equipment used. This specificity poses a challenge in creating large-scale datasets, which are essential for advancing material recognition technologies. The lack of such extensive datasets remains a key hurdle in the field

Alternative approaches such as the 4D light field camera [32]

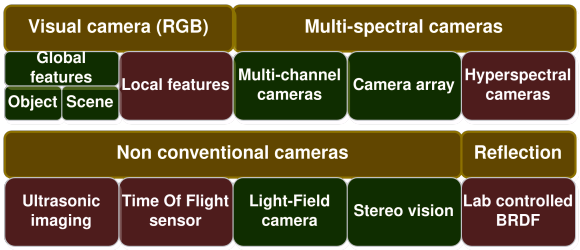


Figure 1: Representation of the state-of-the-art methods. Methods in green are adequate for non-constraint environments. Methods in Red are not adequate.

and stereo vision differential imaging [11] assess reflection variance through differences in aperture and viewing angles, respectively. While these added modalities have proven to be complementary to visual texture, they are not comprehensive in characterizing the full reflective properties of material surfaces, such as those detailed by the BRDF. The primary limitation lies in their inability to account for the effects of geometry and external lighting, which are crucial for perceiving phenomena like self-occlusion and self-shadowing. These phenomena, often encountered on non-planar surfaces, can be easily mistaken for texture, as noted by Guarnera et al. [33], leading to potential inaccuracies in material recognition.

Methods utilizing ToF sensors, ultrasonic sensors, BRDF acquisition, and RGB local texture features, as illustrated in red in figure 1, more or less require specific conditions to be met, making them unsuitable for tackling the complexities of natural environments. Leveraging the strengths of visual methods, which encompass texture and context, alongside the integration of multi-spectral information, an accurate estimation of reflective features would be a valuable complement to visual information. Distinct from methods that depend on analyzing variations in light reflection, such as those employed by light field cameras and stereo vision systems, our approach is innovatively designed to capture detailed reflectivity characteristics. Crucially, it takes into account complex phenomena like self-occlusion and self-shadowing, which are often overlooked. This is achieved through the advanced modalities we have proposed, setting our method apart in its ability to provide a more comprehensive and accurate analysis of reflective surfaces. In this paper, we propose a novel method for material recognition that combines multiple image modalities using a laser dot-projecting depth camera capable of capturing both RGB and NIR information. Our method aims to narrow down the prior information distribution by employing univariate spatial sampling to capture the reflective features of surfaces. Capitalizing on the spectral proximity between NIR and visible light, reflective properties are deduced from the disparity between the visible frame captured under natural illumination and the NIR frame containing structured laser dots. Unlike previous multi-spectral methods that rely on clustering models to process limited datasets, we utilize state-of-the-art image recognition models to extract high-level features and infer the reflective features, while addressing the data shortage issue through methods we

have proposed.

The organization of this paper is structured as follows: Section II provides an overview of the related work in the field, highlighting the methods that are relevant to the proposed approach. Section III details the proposed setup and the theoretical principles it is based on. Section IV is dedicated to discussing the software aspects of the study. Afterwards, Section V showcases the experimental results obtained from testing the efficacy of the proposed method. Finally, a conclusion of the study is formulated.

2. Related works

The innovative process of extracting reflective features, combined with the integration of NIR and RGB-D camera frames, underscores the significance of two crucial areas in modern imaging technology: image fusion and visual representation learning. This combination not only enriches the quality of the captured images but also highlights the critical role these topics play in advancing our understanding and capabilities in accurately interpreting and representing visual information.

2.1. Image fusion

Employing multiple image sensors as input to a single system underscores the importance of fusion methods, aimed at enhancing feature extraction efficiency. In scenarios involving more than one physically separated camera, fusion becomes indispensable for feature transfer. Fusion finds its utility in a myriad of applications including High Dynamic Range (HDR) imaging, color transfer, and infrared-visible fusion [34]. The literature delineates three primary approaches to image fusion [35]: Low Level Fusion (LLF) or early fusion [36, 37], Mid-Level Fusion (MLF) [38, 39], and High Level Fusion (HLF) or late fusion [40]. These approaches are distinguished by the stage at which fusion occurs prior to input, during feature extraction, or post feature extraction. Fusion is executed through various methods, such as CNN dense fusion like DenseFuse [38], and transformer-based fusion like SwinFuse [39].

In the context of vision transformer fusion, the choice between HLF and LLF may not markedly impact performance [41]. Additionally, employing MLF or LLF necessitates alterations in model dimensions, thereby mandating reperforming the extensive pre-training process.

2.2. Visual Representation Learning

Visual Representation Learning is Semi-supervised learning methods used for pre-training of large models. It aims to leverage vast amounts of unlabeled or weakly-labeled data to improve the performance of a model trained with a limited amount of labeled data. Within this realm, two prominent methodologies emerge: Masked Image Modeling (MIM) and Contrastive Learning Modeling. MIM is exemplified by techniques such as masked autoencoders [42], BEiT [43] (Bidirectional Encoder Representations from Transformers for Images), and EVA [44]. In contrast, Contrastive Learning Modeling is characterized by other prominent models, each with its unique approach to learning from contrastive loss applied on combinations of unlabled data [45, 46, 47]. Additionally, there are hybrid methods that effectively combine elements of both MIM and Contrastive Learning, leveraging the strengths of each to create more robust and versatile models [48].

3. Used setup and material surface proprieties

We propose an innovative multi-modal, vision-based system specifically designed for material recognition. This advanced setup integrates an array of sensing technologies, including RGB and Near-Infrared (NIR) cameras, depth sensors, and NIR laser projectors. The core principle of this system lies in analyzing the contrasts between visible surface illumination (captured by the RGB camera) and NIR surface illumination (enhanced with laser projection), as well as leveraging the detailed surface geometry data provided by the depth cameras. By synthesizing these diverse data streams, our vision model, embedded within a sophisticated fusion framework, is adept at deducing the reflective properties of various surfaces. This multi-modal approach represents a significant leap in material recognition technology, offering a more nuanced and accurate analysis than traditional methods. Combining these features with texture features and object context will lead to more accurate material recognition model comparing to the previous state-of-the-art methods.

The methodology would be discussed as follows: The statistical characteristic of the input is discussed. Then, the reflection modeling and the camera setup would be presented in that order.

3.1. Statistical characteristics of texture images

Our proposed methods utilize an input consisting of real world images from the visible-NIR spectral range.

