Titre: Title:	Super-résolution d'images thermiques d'enfants en soins intensifs pédiatriques
Auteur: Author:	Cyprien Arnold
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Arnold, C. (2024). Super-résolution d'images thermiques d'enfants en soins intensifs pédiatriques [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/61696/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/61696/
Directeurs de recherche: Advisors:	Lama Séoud, & Philippe Jouvet
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Super-résolution d'images thermiques d'enfants en soins intensifs pédiatriques

CYPRIEN ARNOLD
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Décembre 2024

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Super-résolution d'images thermiques d'enfants en soins intensifs pédiatriques

présenté par **Cyprien ARNOLD**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Benjamin DE LEENER, président
Lama SÉOUD, membre et directrice de recherche
Philippe JOUVET, membre et codirecteur de recherche
Guillaume-Alexandre BILODEAU, membre

DÉDICACE

À mes parents,
Merci de m'avoir permis d'arriver jusqu'ici ...

REMERCIEMENTS

Ce mémoire est le fruit de deux années de recherche et marque la fin de mes six ans d'études. La démarche d'une maîtrise recherche exige une grande autonomie, tout en nécessitant un encadrement qui offre un cadre propice à l'avancement du projet. C'est pourquoi, je tiens, avant tout, à exprimer ma sincère gratitude envers ma directrice de recherche, Lama Séoud, sans qui ce travail n'aurait pas vu le jour. Tout en m'encourageant avec bienveillance à donner le meilleur de moi-même, elle m'a guidé avec des orientations claires et a su cultiver en moi le goût de la recherche. Cette expérience à ses côtés va sans aucun doute fortement structurer mon futur parcours professionnel.
Je tiens également à remercier le professeur Philippe Jouvet. Il fut un excellent lien entre la partie technique de ce travail et son application dans le contexte hospitalier. Grâce à lui et à Mario Munoz, j'ai pu bénéficier de l'infrastructure du Centre hospitalier Sainte-Justine, ce qui m'a permis d'obtenir des résultats en lien direct avec la réalité des soins intensifs. Je souhaite également remercier Aya ainsi que le personnel de recherche des soins intensifs pour leur aide précieuse dans l'acquisition des images de patients qui figurent dans ce mémoire.
Je souhaite également exprimer ma gratitude envers Philippe Debanné pour ses relectures attentives et ses corrections précieuses, notamment lors de la rédaction de l'article.

Par ailleurs, il est essentiel pour moi de remercier mes camarades de laboratoire : Philippe, Étienne, Hugo, Lauriane, Victor B., Valérie, Gaspar, Sidney, Corentin, Yu-Chi, Clément, Doha, Victor N. et Nathan. Les moments de convivialités partagés et leur soutien ont presque réussi à rendre agréable le travail dans une bureau sans fenêtre.
Un immense merci aussi à mes amis pour leur présence au quotidien et leurs encouragements si précieux dans les moments de doute. Merci à Pierrick, Sungwoo, Léo, Guillaume, David, Célia, Héloïse et Bich-Lien d'avoir été présents au quotidien.
Enfin, je tiens à remercier avec affection mes parents et mon frère pour leur soutien inconditionnel malgré la distance géographique. Ils m'ont toujours fait confiance en m'encourageant à poursuivre les voies qui m'inspirent.

RÉSUMÉ

Ce travail de recherche porte sur la super-résolution d'images thermiques. Il a été développé pour améliorer le monitorage des enfants en soins intensifs pédiatriques. En effet, l'utilisation de caméras thermiques permettrait d'orienter et de faciliter le travail du personnel médical vers les cas les plus sévères, visant ainsi à améliorer l'organisation du travail du personnel soignant dont les effectifs sont très tendus.

Les caméras thermiques ont de nombreuses applications dans le milieu médical. Elles permettent une surveillance non-invasive (qui ne gêne pas le patient) tout en permettant l'accès à sa température ou à la position de ses membres sous une couverture. Cependant, le coût à l'achat d'une caméra thermique de bonne qualité, avec une grande résolution, demeure très élevé. Il existe des caméras thermiques bon marché mais avec une résolution très dégradée.

L'objectif de ce travail de recherche est de pouvoir exploiter les images thermiques en augmentant la qualité des images acquises avec des caméras thermiques de faible résolution. Ce processus d'augmentation de la qualité d'une image est appelé super-résolution (SR). Cette tâche de vision par ordinateur est une tâche qui a été très étudiée ces dernières années et dans ce mémoire nous présentons une nouvelle architecture de réseau de neurones avec des résultats très compétitifs. L'originalité de ce travail de recherche porte sur le fait d'utiliser l'image du spectre du visible dans le but d'améliorer la qualité de l'image thermique tout en assurant une robustesse de la méthode dans le cas où une modalité manquerait.

ABSTRACT

This research focuses on thermal image super-resolution, developed to enhance monitoring in pediatric intensive care units. The use of thermal cameras could guide medical staff toward the most severe cases, thereby improving workflow organization in teams that are often understaffed.

Thermal cameras have numerous applications in the medical field. They enable non-invasive monitoring (avoiding patient discomfort) while providing access to vital information such as body temperature or limb positioning, even under a blanket. However, the cost of high-resolution, high-quality thermal cameras remains prohibitively expensive. More affordable thermal cameras exist but typically suffer from significantly reduced resolution.

The goal of this research is to leverage thermal images by enhancing the quality of images captured by low-resolution thermal cameras. This process of improving image quality is known as super-resolution (SR). This computer vision task has been extensively studied in recent years. In this work, we present a novel neural network architecture that achieves highly competitive results. The originality of this research lies in leveraging visible-spectrum images to improve the quality of thermal images while ensuring the robustness of the method in cases where one modality may be unavailable.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

SIP	Soins Intensifs Pédiatriques
CHUSJ	Centre Hospitalier Universitaire Sainte Justine
SR	Super-Résolution
PICUPE	Pediatric Intensive Care Unit Pose Estimation
RGB	Red-Green-Blue
IR	InfraRouge
EPH	Estimation de Pose Humaine
GAN	Generative Adversarial Network
GTISR	Guided Thermal Image Super Resolution
PSNR	Peak Signal to Noise Ratio
SLP	Simultaneously-collected multi-modal Lying Pose
PBVS	Perception Beyond the Visible Spectrum
SE	Sous-Echantillonné

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

Ce projet de recherche est réalisé en partenariat avec le Centre Hospitalier Universitaire Sainte-Justine (CHUSJ). Le service de soins intensifs pédiatriques (SIP) accueille des enfants qui ont subi une intervention chirurgicale majeure et dont l'état de santé critique nécessite un suivi, souvent à la suite d'interventions chirurgicales majeures. Actuellement, une infirmière est responsable de la surveillance simultanée de deux chambres. Cette surveillance repose sur un suivi des signes vitaux via des moniteurs informatiques, ainsi qu'un contrôle visuel des enfants à travers des vitres. Plusieurs projets, menés par Dr Philippe Jouvet, visent à développer des systèmes d'intelligence artificielle (IA) permettant d'alléger la charge de travail des infirmières. Mon projet de recherche s'inscrit dans cette démarche. L'équipe de Dr Jouvet a notamment étudié comment les données thermiques peuvent contribuer à la détection d'anomalies cardiaques post-chirurgie cardiaque [22] et à l'évaluation de la température corporelle [23]. La surveillance du gradient thermique entre le canthus de l'œil et l'orteil ou le pied a été avancé comme marqueur du débit cardiaque des patients et a un intérêt dans la surveillance post-opératoire de chirurgie cardiaque pour détecter le syndrome de bas débit cardiaque post-opératoire [22]. De plus, les informations thermiques offrent peu de détails sur les textures des objets, rendant impossible l'identification des visages des patients. Par conséquent, l'imagerie thermique garantit la confidentialité, ce qui facilite, par exemple, le partage des images. D'autres travaux ont démontré des applications de la thermographie dans le domaine médical, telles que la détection de la fièvre chez un patient et l'identification de signes d'inflammation [24].

Le projet global Pediatric Intensive Care Unit Pose Estimation (PICUPE) a pour objectif la mise en place d'algorithmes d'estimation de pose dans les chambres du SIP pour monitorer le mouvement des patients dans l'unité des SIP. Dans ce projet, le travail de maîtrise d'Olivier Desclaux [25] consistait à établir un protocole pour la création d'une base de données multimodales dans ces chambres. En effet, pour entraîner un algorithme d'estimation de pose humaine (EPH), il était nécessaire de disposer de données issues de conditions réelles. Ces données multimodales incluent des images dans le visible RGB (Red-Green-Blue), de profondeur, et en infrarouge (IR). Le travail d'Olivier Desclaux a démontré que l'ajout de la modalité IR améliorait les performances de l'EPH, en particulier dans les cas d'occlusions visuelles. Également, le travail de maîtrise de Ghassen Cherni [26] confirme ces résultats : la fusion des modalités IR et profondeur avec la modalité RGB permet d'améliorer les performances de l'EPH unimodale.

Ainsi l'obtention d'une image thermique de l'enfant y compris en cas d'occlusion permettrait une surveillance continue du gradient thermique et l'estimation de pose des enfants alités dans le but d'une meilleure prise en charge.

Cependant, la résolution spatiale d'une image thermique peut varier considérablement en fonction du capteur utilisé. Par exemple, la caméra infrarouge FLIR T1020-45, offre une résolution de 1 024 $\times$ 768 pixels, pour un coût d'environ 60 000 $ CAD chez ITM Instrument, tandis que la caméra FLIR Lepton offre une résolution limitée à 60 $\times$ 80 pixels, pour un coût d'environ 200 $ CAD. Pour des raisons budgétaires, il est irréaliste d'équiper toutes les chambres des SIP avec des caméras onéreuses comme la FLIR T1020-45.

Pour pallier ce problème, nous proposons dans ce mémoire d'explorer des techniques de super-résolution afin d'améliorer la résolution des images en provenance de caméras thermiques plus abordables mais à faible résolution. Acquérir une image de qualité moyenne, puis l'améliorer, est une solution bien plus économique que d'obtenir directement une image haute résolution.

La principale difficulté de la super-résolution réside dans la génération de nouveaux pixels à partir d'une quantité d'information très limitée, en particulier pour des facteurs de super-résolution élevés. Les images infrarouges, souvent de très faible résolution, nécessitent des facteurs de super-résolution élevés (par exemple 8 ou 16) pour obtenir une amélioration notable de leur qualité.

Pour surmonter cette difficulté, des méthodes exploitant la modalité du spectre visible ont émergé. Bien que prometteuses, ces méthodes dites quidées, sont récentes et peu de travaux les ont exploitées. De plus, l'aspect de robustesse en cas d'absence de la modalité guide a été très peu étudié. Cet aspect nous semble crucial, notamment dans un contexte hospitalier où il est essentiel de disposer d'une méthode performante dans toutes les conditions. Au CHUSJ, l'image dans le spectre visible est déjà acquise par une caméra Kinect-Azure, ce qui rend ces méthodes particulièrement adaptées au contexte de notre projet.

Ce travail de recherche vise à proposer une méthode **efficace** tout en garantissant une **robustesse** accrue dans les situations où la modalité guide serait absente.

Dans le **Chapitre 2**, nous proposerons une revue des connaissances sur les caractéristiques des images thermiques ainsi que des capteurs utilisés pour les acquérir. Par la suite, une revue de la littérature sur les différentes méthodes de super-résolution est proposée, à commencer par les techniques appliquées aux images dans le spectre visible, puis celles utilisées dans le spectre infrarouge. Nous mettrons en lumière les approches spécifiques développées pour la SR sur les images IR, en particulier les méthodes guidées qui exploitent les informations RGB

pour reconstruire l'image. Enfin, nous présenterons les métriques afin d'évaluer la tâche de SR ainsi que les différents jeux de données disponibles pour notre étude.

À la lumière de la revue de littérature, le **Chapitre 3** exposera les objectifs du projet de recherche, qui visent à répondre à la problématique soulevée plus haut concernant la faible résolution des images thermiques.

Le **Chapitre 4** rapportera les tests préliminaires avec des méthodes de super-résolution existantes, adaptées aux images thermiques dans des conditions proches de la réalité. Ces expérimentations ont permis de dégager des pistes pour le développement d'une nouvelle architecture de réseau de neurones pour la SR des images thermiques, en tirant profit des informations de la modalité RGB. Cette nouvelle méthode est présentée au **Chapitre 5**, sous la forme d'un article de conférence [27] accepté et présenté en juin 2024, au workshop PBVS (Perception Beyond the Visible Spectrum) de la conférence CVPR (Computer Vision and Pattern Recognition).

Ensuite, nous exposerons dans le **Chapitre 6** les résultats préliminaires de la méthode de SR proposée appliquée à des images de la base de données PICUPE, acquises au CHUSJ à l'été 2024.

Enfin le **Chapitre 7** résumera les travaux réalisés, mettant en avant les principales contributions et les résultats obtenus lors de cette maîtrise. Nous clôturerons ce travail de recherche en identifiant les pistes de recherche prometteuses et les axes d'amélioration possibles.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans cette revue de littérature, nous commencerons par aborder des sujets plus généraux comme **l'imagerie thermique 2.1** et **la super-résolution et ses applications 2.2** puis nous entrerons plus dans les détails techniques en présentant plus spécifiquement les méthodes de **super-résolution dans le visible 2.2.1** et **dans l'infrarouge 2.2.2**. Nous aborderons aussi **la fusion de modalités 2.3** et **les méthodes de super-résolution guidée 2.2.3**, deux sujets très liés.
Nous terminerons ce chapitre par la présentation **des métriques 2.4** utilisées et **des jeux de données 2.5** utilisés dans ce travail de recherche.

2.1 L'imagerie thermique

L'imagerie thermique, également appelée thermographie infrarouge, est une technique non invasive de détection et de visualisation des rayonnements infrarouges émis par les objets en fonction de leur température. Cette méthode repose sur le principe que tout corps ayant une température supérieure au zéro absolu émet un rayonnement électromagnétique dans le spectre infrarouge [28].

Cet article [29] explique de manière approfondie le fonctionnement des caméras thermiques, leurs applications ainsi que les enjeux qui leur sont associés. Les caméras thermiques captent ces émissions infrarouges, généralement dans les longueurs d'onde de 1 à 14 μm, et les convertissent en images visuelles représentant la distribution spatiale des températures à la surface des objets observés. Les trois principales bandes de fréquences infrarouges sont : le SWIR (Short Wavelength Infrared, 0,9–1,7 μm), utilisé pour capturer des images à travers les nuages, la fumée et le brouillard ; le MWIR (Mid Wavelength Infrared, 3,0–5,0 μm), spécialisé dans la détection de fuites de gaz invisibles à l'œil nu ; et le LWIR (Long Wavelength Infrared, 8,0–14,0 μm), largement utilisé pour les inspections thermiques, comme la détection des variations de température dans les bâtiments ou les systèmes électroniques. La plupart des caméras thermiques opèrent dans la bande LWIR, qui sera la plage de longueurs d'onde étudiée dans notre travail.
Les images thermiques générées par les caméras infrarouges sont intrinsèquement monochromatiques, utilisant un seul canal pour représenter les variations de température. Cette caractéristique les distingue des images produites par les caméras optiques conventionnelles, qui utilisent généralement trois canaux distincts RGB pour capturer et reproduire le spectre visible. Toute coloration apparente dans les images thermiques résulte d'un traitement post-

acquisition, où des palettes de pseudo-couleurs 2.1 sont appliquées pour faciliter l'interprétation visuelle des gradients thermiques.

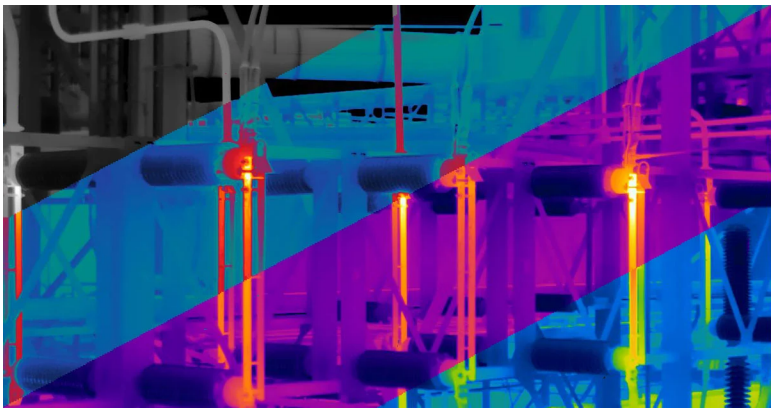


FIGURE 2.1 Exemple de palettes de pseudo-couleurs [1]

Cette technologie permet de mesurer des variations de température avec une précision pouvant atteindre 0,1°C dans certains cas. À titre d'exemple, la caméra thermique FLIR T1020-45 utilisée au laboratoire a une sensibilité de 0,2 °C lors d'une acquisition à une température ambiante de 30 °C (selon le fabricant).

L'utilisation des informations thermiques, obtenues par thermographie infrarouge, présente plusieurs défis méthodologiques et techniques. Premièrement, la précision des mesures est affectée par plusieurs facteurs comme l'émissivité des matériaux observés, qui varie selon leur nature et leur état de surface. Elle est aussi affectée par les conditions environnementales, telles que la température ambiante, l'humidité relative, la pluie et la neige [30]. Cette variabilité peut conduire à des erreurs d'interprétation si elle n'est pas correctement prise en compte.
Pour résoudre ces problèmes liés à la précision des mesures ainsi que les distorsions optiques, il est essentiel de calibrer la caméra de manière à ce que les couleurs de l'image correspondent fidèlement aux températures réelles. Contrairement aux caméras visibles, les méthodes de calibration classiques, comme les damiers imprimés, sont inefficaces dans le spectre infrarouge car elles ne produisent pas de contraste suffisant. Des techniques spécifiques, comme l'utilisation de panneaux actifs ou de matériaux à émissivité variée, sont donc nécessaires pour créer des images thermiques contrastées [31].
Deuxièmement, la résolution spatiale des caméras thermiques, généralement très inférieure à celle des caméras optiques, limite la détection de petites anomalies thermiques ou de détails fins. Ceci s'explique par la limitation sur la taille des éléments détecteurs. Contrairement aux caméras visibles dont les pixels mesurent 1 à 2 µm, les détecteurs thermiques mesurent entre

12 et 17 µm [32]. Cette différence s'explique par le fait que les caméras thermiques doivent capter de l'énergie infrarouge, dont la longueur d'onde est beaucoup plus grande que celle de la lumière visible. Par conséquent, les détecteurs doivent être plus grands pour capter suffisamment de cette énergie, ce qui réduit le nombre de pixels dans un capteur de même taille physique. Résultat : les caméras thermiques ont une résolution plus faible, avec moins de détails dans les images produites.

2.2 La super-résolution et ses applications

La super-résolution est un traitement numérique qui permet d'augmenter la résolution d'une image à basse définition en générant des pixels supplémentaires et en y rajoutant des détails visuels. En termes simples, elle permet de produire une version plus détaillée d'une image de faible résolution. La super-résolution s'inscrit dans le cadre plus général de la restauration d'image, un domaine qui vise à améliorer la qualité visuelle d'images dégradées ou incomplètes.

En fait, la SR est souvent considérée comme un problème inverse, car elle consiste à inverser le processus de dégradation qui a réduit la résolution originale. Pour modéliser artificiellement cette perte d'information visuelle, on utilise la plupart du temps du sous-échantillonnage (bicubique par exemple), puis on peut ajouter du bruit ou du flou. Nous verrons dans les parties **2.2.1** et **2.2.2** qu'il est possible de modéliser différemment cette perte d'information visuelle.

La grande majorité des algorithmes de SR apprennent à reconstruire les détails manquants en étudiant les motifs récurrents dans les images. En super-résolution, un facteur multiplicatif $\times n$ indique que les dimensions de l'image (largeur et hauteur) sont multipliées par n , entraînant une augmentation du nombre total de pixels par un facteur de n^2. Par exemple, un agrandissement $\times 2$ quadruple le nombre de pixels, car chaque dimension est doublée, entraînant 4 fois plus de pixels qu'au départ.

L'intérêt de SR réside dans sa capacité à améliorer la qualité visuelle des images, en particulier dans les situations où il est difficile ou coûteux d'obtenir une haute résolution native, comme en imagerie médicale, en vidéo surveillance, ou encore en photographie. Plutôt que d'augmenter la taille de l'image comme le font les techniques d'interpolation (bicubique, bilinéaire...), la super-résolution ajoute des détails réalistes et essaie de réduire les artefacts visuels (flou, bruit...).

Les applications de la super-résolution sont multiples. Dans le domaine médical, elle permet de rendre plus précises les images issues de scanners ou d'IRM [33], facilitant ainsi la détection précoce de pathologies. Des chercheurs ont directement démontré l'efficacité de la

super-résolution dans une tâche médicale en prouvant que son utilisation améliore la détection de problèmes médicaux chez les nouveau-nés [2] (voir Annexe A pour des résultats visuels).

FIGURE 2.2 Le processus pour catégoriser la bonne santée du nouveau-né [2]

Dans le domaine de la vidéo surveillance, la super-résolution améliore les détails des images capturées par des caméras à basse résolution, permettant une meilleure identification des objets [34] ou des individus [35]. Dans les jeux vidéo, la super-résolution est employée pour améliorer les graphismes tout en minimisant la charge de calcul nécessaire à l'affichage en haute résolution (voir Automatic Super Resolution de Microsoft). Enfin, la SR est aussi utilisée en astronomie [36–38] pour améliorer la qualité des images spatiales.

2.2.1 La super-résolution dans le spectre visible

Les premières approches de super-résolution ont été développées pour le spectre visible. Elles utilisaient des méthodes dites "traditionnelles" [39]. Ces méthodes se concentraient soit sur le domaine fréquentiel, en tentant de modéliser la relation entre les images haute résolution (HR) et basse résolution (BR) à l'aide de modèles mathématiques [40–42], soit sur des méthodes de dictionnaires pour faire correspondre des patchs BR à des patchs HR [43–45].
En 2015, des méthodes basées sur l'apprentissage profond utilisant des réseaux de neurones convolutifs (connus sous l'acronyme CNN pour convolutionnal neural network) ont émergé. Ainsi SRCNN [46], FSRCNN [47] et ESPCN [3] ont introduit les couches de convolution sous-pixel, une nouvelle opération de sur-échantillonnage pour reconstruire l'image.

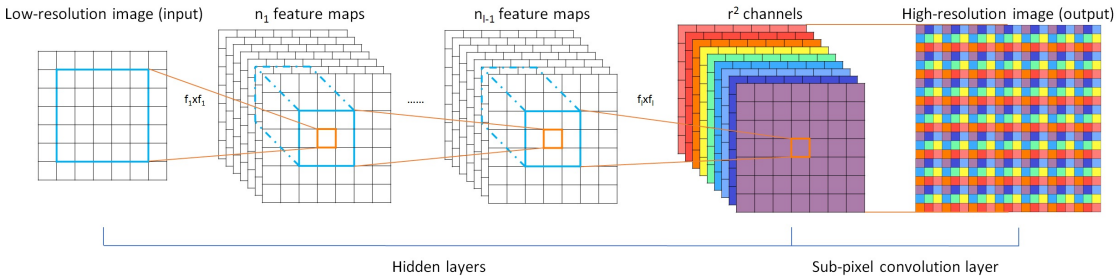
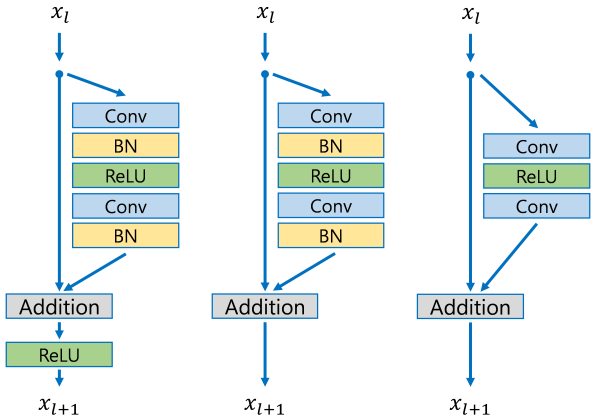


FIGURE 2.3 Visualisation de l'opération de convolution sous-pixel [3]

L'avènement des réseaux résiduels [48] (notamment pour résoudre le problème du gradient évanescent) a conduit à de nouvelles architectures comme VDSR [49], RED [50], SRResNet [51] et EDSR [4], ce dernier proposant une nouvelle connexion résiduelle et remportant le NTIRE 2017 Super-Resolution Challenge [52].