3.1.1. Spectral features

The similarity observed between visible and near-infrared (NIR) vision is primarily attributed to the closely related reflection and absorption characteristics of materials across these two spectral ranges. Notably, popular sensors for capturing RGB (visible light) and NIR images are calibrated for specific wavelength ranges: RGB sensors typically cover the range of 400nm to 700nm, while NIR sensors operate within the 650nm to 900nm spectrum. This overlapping wavelength range partly explains why the visual and NIR responses of materials tend to be locally similar, thereby influencing how materials are perceived and analyzed in these different imaging techniques. This relation is proven by computing correlations of the reflective response of 7000 raw material found in the Splib07 database [49]. The correlations are presented in table 1.

3.1.2. Abstract features

Both high-level and low-level features in natural images need to be extracted for efficient prediction of material classes. The extraction of these feature types occurs in distinct manners:

Table 1: Correlation between reflective response in visible-NIR spectrum. NIR spectrum is [700nm to 900nm] and visible spectrum is [400nm to 700nm].

	700nm	800nm	900nm
400nm	0.42	0.42	0.42
500nm	0.77	0.77	0.77
600nm	0.55	0.51	0.51
700nm	1.0	1.0	1.0

- These encompass color, gradients, texture, etc., and are extracted using models that predict features locally based on the pixel neighborhood [50].
- Real image distributions exhibit non-Gaussian characteristics [51]. They encapsulate higher-level auto-correlations described by the context of the image (e.g., object, scene, lighting), which are crucial for material recognition in non-constrained environments. These features are deduced by aggregating information from multiple non-adjacent ranges of pixels.

3.2. Material reflectance and transmittance proprieties

The BRDF is a 4-Dimensional functions that describes the light behavior on homogeneous material surfaces. As described in [33], it describes the amount of reflected light from every direction in function of the incident light.

3.2.1. Reflection model and distribution function sampling

In this section we explore the interplay between the optical properties of materials and the imagery captured off-the-shelf stereo ir-vision RGB-D cameras providing modalities we suggested. Accurately estimating the Bidirectional Reflectance Distribution Function (BRDF) of a material under normal conditions is an exceedingly challenging task. Typically, obtaining a precise BRDF measurement necessitates the use of specialized laboratory equipment, such as geni-reflectors or hemispherical domes, coupled with controlled lighting conditions. These tools and settings are crucial because they enable detailed analysis of how light interacts with material surfaces under various angles and conditions. However, in the absence of such specialized equipment and specific lighting environments, achieving a reliable estimation of the BRDF becomes significantly more difficult, underscoring the complexity involved in capturing and interpreting material optical characteristics using standard depth cameras. Therefore, to be able to estimate a coarse measure of BRDF, further simplifications were applied on the model equation, stated in equation 1.

$$brdf(\theta_r, \phi_r, \theta_i, \phi_i) = \frac{dL(\theta_r, \phi_r)}{E(\theta_i, \phi_i)\cos\theta_i d\omega_i} \tag{1}$$

L is the reflected illumination coming from a measuring point, θ_r and ϕ_r the zenith and azimuth angles of the view direction. E is the radiance going to a measuring point and θ_i and ϕ_i are the zenith and azimuth angles of the light direction. To further reduce the complexity of the model, the following assumptions had been considered.

- All materials surfaces are considered isotropic (which is the case for most of the real world material surfaces). The BRDF function would be symmetrical around the surface normal axis.
- The light direction and the camera direction are joined, so $\theta_i = \theta_r$ and $\phi_i = \phi_r$.
- The projected laser points are considered collimated laser beams with negligible waist. Thus, the beam spread phenomena won't be considered.

The BRDF equation is reduced to equation 2:

$$brdf(\theta_r) = \frac{dL(\theta_r)}{E(\theta_r)\cos\theta_r d\omega_i} \tag{2}$$

This equation yields three terms to be measured by our setup: E the itensity of our laser beamms,which is constant. θ_r the geometry information sensed by the depth sensor. And dL the reflected light information deduced by the difference between RGB and NIR frames.

The visual aspect of the BRDF could be explained more using the Phong reflection model [52]. This model split the illumination components into three parts, the ambient, diffuse and specular reflections. These physical phenomena could be individually observed by the human eye.

The basic Phong model, with one light source applied at one point on a surface, is stated in equation 3 and equation 4

$$dL(\theta_r, \phi_r, \theta_i, \phi_i) = dL(\vec{Lp}(\theta_i, \phi_i), \vec{V}(\theta_r, \phi_r)) \tag{3}$$

$$dL(\vec{Lp}, \vec{V}) = k_a I_a + k_d I_d(\vec{Lp}.\vec{N}) + k_s I_s(\vec{R}.\vec{V})^n \tag{4}$$

Where $\vec{V}$ is the view direction or the angle of measurement and $\vec{L_p}$ is the light direction. k_a, k_d and k_s are the ambient, diffuse and specular reflection constants respectively. These constants are specific to the material surface proprieties. I_a, I_d and I_s are respectively the diffuse and specular reflections' illumination. The n represents the shininess constant describing the size of the specular highlights. And lastly, $\vec{N}$ represents the surface normal direction.

Applying the same simplifications demonstrated earlier on the current Phong model, the equations are reduced to equation 5.

$$dL(\theta_r) = k_a I_a + k_d I_d \cos(\theta_r) + k_s I_s \cos^n(2.\theta_r) \tag{5}$$

At this point, the angle θ_r could be known under the conditions that the information on the geometry of the surface is given. n, k_a, k_d and k_s are constants characterizing the material surface proprieties. For θ_r and the illuminations I_a, I_d and I_s, they depend on the light source E direction and intensity. The BRDF equation could be now rewritten in terms of measured variables and characterizing constant in equation 6 and equation 7

$$brdf = brdf_{k_a,k_d,k_s,n}(I_a(E), I_d(E), I_s(E), \theta_r, E) \tag{6}$$

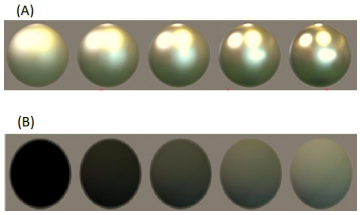


Figure 2: Representation of the BRDF function types appearance in objects generated by Blender: A-Reflection type: diffuse to specular. B-Surface brightness.

$$brdf = \frac{k_a I_a(E) + k_d I_d(E)\cos(\theta_r) + k_s I_s(E)\cos^n(2.\theta_r)}{E(\theta_r)\cos\theta_r d\omega_i} \quad (7)$$

As solid example for the BRDF coarse measure, BRDF distribution narrowness and mean value could be estimated by using its visual aspect, as visualized in figure 2.a and figure 2.b respectively. The two criteria are visually translated by the specular reflections (narrawness of the BRDF distriburion) and the albedo (integral of BRDF distribution). These sensed phenomena could vary across an identified homogenuos material surface further characterizing the BRDF distribution.