(a) Original [48] (b) SRResNet [51] (c) EDSR [4]

FIGURE 2.4 Comparaison des différents blocs résiduels [4]

2014 est l'année d'apparition de la célèbre architecture de réseaux antagonistes génératifs (GAN) [53]. Celle-ci consiste en deux réseaux de neurones qui s'entraînent simultanément : un générateur, qui crée des échantillons de données, et un discriminateur, qui tente de distinguer ces échantillons des données réelles. Le générateur s'améliore en essayant de tromper le discriminateur, tandis que le discriminateur devient plus performant à reconnaître les fausses données, permettant ainsi au GAN de produire des résultats réalistes.
En 2017, SRGAN [54] a obtenu des résultats remarquables en utilisant un GAN à la tâche de super-résolution. Un an plus tard, ESRGAN [55], une version améliorée de SRGAN, a vu le

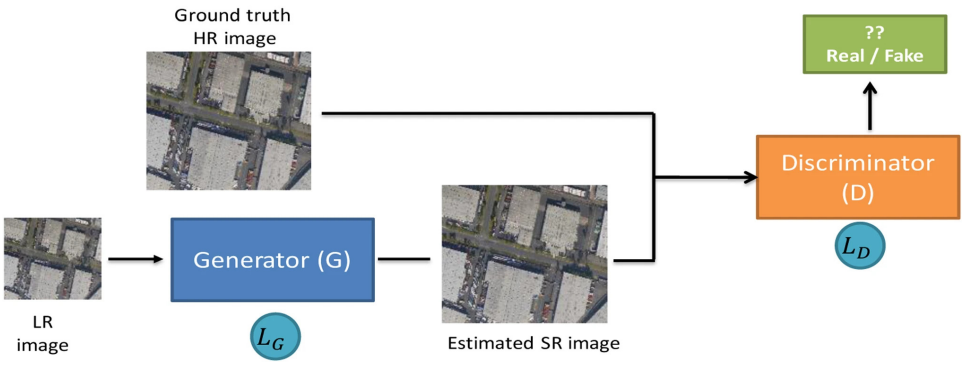


FIGURE 2.5 Exemple de GAN pour la super-résolution [5]

jour. ESRGAN améliore SRGAN avec un nouveau bloc appelé RRDB (Residual-in-Residual Dense Block), qui utilise des connexions denses et un apprentissage résiduel à plusieurs niveaux pour une meilleure performance. Il ajoute également des techniques comme le "residual scaling" [56] et une initialisation des poids plus faible pour stabiliser l'entraînement des réseaux profonds, car les GAN sont réputés pour leurs problèmes de convergence et de stabilité.

Depuis l'avènement des transformers [57], qui ont révolutionné l'apprentissage profond, leur adaptation au champ de la vision par ordinateur avec le Vision Transformer (ViT) [58] a marqué une étape importante. Le Swin Transformer [6] a ensuite résolu le problème de la complexité computationnelle du ViT en introduisant des fenêtres locales. Ces fenêtres, présentées dans SwinIR, appliquent l'attention non plus sur des patches globaux de l'image, comme le faisait ViT, mais sur des sous-régions (fenêtres) plus petites, limitant ainsi la portée de l'attention. Ces fenêtres sont décalées entre les couches pour permettre à l'attention de capturer les relations entre fenêtres adjacentes, tout en réduisant la complexité computationnelle du modèle.

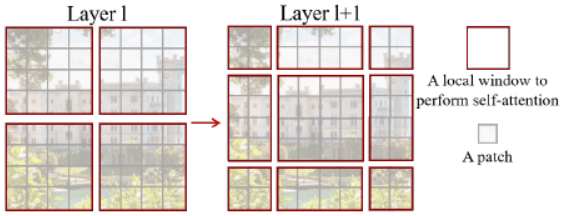


FIGURE 2.6 Illustration du calcul de self attention sur les patchs de l'images [6]

Grâce à cette innovation, SwinIR [59] a pu appliquer l'architecture Swin à la tâche de restauration d'images, surpassant ainsi les meilleures architectures existantes. La force principale

de SwinIR réside dans ses blocs résiduels de Swin Transformer, qui permettent d'extraire des caractéristiques particulièrement pertinentes. Plus récemment, les architectures HAT [60] et SwinFIR [61] ont apporté des améliorations à SwinIR et représentent aujourd'hui l'état de l'art en matière de super-résolution. Un des problèmes des modèles de transformers en SR est le mécanisme d'attention sur une fenêtre qui peut restreindre la capacité du modèle à capturer des dépendances globales essentielles pour une super-résolution efficace [62].

Une autre approche de la super-résolution appelée super-résolution aveugle (Blind SR) se concentre sur la modélisation du processus de dégradation afin de convertir une image haute résolution en une image basse résolution. Elle se distingue en réalisant la super-résolution d'image sans avoir besoin de connaître à l'avance le type de dégradation que l'image basse résolution a subi. Contrairement aux méthodes traditionnelles qui supposent un processus de dégradation pré-défini (très souvent un sous-échantillonnage bicubique), le Blind SR vise à proposer un modèle de dégradation modélisant des dégradations plus complexes et réalistes, ce qui le rend plus adapté aux scénarios du monde réel [7]. Ces méthodes sont censées réduire le "domain gap" lorsque la dégradation ne se limite pas à un simple sous-échantillonnage.

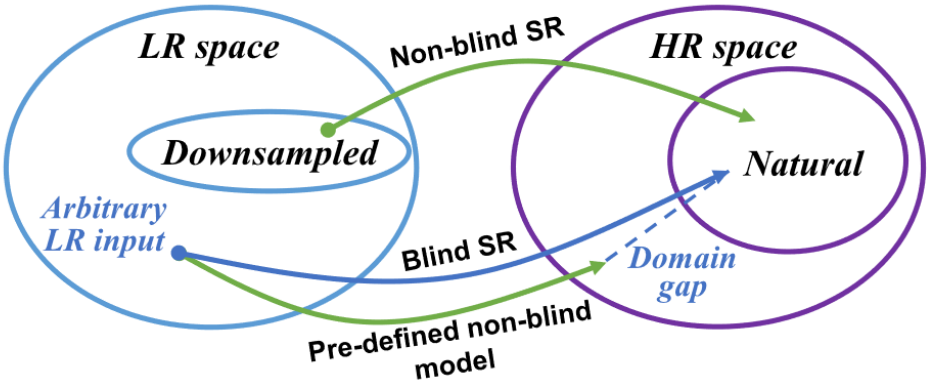


FIGURE 2.7 Illustration du décalage de domaine qu'essaie de palier le Blind SR [7]

De nombreuses méthodes proposent un modèle de dégradation différent. Récemment, [21] a proposé un modèle particulièrement intéressant. Il introduit la conception d'un estimateur probabiliste de dégradation (PDE), qui prédit la dégradation sous forme de distribution plutôt qu'un point fixe, rendant le modèle plus robuste aux erreurs d'estimation. Ces estimateurs probabilistes de dégradation vont prédire les paramètres des distributions du flou gaussien, du bruit et de la compression JPEG. Ces estimateurs sont des réseaux de neurones (Unet [63] et Restnet [48]) qui vont par exemple prédire la moyenne et l'écart type de la distribution. De plus, pour quantifier la différence entre les distributions prédites et réelles, l'article introduit

une nouvelle fonction de perte basée sur IoU (Intersection over Union Loss). L'algorithme pour l'inférence est donné en annexe B.

Ces derniers mois, de nouvelles méthodes innovantes ont émergé et montrent un potentiel prometteur.
Parmi celles-ci, on peut citer l'architecture IPG [8], qui utilise des réseaux de neurones graphes (GNN) pour la super-résolution.Il repose sur une représentation par graphes où chaque pixel est un noeud, plutôt que des patchs, afin d'éviter les problèmes d'alignement. Elle introduit une flexibilité de degré de noeud, assignant des degrés plus élevés aux pixels riches détail (haute fréquence par exemple) contrairement aux approches de graphes traditionnels. Donc pour reconstruire un patch que l'on souhaite reconstruire, on pourra s'appuyer sur un nombre variable et de distance variable de patch autour du patch qu'on veut reconstruire.

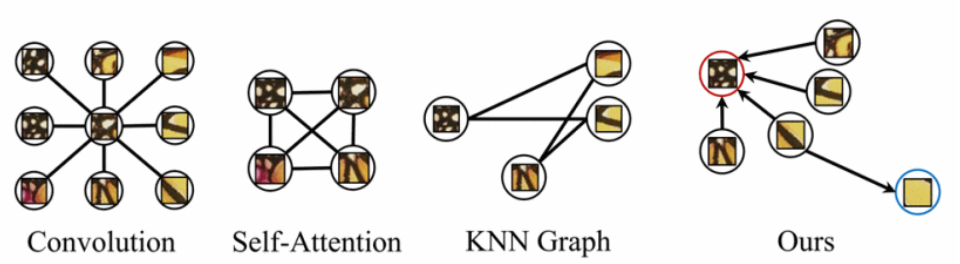


FIGURE 2.8 Visualisation des différentes méthodes d'agrégation pour reconstruire un patch [8]

Les modèles de diffusion commencent aussi à être utilisés en SR [64]. Les méthodes de diffusion sont très utilisées en vision par ordinateur, notamment pour toutes les tâches génératives. Elle consiste à générer une image en inversant un processus de bruitage progressif, où chaque étape d'inférence réduit le bruit ajouté à l'image jusqu'à obtenir une reconstruction nette et détaillée [9].
SinSR [65] par exemple permet de réaliser la super-résolution en un seul pas d'inférence, grâce à un processus de distillation qui convertit le mapping déterministe d'un modèle de diffusion avancé en un modèle rapide et performant. Dans le même style, ResShift [66] propose un modèle de diffusion avec chaîne de Markov pour la SR capable de réduire drastiquement le nombre d'étapes de diffusion à seulement 20, sans sacrifier la qualité des résultats, contrairement aux méthodes actuelles qui nécessitent des centaines d'étapes. Un des problèmes des modèles de diffusion en SR est qu'ils introduisent une part de hasard lors de la génération d'images, ce qui peut entraîner des résultats variables et parfois incohérents, surtout lorsque le nombre d'itérations est limité [67]. Cette variabilité complique l'obtention systématique d'images de haute qualité.

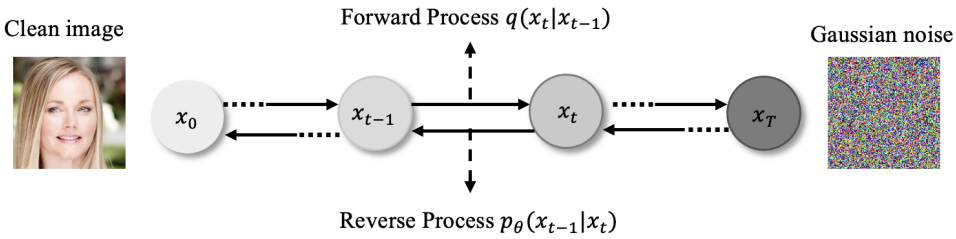


FIGURE 2.9 Principe général des modèles de diffusion par Denoising Diffusion Probabilistic Models (DDPM) [9]

2.2.2 La super-résolution dans le spectre infrarouge

Par rapport aux images RGB, les images IR sont monocanal. Elles présentent de faibles gradients et une "superposition d'informations entre les hautes et basses fréquences" [39]. Pour gérer ces particularités, des architectures spécifiques ont été proposées pour les images IR.
Avant l'avènement de l'apprentissage profond, des solutions basées sur le domaine fréquentiel comme [68] ou des méthodes basées sur des dictionnaires [69] étaient proposées.
Par la suite, inspirés par les techniques utilisées dans le spectre visible, [70] et [71] ont exploité les réseaux CNN et les réseaux résiduels. D'autres architectures ont introduit l'idée d'utiliser les informations du spectre visible (données plus abondantes) pour reconstruire l'image IR. Par exemple, [72] a intégré les informations visibles dans la fonction de perte, tandis que PSRGAN [55] a utilisé un cadre GAN et l'apprentissage par transfert (fine-tuning en anglais) à partir d'images RGB pour entraîner leur algorithme de SR.
Plus récemment, des approches basées sur les transformers ont émergé, telles que LKFormer [73] et DASR [74], qui exploitent conjointement une attention spatiale et une attention au niveau des canaux.
De même que pour le spectre visible, les chercheurs ont essayé des modèles de dégradation pour le spectre IR. Citons SwinIBSR [10], qui propose un modèle relativement simple, mais qui semble générer des images plus réalistes que les méthodes "non blind". Le modèle de dégradation consiste en plusieurs portes de dégradation. Chaque porte représente une forme de dégradation possible, telle que le flou ou le bruit, et elle peut être contournée selon une certaine probabilité.
Le réseau est alors entraîné à partir de ces images basse résolution générées avec ce modèle de dégradation, ce qui lui permet de s'entraîner sur des types de dégradations plus variés, le rendant ainsi plus robuste face à des images basse résolution dans des conditions réelles.Malheureusement, peu de travaux se penchent sur les dégradations spécifiques au spectre infrarouge. De plus, les approches existantes, qui tentent de les modéliser, demeurent

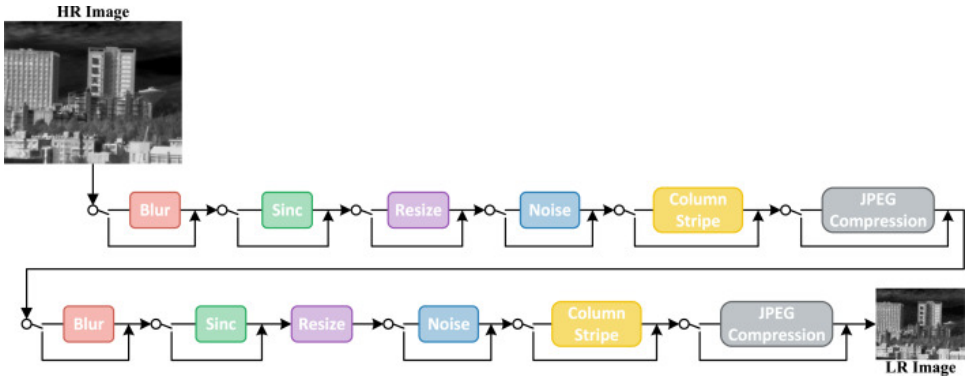


FIGURE 2.10 Processus de dégradation pour générer l'image basse résolution [10]

souvent simplistes et éloignées des résultats réalistes que produirait une véritable caméra basse résolution.

2.2.3 La super-résolution guidée dans le spectre infrarouge

Contrairement aux méthodes présentées ci-dessus, les méthodes guidées utilisent deux images appariées en entrée : une image thermique basse résolution (image cible) et une image de référence de plus haute résolution pour faciliter la tâche de super-résolution (SR). En pratique, l'image de référence est une image RGB, car il s'agit d'une modalité peu coûteuse à acquérir et qui offre une grande richesse en détail dans les hautes fréquences. Ces méthodes dirigées permettent d'obtenir des super-résolutions avec des facteurs multiplicateurs plus élevés que les méthodes classiques, atteignant par exemple un facteur de $\times 8$ ou $\times 16$.
Ces méthodes sont relativement récentes et suscitent un intérêt croissant depuis l'année dernière, notamment grâce à la compétition [75], qui se tient chaque année dans le cadre du worshop PBVS dans le cadre de CVPR. En 2023, une nouvelle catégorie dédiée aux méthodes guidées a été ajoutée à la compétition, ce qui a encouragé de nombreux chercheurs à explorer davantage ces approches.
L'un des premiers travaux sur la super-résolution guidée par image thermique (GTISR) a été introduit dans [76], où un réseau résiduel à double chemin a été utilisé pour fusionner les caractéristiques des domaines visibles et infrarouges. Plus récemment, CoRefusion [77] a été proposé. Son architecture se compose de deux U-Nets [63] avec des connexions résiduelles pour fusionner les deux modalités. Un terme contrastif est ajouté à la fonction de perte, ce qui améliore les performances. La principale limitation de cette architecture réside dans la simplicité de sa conception : les U-Nets extraient des caractéristiques limitées, ce qui restreint les performances de reconstruction. En revanche, les méthodes basées sur des transformers permettent d'extraire des caractéristiques plus profondes, offrant ainsi des performances su-

périeures. C'est pourquoi les approches par transformers sont davantage étudiées que celles reposant uniquement sur des CNN.
Les auteurs de [11] ont présenté comment la transformation de l'image guide RGB en une "image semblable à une image thermique" améliore les performances des méthodes guidées. Ils montrent que cette substitution permet d'augmenter les performances de quelques points de pourcentage dans différentes architectures de super-résolution guidée.

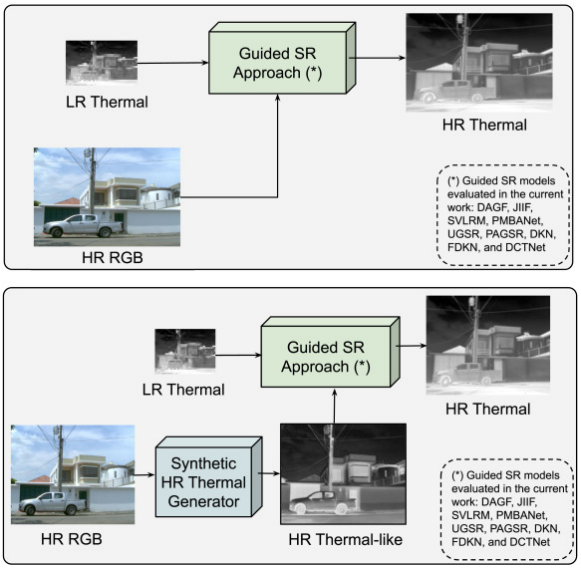


FIGURE 2.11 Illustration des méthodes guidées traditionnelles (en haut) et de la méthode proposée par [11] (en bas)

Cette année 2024 a été très propice au développement de méthodes guidées, notamment grâce au workshop de CVPR intitulé Perception Beyond the Visible Spectrum (PBVS) et à sa compétition, dans laquelle l'article présenté au **Chapitre 5** a obtenu la 3e place en termes de performance parmi 76 équipes [17].
Notons que le gagnant de cette compétition a une solution inédite en utilisant des mélanges d'experts (MoE [78]) et divise la tâche en sous-parties, chaque expert se spécialisant dans une fonction spécifique. Elle utilise des experts pour traiter les images sous différents angles et un autre expert pour intégrer les résultats (voir résultats visuels **Annexe C**). Le principal inconvénient de cette méthode est son grand nombre de paramètres, 600 millions de paramètres, environ 200 fois plus que l'architecture présentée au **Chapitre 5**.
De nombreuses autres architectures de super-résolution guidées ont été proposées lors de cette conférence [79–81]. Ce dernier article [81] adapte des modèles de diffusion au problème de super-résolution d'images thermiques, tout en exploitant leur capacité à estimer l'incertitude aléatoire pour prédire l'erreur par pixel, même en l'absence de vérité-terrain haute résolution.

Le travail de [19] présente une approche de super-résolution guidée sans recours à l'apprentissage automatique. En supposant que l'image infrarouge haute résolution peut être exprimée comme une combinaison linéaire de l'image guide RGB, l'auteur propose un filtre permettant d'exploiter les informations de l'image guide. Cette méthode est particulièrement intéressante car elle se différencie des approches traditionnelles basées sur le machine learning, généralement utilisées pour la SR. Cependant, comme nous le verrons dans le Chapitre 4, cette méthode s'apparente davantage à une approche de fusion de modalités qu'à une véritable méthode de super-résolution, car elle intègre l'image RGB dans l'image infrarouge de manière peu nuancée. Cela met en lumière l'une des principales limites des approches sans apprentissage : leur difficulté à fusionner efficacement les informations provenant du spectre visible avec celles du spectre infrarouge. Cette fusion, en raison de sa grande complexité et de sa nature difficilement modélisable, demeure un défi majeur.

L'utilisation d'une modalité guide provenant du spectre visible pour la super-résolution d'images thermiques s'apparente fondamentalement de la fusion d'images multimodales, notamment RGB et IR. La section suivante fait un survol rapide de ce champs de recherche qui inspirera la méthode de super-résolution proposée dans ce travail.

2.3 La fusion de modalités

La fusion d'images multimodales vise à combiner les informations pertinentes provenant d'images capturées par différents capteurs en une seule image. Les méthodes de fusion basées sur l'apprentissage profond se divisent en trois catégories : la fusion précoce, la fusion tardive et la fusion hybride. La première fusionne les caractéristiques avant les couches de neurones liées à la tâche, la deuxième applique ces couches de neurones sur chaque modalité avant d'agréger les informations, tandis que la dernière combine les deux approches [26, 82].
SwinFusion [12] utilise une fusion hybride pour fusionner des images provenant de deux modalités en utilisant la fusion inter-domaines guidée par l'attention. Inspiré par le bloc Swin Transformers [6], ce modèle fusionne les informations des deux branches de modalités via des modules alternants de « fusion intra-domaine basée sur l'auto-attention » et de « fusion inter-domaine basée sur l'attention croisée ». Cette architecture (voir Figure 2.12) a été une grande source d'inspiration pour l'article présenté au **Chapitre 5**.
Ce modèle a été utilisé pour initier le projet Undercover. L'objectif de celui-ci est d'utiliser l'information IR pour générer une image RGB sans couverture, facilitant ainsi l'estimation de pose (voir détails et résultats en Annexe F).
Dans des applications concrètes, il est essentiel de concevoir des systèmes robustes, capables de maintenir des performances élevées, même en cas de défaillance de certains capteurs. Il

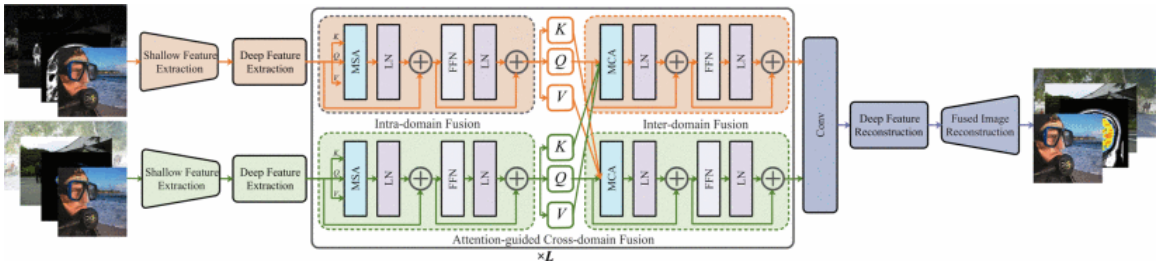


FIGURE 2.12 Architecture pour de la fusion multimodale [12]

est notamment souhaitable qu'en l'absence d'une modalité, le système soit toujours en mesure de produire un résultat satisfaisant à partir d'une seule modalité en entrée, même si la performance peut être légèrement inférieure à celle obtenue avec l'ensemble des modalités disponibles. Cette capacité d'adaptation garantit la fiabilité du système dans des conditions d'acquisition d'information partiellement dégradées.

Peu de travaux ont étudié l'impact d'une modalité manquante sur le système. Dans l'étude [83], les auteurs ont évalué l'impact du type d'architecture, de l'augmentation de données et de la technique de fusion d'images sur les performances de reconnaissance d'actions en cas d'absence d'une modalité. Ils ont conclu que la fusion basée sur les transformers est plus robuste dans ce contexte que la somme ou la concaténation des caractéristiques.

Parallèlement, [84] a examiné l'impact de l'absence d'une modalité (texte, audio ou image) lors de l'entraînement ou des tests d'un modèle GAN ou autoencodeur. Les auteurs ont proposé un cadre de méta-apprentissage bayésien afin de mieux gérer les modalités manquantes.

2.4 Les métriques d'évaluations en restauration d'image

Une fois que l'image haute résolution est générée par le système, nous souhaitons obtenir des métriques quantifiant à quel point elle s'avère réaliste. Il existe alors deux types de métriques.

2.4.1 Métriques avec référence

Les métriques avec référence sont des métriques qui s'appuient sur une image vérité terrain. Le but de ces métriques est de proposer des calculs pour comparer à quel point l'image générée est proche de l'image vérité terrain.

Peak Signal to Noise Ratio

Le Peak Signal to Noise Ratio (PSNR) a été la première métrique utilisée en super-résolution et demeure l'une des plus simples à appliquer. Elle est définie par la formule suivante :

$$PSNR(I_g, I_r) = 10 \cdot \log_{10}\left(\frac{255^2}{EQM}\right)$$

avec EQM, l'erreur quadratique moyenne entre l'image de référence et l'image générée

$$EQM = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}(I_g(i,j) - I_r(i,j))^2$$

Les variables m et n représentent respectivement le nombre de pixels en largeur et en longueur, I_g est l'image générée et I_r l'image de référence.
Plus le PSNR est haut et plus l'image générée est proche de l'image de référence

Structural SIMilarity

La métrique Structural SIMilarity (SSIM) [13] est une des deux métriques les plus utilisées en SR avec PSNR. L'idée est que l'œil humain est plus sensible aux changements de structure qu'aux différences pixel par pixel. SSIM vérifie et essaie de quantifier à quel point des changements structurels existent dans l'image générée par rapport à l'image de référence
Pour des blocs de pixels x,y dans l'image (ou patch en anglais), la formule de SSIM est donnée par :

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

avec μ_x, μ_y les moyennes des pixels des blocs x et y, σ_x, σ_y les écarts-types, σ_{xy} la covariance de x et y et c_1 et c_2 des constantes pour stabiliser la division.
Tous ces calculs sont effectués sur la luminance de l'image et on fait la moyenne sur les différents blocs (en général des blocs de 8×8 pixels).
SSIM est compris entre -1 et 1, plus la valeur absolue est proche de 1 et plus l'image générée est proche de l'image de référence.
Une variante de SSIM nommée MS-SSIM [85] où SSIM est calculé à plusieurs échelles grâce à un sous-échantillonnage à plusieurs niveaux.
La figure 2.13 permet de comparer les valeurs de SSIM selon les différents types de dégradation. À noter que les 5 images dégradées ont le même PSNR, ce qui montre les limitations de cette métrique, car l'image floue semble beaucoup plus dégradée que celle avec le contraste.

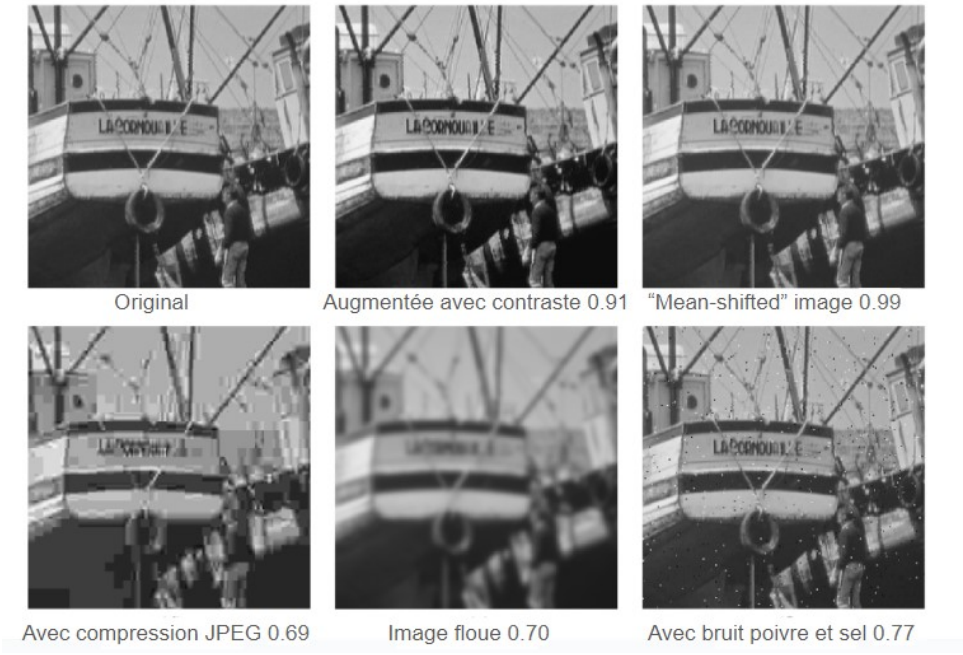


FIGURE 2.13 Effet des différentes dégradations sur les métriques [13]

2.4.2 Métriques perceptuelles

Les métriques perceptuelles évaluent la similarité visuelle en tentant de modéliser la perception humaine. Les méthodes classiques comme le PSNR et le SSIM, bien qu'efficaces, échouent à capturer les nuances de cette perception. Récemment, des caractéristiques profondes issues de réseaux de neurones (comme VGG [86]) se sont révélées bien plus performantes pour mesurer la similarité perceptuelle [14].
Pour une image référence x_0 et l'image générée x, la formule est donnée par :

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})\|_2^2$$

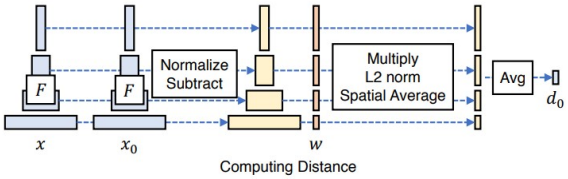


FIGURE 2.14 Schéma de calcul d'une distance perceptuelle [14]

Avec l la couche du réseau, H_l et W_l les dimensions spatiales de la couche, $\hat{y}^l_{hw}$ les caracté-

ristiques profondes de l'image générée à la couche l et $\hat{y}^l_{0hw}$ les caractéristiques profondes de l'image référence à la couche l.
Pour un réseau de neurones donné, cela revient à calculer les caractéristiques profondes, à normaliser les activations par canal, les mettre à l'échelle avec un vecteur w, à prendre la distance L_2 au carré, puis à faire la moyenne sur les dimensions spatiales et à travers toutes les couches.
Les métriques perceptuelles sont rarement utilisées en super-résolution d'images thermiques, car les réseaux de neurones qui les exploitent sont généralement entraînés sur des images issues du spectre visible. Étant donné la différence significative entre le spectre visible et le spectre infrarouge, rien ne garantit que les caractéristiques extraites par ces réseaux soient pertinentes pour les images thermiques.

Métriques sans référence

Dans le cas où nous n'avons pas accès à une image haute résolution avec laquelle comparer l'image générée, les métriques sans référence permettent d'apporter une quantification de la qualité de l'image générée. Elles essaient de prédire automatiquement la qualité des images générées telle qu'elle serait perçue par un être humain moyen.
L'indicateur le plus connu et le plus utilisé est NIQE [87] (Natural Image Quality Evaluator). C'est une métrique qui évalue la qualité des images en mesurant les écarts par rapport aux régularités statistiques des scènes naturelles. Elle utilise des caractéristiques statistiques basées sur un modèle simple de scènes naturelles. Ces caractéristiques sont dérivées d'un ensemble d'images naturelles non dégradées.

2.5 Jeux de données disponibles

L'utilisation de méthodes d'apprentissage requiert des données d'entraînement. Dans le cadre des méthodes de super-résolution, l'idéal est de disposer d'une image basse résolution ainsi que d'une image haute résolution de référence. Si l'image basse résolution n'est pas disponible, il est facile de la générer en appliquant une méthode de sous-échantillonnage ou une dégradation plus complexe, comme celles que l'on rencontre avec les méthodes de super-résolution aveugle (Blind SR).Le problème du sous-échantillonnage réside dans le fait qu'il génère des résultats souvent éloignés de ceux obtenus avec une image directement capturée par une caméra de basse résolution. Pour répondre à cette limitation, les méthodes de Blind Super-Resolution proposent des techniques plus avancées afin de modéliser de manière réaliste la dégradation de l'image. Cependant, par souci de simplicité, les ensembles de données utilisés en super-résolution adoptent généralement une dégradation bicubique, à laquelle on ajoute

parfois un flou pour accroître la difficulté de la tâche. Les chercheurs sont bien conscients des limites de ces modèles de dégradation et des implications que cela peut avoir sur l'application des méthodes en conditions réelles.
Pour les méthodes de super-résolution d'images thermiques guidées, il est nécessaire de disposer de l'image haute résolution infrarouge ainsi que de l'image guide haute résolution RGB. Il existe peu de jeux de données de ce type, car l'acquisition de caméras thermiques capables de produire des images de haute qualité est coûteuse, comme mentionné en introduction. De plus, pour la plupart des algorithmes, il est essentiel que les images des deux modalités soient alignées (registered images), c'est-à-dire qu'il ne doit y avoir ni décalage ni distorsion spatiale entre les deux images.

2.5.1 Flir dataset

Le Teledyne FLIR ADAS Dataset [88] propose 26,442 paires d'images thermiques et visibles, annotées et alignées, idéales pour le développement d'algorithmes de super-résolution (SR) guidée. Il contient plus de 520,000 annotations réparties sur 15 catégories d'objets, comme les piétons et les véhicules. Les images thermiques, capturées en 640×512 avec une caméra FLIR Tau 2, sont complétées par des images visibles obtenues avec une caméra FLIR Blackfly S, permettant l'entraînement et la validation des modèles SR sur des données multi-modales (RGB et thermique). Ces images ont été acquises depuis le tableau de bord d'une voiture et visent à encourager la recherche sur la fusion d'informations thermiques et RGB, ainsi que sur les systèmes d'assistance à la conduite (ADAS).

FIGURE 2.15 Exemple d'images RGB et infrarouge issue du jeu de donnée Flir

2.5.2 SLP dataset

Le SLP dataset (Simultaneously-collected multimodal Lying Pose) [15, 16] est un ensemble de données comprenant des images de pose de personnes allongées, capturées avec plusieurs modalités : RGB, infrarouge, profondeur et carte de pression. Pour chaque image, une annotation d'estimation de pose sous forme de squelette est associée.
De plus, diverses conditions de couverture sont également présentes, telles qu'une couverture fine, épaisse ou l'absence de couverture. Ces conditions permettent une meilleure utilisation des images thermiques, qui révèlent les membres sous les couvertures. Le jeu de données comprend 109 sujets adoptant différentes poses, pour un total de près de 15 000 images. Il s'agit du jeu de données offrant les images les plus représentatives des situations réelles en milieu hospitalier.
SLP gère l'alignement entre différentes modalités d'imagerie en utilisant des repères partagés, qui sont visibles dans toutes les modalités. Les marqueurs utilisés pour l'alignement dans SLP sont conçus pour être détectés simultanément dans toutes les modalités. Chaque marqueur est constitué d'un bocal cylindrique facilement repérable en imagerie RGB. Sur le dessus du bocal, une plaque thermique est ajoutée pour générer une signature en imagerie thermique (LWIR), alimentée par des batteries internes. Le bocal est également lesté pour augmenter le profil de pression, visible dans la carte de pression (PM), et sa hauteur d'environ 10 cm modifie la distance perçue en imagerie de profondeur (D). Ces marqueurs permettent d'estimer l'homographie nécessaire au recalage des différentes modalités.

2.5.3 Jeu de donnée PBVS

Dans le cadre de la compétition PBVS [17], les organisateurs ont mis à disposition un nouveau jeu de données multispectrales. Il offre aux chercheurs en super-résolution guidée un ensemble de données leur permettant de comparer leurs méthodes, devenant ainsi le jeu de données de référence actuel dans ce domaine.
Il se compose de 700 échantillons d'entraînement et de 200 échantillons de validation, chaque échantillon étant une image IR de taille 640×448, ainsi que sa version sous-échantillonnée par un facteur de 8 (et 16) et son image RGB 640×448 appariée. Les 100 échantillons de test sont fournis sans les vérités de terrain haute résolution. Ces images enregistrées ont été acquises par des caméras Balser (pour le RGB) et Flir TAU2 (pour l'IR) et représentent des images de scènes urbaines extérieures.

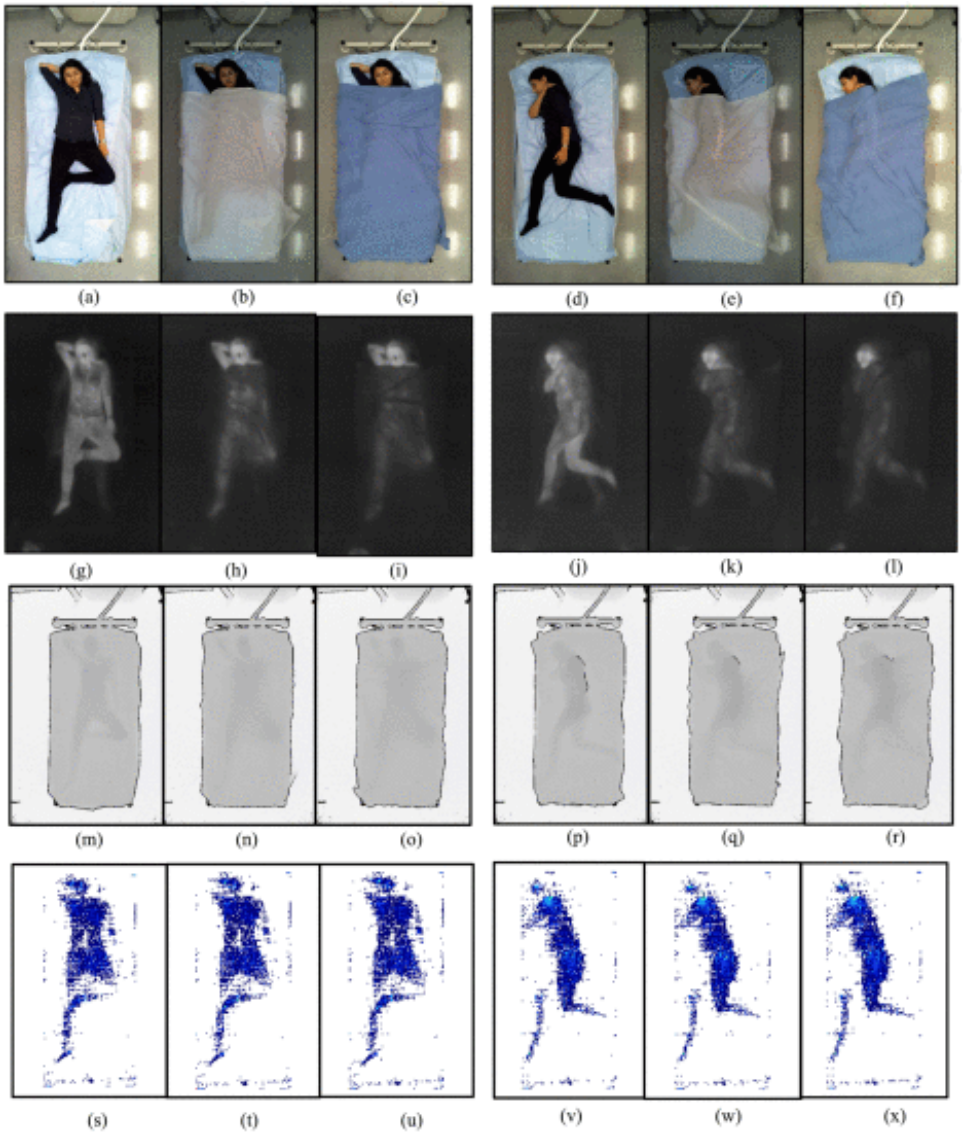


FIGURE 2.16 Images de SLP [15, 16]. (a-f) RGB, (g-l) infrarouge, (m-r) profondeur, (s-x) carte de pression

2.5.4 Jeu de donnée LWIRpose

Le dataset LWIRpose [18] est un ensemble d'images thermiques (LWIR) et RGB appariées, comprenant plus de 2 400 images annotées avec des poses humaines en 2D. Capturé à partir de sept acteurs effectuant des activités quotidiennes variées (marcher, s'asseoir, manger, etc.), il est conçu pour traiter la tâche d'estimation de pose humaine sur des images thermiques, telles que l'occlusion, la variabilité des formes corporelles et les différences de vêtements.
Ce dataset est utile car il utilise des images thermiques de moyenne résolution (640×480) et fournit les images RGB haute résolution associées, ce qui permet de réaliser de la SR guidée.

FIGURE 2.17 Couples d'image RGB et infrarouge issues du dataset PBVS [17]

En raison de sa sortie récente, ce dataset n'a pas encore pu être utilisé, mais il présente un réel potentiel pour démontrer l'impact de la super-résolution sur des tâches annexes, comme l'estimation de pose humaine.

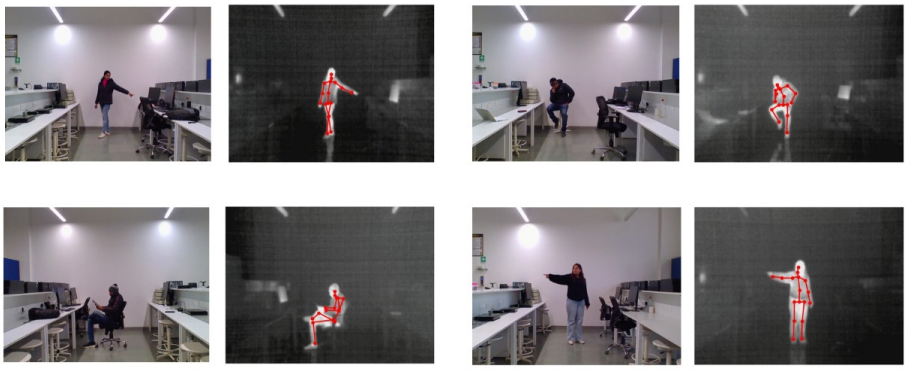


FIGURE 2.18 Exemples d'images et d'annotations du jeu de données LWIRpose [18]

CHAPITRE 3 OBJECTIFS DE RECHERCHE

Dans l'introduction, nous avons mis en évidence l'importance de l'information thermique dans le contexte médical. Nous avons également souligné la problématique liée à la faible résolution des images thermiques. La revue de littérature a montré que la super-résolution constitue une solution prometteuse pour améliorer la qualité des images thermiques. Nous avons examiné plusieurs approches de super-résolution, en particulier celles adaptées aux images infrarouges. En outre, nous avons souligné que les méthodes guidées ont suscité un intérêt croissant ces derniers mois, en raison de leur capacité à améliorer significativement les performances de la super-résolution dans le domaine de l'infrarouge. Cela renforce l'idée que les méthodes guidées s'avèrent un axe prometteur à explorer pour avoir une méthode des plus performantes. Exploiter l'information RGB et infrarouge nécessite une approche multi-modale. Pour garantir la robustesse de notre système, il est essentiel qu'il fonctionne même en l'absence d'informations RGB. Ce cas de figure peut survenir en raison d'une défaillance de la caméra RGB, d'un mauvais recalage entre les images des deux modalités, ou encore dans des conditions d'obscurité totale. Ces scénarios, potentiellement fréquents en milieu critique, comme dans une chambre de soins intensifs pédiatriques, soulignent la nécessité de développer une approche robuste et résiliente, capable de maintenir des performances acceptables même en l'absence d'informations RGB.

Cela nous conduit à formuler l'objectif principal de cette recherche :

Objectif principal du projet de maîtrise :
Proposer et évaluer une solution de super-résolution guidée robuste pour des images thermiques dans le contexte des soins intensifs pédiatriques.

Cela nous amène à poser les sous-objectifs de recherche suivants :

O1 : Implémenter et comparer différentes méthodes existantes de super-résolution pour les images infrarouges et identifier les forces et faiblesses de ces méthodes.

O2 : Concevoir et évaluer une nouvelle architecture de réseau de neurones pour la super-résolution d'images infrarouges robuste guidée par l'imagerie dans le visible.

O3 : Évaluer la généralisation du modèle proposé en (**O2**) sur des données en conditions

réelles acquises dans l'unité de soins intensifs pédiatriques du CHU Sainte-Justine.

Dans le **Chapitre 4**, nous répondrons à **O1** en évaluant différentes méthodes de super-résolution sur les jeux de données présentés dans la revue de littérature.

Dans le **Chapitre 5**, nous présentons l'article publié lors du workshop PBVS de CVPR 2024 qui propose une nouvelle architecture de super-résolution d'images thermiques guidée. Ceci répond à **O2**. La méthode est évaluée quantitativement sur des jeux de données publiques, notamment sur les données d'un challenge proposé au workshop PBVS et pour lequel notre méthode s'est classée 3e sur plus de 76 solutions proposées. L'architecture proposée est à la fois légère et robuste, même en l'absence de la modalité guide.

Enfin, pour répondre à l'objectif **O2**, nous présenterons dans le **Chapitre 6** les résultats obtenus avec l'architecture proposée sur des données acquises au CHUSJ en conditions réelles.

CHAPITRE 4 ÉVALUATIONS PRÉLIMINAIRES DE MÉTHODES DE SUPER-RÉSOLUTION EXISTANTES

Dans ce premier chapitre, nous évaluerons différentes méthodes de super-résolution (SR) afin d'identifier les principaux défis liés à la SR d'images thermiques et d'examiner l'efficacité des solutions existantes. Cette analyse nous permettra de mieux comprendre les limites actuelles des techniques disponibles et d'orienter les recherches futures vers des approches plus adaptées aux spécificités des images thermiques.
Pour afficher les images de résolutions différentes et calculer des métriques, l'image sous-échantillonnée (SE) sera mise à la même résolution que les images super-résolues avec une interpolation bicubique.

4.1 Super-résolution d'images thermiques du jeu de données FLIR avec PSRGAN

Comme nous l'avons vu dans le **Chapitre 2**, une des premières architectures proposées pour la super-résolution d'images thermiques est PSRGAN. Cette architecture permet de faire de la super-résolution avec des facteurs de 2 ($\times 2$) et de 4 ($\times 4$).

Nous avons testé PSRGAN sur des images du dataset FLIR. À partir d'une image pleine résolution (résolution de 640×512) considérée comme référence 4.1b, nous la sous-échantillonnons d'un facteur 2 (donc 4 fois moins de pixels) pour obtenir l'image sous-échantillonnée 4.1c. Nous utilisons alors PSRGAN afin d'effectuer la SR d'un facteur 2 avec, en entrée, l'image sous-échantillonnée. Nous obtenons alors l'image super-résolue 4.1a. Les métriques de qualité PSNR et SSIM sont alors calculées entre l'image super-résolue et l'image de référence.

(a) Image après SR (b) Image Référence (c) Image SE

FIGURE 4.1 Résultat d'une SR $\times 2$ avec PSRGAN, sur une image du dataset FLIR. Les résultats quantitatifs correspondants sont les suivants : PSNR : 32.73 db SSIM : 0.7429

Nous remarquons que même si les métriques montrent que les images sont différentes (PSNR différent de l'infini et SSIM différent de 1), les trois images semblent identiques à l'œil nu. Cela est dû au fait que les différences sont imperceptibles lorsque les images sont affichées à petite échelle, surtout pour des facteurs de super-résolution faibles ($\times 2$ et $\times 4$). Nous pouvons toutefois afficher une carte de chaleur de l'erreur de pixels en valeur absolue entre l'image de référence et l'image après SR. Pour l'exemple de l'image à la Figure 4.1a, la carte de chaleur de l'erreur semble uniforme et aucune zone ne semble avoir plus d'erreurs que d'autres (voir Figure 4.2).