3.2.2. Camera setup for 2D image BRDF sampling

Since the camera and the light source directions are joined, it is evident that the sampling of the BRDF is a function to the zenith angle θ_r. This model is referred as uni-variate spatial sampling, which is estimating a distribution by capturing multiple samples from a homogenous surface, in function of one variable, from an image. This operation is done as follows:

- Using laser projectors light, multiple laser beams are emitted with many directions towards the material surface, in the near-infrared spectrum.
- The created reflection illumination is estimated using the comparison between the NIR frame containing the created laser beams reflection and the visible frame containing only natural scene illumination. It is demonstrated in section III.A that materials have almost same reflective and absorbtive proprieties between visible and NIR spectrum. A representation of this operation is shown in figure 3.
- The surface normals are deduced from the depth camera frame.
- The collocated depth camera, NIR and RGB cameras, laser projectors setup is, then, able to sense multiple data points as a function of incident illumination and surface normal, as depicted in figure 4.
- The homogeneous surface boundaries could be automatically learned by the network. Thus, BRDF distribution could be estimated in every region of a surface, if adequate conditions are met. To gather maximum information from one reflection image the image patch must contain sufficient amount of spatial resolution, as well as not containing excessive non-homogeneous appearances.

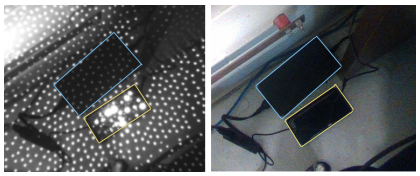


Figure 3: RGB and NIR images of two visually similar materials in the visible spectrum. Surface inside the blue and yellow rectangles are, respectively, plastic and glass. Discriminative reflective features could be seen in the NIR frame.

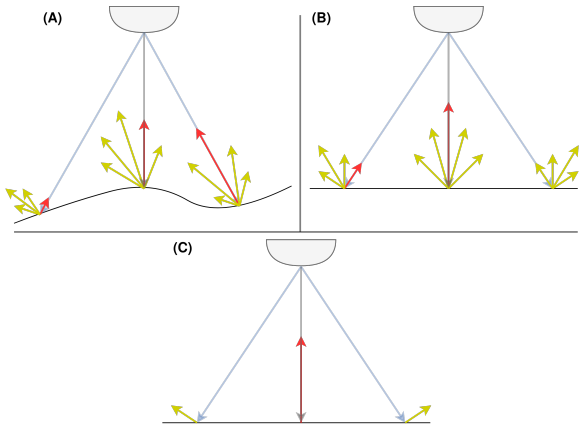


Figure 4: Representation of uni-variante spatial sampling from one 2D image to capture a BRDF estimation.

Returning to equation 7, which is a function with one variable θ_r and 5 constants n, k_a, k_d, E and k_s. It would be possible to estimate the characterizing constants, by estimating the BRDF distribution through uni-variate spacial sampling, as shown in figure 4, and controlling the variables using our setup. Although, the simplified method isn't enough to directly acquire a precise measurement of the BRDF, the model could estimate a coarse measurement of the distribution. Therefore, this data would offer more discriminating information to our model as a complement to other modalities (texture and context).

4. Proposed model and Training framework

In this section, we delve into deep learning models, specifically Vision Transformers and CNNs, which integrate both low-level and high-level features to enhance prediction accuracy. We also explore the concept of image fusion, a critical aspect when dealing with multiple modalities of frames sourced from various cameras. To address the challenges posed by limited-size datasets, we implemented strategies such as semi-supervised training techniques. Additionally, leveraging large-scale RGB public datasets has been employed as a means

of regularization, further mitigating the issues associated with smaller datasets.

4.1. Classification model

Deep learning models are pivotal in extracting high-level features that effectively encapsulate the context within images. Among these models, Transformers especially excel due to their attention mechanisms. These mechanisms enable them to capture extensive contextual information efficiently in a single operation, making them adept at understanding complex image relationships and structures. In contrast, Convolutional Neural Networks (CNNs) tend to extract high-level features in their deeper layers, where each pixel in these layers is associated with a specific receptive field. This process is more gradual, as CNNs build up from simple to complex features layer by layer. However, CNNs possess a distinct advantage in extracting low-level features like textures and edges. Their convolutional architecture is inherently suited for this task, allowing for the efficient and effective identification of basic patterns and details in the initial layers. This fundamental difference underscores the unique strengths and applications of both Transformers and CNNs in the realm of image processing and analysis.

Low-level features play a pivotal role in predicting texture attributes, which are crucial for material recognition. Hence, it is advantageous if the deployed models facilitate a streamlined pathway for these features in terms of transformations (convolution kernels or transformer blocks) towards the output. This is effectively served by residual connections, which are integrated into almost every modern network architecture. On the other hand, high-level features, such as object and scene context within the feature manifold of natural images, cannot be deduced by early transformations. Therefore, a more sophisticated model is required to learn the function that maps all these modalities. This necessitates the employment of Transformer and deep Convolutional Neural Network (CNN) models, which are adept at handling such complex mapping tasks.

The standard benchmark for evaluating object classification models is typically the top-1 accuracy on the ImageNet-1k dataset. However, it's crucial to consider additional metrics when assessing these models. For instance, zero-shot accuracy offers insights into a model's ability to establish abstract connections and adapt to recognizing previously unseen objects. Conversely, top-5 accuracy, or top-k accuracy in general, while somewhat less crucial, demonstrates a model's capacity to differentiate between classes, especially when the differences are subtle. To validate the choice of the model utilized in this research, we've curated a list of top-performing state-of-the-art recent models on the ImageNet-1k dataset, presented in [53]. We deliberately included both Convolutional Neural Network (CNN) and Transformer-based models in this selection for specific reasons: Vision transformers have demonstrated remarkable proficiency in zero-shot learning, underscoring their ability to swiftly adapt to new tasks and recognize unseen objects. Conversely, CNNs were deliberately chosen for their inherent bias toward capturing local features, a crucial characteristic for extracting texture-related information.

Table 2: Metrics of state-of-the-art models across various model families, on the Imagenet-1k dataset.