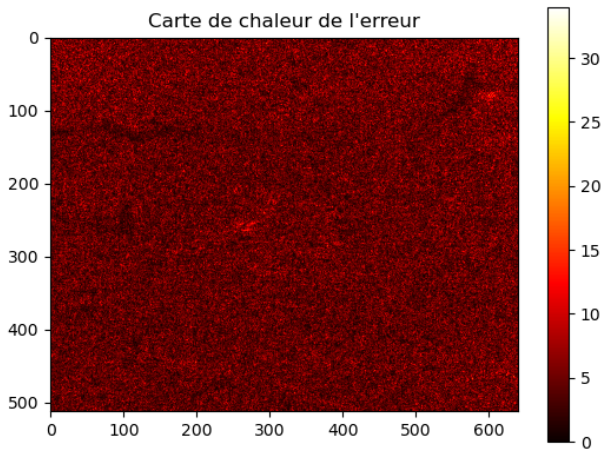
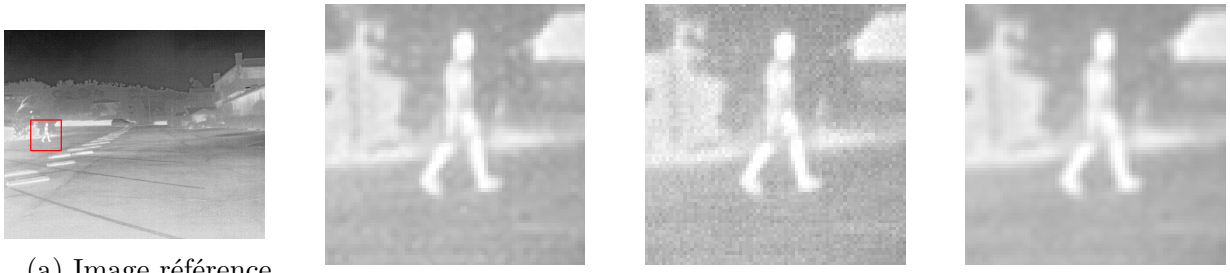


FIGURE 4.2 Carte de chaleur de l'erreur de reconstruction entre l'image après SR 4.1a et l'image de référence de la Figure 4.1.

Pour observer des différences visuelles pour ces bas facteurs de résolution, il faut zoomer sur des sections de l'image. Lorsque nous zoomons sur une personne présente dans l'image (voir Figure 4.3a), nous remarquons que l'image de référence 4.3c semble initialement bruitée, ce qui est souvent le cas pour les images thermiques observées de près. On peut également remarquer que l'image super-résolue 4.3b est plus nette que l'image SE 4.3d, ce qui démontre l'efficacité de cette méthode.



(a) Image référence (b) Section après SR (c) Section référence (d) Section SE

FIGURE 4.3 Résultat d'une SR ×2 avec PSRGAN, sur une image du dataset FLIR avec zoom sur une région locale.

Étudions, par la suite, une super-résolution de facteur 4 sur la même image du jeu de données FLIR (voir Figure 4.4b). Pour ce facteur de SR, on peut plus aisément remarquer les différences visuelles. En effet, l'image SE 4.4c et l'image super-résolue 4.4a sont plus floues que l'image de référence 4.4b. Les métriques PSNR et SSIM sont plus basses que dans l'exemple de la Figure 4.1 avec un facteur de 2, ce qui confirme ce qu'on peut constater visuellement.

(a) Image après SR (b) Image Référence (c) Image SE

FIGURE 4.4 Résultat d'une SR ×4 avec PSRGAN, sur une image du dataset FLIR. Les résultats quantitatifs correspondants sont les suivants : PSNR : 31.27 db SSIM : 0.6430

Intéressons-nous maintenant à la carte de chaleur de l'erreur sur l'exemple de la Figure 4.4a en comparaison avec une visualisation des hautes fréquences résultant d'un filtrage Laplacien de l'image de référence. Nous remarquons tout d'abord que l'ordre de grandeur de l'erreur a augmenté, atteignant localement une valeur de 60. De plus, contrairement à l'exemple précédent, certaines zones très spécifiques de l'image présentent des erreurs de reconstruction plus élevées. Il s'agit principalement des zones contenant des hautes fréquences (Figure 4.5). En effet, lorsque nous perdons des pixels dans les zones de hautes fréquences lors du sous-échantillonnage, il devient plus difficile de les reconstruire, car ces pixels diffèrent de ceux qui les entourent. Il est donc plus compliqué d'utiliser l'information autour pour les reconstruire.

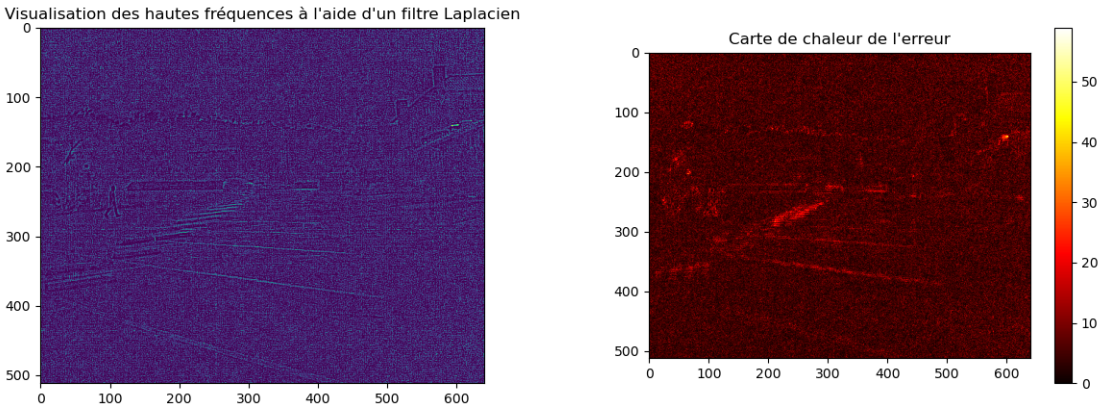
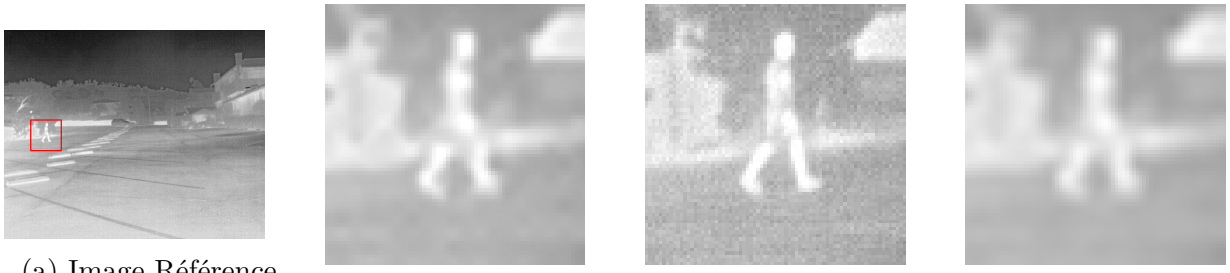


FIGURE 4.5 Comparaison entre l'erreur de reconstruction et les hautes fréquences de l'image de référence illustrée à la Figure 4.4b

Cette observation est l'une des raisons pour lesquelles l'utilisation d'une image guide semble pertinente pour la tâche de super-résolution. Les images RGB en haute résolution contiennent des hautes fréquences qui peuvent aider à reconstruire ces zones dans les images infrarouges associées. On remarque que, par rapport à la SR avec un facteur de 2, la reconstruction de



(a) Image Référence (b) Après SR (c) Image Référence (d) Image SE

FIGURE 4.6 Résultat d'une SR ×4 avec PSRGAN, sur une image du dataset FLIR avec zoom sur une région locale.

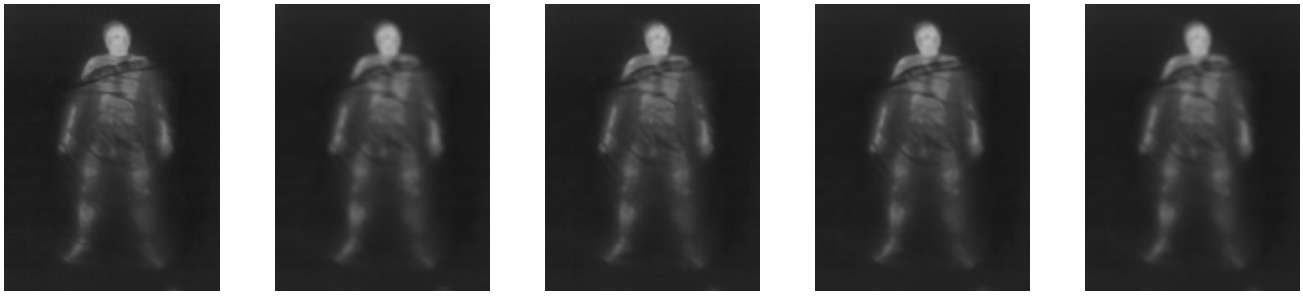
l'image avec un facteur de 4 (voir Figure 4.6b) est moins performante et l'image apparaît plus floue. Cependant, la super-résolution semble tout de même améliorer visuellement l'image par rapport à l'image SE initiale (voir Figure 4.6d).

Conclusion : Dans cette première expérience, nous avons appliqué le modèle PSRGAN pour la SR d'images infrarouge du jeu de données FLIR. Nous avons constaté que les zones les plus difficiles à reconstruire sont celles contenant des hautes fréquences. Toutefois, même pour des facteurs de SR allant jusqu'à 4, cette méthode permet d'ajouter des détails par rapport à l'image dégradée.

4.2 Comparaison de différentes méthodes de SR sur le jeu de données SLP

Le jeu de données SLP étant le plus proche de l'application finale du projet, une comparaison de différentes méthodes de SR a été effectuée sur ce jeu de données. Pour cette comparaison, nous avons sélectionné SwinIR [59], architecture de référence en SR dans le visible utilisée par de nombreux benchmarks, DASR [74] l'une des dernières architectures de SR pour le spectre IR et PSRGAN [55] l'une des premières architectures pour la SR dans l'IR et considérée comme référence dans le domaine. Pour les expérimentations, les poids du réseau fournis par les auteurs ont été utilisés tels quels, et aucun fine-tuning n'a été appliqué.

Le défi du jeu de données SLP est la taille originale très faible des images thermiques (120×160 pixels). Si on veut pouvoir calculer des métriques pour quantifier l'effet de la SR, on doit sous-échantillonner des images originellement de basse résolution, ce qui peut limiter la portée des résultats. Nous avons toutefois réalisé cette expérimentation sur des images SLP sous-échantillonnées d'un facteur 2. Les résultats sur une image en particulier sont présentés à la Figure 4.7. Les métriques PSNR et SSIM moyennes sur un ensemble de 45 images du jeu SLP sont rapportées au Tableau 4.1.



(a) Référence (b) SE /2 (c) PSRGAN (d) SwinIR (e) DASR

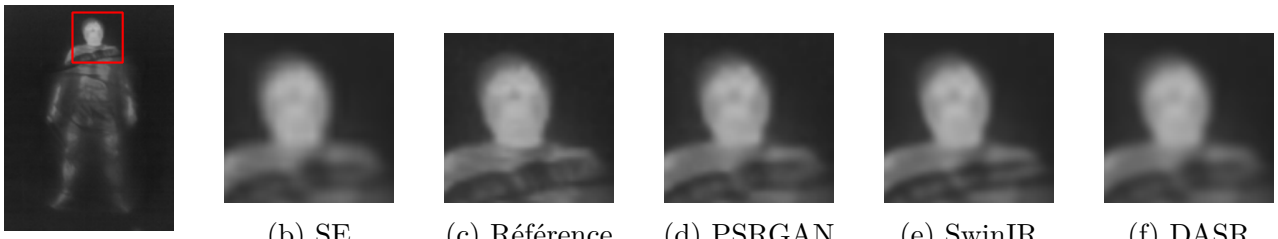
PSNR	43.28	45.53	**47.04**	44.71
SSIM	0.984	0.985	**0.987**	0.983

FIGURE 4.7 Comparaison d'une SR ×2 avec trois méthodes (PSRGAN, SwinIR et DASR) sur une image du jeu SLP sous-échantillonnée (SE/2).

TABLEAU 4.1 Comparaison des métriques PSNR et SSIM pour PSRGAN, SwinIR et DASR

Sur 45 images	SE /2	PSRGAN	SwinIR	DASR
PSNR	44.63	45.54	**46.92**	44.63
SSIM	0.984	0.984	**0.987**	0.963

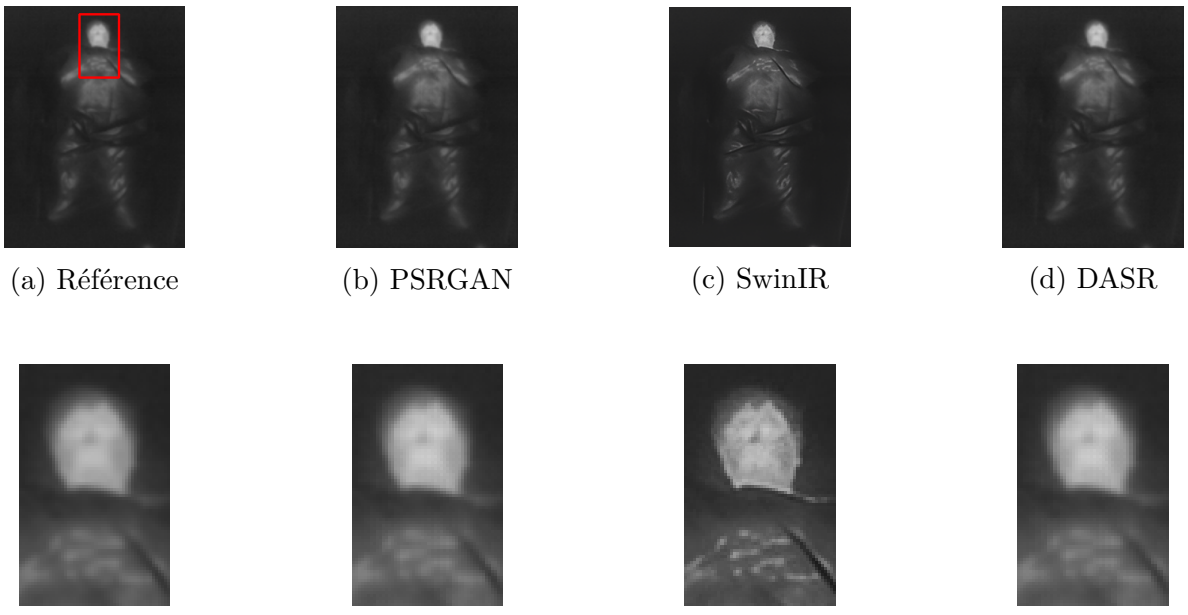
Visuellement, on peut voir que l'image SE (Figure 4.7b) est plus floue que l'image de référence (Figure 4.7a). La SR par DASR (Figure 4.7e) semble aussi floue que l'image SE d'entrée. SwinIR (Figure 4.7d) apparaît qualitativement et quantitativement comme la solution la plus adaptée pour améliorer l'image.



(a) Référence (b) SE (c) Référence (d) PSRGAN (e) SwinIR (f) DASR

FIGURE 4.8 Zoom sur une SR ×2 avec différentes architectures

La Figure 4.8 offre une vue plus locale de l'image afin d'apprécier plus en détail les différences. Nous pouvons constater que seules PSRGAN et SwinIR semblent améliorer l'image SE. En effet, la tête est moins floue avec ces deux solutions et les détails sont plus nets comparés à l'image SE 4.8b. Regardons maintenant les résultats qualitatifs de ces trois méthodes de SR sur l'image de référence sans la sous-échantillonner. Il n'y aura donc pas de métriques calculables car aucune image de référence ne sera disponible.



(a) Référence (b) PSRGAN (c) SwinIR (d) DASR

(e) Référence (f) PSRGAN (g) SwinIR (h) DASR

FIGURE 4.9 Résultat d'une SR ×2 sans référence sur SLP

Conclusion : Les résultats de la SR avec et sans référence sont assez clairs. SwinIR est la seule architecture qui semble améliorer l'image en y ajoutant des détails. Ce qui est assez surprenant, c'est que parmi les trois architectures, elle est la seule à ne pas avoir été entraînée sur des images thermiques. Cela démontre l'efficacité de cette architecture, source d'inspiration pour la méthode que nous proposerons dans le **Chapitre 5**.

4.3 Méthode de SR guidée, sans apprentissage, sur le jeu de données SLP

Comme vu en introduction, une méthode de super-résolution d'images infrarouges s'appuyant sur la modalité RGB pour guider la super-résolution a été proposée par [19]. Ils se sont inspirés d'un article de 2013 [89] pour l'adapter au domaine de l'infrarouge. Son originalité par rapport aux autres méthodes vient du fait qu'elle n'utilise ni réseau de neurones, ni apprentissage. Cette méthode fait deux hypothèses. La première est qu'un pixel i de l'image super-résolue $\hat{X}$ est une combinaison linaire des pixels de l'image guide G,

$$\hat{X}_i = a_k G_i + b_k, \forall i \in \omega_k, \tag{4.1}$$

où (a_k, b_k) sont des coefficients linéaires qui sont constants dans la fenêtre ω_k. La deuxième hypothèse est que l'image super-résolue $\hat{X}$ est égale à l'image approximative $\check{X}$ (générée avec une interpolation en partant de l'image basse résolution $\tilde{X}$) à laquelle on a soustrait certaines distorsions η comme du bruit :

$$\hat{X}_i = \check{X}_i - \eta_i. \tag{4.2}$$

Une fois ces deux hypothèses faites, si on veut que le bruit η_i 4.2 soit le plus faible possible tout en respectant l'hypothèse de linéarité 4.1, le problème peut se formuler sous cette forme :

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left((a_k G_i + b_k - \check{X}_i)^2 + \epsilon a_k^2 \right) \tag{4.3}$$

où ϵ est un paramètre de régularisation qui "contrôle l'influence des grands a_k", et $E(a_k, b_k)$ est le problème à minimiser. Ce problème d'optimisation est connu sous le nom de "linear ridge regression" [90, 91], et pour une fenêtre w_k et un régulariseur ϵ donnés, une formulation de a_k et b_k existe. Le schéma général de la méthode est présenté à la Figure 4.10.

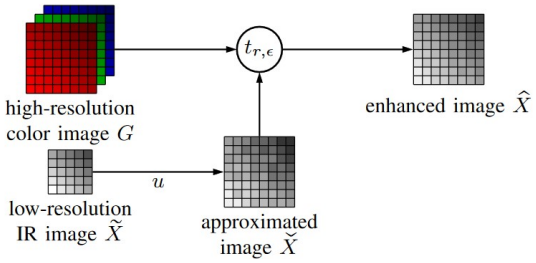


FIGURE 4.10 Schéma du fonctionnement de la méthode de [19]

La première étape pour utiliser des méthodes guidées est de s'assurer que les images soient recalées, c'est-à-dire correctement alignées spatialement. Dans le cas du jeu de données SLP,

les homographies entre images RGB et IR sont fournies. Un exemple illustratif est présenté à la Figure 4.11.

(a) IR alignée spatialement (b) Image RGB (c) IR non alignée spatialement

FIGURE 4.11 Exemple d'image du jeu de données SLP avant et après recalage de la modalité IR sur la modalité RGB

Notons qu'avec la méthode présentée dans la figure 4.12, l'image RGB doit être transformée en image en noir et blanc, car le filtre accepte en entrée des images ayant le même nombre de canaux.

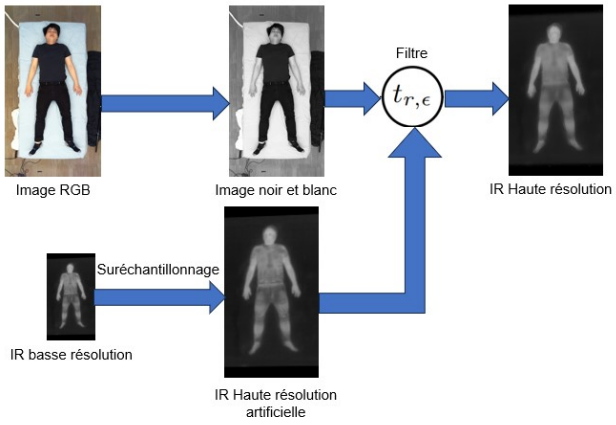
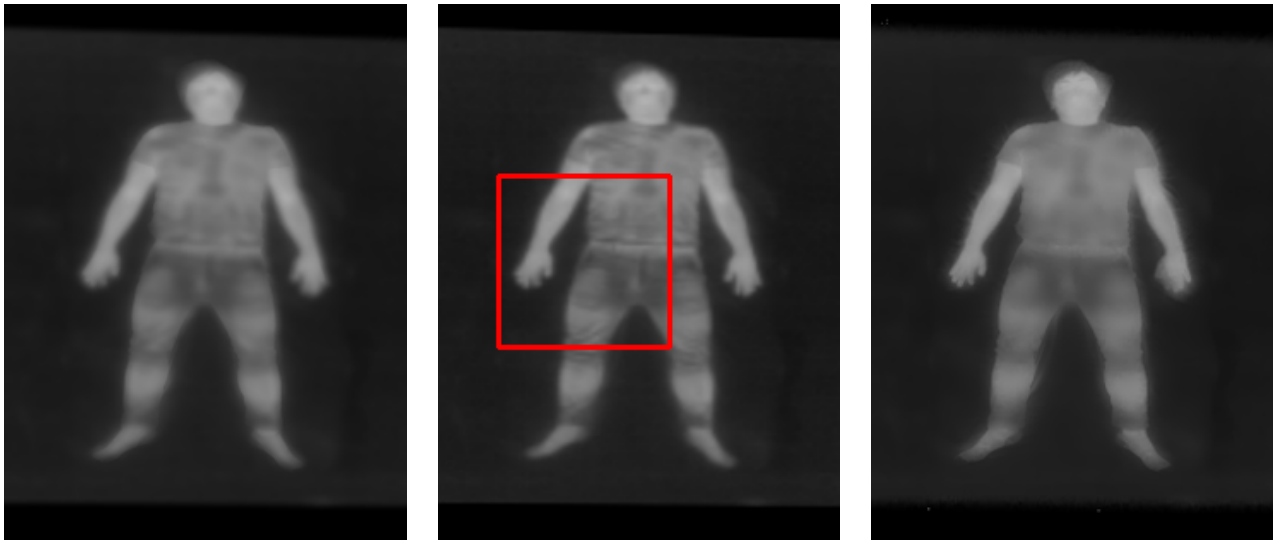
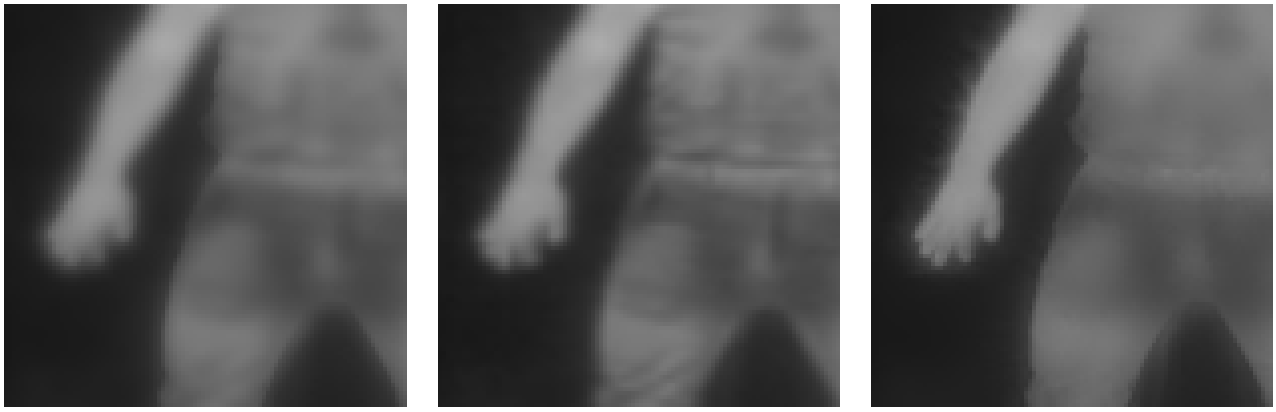


FIGURE 4.12 Méthode [19] sur le jeu de données SLP

Comme nous l'avons vu dans la description de la méthode, le filtre prend en entrée, deux paramètres, le rayon de la fenêtre r et le paramètre de régularisation ϵ. Dans l'article, ils proposent les valeurs $r = 6, \epsilon = 0.0001$ comme valeurs de référence. Regardons les résultats visuels de cette méthode.