Model	Top-1 test accuracy	Zero-Shot accuracy	Top-5 test accuracy	Family
CoatNet-6	90.45%	84.2%	99.3%	Hybrid
Moat-4	89.1%	–	–	Hybrid
Eva-02-L	90.0%	80.4%	99.0%	Transformer
VIT-G/14	91.1%	88.3%	99.1%	Transformer
Swin-v2-G	90.2%	84.0%	97.2%	Transformer
DaVit-G	90.4%	–	–	Transformer
ResNet-152	85.6%	24.1%	97.2%	Convolution
EffNet-L2	88.6%	–	98.1%	Convolution

Given that our proposed method involves three input modalities, we find it necessary to incorporate image fusion techniques to effectively combine these modalities:

- For convolution-based models, three fusion methods will be experimented with. In the case of early fusion, only the dimension of the input layer will be altered. For mid-level fusion, DenseFuse will be integrated into the architecture. Regarding late fusion, three separate networks will be stacked together, with a classification head added at the end.
- For transformers, only the late fusion option is available, as any alteration in the architecture necessitates the repetition of extensive pre-training.

4.1.1. Pre-training on Large-scale RGB datasets

Given the vast repositories of RGB-labeled datasets, we can effectively train the RGB-dedicated segment of the network. During this initial pre-training phase, it is crucial to keep the weights pertaining to other modalities static, thereby ensuring that the network's learning is concentrated on the RGB information. This pre-training regimen unfolds in a sequential two-step process. Initially, the network's RGB pathway is trained exclusively with RGB data to establish a robust foundational understanding of this modality. Subsequently, the pre-training expands to encompass all modalities, employing the full range of available data inputs. This inclusive approach facilitates a holistic learning experience, enabling the network to integrate and interpret the entire gamut of sensory inputs

The initial phase of the operation capitalizes on a large-scale dataset to fine-tune the model's bias, bringing it closer to the true distribution of data while keeping variance minimal. This is achieved by deactivating or freezing other pathways in the network. In case of mid-level fusion zero tensors are assigned to branches not being trained, effectively silencing any activation signals for these masked branches. The use of the ReLU (Rectified Linear Unit) activation function, devoid of bias, ensures that the corresponding neurons remain inactive, thereby preserving their weights during the back-propagation process and preventing any unintended changes to the model's predictions.

4.2. Visual representation learning

In the subsequent phase of pre-training, the complete set of input data is employed, aiming to further narrow the gap between the model's learned bias and the true underlying bias of

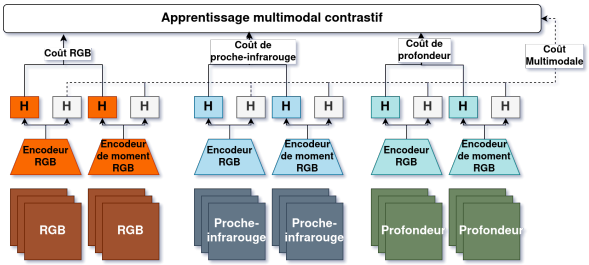


Figure 5: Multi-modal contrastive visual representation learning framework. All used losses aim to maximize agreement between logits predicted using similar datapoints and disagreement between negatively similar datapoints. For multimodal loss, the presence of all modalities is not necessary.

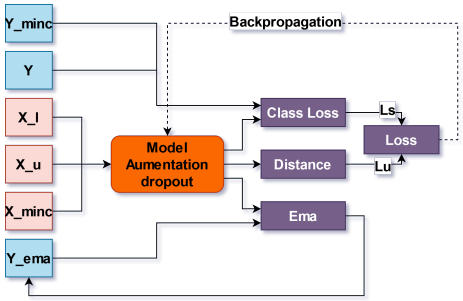


Figure 6: Main building blocks of our training process. Block A: RGB back bone training with large scale RGB dataset aiming for visual representation learning. Block B: Multimodal Contrastive learning pre-training for inter-modal visual representation learning. Block C: Supervised training process representation: Class loss: recognition cross-entropy loss. Distance: It is the unsupervised loss, which is literally the MSE distance between 2 stochastic predictions of the same data point but with different augmentations.. EMA: the exponential moving average of the predicted label. X_u: unlabeled data. X_l, Y: labeled data-point. The blocks are annotated according to the sequential order of their execution.

the data, situated within an abstract feature space. This step is crucial as it not only decreases variance but also substantially boosts the model's predictive accuracy. By integrating all available sensory information, the network refines its parameters to more accurately reflect the complexity and nuances of the input data.

Deep learning algorithms, including Convolutional Neural Networks (CNNs) and vision transformers, typically necessitate large-scale datasets to achieve optimal convergence and robust generalization capabilities. This requirement often poses a challenge in our scenario, where there is a limited availability of labeled data. In such cases, semi-supervised pre-training has proven to be an effective strategy. This approach leverages a combination of a small amount of labeled data and a larger pool of unlabeled data. By doing so, it allows the model to learn meaningful representations from the extensive unlabeled dataset, while the labeled data guides the learning towards more accurate and specific outcomes. This method not only enhances the model's performance in data-scarce environments but also contributes to more efficient and versatile machine learning applications

4.2.1. Pre-training on large-scale multimodal datasets

The methodology for learning visual representations in multimodal data networks, as demonstrated in Yuan et al. (2021) [45], is underpinned by a twofold objective (figure 5): the preservation of both intra-modal and inter-modal similarities. This technique exploits the unique characteristics intrinsic to each modality, while simultaneously capturing the rich semantic information that emerges from correlations between different modalities. Such a strategy considerably improves the sophistication and accuracy of the visual representations that the network is able to learn. This simultaneous exploitation of both within-modality characteristics and between-modality relationships allows for a more robust and semantically rich representation of the data.

4.2.2. Training with consistency regularization

To optimize data use and mitigate the risk of overfitting, semi-supervised learning techniques with consistency regularization are applied during the training across all modalities. Consistency regularization is a technique that enhances model stability and increases resistance to noise in the training data. Specifically, we utilize the temporal ensembling method [54], which relies on a collection of models with identical architectures and hyperparameters but differing initializations and input augmentations. During training, each model in the ensemble is exposed to varied inputs, and their predictions are aggregated to form a collective output, as illustrated in figure 6.C. Consistency regularization imposes a penalty on the discrepancies between individual model outputs and the aggregated ensemble output. Additionally, temporal averaging is employed at each data point to bolster the stability of predictions over successive training iterations. The Exponential Moving Average (EMA) block, depicted in figure 6.C, serves to temper any sudden shifts in consistency throughout the training phase. Its objective is to ensure steady and reliable predictions at each inference juncture, guided by the equation 8.

$$y_{ema} = \alpha y_{ema} + (1 - \alpha) y_{pred} \tag{8}$$

5. Experiments and Results

A comprehensive explanation of the proposed procedures is provided. These procedures include the creation of the dataset, the selection and training of fusion and classification models. Finally, the integration of the reflection modality is presented.

5.1. Datasets overview

The followed methodology includes specific criteria to be met: size, diversity, well-sampling, and number of categories. As stated in [20] with some included changes:

- Size: solving the issue of the lack of data and the difficulty of data acquisition, semi-supervised methods are employed. Thus, the gathered data must be sufficient for the

semi-supervised training, which requires less size than the data needed by supervised training in order of one or two magnitudes.

- Number of categories: Same categories as [20] are taken, while dropping categories that are not present in indoor environments and merging more similar categories to minimize the size of data requirements.

5.1.1. Dataset acquisition

Since no publicly available dataset contains adequate input information for our proposed method, it is then required to create our own dataset employing the data acquisition procedure and manual labeling. Adding to that our method is setup-dependent. So, if the setup assembly differs, the data profile would obviously change. Our dataset was gathered using the Intel RealSense d435i depth camera, which features an 87° x 58° horizontal and vertical field of view for the depth and near-infrared sensors, and a 69° x 42° field of view for the RGB sensor. Additionally, we configured the device to align the depth frame with the RGB frame, and to generate depth frames with high-density mode. It is also noted that the Near-infrared sensor and the projected lasers operate within the 850nm wavelength. We propose a system, described in figure 7, to automatically extract structurally different frames with the minimum of motion blur from a multi-modal video stream. Frame choice from the video streams is based on 3 criteria:

- The difference of color-histogram distance between the instant frame I_t and the last chosen frame I_{i-1} exceeding a threshold, as in equation 9
- The time distance between I_t and I_{i-1}, described in equation 10.
- The variance of the frame I_t Laplacien exceeding a threshold to eliminate motion-blurd frames, described in equation 11.

When these three conditions are met, the instant frame I_t is added to the chosen frames array as I_i.

$$\|H_c(I_{i-1}) - H_c(I_t)\| > H_{th} \tag{9}$$

$$T(I_t) - T(I_{i-i}) > T_{th} \tag{10}$$

$$Var(\triangle I_t) > v_{th} \tag{11}$$

H_c denotes the color histogram vector. I_t the instant frame, I_i and I_{i-1} are instants of the chosen frames array. $Var()$ is the statistical variance and $\triangle$ is the Laplacian operator.

After extracting the frames, the final image patches would be extracted using the click methods as described in [20].

5.1.2. Data collection, sampling and annotation

Post the dataset acquisition procedure, 7,500 input features from 1300 scene images were produced. Rather than a random splitting, data was split into training and test sets based on the different scenes captured, aiming for more accurate test metrics. Furthermore, 7,500 patches were randomly extracted

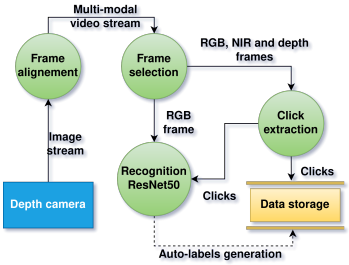


Figure 7: Dataflow of the data acquisition procedure.

from raw frames and pseudo-labeled using an ensemble of models (CoatNet-6, Eva-02, and EffNet-L2) trained on RGB large-scale data, as illustrated in Figure 7. These models were pre-trained on both MINC and the initial patches for pseudo-labeling. Additionally, 10,000 more patches were randomly harvested from the raw data to form an unlabeled dataset, facilitating the semi-supervised learning process.

In the RGB branch training of our proposed networks, we initially employed the complete set of 23 categories offered by MINC. Then, we reduced the number of output categories to 15 as explained below:

- Ceramic and Polished stone are combined as they are rare and hard to collect, and they have similar reflective and visual features.
- Carpet is merged with Fabric.
- Foliage, Food, Hair, Skin, Sky and Water are merged with others, as they could not be considered as materials.
- Wallpaper is also merged with plastic, paper or painted according to the consistent material. When training on MINC this category is discarded.

Situations where data is missing or incomplete are common to encounter. A simple approach is to simply exclude the incomplete samples from the analysis, but this can lead to a loss of valuable information, in such case where data is scarce. Alternatively, considering missing data as a separate category may be appropriate when it comes to avoiding false predictions in situations where depth or IR laser dots are missing.

5.1.3. Pre-processing and frames alignment

Different cameras are used to capture RGB and NIR frames, resulting in a misalignment between the two images. To address this issue, it is necessary to reduce the misalignment between the two images before they are fed to the fusion stage. As previously mentioned, the depth frame is aligned with the RGB frame, thus the alignment must be applied on the NIR image. Thus, to achieve alignment, we selected the best-fit resized crops from the NIR frame for each multimodal image based on the chamfer distance and measured the mean depth of the same frame. Subsequently, we conducted linear regression on the crop center and size as a function of mean depth. By applying

these functions on the NIR frames, we got negligible misalignment, as shown in Figure 8.

Figure 8: Superposed frames with alignment on the left and without alignment on the right.

5.1.4. Large-scale RGB datasets

Several publicly available datasets are accessible online, including FMD, CUReT [55], KTH-TIPS [56], and MINC. Among these datasets, the MINC dataset stands out with its impressive scale, boasting 1.2 million data points. This large-scale dataset aligns perfectly with the requirements of modern state-of-the-art classification models and is particularly well-suited for challenging recognition tasks in real-world scenarios. As a result, we made the deliberate choice to employ the MINC dataset for pre-training purposes. This decision serves the dual purpose of enhancing the robustness of our model's predictions in the RGB branch during training and mitigating the potential risks of overfitting.

5.2. Training process

Our training methodology is divided into three distinct phases: supervised training using extensive RGB datasets, pre-training with multimodal data, and semi-supervised training incorporating data from all modalities. Specific augmentations are tailored to each phase. For the RGB data training, we implement the same augmentations as those detailed in [20]. In the temporal ensembling framework, we employ texture-preserving augmentations, while also introducing various noise augmentations such as the application of blurs and white noise to the RGB frames. Additionally, positional augmentations, including flipping and resized cropping, are applied uniformly across all frames. The use of augmentations during the supervised phase has been strategically chosen to reduce overfitting. These augmentations mirror those utilized in the temporal ensembling to ensure consistency in the training regimen. The entire training process is systematically illustrated in figure 6, and the loss function we adopted is delineated in equation 12.

$$L_{M,Aug}(x, y) = C_E(M_{Aug}(X_l), Y_l) + MSE(M_{Aug}(X_u), Y_{ema}(X_u)) \tag{12}$$

C_E, MSE and M_{Aug} are respectively the cross entropy, the mean square error losses and the model with stochastic augmentation and dropout.