(a) $r = 2, \epsilon = 0.0001$ (b) Image originale (c) $r = 6, \epsilon = 0.0001$

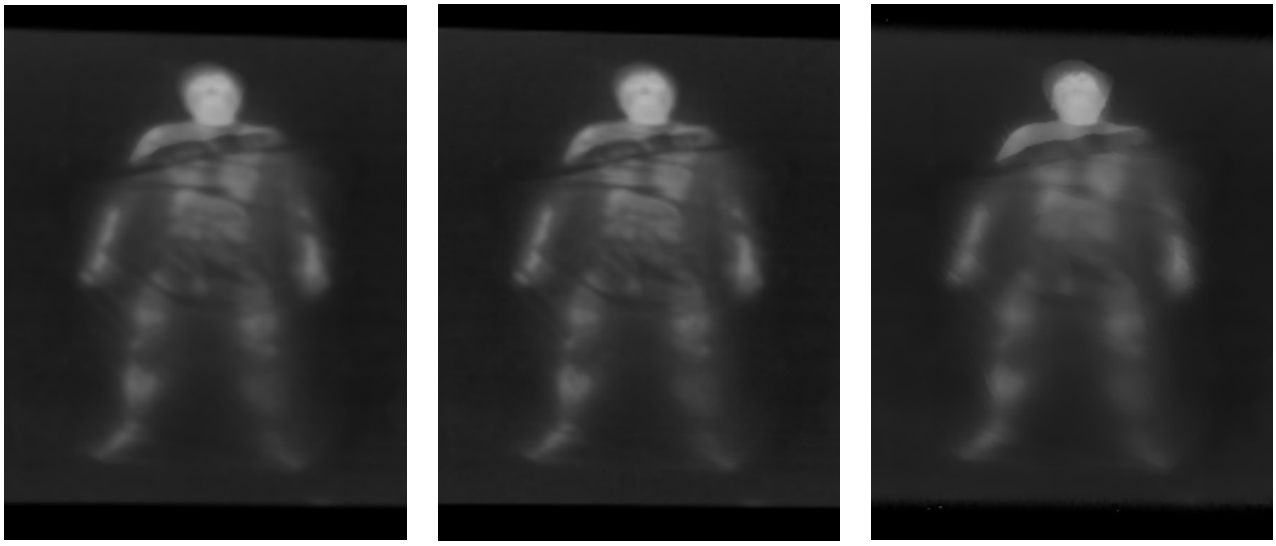


(d) $r = 2, \epsilon = 0.0001$ (e) Image originale (f) $r = 6, \epsilon = 0.0001$

FIGURE 4.13 Exemples de SR guidée [19] sur SLP

On observe que les hautes fréquences, notamment les contours de la silhouette de la personne (voir Figure 4.13f), apparaissent plus nettement. Les détails de texture, pratiquement absents dans le spectre infrarouge, commencent également à se manifester, en particulier au niveau des cheveux (voir Figure 4.13c). Cependant, on peut remarquer de légers artefacts au niveau des contours.

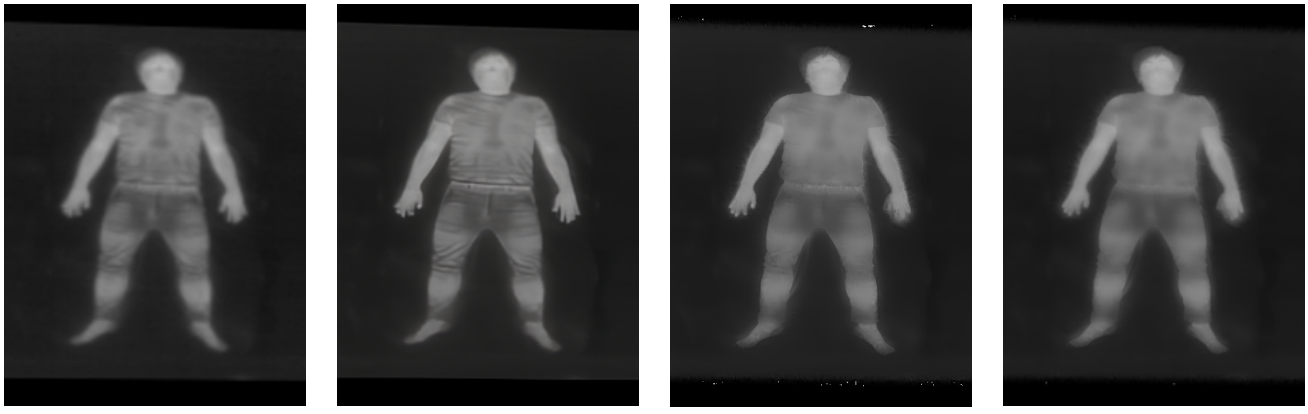
Considérons un cas où une couverture est présente (voir Figure 4.14), limitant l'apport d'informations en RGB pour ajouter des détails, car elle ne permet pas de percevoir ce qui se trouve en dessous.



(a) $r = 2, \epsilon = 0.0050$ (b) Image originale (c) $r = 6, \epsilon = 0.0001$

FIGURE 4.14 Exemples de SR guidée [19] sur des personnes sous des couvertures

Comme précédemment, les zones au-dessus de la couverture présentent des contours et des hautes fréquences plus nets. En revanche, les zones sous la couverture apparaissent plus floues. En effet, la méthode ajoute du bruit à cette partie, car les informations RGB n'apportent aucun détail supplémentaire.
Nous avons essayé d'utiliser une méthode de SR à la place du suréchantillonnage (voir le schéma à la Figure 4.10). Voyons les résultats si on utilise SwinIR :



(a) Image originale (b) SR avec SwinIR (c) [19] avec SwinIR (d) [19] sans SwinIR

FIGURE 4.15 Exemples de SR guidée [19] en utilisant SwinIR

On constate que la méthode utilisant SwinIR (voir Figure 4.15c) permet de mieux rendre

les textures des vêtements par rapport à la méthode sans SwinIR (voir Figure 4.15d). De plus, l'application de SwinIR améliore les contours de l'image super-résolue, en particulier les détails des doigts, qui apparaissent beaucoup plus nets.

Conclusion : Cette partie présente une première approche guidée, utilisant l'information RGB pour améliorer la qualité de l'image infrarouge. Elle permet d'affiner les hautes fréquences de l'image IR lorsque celles-ci sont présentes dans les données RGB, mais introduit du bruit dans les zones où ces informations RGB font défaut. La vitesse de traitement de cette méthode est de $1,25$ secondes par couple d'images IR/RGB.
L'inconvénient principal de cette méthode réside dans son incapacité à déterminer automatiquement la pertinence de l'information RGB, contrairement aux méthodes d'apprentissage automatique qui offrent davantage d'adaptabilité. De plus, bien que l'image soit visuellement améliorée, rien ne garantit que l'image générée se rapproche fidèlement de ce qu'aurait été une image infrarouge en haute résolution. Cette approche s'apparente davantage à une méthode de fusion de modalités qu'à une véritable technique de super-résolution, car elle repose principalement sur l'intégration d'informations RGB sans véritable reconstruction des détails infrarouges manquants.

CHAPITRE 5 ARTICLE 1 SWINFUSR : AN IMAGE FUSION-INSPIRED MODEL FOR RGB-GUIDED THERMAL IMAGE SUPER-RESOLUTION

Cet article [27] a été accepté et présenté en juin 2024 au 20th IEEE Workshop on Perception Beyond the Visible Spectrum, un workshop de la conférence CVPR. La méthode proposée s'est classée à la 3e place en termes de performance parmi 76 équipes participantes au challenge de super-résolution guidée [17].
Authors :

— *Cyprien Arnold*
Polytechnique Montreal
Montreal, Canada
cyprien.arnold@polymtl.ca

— *Philippe Jouvet*
CHU Sainte Justine
Montréal, Canada
philippe.jouvet.med@ssss.gouv.qc.ca

— *Lama Seoud*
Polytechnique Montreal
Montreal, Canada
lama.seoud@polymtl.ca

5.1 Abstract

Thermal imaging plays a crucial role in various applications, but the inherent low resolution of commonly available infrared (IR) cameras limits its effectiveness. Conventional super-resolution (SR) methods often struggle with thermal images due to their lack of high-frequency details. Guided SR leverages information from a high-resolution image, typically in the visible spectrum, to enhance the reconstruction of a high-res IR image from the low-res input. Inspired by SwinFusion, we propose SwinFuSR, a guided SR architecture based on Swin transformers. In real world scenarios, however, the guiding modality (e.g. RGB image) may be missing, so we propose a training method that improves the robustness of the model in this case. Our method has few parameters and outperforms state of the art

models in terms of Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM). In Track 2 of the PBVS 2024 Thermal Image Super-Resolution Challenge, it achieves 3rd place in the PSNR metric. Our code and pretained weights are available at https ://github.com/VisionICLab/SwinFuSR.

5.2 Introduction

Improving the quality of digital images is crucial in numerous fields, from mobile photography [92] and healthcare [33, 93, 94] to law enforcement [34]. Super-resolution (SR) has emerged as a promising technique to achieve this goal, allowing the reconstruction of high-resolution (HR) images from their low-resolution (LR) counterparts. In the realm of RGB images, SR has witnessed significant advancements in recent years. Diverse techniques, ranging from traditional methods to deep learning, have been developed to exploit information within LR images and generate realistic detailed HR images.

Infrared (IR) images, capturing the heat emitted by objects, enable night vision and the ability to detect features invisible to the naked eye. The IR modality is used for continuous and contactless monitoring of patients' vital signs in the intensive care units (ICU) [22,23] and to integrate this information into clinical decision support systems [95]. To achieve this, IR acquisitions are often combined with RGB images, and even 3D images [96]. High definition IR sensors with spatial resolution of up to 1,024 $\times$ 768 pixels are commercially available, but can cost tens of thousands of dollars. Hence, lower resolution IR sensors tend to be used instead in ICU rooms.

Thermal image super-resolution (TISR) tackles this challenge by increasing image resolution and revealing details obscured in the LR image. This topic is increasingly studied because of its many applications [39] including in medical science [2,97], agricultural management [98,99] or even space studies [36, 37]. Several challenges remain in fully realizing the potential of IR super-resolution. One key challenge lies in the inherent differences between IR and RGB images. IR images exhibit higher noise and poorer texture information [39], making HR reconstruction more complex.

Guided thermal image super resolution (GTISR) presents itself as a particularly promising approach for IR image reconstruction. By relying on an HR reference image as input, such as a corresponding visible spectrum image, guided SR can improve the accuracy and consistency of the reconstruction. In effect, HR RGB images are cost-effective to obtain and have higher frequencies than IR images. To encourage researchers to innovate in this little-explored field, the 19th IEEE Workshop on Perception Beyond the Visible Spectrum introduced a challenge

track [75] in 2023 to generate x8 super-resolution thermal images by using visible HR images as guidance. Candidates are ranked according to Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) between images produced from the test set and the (non-public) ground truth HR IR images.

In this paper, we draw inspiration from multimodal image fusion based on Swin Transformers to propose SwinFuSR, a novel method for RGB-guided thermal image super-resolution. Our contributions are two-fold :

— a lightweight transformer-based model that outperforms other state of art GTISR methods.

— a modified training strategy that improves the robustness of GTISR in the absence of the guiding modality.

5.3 Related Works

5.3.1 Visible image super-resolution

The first approaches to super resolution employed so-called “traditional” approaches [39]. These methods either focused on the frequency domain, trying to model the relationship between the HR and LR images using mathematical models [40–42], or used dictionaries methods to map LR patches to HR patches [43–45].

2015 saw the emergence of deep learning-based methods using convolutional neural networks (CNNs) such as SRCNN [46], FSRCNN [47] and ESPCN [3], which introduced subpixel convolutional layers, a new upsampling operation. The advent of residual networks [48] (to solve the vanishing gradient problem in particular) led to new architectures like VDSR [49], RED [50] and EDSR [4], the latter proposing a new residual connection and winning the NTIRE2017 Super-Resolution Challenge [52].

In 2017, SRGAN [54] achieved remarkable results by applying a generative adversarial network (GAN) to the SR task. One year later appeared ESRGAN [100], an enhanced version of SRGAN.

Since transformers [101] have been adapted to the field of computer vision with the Vision Transformer ViT [58], the Swin transformer [6] resolved the computational complexity problem of ViT by using shifted windows. With this mechanism, SwinIR [59] applied the Swin architecture to the image reconstruction task and outperformed the best existing architectures. SwinIR’s main strength lies in its Residual Swin Transformer Blocks, which extract highly relevant features. More recently, the HAT [60] and SwinFIR [61] architectures have proposed improvements to SwinIR and represent the current state of the art in SR.

5.3.2 Thermal image super-resolution

Compared to RGB images, IR images are single channel, have low gradients and “overlapping information between high and low frequencies” [39]. To manage these characteristics, specific architectures have been proposed for IR images.

Before deep learning, frequency domain-based solutions like [68] or dictionary-based methods [69] were proposed. Then, inspired by the methods used in the visible spectrum, [70] and [71] exploited CNNs and residual networks. Other architectures have come up with the idea of using visible information (more abundant data) to reconstruct the IR image. For example, [72] used visible information in the loss function, while PSRGAN [55] used a GAN framework and transfer learning from RGB images to train their SR algorithm.

More recently, approaches using transformers have appeared, namely DASR [74] that exploits spatial and channel attention. In the same spirit, [102] dynamically reweights the output of attention and non-attention branches to improve the resolution and restore high-frequency details, offering a lightweight structure suitable for edge device deployment.

5.3.3 Guided thermal super-resolution

Unlike the methods presented above, guided methods take two paired images as input : an LR thermal image (or target image) and a higher-resolution guide image to help with the SR task. One of the first GTISR works was introduced in [76] ; it used a dual-path residual network to merge features from the visible and IR domains. More recently, CoRefusion [77] has been proposed. Its architecture is composed of two U-Nets [63] with residual connections to fuse both modalities. A contrastive term is added to the loss function and yields improved performance.

CoreFusion was part of the first GTISR track in the 2023 PBVS competition [75]. However, the winner of that competition was GuidedSR [75] ; this latter approach concatenates RGB and IR features from the shallow feature extraction layers and uses Non-linear Activation Free (NAF) blocks [103] to fuse RGB and IR information.

More recently, the authors of [11] described several SR guided methods applied to thermal images and how transforming the RGB guide image into a "thermal-like image" improves performance. They show that this substitution boosts performance by a few percentage points in different super-resolution guided architectures.

5.3.4 Multimodal image fusion

Multimodal image fusion aims at combining relevant information from images acquired with different sensors into a single image. DL-based fusion methods can be divided into three categories : early fusion, late fusion and hybrid fusion [82]. The first one merges features before task related layers, the second one uses task related layers on each modality before aggregating the information, while the last one combines the first two approaches.

SwinFusion [12] proposed to fuse images from two modalities using Attention-guided Cross-domain Fusion (ACF). Inspired by the Swin Transform block, this model merges information from the two modality branches via alternating modules of "self-attention-based intra-domain fusion" and "cross-attention-based inter-domain fusion" units.

The work in [104] proposed a Target-aware Dual Adversarial Learning for object detection. The idea is to exploit structural information in the IR image and textural details from the visible image to improve object detection. This is made possible by means of a generator and two discriminators that seek to retain relevant information from the two modalities.

5.3.5 Robustness to missing imaging modality

GTISR is subject to degraded performance when one of the inputs, e.g. the guiding RGB image, is missing at inference time. Little work in the literature addresses this issue directly for the thermal image SR task, but some studies have examined it in other application areas.

In [83], the authors evaluated the impact of the type of architecture, data augmentation and image fusion technique on action recognition performance in the case of a missing modality. They concluded that transformer-based fusion is more robust in this situation than feature summation or concatenation. Meanwhile, [84] studied the impact of a missing modality (text, audio, or image) in training or testing a GAN or autoencoder model. They proposed a Bayesian meta-learning framework to better manage missing modalities.

5.4 Method

In this paper, we propose a novel architecture, named SwinFuSR, as a contender for the PBVS 2024 TISR Track 2 challenge. The aim of this competition is to obtain a high-resolution (x8) infrared image from a low-resolution IR image and a medium-resolution RGB image.

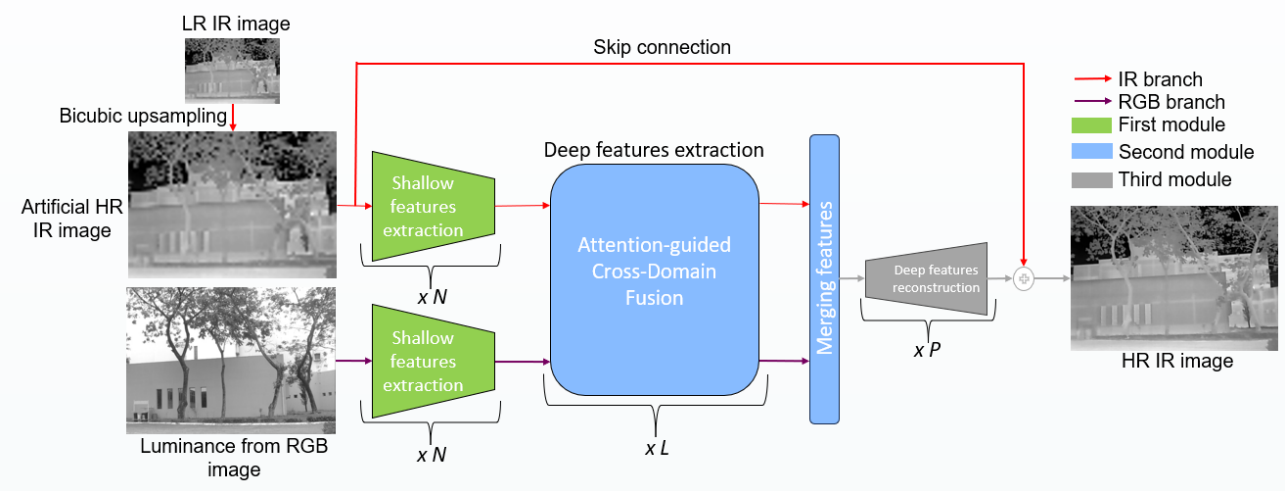


FIGURE 5.1 Architecture of the proposed SwinFuSR model.

5.4.1 Proposed architecture

As in many other super resolution transformer architectures [59, 60, 75], our own, illustrated in 5.1, is composed of three modules.

The first module extracts shallow features using convolutional layers followed by N Swin Transformer (STL) layers. The second module focuses on deep feature extraction. Its role is to extract characteristics that are useful to reconstruct the image by combining IR and RGB features. L Attention-guided Cross-domain Fusion (ACF) blocks are used to extract useful information from RGB and IR features. Then, concatenation and convolution are performed to merge the two branches. The third module carries out deep feature reconstruction. It is composed of P Swin Transformer layers to refine the merged features and three convolution layers to return to image space.

In the first two modules, the architecture is divided into two branches, similarly to SwinFusion [12] : one dedicated to the RGB image and the other to the IR image. A bicubic interpolation is performed on the IR image so that its dimensions (height h and width w) match those of its paired RGB image. Inspired by [59,71,100], a skip connection from the interpolated IR image to the reconstructed image is introduced for faster convergence and better performance. This gives the network an initial solution to improve upon.

5.4.2 Loss function

As a loss function, we use a combination of two differentiable pixel losses commonly used to measure the similarity between two images :

— An L_1 loss (or MAE) allows for relatively stable convergence and avoids gradient explosion [105] :

$$L_1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

with n the number of pixels, y_i the value of the i^{th} pixel in the ground-truth (GT) image and $\hat{y}_i$ the value of the i^{th} pixel in the reconstruction.

— An L_2 loss (or MSE) is more sensitive to higher reconstruction errors but can make the reconstruction smoother at the expense of valuable high-frequency details :

$$L_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The lower these two metrics are, the closer the reconstruction is to the GT.

We use the loss strategy proposed by GuidedSR, the winning solution in the 2023 PBVS challenge, described in [75].

$$Loss = \begin{cases} L_1 & \text{for the first } T \text{ epochs} \\ L_2 & \text{after} \end{cases}$$

This strategy allows us to obtain good convergence properties with an L_1 loss, then refine the optimization with an L_2 loss.

5.4.3 Training strategy

Specific training strategies can help build missing modality robustness into the model. The literature proposes two main ways to handle this. The first one is to remove the entire portion of the network dealing with the missing information ; in that case, the modalities must be processed independently as in CoRefusion [77]. The second and simpler method is to arbitrarily set the corresponding input values to the network to a fixed value such as zero.

To reduce performance degradation in the case of a missing modality, we propose a new model training regime that consists in randomly removing the training RGB images. More formally, at each training iteration, the input I to the network is given by :

$$I = \begin{cases} (I^{IR}_{h,w}, I^{RGB}_{h,w}) & \text{if } p < p_{th} \\ (I^{IR}_{h,w}, O_{h,w}) & \text{otherwise} \end{cases}$$

with $I_{h,w}^{RGB}$ the RGB image, $I_{h,w}^{IR}$ its corresponding IR image after bicubic interpolation, $O_{h,w}$ an all-zero image, p a random probability following a uniform distribution $\mathcal{U}(0,1)$ and p_{th} a fixed threshold between 0 and 1.

5.5 Experiments

5.5.1 Implementation details

To train our model, we used the dataset provided for the second track of the PBVS 2024 TISR Challenge. It is composed of 700 training samples and 200 validation samples, each sample being a 640x448 IR image, along with its downsampled version by a factor of 8 and its paired 640x448 RGB image. The 100 testing samples are provided without the HR ground truths. These registered images were acquired by Balser (for RGB) and TAU2 (for IR) cameras and represent images of outdoor urban scenes. We evaluated our model's performance on the training and validation sets using the PSNR and SSIM metrics.

Following common practice for training transformer architectures [12, 59, 61, 74], we used patches rather than the entire image as input. The patch size used was 128x128 and batch size was 16. The input patches were augmented with random horizontal and vertical flips and random rotations. Pixel values were normalized between 0 and 1.

The number of heads, the window size and the embedding dimensions were set to 6, 9 and 60 respectively. We set the network module depths to $N = 2$, $L = 3$ and $P = 3$, according to the study detailed in 5.5.2 below.

For the training, the learning rate was set to 4×10^{-4} until $T = 3300$, then reduced to 1×10^{-4} for the remainder. We used the Adam optimizer. The run lasted 72 hours (4300 epochs) on two Tesla V100 GPUs with 32.0 GB of VRAM each.

5.5.2 Ablation study

Effect of the number of modules

To study the effect of the number of STL blocks (N), ACF blocks (L) and STL blocks (P) in the extraction, fusion and reconstruction modules respectively, we set as a baseline $N = 1, L = 2, P = 1$ as in the original SwinFusion paper [12]. Then, we increased for each module separately these values by 1 and by 2 and observed the effect on performance (PSNR and SSIM) (see 5.2).

We can see that the increase in performance is most visible in the reconstruction module,

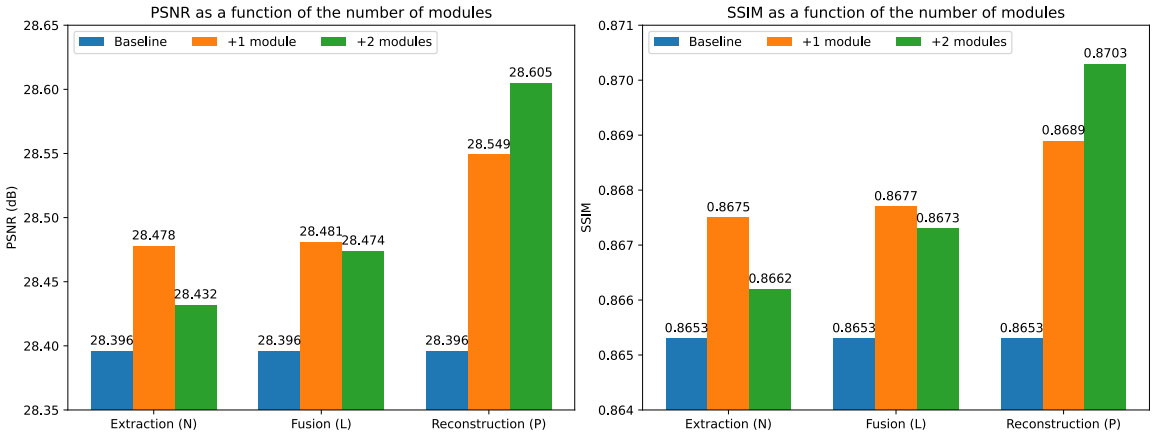


FIGURE 5.2 Effect of module depth on overall performance.

suggesting that the latter is the network bottleneck. Increasing the number of modules in the extraction and fusion modules by 1 each also improves performance, but to a lesser degree. Based on these results, we set the numbers of modules to N = 2, L = 3 and P = 3 for the experiments in 5.6.1 below. For the remaining experiments, we set them to N = 1, L = 2 and P = 1 to limit required resources.

Effects of skip connection

In SR, it is common to use skip connections between an artificial upsampling or early feature extraction layer and the end of the network. We trained our model with and without the skip connection in our SwinFuSR model (5.1). 5.3 shows the difference in performance.

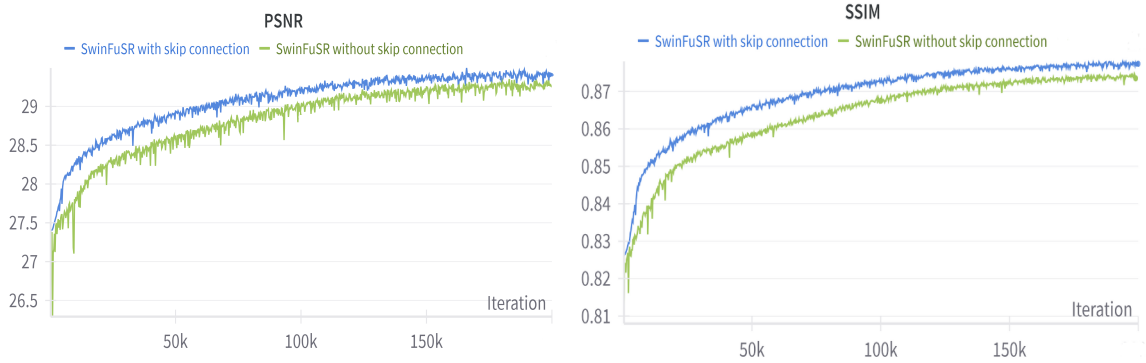


FIGURE 5.3 Performance with (blue) and without skip connection (green).

The results demonstrate that using a skip connection improves the convergence speed of the model and improves final performance by 0.3%. It is important to note that this performance

enhancement does not come at the cost of additional parameters in the model.

5.6 Results and discussion

5.6.1 RGB guided thermal image super-resolution

For a fair comparison between our solution and the existing methods GuidedSR and CoReFusion, we retrained the latter two models on the PBVS24 Track 2 dataset, using the same training setup as originally described in their respective papers [75, 77] (no pre-trained weights were available). Quantitative results are provided in 5.1.

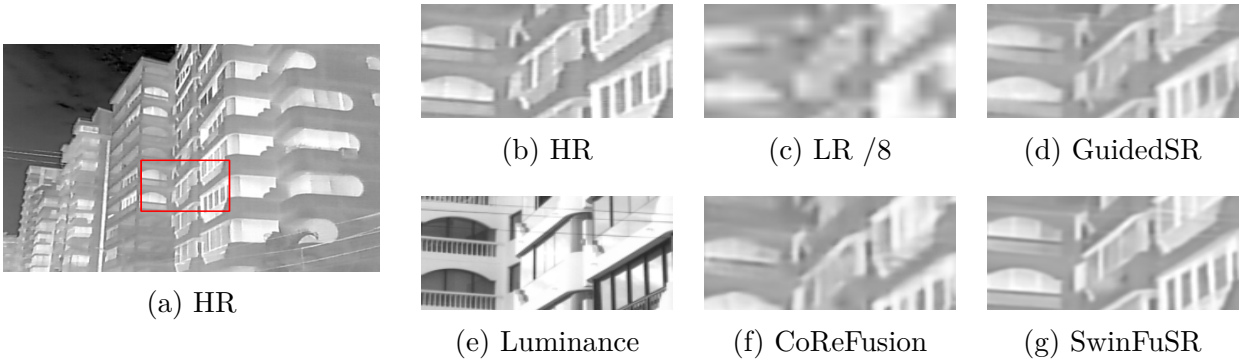
Method	PSNR	SSIM	#parameters
Bicubic	25.17	0.774	∅
CoReFusion	27.27	0.835	46.31M
GuidedSR	27.22	0.834	116.35M
SwinFuSR (ours)	**28.96**	**0.878**	**3.30M**

Tableau 5.1 PSNR and SSIM on validation set.

5.4 provides some qualitative results for guided SR on an image from the PBVS2024 challenge dataset. We can notice that SwinFuSR offers the closest output to the GT and seems clearer than the other 2 reconstructions.
To test our solution on different kinds of images and to verify generalization capabilities, we applied guided SR on images from the Simultaneously-collected multimodal Lying Pose (SLP) dataset [15]. This dataset is composed of low-resolution (120x160) infrared and RGB image pairs of adult subjects lying down in a hospital bed. 5.5 shows the results of the x8 guided SR of an image from this dataset. Unfortunately no GT IR images of higher resolution are provided in SLP. Thus, we used the available images as is but could not compare the SR results to reference HR images. Qualitatively, all three SR solutions enhance the very low-quality original image. Nevertheless, the details of the hand generated by SwinFuSR seem to be the most accurate, even if the shape of the hand seems unrealistic.

5.6.2 Robustness to missing RGB modality

To evaluate our proposed training regime to improve robustness to missing RGB input, we trained our network five times, each with a different probability threshold p_{th}. 5.7 illustrates the performance with and without RGB guide images at inference on the PBVS24 validation set. $p_{th} = 0$ means that during the training, all RGB guide images were used during the training.

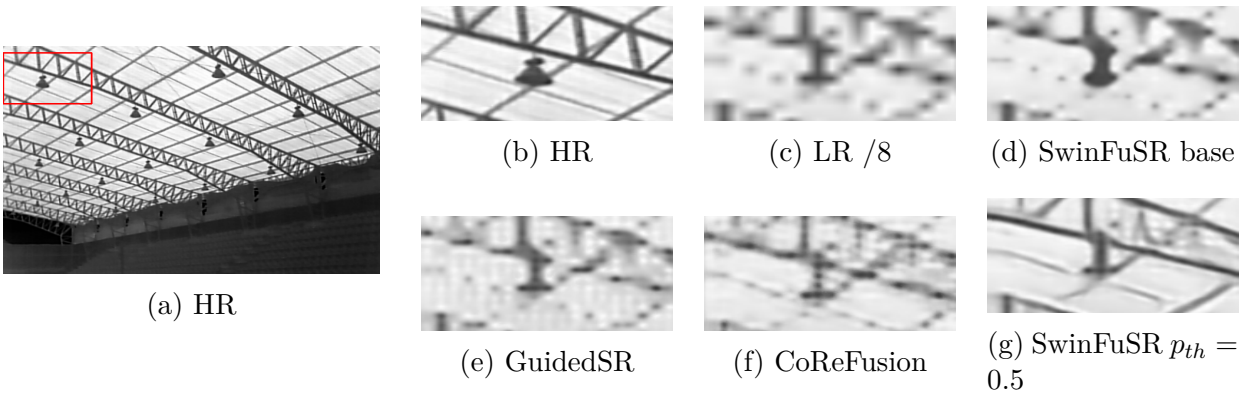


(a) HR (b) HR (c) LR /8 (d) GuidedSR (e) Luminance (f) CoReFusion (g) SwinFuSR

FIGURE 5.4 GTISR on image 292_01_D4 from PBVS 2024 Track- dataset.

(a) Original LR image (b) LR (c) GuidedSR (d) CoReFusion (e) SwinFuSR

FIGURE 5.5 GTISR on sample image from SLP dataset [15].



(a) HR (b) HR (c) LR /8 (d) SwinFuSR base (e) GuidedSR (f) CoReFusion (g) SwinFuSR $p_{th} = 0.5$

FIGURE 5.6 Unguided super resolution on image 044_02_D1 from PBVS 2024 Track 2 dataset.

First of all, we note that GuidedSR and CoReFusion have a smaller drop in performance than SwinFuSR when removing the RGB guide images. We can explain this by the fact that

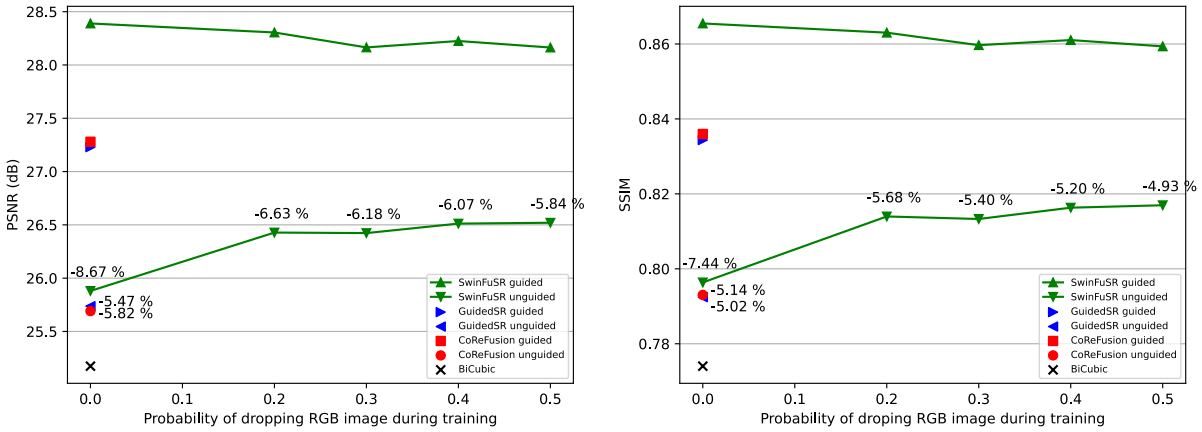


FIGURE 5.7 Effect of training parameter p_{th} on performance with (SwinFuSR guided) and without (SwinFuSR unguided) RGB input images at inference on the PBVS24 validation set.

their baseline performance in guided SR is much lower than SwinFuSR.

Second, we see that increasing p_{th} substantially improves the performance of SwinFuSR when no guide image is used for inference (from -8.67% to -6.63% for PSNR and from -7.44% to -5.68% for SSIM) when p_{th} goes from 0 to 0.2, while only slightly reducing performance for guided SR in terms of both metrics. This result suggests that dropping RGB images during training with a certain probability enables a trade-off between maintaining good performance in guided SR and improving results in the absence of guide images.
5.6 confirms visually that this training strategy can increase performance in unguided SR. Indeed, 5.6g is much clearer than 5.6d.

5.6.3 Discussion

Our model is much smaller in terms of parameters than the two competing methods (CoReFusion and GuidedSR), but is slower at inference (1.3s to go from 80x56 to 640x448 running on a PC equipped with an RTX 3080 GPU and 12 GB of VRAM). This limitation restricts the use of SwinFuSR for real-time inference. Moreover, model selection (varying the number of blocks) is costly in terms of VRAM usage, and required us to run those experiments on a GPU cluster. These two drawbacks are probably due to the high proportion of transformers in the network, which are known to be particularly resource-hungry.
Another aspect to consider in order to efficiently use the proposed architecture on other datasets is the fact that the IR and RGB images must be registered. For this purpose, several algorithms are available, such as the one proposed in [106] or Elastix [107], the method used

in the PBVS competition. In future work, we will study the robustness of the proposed model to IR-RGB registration errors.

5.7 Conclusion

This article proposes a new method for RGB guided thermal image super resolution. Our solution, named SwinFuSR, was submitted to Track 2 of the PBVS 2024 Thermal Image Super-Resolution Challenge and achieved better qualitative and quantitative results than other state-of-the-art architectures. We also present a novel training strategy that improves robustness to missing guide images at inference time. By randomly dropping a portion of the RGB images during training, the model's performance in unguided SR improves significantly compared to the guided SR baseline.
In future work, we will explore how to make better use of the RGB image data, for instance by generating pseudo-IR images. In addition, we will examine how super resolution can improve the performance of related tasks such as estimating in-bed human pose.

5.8 Acknowledgments

We thank Philippe Debanné for his valuable help in editing this paper. The project was supported by L. Seoud's NSERC Discovery grant. This research was enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca).

CHAPITRE 6 ÉVALUATION PRÉLIMINAIRE DE SWINFUSR SUR DES IMAGES DU CHUSJ EN CONDITIONS RÉELLES

6.1 Contexte de l'acquisition

Les acquisitions de vidéos multimodales (profondeur, IR et RGB) ont débuté en juillet 2024 dans le cadre de la collecte de données de patients hospitalisés en service de soins intensifs pédiatriques. Ces enregistrements ont été réalisés par une stagiaire de recherche, Aya Chetto, en suivant le protocole élaboré par Olivier Desclaux [25], qui propose une configuration optimisée pour le positionnement des caméras, ainsi qu'un logiciel dédié pour la calibration, l'acquisition et le stockage des images. Ce protocole d'acquisition (voir détails en Annexe D) a reçu l'approbation du comité d'éthique du CHUSJ (2022-3505).

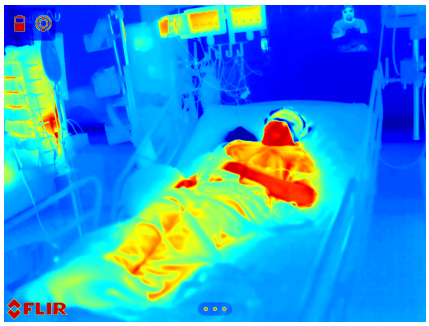
FIGURE 6.1 Installation pour l'acquisition [20]

Le montage des caméras est illustré à la Figure 6.1. En haut à droite, se trouve la caméra infrarouge haute résolution FLIR T1020, et à gauche, une caméra Kinect Azure permettant de capturer les images de profondeur et RGB. On remarque que les objectifs des deux caméras sont espacés d'environ trente centimètres, rendant une bonne calibration primordiale. Au centre, se trouve l'ordinateur équipé du logiciel d'acquisition où toutes les vidéos seront stockées avant d'être copiées sur le serveur sécurisé de l'hôpital. En dessous, on voit une grille pour effectuer la calibration entre les caméras ainsi qu'un sèche-cheveux utilisé pour chauffer la grille. En effet, comme vu dans la revue de littérature, une différence de température est nécessaire pour calibrer la caméra thermique. Les vidéos sont acquises dans des chambres de l'USI et durent en moyenne 3 minutes. Lors du stage recherche de Aya Chetto, 15 acquisitions ont été réalisées, puis le personnel de recherche a continué les acquisitions pour atteindre, en

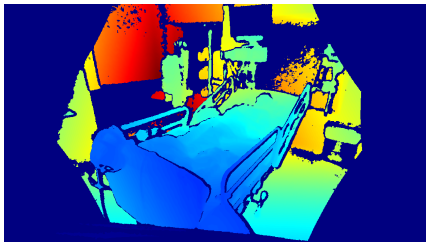
date du 6 novembre, un total de 70 patients.

Des exemples d'images tirées de ces acquisitions sont proposés aux Figures 6.3 et 6.2. Les visages ont été floutés sur les images dans le spectre visible pour désidentifier les sujets.

(a) Image RGB



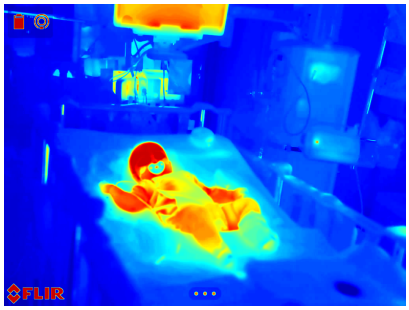
(b) Image IR



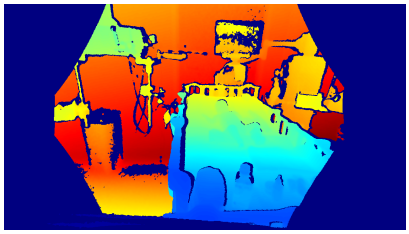
(c) Image de profondeur

FIGURE 6.2 Images issues du patient 16

(a) Image RGB



(b) Image IR

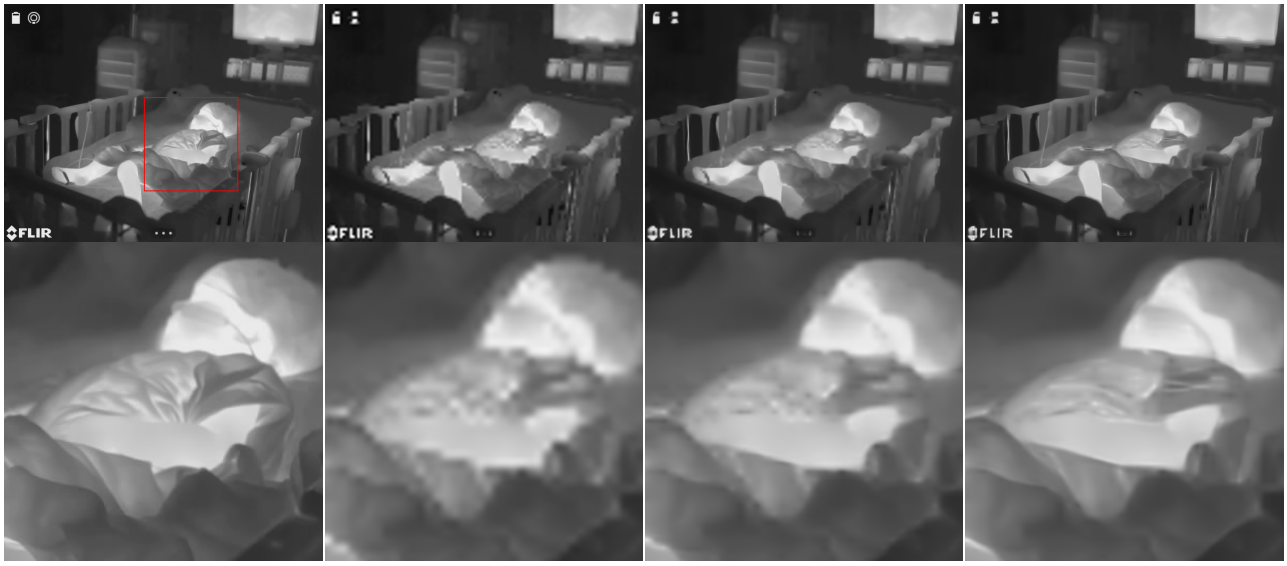


(c) Image de profondeur

FIGURE 6.3 Images issues du patient 34

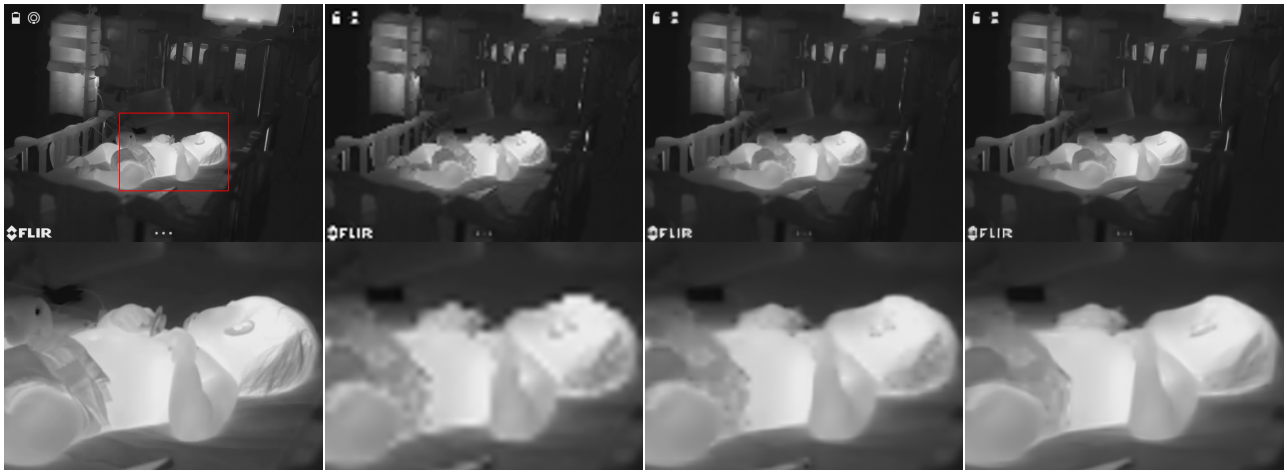
6.2 Évaluation de SwinFuSR sans modalité guide

Pour tester la solution SwinFuSR proposée dans le Chapitre 5, nous commencerons par l'utiliser en mode non guidé, c'est-à-dire sans l'aide de l'image RGB pour la SR. Nous testerons deux modèles SwinFuSR différents ayant la même architecture. La seule différence résidera dans la manière dont les modèles ont été entraînés. Pour le premier modèle, nous fixons une probabilité $p_{th} = 0.5$ que le modèle n'ait pas accès à l'image guide lors de l'entraînement, afin d'améliorer sa robustesse dans les cas où la modalité guide est manquante. Pour le second modèle, nous fixons une probabilité $p_{th} = 0$ que le modèle n'ait pas accès à l'image guide lors de l'entraînement, c'est à dire que ce modèle est entraîné dans tous les cas sur des paires d'images IR et RGB. Les poids des modèles proviennent de l'article mentionné au chapitre 5, et par conséquent, les modèles n'ont pas été entraînés sur des images acquises au CHUSJ. Les Figures 6.4 et 6.5 présentent les résultats des deux modèles sur deux images de référence. À titre indicatif, les images IR de référence ont une résolution originale de 1024×768 pixels. Pour une super-résolution avec un facteur de $\times 8$, l'image basse résolution obtenue par sous-échantillonnage a une résolution de 128×96 pixels.



(a) Référence	(b) SE / 8	(c) SR avec $p_{th} = 0$	(d) SR avec $p_{th} = 0.5$
PSNR :	**29.05**	28.92	28.74
SSIM :	0.945	0.953	**0.956**

FIGURE 6.4 Résultats d'une SR non guidée par SwinFuSR entrainé avec $p_{th} = 0$ et $p_{th} = 0.5$ pour le patient 3



(a) Référence	(b) SE / 8	(c) SR avec $p_{th} = 0$	(d) SR avec $p_{th} = 0.5$
PSNR :	28.64	**28.87**	28.72
SSIM :	0.941	0.951	**0.954**

FIGURE 6.5 Résultats d'une SR non guidée par SwinFuSR entrainé avec $p_{th} = 0$ et $p_{th} = 0.5$ pour le patient 32

Lorsqu'on effectue un zoom sur l'image sous-échantillonnée d'un facteur 8 (voir Figure 6.5b et 6.4b), on observe une importante pixellisation, indiquant une grande perte d'information par rapport à l'image de référence (voir Figure 6.5a et 6.4a). Visuellement, on constate que les deux modèles améliorent la qualité de l'image d'origine. Cependant, comme on pouvait s'y attendre, le modèle avec une probabilité $p_{th} = 0.5$ produit de meilleurs résultats visuels, confirmant ainsi l'efficacité de cette méthode pour augmenter la robustesse du modèle dans les cas où les images guides sont absentes.