5.3. Selection of classification model

To validate the model selection strategies outlined in Section 4.1, especially after their post-evaluation performance on ImageNet-1k, we conducted further assessments on the MINC dataset. This evaluation focused on the models' capabilities in material recognition within a 'wild task' context, which involves integrating both texture and object context in a non constrained envirement. Our selection criteria for the models were based on their performance within their respective families. From the CNN family, we chose ResNet-152 and EfficientNet-L2, while CoatNet-6 was selected from the Hybrid family. For the Transformer family, we opted for ViT-G, EVA-02-L, and Swin-V2-G, taking into consideration their strong zero-shot accuracy as a deciding factor. Additionally, GoogLeNet, a state-of-the-art (SOTA) model, was also included in our test lineup.

We trained all networks on 10 epoch on MINC dataset using Adam optimizer with batch-size 8 and a base learning rate of 10^{-4} dropping by 25% every 100000 iterations with an input patch size of 256, as well as a cosine learning rate decay. Evaluations is depicted in table 3. Results shows that CoatNet-6 is the best performing. Also, Eva-02 could be chosen as a trade-off between

Table 3: Test accuracy of multiple state-of-the-art models on MINC dataset.

Model	Top 1 test accuracy	Top 5 test accuracy
CoatNet-6	**90.5%**	**99.4%**
Eva-02-L	89.2%	99.0%
ResNet101	87.5%	98.9%
EfftNet-L2	85.5%	98.7%
Swin-v2-L	88.2%	98.9%
ViT-G/14	71.1%	93.9%
GoogLeNet	85.5%	98.1%

5.4. Modality fusion

In our study, we present a comparative analysis of our approach with the state-of-the-art method proposed by Bell et al. [20]. For this comparison, we selected three models—CoatNet-6, EVA-02-L, and EfficientNet-L2—and trained them on our dataset. These models were chosen to serve as the backbones for our constructed fusion model. We explored different fusion methods, ranging from late to early

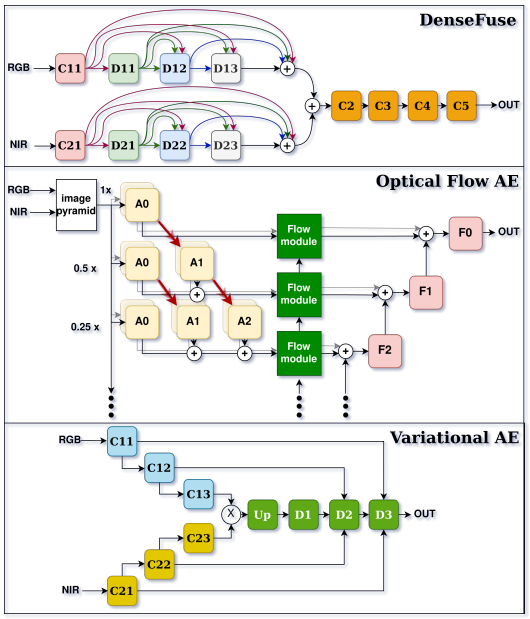


Figure 9: Representation of fusion methods across different backbones.

fusion, as shown in figure 9, across all backbones. Additionally, we incorporated an integrated Mid-level fusion with DenseFuse feature extractor, for the EfficientNet-L2 and CoatNet-6 models. This allowed us to assess the impact of various fusion strategies and the effectiveness of the feature extraction in enhancing the performance of the multimodal learning framework. Results in table 4 shows the supremacy of late fusion (High level fusion) in comparaison to the rest. Late fusion maintain the structure of the backbone, which preserve the meaning of pre-trained weights.

5.4.1. *Pre-training and visual representation learning*

We conducted pre-training sessions using three networks—CoatNet-6, EVA-02-L, and EfficientNet-L2—on

Table 4: Comparison between fusion methods.

Architecture	Our dataset Top-1 accuracy
$EffNetL2_{RGB}$	79.9%
$EffNetL2_{Late}$	**83.3%**
$EffNetL2_{Early}$	75.3%
$EffNetL2_{Mid}$	76.8%
$EVA02_{RGB}$	80.2%
$EVA02_{Late}$	**86.4%**
$EVA02_{Early}$	65.5%
$CoatNet6_{RGB}$	82.7%
$CoatNet6_{Late}$	**88.9%**
$CoatNet6_{Early}$	80.2%
$CoatNet6_{Mid}$	81.0%

Table 5: Zero-shot average accuracy with and without CVRL

Architecture	Zero-shot accuracy with CVRL	Zero-shot accuracy without CVRL
$EffNetL2_{Late}$	24.2%	25.9%
$EVA02_{Late}$	62.1%	69.5%
$CoatNet6_{Late}$	61.0%	68.8%

the MINC dataset, each for a duration of 5 epochs. The pre-training utilized the Adam optimizer with a batch size of 8 and an initial learning rate of 10^{-4}. A cosine learning rate scheduler was implemented to gradually decrease the rate during the training period. Each network processed input data comprising patches of 256 pixels. The results, depicted in figure 3, revealed that all networks exhibited comparable performance levels when integrated with fusion techniques.

Notably, Contrastive Visual Representation Learning (CVRL) was applied exclusively to the late fusion. To facilitate cross-modality pattern learning, we selected specific publicly available datasets for backbone pre-training. For RGB-Depth, ScaNet and SUN RGB-D datasets were used, containing 12,000 and 10,335 data points, respectively. For RGB-NIR, the EPFL dataset, comprising 1,900 data points, was employed. The results, presented in table 5, affirm the effectiveness of this approach in enhancing cross-modal visual representation learning.

5.4.2. *Cross-modality evaluation and state of the art comparaison*

In our initial experiment, we evaluated the three chosen models—CoatNet-6, EVA-02-L, and EfficientNet-L2—using both late and early fusion methods. The second experiment adopted a more structured approach to assess model performance under varying conditions. Initially, we tested the models using only RGB data to establish a baseline. This was followed by a comprehensive test using all available modalities. In a subsequent phase, we excluded depth information to focus on the RGB and NIR modalities. The results, presented in table 6, demonstrate that our proposed models surpass existing state-of-the-art methods on the MINC RGB dataset. Notably, their performance further improves when trained with all modalities, as compared to training with just RGB or RGB-NIR data. Among the models, CoatNet-6 with late fusion achieved the highest performance, as per our findings.