TABLEAU 6.1 Résultats quantitatifs d'une SR non guidée

Sur 31 images	SE /8	SwinFuSR $p_{th} = 0$	SwinFuSR $p_{th} = 0.5$
PSNR	29.77	**29.91**	29.71
SSIM	0.952	0.959	**0.961**

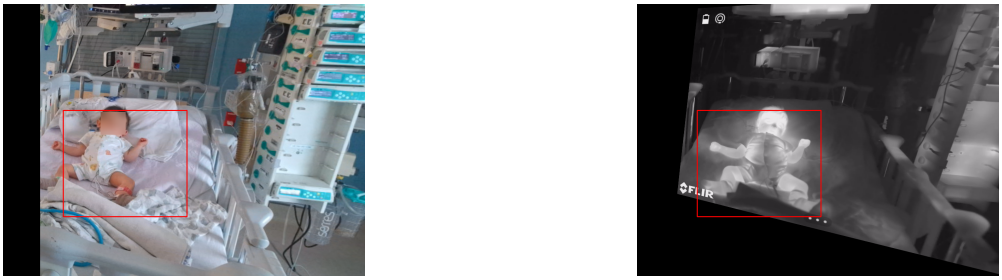
Pour calculer les métriques de la Figure 6.1, nous avons sélectionné les images parfaitement nettes (représentant environ 50% des patients disponibles), puis nous avons calculé les métriques entre l'image de référence et l'image générée. Le PSNR contredit les observations visuelles, mais le SSIM confirme que le modèle avec $p_{th} = 0.5$ est le plus efficace. On note que le PSNR du modèle avec $p_{th} = 0.5$ est plus faible que celui de l'image sous-échantillonnée, ce qui souligne les limites, au moins dans le cas de l'IR, du PSNR en tant que métrique totalement fiable.

Conclusion : Même sans image guide, nous avons montré que la solution SwinFuSR présentée au chapitre 5 donne de bons résultats. Nous avons également confirmé que la méthode d'entraînement visant à augmenter la robustesse du modèle en l'absence de la modalité guide améliore au moins qualitativement et sur la métrique SSIM, les résultats.

6.3 Évaluation de SwinFuSR avec modalité guide

Dans cette section, nous allons exploiter les informations issues de la modalité RGB pour améliorer la qualité de l'image thermique en utilisant SwinFuSR de façon guidée. Cela soulève deux défis. Premièrement, il est essentiel que la modalité RGB apporte une information complémentaire significative. Par exemple, si l'image RGB est trop sombre ou mal exposée, elle ne contribuera pas efficacement à l'amélioration de l'image thermique. Deuxièmement, il est nécessaire que les deux modalités soient spatialement alignées. En d'autres termes, les mêmes pixels de l'image RGB et de l'image IR doivent correspondre à un même point de la scène.

Dans le cas des données acquises au CHUSJ, les images RGB sont claires et fournissent des informations détaillées sur la scène. En revanche, les images RGB et IR ne sont pas recalées. Par contre, moyennant les calibrations intrinsèques et extrinsèques, nous avons pu au mieux appliquer une rectification aux paires d'images, assurant au minimum un alignement vertical entre les images.
La Figure 6.6 présente les calibrages pour le patient 13. Pour tout de même aligner les images, nous allons décaler "à la main" les pixels de l'image RGB afin qu'elle soit alignée avec l'image IR. Par exemple, si on décale les pixels de l'image RGB de 170 pixels vers la droite, on obtient l'image 6.6e. Bien que cette approche ne soit pas rigoureuse, elle permet d'obtenir des images "presque" recalées.
‘



(a) Image RGB réctifiée

(b) Image IR réctifiée

(c) Zoom RGB

(d) Zoom IR

(e) Zoom RGB décalée

(f) Zoom IR

FIGURE 6.6 Images réctifiées issues du patient 13

Présentons sur la Figure 6.7 les résultats obtenus en utilisant le modèle $p_{th} = 0$ pour trois images de patients.

Visuellement, nous pouvons remarquer que les images super-résolues 6.7d et 6.7e ajoutent bien des détails à l'image basse résolution 6.7c. Lorsque nous zoomons, les images avec guide présentent moins d'artefacts que les images sans guide.

D'un point de vue des métriques, les deux méthodes, avec et sans guide, améliorent l'image basse résolution et affichent des performances quasiment identiques. Cependant, le faible nombre d'images utilisées pour calculer ces métriques, combiné à la proximité des résultats obtenus, empêche de tirer des conclusions significatives basées uniquement sur ces mesures.

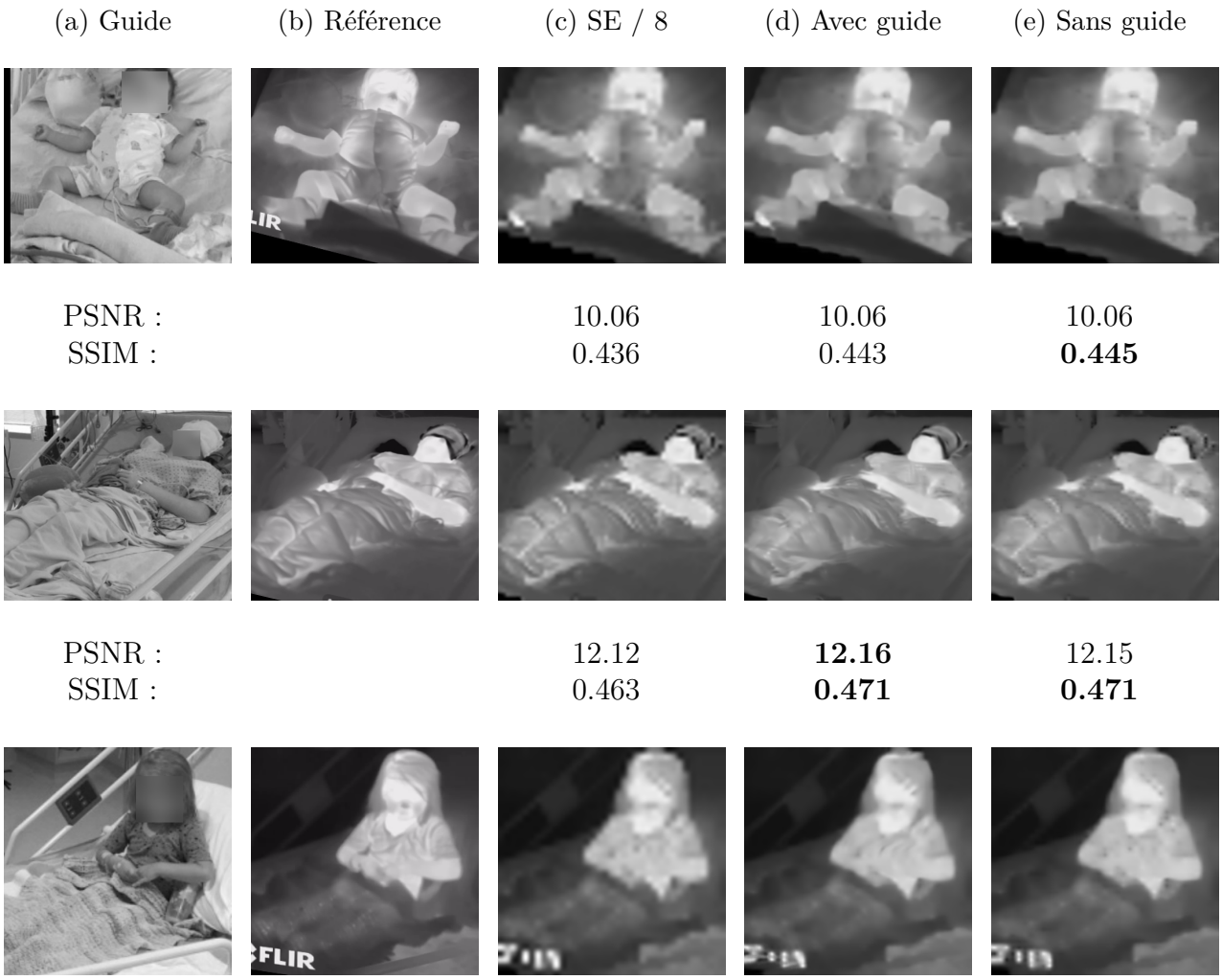


FIGURE 6.7 Super résolution des images des patients 13, 16 et 35

Conclusion : Dans cette section, nous avons mis en évidence les défis liés à l'utilisation d'images multimodales, notamment en ce qui concerne les aspects de recalage. Toutefois, nous avons observé qu'en utilisant des images presque recalées, il est possible d'obtenir des résultats de super-résolution guidée légèrement supérieurs à ceux de la super-résolution non guidée (du moins visuellement). Ce constat confirme notre intuition selon laquelle la modalité visible peut fournir des informations pertinentes pour améliorer la super-résolution d'images infrarouges, en particulier dans un contexte hospitalier.

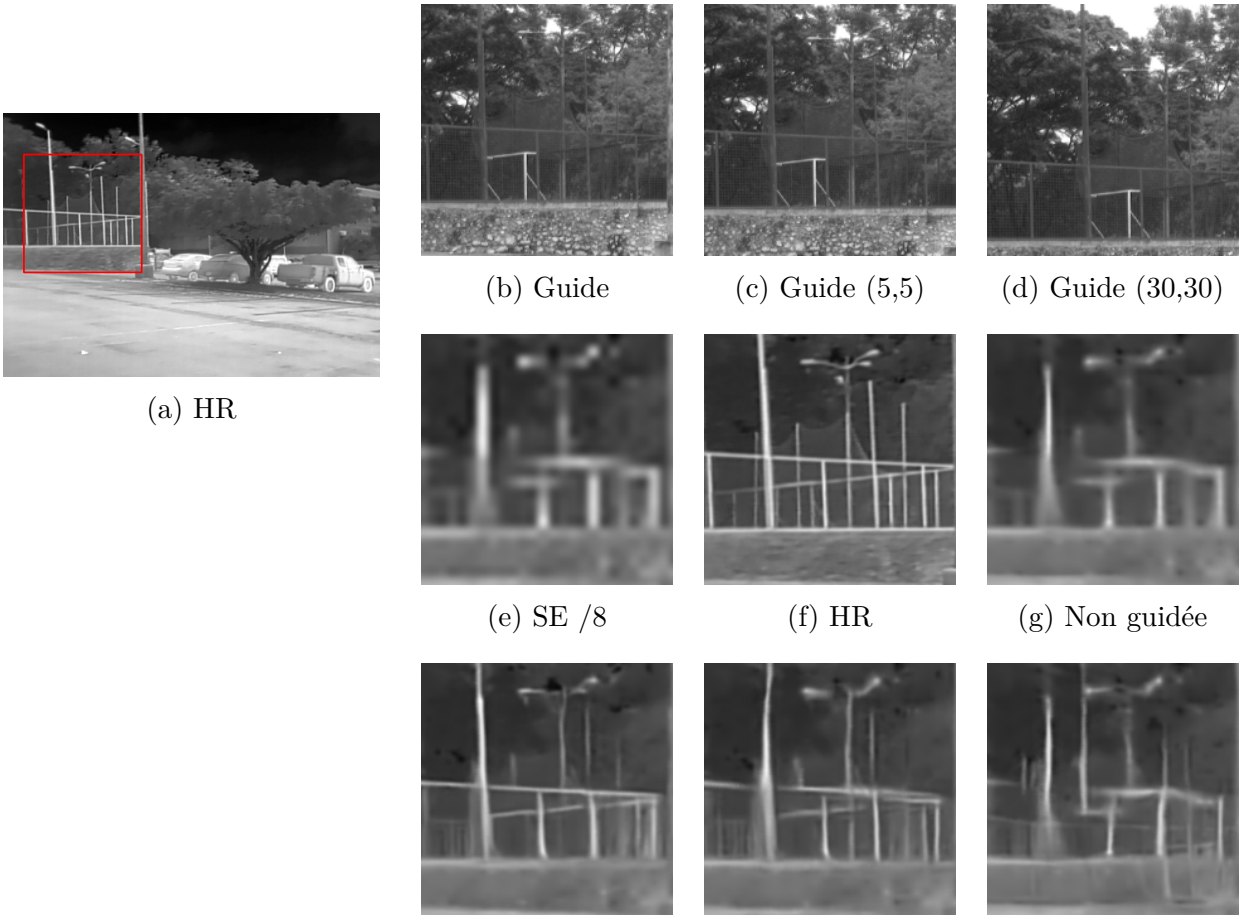
6.4 Évaluation de l'influence du recalage inter-modalité sur les résultats de SwinFuSR avec guide

Comme nous l'avons vu dans la section précédente, travailler avec une image guide pose plusieurs défis, notamment en ce qui concerne le recalage.

Dans cette section, nous évaluerons l'impact d'un mauvais recalage entre les deux modalités. Nous supposerons que les deux images sont déjà dans le même plan, mais qu'un décalage spatial subsiste entre leurs pixels.

Pour cela, nous utiliserons le jeu de données PBVS, présenté dans l'article du Chapitre 5. Ce jeu de données inclut des images RGB et IR "registered", ce qui permet de réaliser des expériences avec une vérité terrain. Cela contraste avec le jeu de données du CHUSJ, qui ne propose pas de couples d'images RGB/IR parfaitement alignés.

Pour ces expérimentations, nous décalerons les pixels de l'image guide de $décalage\_x$ pixels vers la droite (axe horizontal) et de $décalage\_y$ vers le bas (axe vertical). Puis nous utiliserons l'image guide décalée pour faire la SR. Ces expérimentations ont été effectuées sur le modèle SwinFuSR avec $p_{th} = 0$.
Les Figures 6.8 et 6.9 présentent les résultats visuels et les Tableaux 6.2 et 6.3 affichent les métriques associées. Sur ces figures, les parenthèses à côté de "Guide" indiquent le nombre de pixels décalés spatialement pour l'image guide ($décalage\_x$, $décalage\_y$).



(a) HR (b) Guide (c) Guide (5,5) (d) Guide (30,30)

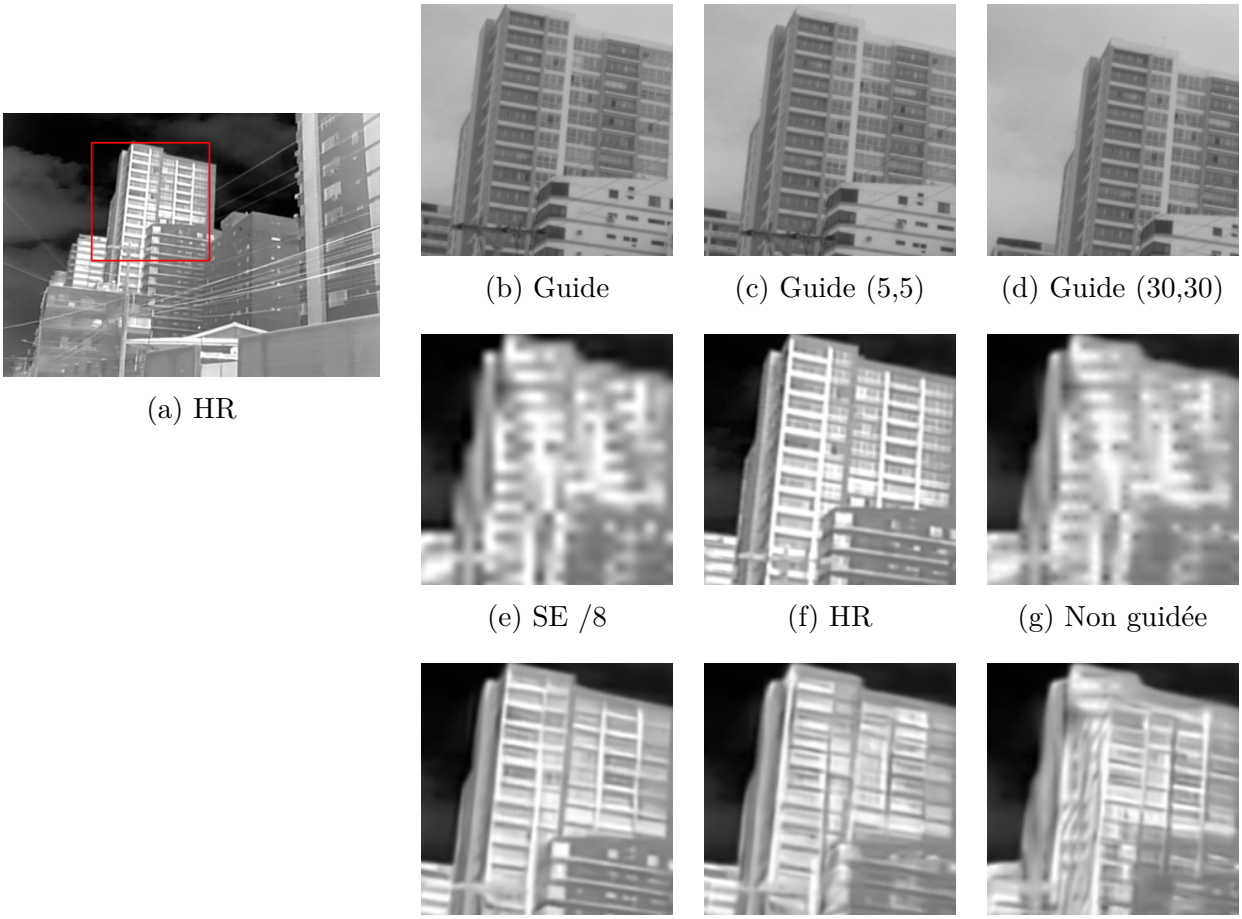
(e) SE /8 (f) HR (g) Non guidée

(h) Guidée (i) Guidée (5,5) (j) Guidée (30,30)

FIGURE 6.8 Résultats visuels de l'effet d'un décalage spatiale de l'image guide pour SwinFuSR (image 002_01_D2)

TABLEAU 6.2 Résultats quantitatifs de l'effet d'un décalage spatiale de l'image guide pour SwinFuSR (image 002_01_D2)

Sur l'image 002_01_D2	SE /8	Non guidée	Guidée	Guidée (5,5)	Guidée (30,30)
PSNR	24.14	24.80	**26.83**	25.07	24.34
SSIM	0.772	0.791	**0.847**	0.803	0.778



(a) HR (b) Guide (c) Guide (5,5) (d) Guide (30,30)

(e) SE /8 (f) HR (g) Non guidée

(h) Guidée (i) Guidée (5,5) (j) Guidée (30,30)

FIGURE 6.9 Résultats visuels de l'effet d'un décalage spatiale de l'image guide pour SwinFuSR (image 029_01_D4)

TABLEAU 6.3 Résultats quantitatifs de l'effet d'un décalage spatiale de l'image guide pour SwinFuSR (image 029_01_D4))

Sur l'image 002_01_D2	SE /8	Non guidée	Guidée	Guidée (5,5)	Guidée (30,30)
PSNR	22.50	23.11	**27.05**	24.28	23.17
SSIM	0.689	0.719	**0.864**	0.783	0.734

Tout d'abord, nous pouvons constater que, comme prévu, un mauvais recalage de l'image guide réduit la performance de SwinFuSR pour l'approche guidée, que ce soit de manière qualitative ou quantitative.

Visuellement, cela entraîne des déformations géométriques : les lignes droites deviennent courbées et des détails sont perdus (voir Figures 6.8j et 6.9j). Cependant, dans les deux exemples présentés ici, même avec un décalage important (30 pixels de translation, à la fois

verticalement et horizontalement), nous parvenons tout de même à améliorer l'image basse résolution (voir Figures 6.8e et 6.9e).

Intéressons-nous maintenant à l'évolution de la performance lorsqu'on fait varier les différentes valeurs de décalage de pixels verticalement et horizontalement. Pour la Figure 6.10 et le Tableau 6.4, les valeurs des métriques sont des moyennes des métriques calculées sur tout l'ensemble de validation du jeu de données PBVS (200 échantillons).

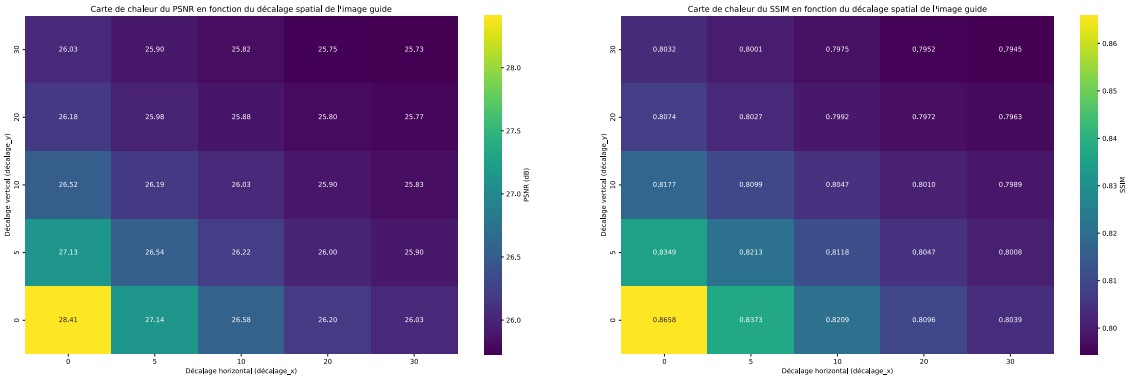


FIGURE 6.10 Cartes de chaleur de l'évolution des performances en fonction du décalage spatial de l'image guide

TABLEAU 6.4 Résultats quantitatifs de SwinFuSR sur tout l'ensemble de validation de PBVS

Sur 200 images	SE /8	Non guidée	Guidée (0,0)
PSNR	25.17	25.88	**28.41**
SSIM	0.774	0.796	**0.865**

La Figure 6.10 nous indique que les performances baissent assez rapidement lorsqu'on décale les pixels. En effet, si on décale les pixels de 5 valeurs vers la droite de l'image guide, nous perdons 1,27 de PSNR et 0,028 de SSIM. Plus on augmente la valeur de décalage et plus la baisse est faible (de 20 à 30 pixels verticalement, nous perdons 0,15 de PSNR et 0,0042 de SSIM). Ce résultat démontre la forte sensibilité de la méthode guidée SwinFuSR au bon alignement des pixels entre les deux modalités.

Nous remarquons que SwinFuSR est légèrement plus sensible à un désalignement vertical qu'horizontal. Nous pouvons noter cependant que pour ce modèle, même pour un désalignement spatial important (exemple (30,30)), le modèle permet quand même d'améliorer l'image basse résolution (25.73 db vs 25.17 en PSNR et 0.794 vs 0.774 en SSIM). Nous pouvons aussi remarquer que la méthode non guidée peut présenter de meilleures performances que la méthode guidée dans le cas où il y a désalignement vertical et horizontal important (

dans nos expériences, à partir d'un décalage absolu de 40 pixels). Cette valeur seuille trouvée ici doit vraisemblablement dépendre de l'architecture, du jeu de données et du facteur de super-résolution.

Conclusion : Dans cette section, nous avons pu évaluer à quel point un désalignement spatial de l'image guide par rapport à l'image IR fait varier la performance. Nous avons pu constater que le modèle $p_{th} = 0$ de SwinFuSR était très sensible au recalage des deux modalités et à partir de certaines valeurs de décalage spatial, il peut être plus intéressant d'utiliser la méthode non guidée.