Adding more modalities to the network (RGB, then RGB-NIR, then full modalities) gradually improves the recognition accuracy, proving that the selected modalities are complementary. The performance dropped when eliminating depth proves the importance of the geometry information to extract more precise reflection proprieties.

After conducting a comparative analysis of our proposed algorithm, which combines CoatNet6 and late fusion, against state-of-the-art methods for each category, as detailed in Table 7, we observed notable improvements in term of accuracy across almost all categories. However, there was a marked increase for certain categories of materials such as Mirror, Metal, Tile and Plastic which could be explained by the ability of re-

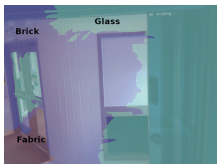
Table 6: Comparison between our method and state of the art. Only categories we have chosen are conserved in MINC dataset during this test.

Architecture	MINC top 1 accuracy	MINC top-5 accuracy	Our dataset top-1 accuracy
$GoogLeNet_{RGB}$	81.1%	97.2%	75.6%
$EffNetL2_{RGB}$	84.5%	98.7%	81.2%
$EffNetL2_{RGB-NIR-Late}$	—	—	82.0%
$EffNetL2_{RGB-NIR-Depth-Early}$	—	—	85.3%
$EffNetL2_{RGB-NIR-Depth-Late}$	—	—	87.3%
$EVA02_{RGB}$	86.5%	99.3%	87.2%
$EVA02_{RGB-NIR-Late}$	—	—	88.4%
$EVA02_{RGB-NIR-Depth-Early}$	—	—	65.5%
$EVA02_{RGB-NIR-Depth-Late}$	—	—	91.4%
$CoatNet6_{RGB}$	89.7	99.4%	89.1%
$CoatNet6_{RGB-NIR-Late}$	—	—	91.7%
$CoatNet6_{RGB-NIR-Depth-Early}$	—	—	84.3%
$CoatNet6_{RGB-NIR-Depth-Late}$	—	—	**94.3%**

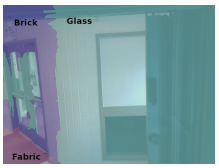
Table 7: Comparison between our method and state of the art in term of per-category accuracy. The chosen model is CoatNet6 with late full modality fusion for tested on our dataset and MINC dataset.

Category	RGB accuracy on MINC	RGB accuracy on our dataset	Multimodal accuracy on our dataset
Brick	89.4%	88.1%	92.3%
Pol. stone	89.7%	89.4%	92.2%
Fabric	95.1%	79.8%	84.3%
Glass	90.5%	83.4%	84.9%
Leather	77.3%	88.6%	90.1%
Metal	87.9%	**58.2%**	**77.8%**
Mirror	73.8%	**23.5%**	**73.0%**
Other	89.4%	83.2%	87.2%
Painted	90.1%	92.7%	83.8%
Paper	68.6%	84.3%	89.3%
Plastic	66.2%	**71.7%**	**80.8%**
Stone	74.8%	91.4%	96.0%
Tile	84.2%	**72.4%**	**89.3%**
Wood	94.5%	96.5%	97.7%

(a) Input image.



(b) CoatNet6 trained on RGB.



(c) CoatNet6 trained on all modalities.

Figure 10: Segmentation experiment using CRF: comparaison between the results of CoatNet6 trained on RGB data and full modality data with late fusion.

flective features to distinguish certain materials,especially when their object context is ambiguous.

5.4.3. Segmentation experiment

In addition to patch-based classification, the recognition model could be used for a segmentation experiment by using it as a sliding window on the entire image to perform dense prediction. However, patch-based training methods may not have local consistency in their output, so a conditional random field (CRF) was attached at the back-end to generate the final output with higher accurqacy, as demonstrated in [20]. Figure 10 presents the results. The figure shows that our proposed model trained on all modalities gives more visually consistent results. It shows more accuracy in detecting plastic and metal, which other methods may struggle with.

6. Conclusion

We propose an innovative material recognition method that uniquely integrates multi-spectral texture with surface reflectivity characteristics across the visual-infrared spectrum. Our approach is structured in two phases: the first involves multi-modal camera fusion to consolidate features and address camera misalignment, and the second phase is a classification model specifically designed for material recognition. This method stands out for its ability to accurately identify material surface classes in non-constrained environments by combining proposed reflectivity features and material texture from the NIR-visible spectrum. We utilize NIR laser projectors to generate reflections in the NIR frame, which are then modeled using both the analytical and Phong models. Our findings indicate that by estimating reflected light from the disparity between RGB and NIR frames, and deducing surface geometry from depth data, we can achieve a reliable approximation of reflection features.

The efficacy of our method is further bolstered by implementing solutions to address the data limitation issues commonly found in previous multi-spectral and hyper-spectral approaches. These solutions include the adoption of a semi-supervised learning framework and the strategic use of large-scale RGB datasets for regularization. Comparative evaluations of our model against existing state-of-the-art methods on the MINC dataset and our proprietary dataset revealed that our approach achieves comparable prediction accuracy using only RGB data. More notably, there is a significant 6% increase in test accuracy when all input modalities are employed on our dataset.

Using multi-spectral information in the visual-infrared range spectrum is possibly expandable by using other spectral ranges sensors like near-UV or short-wave infrared cameras to enrich the model prior information with more multi-spectral texture, as discriminating features.

References

[1] Z. Erickson, E. Xing, B. Srirangam, S. Chernova, and C. C. Kemp, "Multimodal material classification for robots using spectroscopy and high resolution texture imaging," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10452–10459, IEEE, 2020.

[2] S. M. Zainab, K. Khan, A. Fazil, and M. Zakwan, "Foreign object debris (fod) classification through material recognition using deep convolutional neural network with focus on metal," *IEEE Access*, vol. 11, pp. 10925–10934, 2023.

[3] H. He, D.-W. Sun, Z. Wu, H. Pu, and Q. Wei, "On-off-on fluorescent nanosensing: Materials, detection strategies and recent food applica-

tions," *Trends in Food Science & Technology*, vol. 119, pp. 243–256, 2022.

[4] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[5] H. Yu, F. Zhang, L. Wei, Y. Huang, and W. Hu, "Deep convolutional neural network for hyper spectral sensing classification," *J. Sens*, vol. 2, pp. 1–12, 2015.

[6] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, 2017.

[7] L. Sun, Z. Wu, J. Liu, L. Xiao, and Z. Wei, "Supervised spectral–spatial hyperspectral image classification with weighted markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1490–1503, 2014.

[8] A. Porebski, M. Alimoussa, and N. Vandenbroucke, "Comparison of color imaging vs. hyperspectral imaging for texture classification," *Pattern Recognition Letters*, vol. 161, pp. 115–121, 2022.

[9] O. Wang, P. Gunawardane, S. Scher, and J. Davis, "Material classification using brdf slices," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2805–2811, IEEE, 2009.

[10] C. Hahlweg and H. Rothe, "Classification of optical surface properties and material recognition using mutlispectral brdf data measured with a semihemispherical spectro-radiometer in vis and nir," in *Optical Fabrication, Testing, and Metrology II*, vol. 5965, pp. 150–161, SPIE, 2005.

[11] J. Xue, H. Zhang, K. Dana, and K. Nishino, "Differential angular imaging for material recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 764–773, 2017.

[12] C. Liu and J. Gu, "Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral brdf," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 86–98, 2013.

[13] M. Weinmann, J. Gall, and R. Klein, "Material classification based on training data synthesized using a btf database," in *European Conference on Computer Vision*, pp. 156–171, Springer, 2014.

[14] X. Zhang and J. Saniie, "Material texture recognition using ultrasonic images with transformer neural networks," in *2021 IEEE International Conference on Electro Information Technology (EIT)*, pp. 1–5, IEEE, 2021.

[15] J. Kim, H. Lim, S. C. Ahn, and S. Lee, "Rgbd camera based material recognition via surface roughness estimation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1963–1971, IEEE, 2018.

[16] D. Tafone, L. McEvoy, Y. M. Sua, P. Rehain, and Y. Huang, "Surface material recognition through machine learning using time of flight lidar," *Optics Continuum*, vol. 2, no. 8, pp. 1813–1824, 2023.

[17] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2407–2416, 2016.

[18] J. DeGol, M. Golparvar-Fard, and D. Hoiem, "Geometry-informed material recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1554–1562, 2016.

[19] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 103–110, 2007.

[20] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3479–3487, 2015.

[21] P. Vácha and M. Haindl, "Texture recognition under scale and illumination variations," *Journal of Information and Telecommunication*, pp. 1–19, 2023.

[22] V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters*, vol. 84, pp. 63–69, 2016.

[23] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks for texture classification," *arXiv preprint arXiv:1707.07394*, 2017.

[24] H. Zhang, J. Xue, and K. Dana, "Deep ten: Texture encoding network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 708–717, 2017.

[25] O. Adedeji and Z. Wang, "Intelligent waste classification system using deep learning convolutional neural network," *Procedia Manufacturing*, vol. 35, pp. 607–612, 2019.

[26] W. Lu, J. Chen, and F. Xue, "Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach," *Resources, Conservation and Recycling*, vol. 178, p. 106022, 2022.

[27] X. Xie, L. Yang, and W.-S. Zheng, "Learning object-specific dags for multi-label material recognition," *Computer Vision and Image Understanding*, vol. 143, pp. 183–190, 2016.

[28] M. S. Drehwald, S. Eppel, J. Li, H. Hao, and A. Aspuru-Guzik, "One-shot recognition of any material anywhere using contrastive learning with physics-based rendering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23524–23533, 2023.

[29] F. Xu, M. S. Wong, R. Zhu, J. Heo, and G. Shi, "Semantic segmentation of urban building surface materials using multi-scale contextual attention network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 158–168, 2023.

[30] S. T. Namin and L. Petersson, "Classification of materials in natural scenes using multi-spectral images," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1393–1398, IEEE, 2012.

[31] N. Salamati, C. Fredembach, and S. Süsstrunk, "Material classification using color and nir images," in *Color and Imaging Conference*, pp. 216–222, Society for Imaging Science and Technology, 2009.

[32] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4d light-field dataset and cnn architectures for material recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 121–138, Springer, 2016.

[33] D. Guarnera, G. C. Guarnera, A. Ghosh, C. Denk, and M. Glencross, "Brdf representation and acquisition," in *Computer Graphics Forum*, vol. 35, pp. 625–650, Wiley Online Library, 2016.

[34] L. Ren, Z. Pan, J. Cao, and J. Liao, "Infrared and visible image fusion based on variational auto-encoder and infrared feature compensation," *Infrared Physics & Technology*, vol. 117, p. 103839, 2021.

[35] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[36] D. Shen, M. Zareapoor, and J. Yang, "Infrared and visible image fusion via global variable consensus," *Image and Vision Computing*, vol. 104, p. 104037, 2020.

[37] D. Shen, M. Zareapoor, and J. Yang, "Multimodal image fusion based on point-wise mutual information," *Image and Vision Computing*, vol. 105, p. 104047, 2021.

[38] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[39] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "Swinfuse: A residual swin transformer fusion network for infrared and visible images. arxiv 2022," *arXiv preprint arXiv:2204.11436*, 2022.

[40] J. Kolluri and R. Das, "Intelligent multimodal pedestrian detection using hybrid metaheuristic optimization with deep learning model," *Image and Vision Computing*, p. 104628, 2023.

[41] G. Tziafas and H. Kasaei, "Early or late fusion matters: Efficient rgb-d fusion in vision transformers for 3d object recognition," *Arxiv*, 2023.

[42] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," pp. 16000–16009, 2022.

[43] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers. arxiv 2021," *arXiv preprint arXiv:2106.08254*, 2021.

[44] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva-02: A visual representation for neon genesis," *arXiv preprint arXiv:2303.11331*, 2023.

[45] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.

[46] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum con-

trast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.
[47] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.
[48] Q. Zhou, C. Yu, H. Luo, Z. Wang, and H. Li, "Mimco: Masked image modeling pre-training with contrastive teacher," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4487–4495, 2022.
[49] USGS, "Spectral library version 7." `https://crustal.usgs.gov/speclab/QueryAll07a.php`. Accessed: 2023-02-20.
[50] G. Winkler, *Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction*, vol. 27. Springer Science & Business Media, 2003.
[51] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of mathematical imaging and vision*, vol. 18, no. 1, pp. 17–33, 2003.
[52] B. T. Phong, "Illumination for computer generated pictures," *Communications of the ACM*, vol. 18, no. 6, pp. 311–317, 1975.
[53] paperswithcode, "Image classification on imagenet." `https://paperswithcode.com/sota/image-classification-on-imagenet`. Accessed: 2023-09-20.
[54] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.
[55] CUReT, "Columbia-utrecht reflectance and texture database." `https://www.cs.columbia.edu/CAVE/software/curet/`. Accessed: 2023-11-20.
[56] KTH-TIPS, "The kth-tips and kth-tips2 image databases." `https://www.csc.kth.se/cvap/databases/kth-tips/index.html`. Accessed: 2023-11-20.