CHAPITRE 7 CONCLUSION

7.1 Synthèse des travaux

Grâce à une revue de la littérature et à une évaluation des solutions existantes dans les **chapitre 2 et 4**, nous avons démontré l'efficacité de l'architecture SwinIR ainsi que des méthodes de super-résolution guidées. Sur cette base, dans le **chapitre 5**, nous avons proposé SwinFuSR, une nouvelle architecture de super-résolution guidée, légère et offrant des performances supérieures aux architectures existantes sur des bases de données publiques. Elle s'est notamment classée dans le top-3 sur le challenge de SR thermique guidée PBVS 2024. De plus, une nouvelle méthode d'entraînement a été introduite pour essayer de pallier une éventuelle perte de performance en cas d'absence de la modalité guide. Enfin, dans le chapitre 6, nous avons présenté les images RGB et IR acquises à l'unité de soins intensifs pédiatriques du CHUSJ. Nous avons aussi démontré, sur un ensemble certes restreint, la performance de SwinFuSR ainsi que la méthode d'entraînement sur ces données. En effet, il n'était pas évident que ce modèle qui a été entraîné sur des images thermiques mais dans un contexte différent (scènes extérieures) puisse généraliser en inférence sur les images du CHUSJ. Finalement, nous avons évalué l'impact d'une mauvaise rectification entre les deux modalités et souligné à quel point la méthode guidée de SwinFuSR est sensible à un bon alignement spatial des deux modalités. Ainsi, nous disposons d'une solution prometteuse sur les images du CHUSJ et qui se montre robuste face aux différents événements pouvant empêcher l'utilisation des informations du spectre visible.

7.2 Limitations de la solution proposée

L'architecture SwinFuSR est comme beaucoup d'autres architectures de réseaux de neurones, en particulier celles basées sur des transformers, coûteuse en termes de mémoire sur la carte graphique (VRAM). Pour l'inférence d'une super-résolution $\times 8$ d'une image de taille 128×96, cela nécessite 6.8 Go de mémoire. De plus, le temps de calcul est d'environ 1.3 secondes, ce qui peut limiter son utilisation pour une application en temps réel.

Les images acquises au CHUSJ sont en majorité de bonne qualité. Cependant, le système prototype actuellement en place nécessite un effort important de la part du personnel pour obtenir une calibration précise et un ajustement satisfaisant des caméras et de leur système optique. Comme nous l'avons constaté dans la dernière section, le recalage des deux modalités RGB et IR est nécessaire pour tirer plein profit de la guidance sur la tâche de super-résolution.

Un autre point à souligner est le manque de diversité des images dans les vidéos du jeu de données du CHUSJ. En effet, pour chaque patient, le système capture en continu la scène pendant environ 3 minutes. Or, très peu d'éléments évoluent durant ce laps de temps, ce qui fait que la plupart des images de la vidéo sont sensiblement identiques.

7.3 Travaux futurs

Pour améliorer les performances de SwinFuSR sur les images du CHUSJ, il serait pertinent d'effectuer un "fine-tuning" sur ces données. En effet, l'accès aux images infrarouges haute résolution permettrait de réaliser un tel entraînement. Cependant, cela nécessiterait un nombre important d'images et une plus grande diversité que ce qui est actuellement disponible.

De plus, il serait pertinent de mieux modéliser la dégradation qui transforme une image thermique haute résolution en une image basse résolution. Actuellement, n'ayant pas accès à des images directement acquises par une caméra thermique basse résolution, nous utilisons un simple sous-échantillonnage pour simuler cette dégradation. Cependant, en adoptant une méthode de super-résolution aveugle (voir section 2.2.1), nous pourrions obtenir des images d'entrée basse résolution qui se rapprochent davantage de celles attendues lors des acquisitions au CHUSJ.

Un autre axe d'amélioration concerne le recalage multimodal. À ce jour, nos expérimentations préliminaires se basent sur une rectification puis un recalage vertical moyennant les matrices de calibration intrinsèque et extrinsèque des caméras. Il serait intéressant d'explorer des solutions pour un recalage automatique des deux modalités.

Pour limiter l'impact d'un mauvais recalage entre les deux modalités, nous pourrions introduire une nouvelle forme d'augmentation lors de l'entraînement du modèle. Cette approche consisterait à décaler aléatoirement les pixels de l'image guide, obligeant ainsi le modèle à intégrer des informations provenant de pixels plus éloignés pour reconstruire l'image infrarouge.

À l'avenir, il serait intéressant d'envisager l'utilisation de la super-résolution dans le cadre d'autres tâches de vision par ordinateur, telles que la détection de pathologies [2, 108] ou l'estimation de pose (voir Annexe E.1).

Le défi final de ce projet est de mettre en œuvre une solution efficace, extensible et robuste, exploitant l'information thermique pour améliorer la prise en charge des patients et faciliter le quotidien du personnel des soins intensifs pédiatriques.

RÉFÉRENCES

[1] >. Page. (2021) Picking a Thermal Color Palette. Accessed : 2024-10-03. [En ligne]. Disponible : https://www.flir.com/discover/industrial/picking-a-thermal-color-palette/

[2] F. M. Senalp et M. Ceylan, "A new approach for super-resolution and classification applications on neonatal thermal images," *Quantitative InfraRed Thermography Journal*, p. 1–18, févr. 2023. [En ligne]. Disponible : https://www.tandfonline.com/doi/full/10.1080/17686733.2023.2179282

[3] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert et Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 1874–1883.

[4] B. Lim, S. Son, H. Kim, S. Nah et K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," dans *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, p. 136–144.

[5] M. S. Moustafa et S. A. Sayed, "Satellite Imagery Super-Resolution Using Squeeze-and-Excitation-Based GAN," *International Journal of Aeronautical and Space Sciences*, vol. 22, nº. 6, p. 1481–1492, déc. 2021. [En ligne]. Disponible : https://doi.org/10.1007/s42405-021-00396-6

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin et B. Guo, "Swin transformer : Hierarchical vision transformer using shifted windows," dans *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, p. 10 012–10 022.

[7] A. Liu, Y. Liu, J. Gu, Y. Qiao et C. Dong, "Blind Image Super-Resolution : A Survey and Beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–19, 2022. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/9870558/

[8] Y. Tian, H. Chen, C. Xu et Y. Wang, "Image processing gnn : Breaking rigidity in super-resolution," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, p. 24 108–24 117.

[9] J. Ho, A. Jain et P. Abbeel, "Denoising diffusion probabilistic models," dans *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan et H. Lin, édit., vol. 33. Curran Associates, Inc., 2020, p. 6840–6851. [En ligne]. Disponible : https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[10] Y. Shi, N. Chen, Y. Pu, J. Zhang et L. Yao, “SwinIBSR : Towards real-world infrared image super-resolution,” *Infrared Physics & Technology*, vol. 139, p. 105279, juin 2024. [En ligne]. Disponible : https://linkinghub.elsevier.com/retrieve/pii/S1350449524001634

[11] P. L. Suárez, D. Carpio et A. D. Sappa, “Enhancement of guided thermal image super-resolution approaches,” *Neurocomputing*, vol. 573, p. 127197, mars 2024. [En ligne]. Disponible : https://linkinghub.elsevier.com/retrieve/pii/S0925231223013206

[12] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei et Y. Ma, “Swinfusion : Cross-domain long-range learning for general image fusion via swin transformer,” *IEEE/CAA Journal of Automatica Sinica*, vol. 9, nº. 7, p. 1200–1217, 2022.

[13] Z. Wang, A. Bovik, H. Sheikh et E. Simoncelli, “Image Quality Assessment : From Error Visibility to Structural Similarity,” *IEEE Transactions on Image Processing*, vol. 13, nº. 4, p. 600–612, avr. 2004. [En ligne]. Disponible : http://ieeexplore.ieee.org/document/1284395/

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman et O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT : IEEE, juin 2018, p. 586–595. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/8578166/

[15] S. Liu, X. Huang, N. Fu, C. Li, Z. Su et S. Ostadabbas, “Simultaneously-collected multimodal lying pose dataset : Towards in-bed human pose monitoring under adverse vision conditions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[16] S. Liu et S. Ostadabbas, “Seeing under the cover : A physics guided learning approach for in-bed pose estimation,” *22nd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI2019), Shenzhen, China.*, 2019.

[17] R. E. Rivadeneira, A. D. Sappa, C. Wang, J. Jiang, Z. Zhong, P. Chen et S. Wang, “Thermal image super-resolution challenge results - pbvs 2024,” dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, p. 3113–3122.

[18] A. Upadhyay, B. Dhupar, M. Sharma, A. Shukla et A. Abraham, “Lwirpose : A novel lwir thermal image dataset and benchmark,” 2024.

[19] M. Trammer, N. Genser et J. Seiler, “RGB-Guided Resolution Enhancement of IR Images,” dans *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, juin 2023, p. 1–5, iSSN : 2157-8702.

[20] A. Chetto, “Rapport de stage,” 2024.

[21] S. Li, G. Zhang, Z. Luo, J. Liu, Z. Zeng et S. Zhang, “Degradation regression with uncertainty for blind super-resolution,” *Neurocomputing*, vol. 582, p. 127486, mai 2024. [En ligne]. Disponible : https://linkinghub.elsevier.com/retrieve/pii/S0925231224002571

[22] A. Bridier, M. Shcherbakova, A. Kawaguchi, N. Poirier, C. Said, R. Noumeir et P. Jouvet, “Hemodynamic assessment in children after cardiac surgery : A pilot study on the value of infrared thermography,” *Frontiers in Pediatrics*, vol. 11, 2023. [En ligne]. Disponible : https://www.frontiersin.org/articles/10.3389/fped.2023.1083962

[23] M. Shcherbakova, R. Noumeir, M. Levy, A. Bridier, V. Lestrade et P. Jouvet, “Optical thermography infrastructure to assess thermal distribution in critically ill children,” *IEEE Open Journal of Engineering in Medicine and Biology*, vol. PP, p. 1–1, 12 2021.

[24] E. F. J. Ring et K. Ammer, “Infrared thermal imaging in medicine,” *Physiological Measurement*, vol. 33, n°. 3, p. R33–R46, mars 2012. [En ligne]. Disponible : https://iopscience.iop.org/article/10.1088/0967-3334/33/3/R33

[25] O. Desclaux, “Protocole pour la création d’une base de données multimodale d’estimation de pose 3d d’enfants en salle de soins intensifs,” Mémoire de maîtrise, Polytechnique Montréal, juillet 2022. [En ligne]. Disponible : https://publications.polymtl.ca/10538/

[26] G. Cherni, “Apprentissage profond multimodal pour l’estimation de pose d’humains alités,” Mémoire de maîtrise, Polytechnique Montréal, août 2022. [En ligne]. Disponible : https://publications.polymtl.ca/10517/

[27] C. Arnold, P. Jouvet et L. Seoud, “SwinFuSR : an image fusion-inspired model for RGB-guided thermal image super-resolution,” dans *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA : IEEE, juin 2024, p. 3027–3036. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/10678260/

[28] Thermography - Wikipedia. Accessed : 2024-10-03. [En ligne]. Disponible : https://en.wikipedia.org/wiki/Thermography

[29] R. Gade et T. B. Moeslund, “Thermal cameras and applications : a survey,” *Machine Vision and Applications*, vol. 25, n°. 1, p. 245–262, janv. 2014. [En ligne]. Disponible : http://link.springer.com/10.1007/s00138-013-0570-5

[30] >. l. page, “Importance de la sensibilité thermique pour la précision de la détection,” 2021, accessed : 2024-10-03. [En ligne]. Disponible : https://www.flir.fr/discover/security/perimeter-protection/the-importance-of-thermal-sensitivity-for-detection-accuracy/

[31] I. N. Swamidoss, A. Bin Amro et S. Sayadi, “Systematic approach for thermal imaging camera calibration for machine vision applications,” *Optik*, vol. 247, p. 168039, 2021. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0030402621015989

[32] >. Page, “Choosing the Right Thermal Imager for Your Integrated Project,” 1, accessed : 2024-10-03. [En ligne]. Disponible : https://www.flir.com/discover/cores-components/choosing-the-right-thermal-imager-for-your-integrated-project/

[33] Y. Gu, Z. Zeng, H. Chen, J. Wei, Y. Zhang, B. Chen, Y. Li, Y. Qin, Q. Xie, Z. Jiang et Y. Lu, “MedSRGAN : medical images super-resolution using generative adversarial networks,” *Multimedia Tools and Applications*, vol. 79, n°. 29-30, p. 21 815–21 840, août 2020. [En ligne]. Disponible : https://link.springer.com/10.1007/s11042-020-08980-w

[34] S. Pan, S.-B. Chen et B. Luo, “A super-resolution-based license plate recognition method for remote surveillance,” *Journal of Visual Communication and Image Representation*, vol. 94, p. 103844, 2023. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S1047320323000949

[35] F. Zhou, B. Wang et Q. Liao, “Super-resolution for facial image using multilateral affinity function,” *Neurocomputing*, vol. 133, p. 194–208, 2014. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0925231214000198

[36] M. Yamaguchi, K. Akiyama, T. Tsukagoshi, T. Muto, A. Kataoka, F. Tazaki, S. Ikeda, M. Fukagawa, M. Honma et R. Kawabe, “Super-resolution imaging of the protoplanetary disk hd 142527 using sparse modeling,” *The Astrophysical Journal*, vol. 895, n°. 2, p. 84, may 2020. [En ligne]. Disponible : https://dx.doi.org/10.3847/1538-4357/ab899f

[37] P. M. Harvey, J. D. Adams, T. L. Herter, G. Gull, J. Schoenwald, L. D. Keller, J. M. De Buizer, W. Vacca, W. Reach et E. E. Becklin, “First Science Results from SOFIA/-FORCAST : Super-resolution Imaging of the S140 Cluster at 37 μm,” *The Astrophysical Journal Letters*, vol. 749, n°. 2, p. L20, avr. 2012.

[38] L. Li, Q. Yu, Y. Yuan, Y. Shang, H. Lu et X. Sun, “Super-resolution reconstruction and higher-degree function deformation model based matching for Chang’E-1 lunar images,” *Science in China Series E : Technological Sciences*, vol. 52, n°. 12, p. 3468–3476, déc. 2009. [En ligne]. Disponible : https://doi.org/10.1007/s11431-009-0334-7

[39] Y. Huang, T. Miyazaki, X. Liu et S. Omachi, “Infrared Image Super-Resolution : Systematic Review, and Future Trends,” déc. 2022, arXiv :2212.12322 [cs, eess]. [En ligne]. Disponible : http://arxiv.org/abs/2212.12322

[40] K. Aizawa, T. Komatsu et T. Saito, “Acquisition of very high resolution images using

stereo cameras," dans *Visual Communications and Image Processing'91 : Visual Communication*, vol. 1605. SPIE, 1991, p. 318–328.

[41] S. Rhee et M. G. Kang, "Discrete cosine transform based regularized high-resolution image reconstruction algorithm," *Optical Engineering*, vol. 38, n°. 8, p. 1348–1356, 1999.

[42] H. Shen, L. Zhang, B. Huang et P. Li, "A map approach for joint motion estimation, segmentation, and super resolution," *IEEE Transactions on Image processing*, vol. 16, n°. 2, p. 479–490, 2007.

[43] S. Wang, L. Zhang, Y. Liang et Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," dans *2012 IEEE Conference on computer vision and pattern recognition*. IEEE, 2012, p. 2216–2223.

[44] J. Yang, Z. Wang, Z. Lin, S. Cohen et T. Huang, "Coupled dictionary training for image super-resolution," *IEEE transactions on image processing*, vol. 21, n°. 8, p. 3467–3478, 2012.

[45] H. Zhang, Y. Zhang et T. S. Huang, "Efficient sparse representation based image super resolution via dual dictionary learning," dans *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011, p. 1–6.

[46] C. Dong, C. C. Loy, K. He et X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, n°. 2, p. 295–307, 2015.

[47] C. Dong, C. C. Loy et X. Tang, "Accelerating the super-resolution convolutional neural network," dans *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, p. 391–407.

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

[49] J. Kim, J. K. Lee et K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, p. 1646–1654.

[50] X. Mao, C. Shen et Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," dans *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon et R. Garnett, édit., vol. 29. Curran Associates, Inc., 2016. [En ligne]. Disponible : https://proceedings.neurips.cc/paper_files/paper/2016/file/0ed9422357395a0d4879191c66f4faa2-Paper.pdf

[51] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI : IEEE, juill. 2017, p. 105–114. [En ligne]. Disponible : http://ieeexplore.ieee.org/document/8099502/

[52] R. Timofte, E. Agustsson, L. V. Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, X. Wang, Y. Tian, K. Yu, Y. Zhang, S. Wu, C. Dong, L. Lin, Y. Qiao, C. C. Loy, W. Bae, J. Yoo, Y. Han, J. C. Ye, J.-S. Choi, M. Kim, Y. Fan, J. Yu, W. Han, D. Liu, H. Yu, Z. Wang, H. Shi, X. Wang, T. S. Huang, Y. Chen, K. Zhang, W. Zuo, Z. Tang, L. Luo, S. Li, M. Fu, L. Cao, W. Heng, G. Bui, T. Le, Y. Duan, D. Tao, R. Wang, X. Lin, J. Pang, J. Xu, Y. Zhao, X. Xu, J. Pan, D. Sun, Y. Zhang, X. Song, Y. Dai, X. Qin, X.-P. Huynh, T. Guo, H. S. Mousavi, T. H. Vu, V. Monga, C. Cruz, K. Egiazarian, V. Katkovnik, R. Mehta, A. K. Jain, A. Agarwalla, C. V. S. Praveen, R. Zhou, H. Wen, C. Zhu, Z. Xia, Z. Wang et Q. Guo, "Ntire 2017 challenge on single image super-resolution : Methods and results," dans *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, p. 1110–1121.

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville et Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 06 2014.

[54] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 4681–4690.

[55] Y. Huang, Z. Jiang, R. Lan, S. Zhang et K. Pi, "Infrared Image Super-Resolution via Transfer Learning and PSRGAN," *IEEE Signal Processing Letters*, vol. 28, p. 982–986, 2021, conference Name : IEEE Signal Processing Letters.

[56] C. Szegedy, S. Ioffe, V. Vanhoucke et A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," dans *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 4278–4284.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser et I. Polosukhin, "Attention is all you need," dans *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, édit., vol. 30. Curran Associates, Inc.,

2017. [En ligne]. Disponible : https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit et N. Houlsby, "An image is worth 16x16 words : Transformers for image recognition at scale," dans *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [En ligne]. Disponible : https://openreview.net/forum?id=YicbFdNTTy

[59] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool et R. Timofte, "Swinir : Image restoration using swin transformer," dans *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, p. 1833–1844.

[60] X. Chen, X. Wang, J. Zhou, Y. Qiao et C. Dong, "Activating more pixels in image super-resolution transformer," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 22 367–22 377.

[61] D. Zhang, F. Huang, S. Liu, X. Wang et Z. Jin, "SwinFIR : Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution," sept. 2023, arXiv :2208.11247 [cs]. [En ligne]. Disponible : http://arxiv.org/abs/2208.11247

[62] L. Zhang, Y. Li, X. Zhou, X. Zhao et S. Gu, " Transcending the Limit of Local Window : Advanced Super-Resolution Transformer with Adaptive Token Dictionary ," dans *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA : IEEE Computer Society, juin 2024, p. 2856–2865. [En ligne]. Disponible : https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00276

[63] O. Ronneberger, P. Fischer et T. Brox, "U-net : Convolutional networks for biomedical image segmentation," dans *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells et A. F. Frangi, édit. Cham : Springer International Publishing, 2015, p. 234–241.

[64] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang et Z. Chen, "Diffusion models for image restoration and enhancement – a comprehensive survey," 2023. [En ligne]. Disponible : https://arxiv.org/abs/2308.09388

[65] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot et B. Wen, "Sinsr : Diffusion-based image super-resolution in a single step," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, p. 25 796–25 805.

[66] Z. Yue, J. Wang et C. C. Loy, "Resshift : Efficient diffusion model for image super-resolution by residual shifting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[67] Y. Ma, H. Yang, W. Yang, J. Fu et J. Liu, "Solving diffusion odes with optimal boundary conditions for better image super-resolution," dans *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. [En ligne]. Disponible : https://openreview.net/forum?id=BtT6o5tfHu

[68] J. Wang, J. F. Ralph et J. Y. Goulermas, "An analysis of a robust super resolution algorithm for infrared imaging," dans *2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*. IEEE, 2009, p. 158–163.

[69] Deng Cheng-Zhi, Tian Wei, Chen Pan, Wang Sheng-Qian, Zhu Hua-Sheng et Hu Sai-Feng, "Infrared image super-resolution via locality-constrained group sparse model," *Acta Physica Sinica*, vol. 63, nº. 4, p. 044 202–044 202, 2014. [En ligne]. Disponible : https://wulixb.iphy.ac.cn/en/article/doi/10.7498/aps.63.044202

[70] K. Fan, K. Hong et F. Li, "Infrared image super-resolution via progressive compact distillation network," *Electronics*, vol. 10, nº. 24, p. 3107, 2021.

[71] Y. Zou, L. Zhang, C. Liu, B. Wang, Y. Hu et Q. Chen, "Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections," *Optics and Lasers in Engineering*, vol. 146, p. 106717, 2021.

[72] Y. Yang, Q. Li, C. Yang, Y. Fu, H. Feng, Z. Xu et Y. Chen, "Deep networks with detail enhancement for infrared image super-resolution," *IEEE Access*, vol. 8, p. 158 690–158 701, 2020.

[73] F. Qin, K. Yan, C. Wang, R. Ge, Y. Peng et K. Zhang, "LKFormer : large kernel transformer for infrared image super-resolution," *Multimedia Tools and Applications*, févr. 2024. [En ligne]. Disponible : https://link.springer.com/10.1007/s11042-024-18409-3

[74] S. Liang, K. Song, W. Zhao, S. Li et Y. Yan, "DASR : Dual-Attention Transformer for infrared image super-resolution," *Infrared Physics & Technology*, vol. 133, p. 104837, sept. 2023. [En ligne]. Disponible : https://linkinghub.elsevier.com/retrieve/pii/S1350449523002955

[75] R. E. Rivadeneira, A. D. Sappa, B. X. Vintimilla, J. Kim, D. Kim, Z. Li, Y. Jian, B. Yan, L. Cao, F. Qi, H. Wang, R. Wu, L. Sun, Y. Zhao, L. Li, K. Wang, Y. Wang, X. Zhang, H. Wei, C. Lv, Q. Sun, X. Tian, Z. Jia, J. Hu, C. Wang, Z. Zhong, X. Liu et J. Jiang, "Thermal Image Super-Resolution Challenge Results - PBVS 2023," dans *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans, LA, USA : IEEE, juin 2022, p. 417–425. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/9857432/

[76] Y. Zou, L. Zhang, Q. Chen, B. Wang, Y. Hu et Y. Zhang, "An infrared image super-resolution imaging algorithm based on auxiliary convolution neural network," dans *Other Conferences*, 2020. [En ligne]. Disponible : https://api.semanticscholar.org/CorpusID:225966483

[77] A. Kasliwal, P. Seth, S. Rallabandi et S. Singhal, "CoReFusion : Contrastive Regularized Fusion for Guided Thermal Super-Resolution," dans *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, BC, Canada : IEEE, juin 2023, p. 507–514. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/10208919/

[78] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton et J. Dean, "Outrageously large neural networks : The sparsely-gated mixture-of-experts layer," dans *International Conference on Learning Representations*, 2017. [En ligne]. Disponible : https://openreview.net/forum?id=B1ckMDqlg

[79] R. S. Puttagunta, B. Kathariya, Z. Li et G. York, "Multi-scale feature fusion using channel transformers for guided thermal image super resolution," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, p. 3086–3095.

[80] H. Jiang et Z. Chen, "Flexible window-based self-attention transformer in thermal image super-resolution," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, p. 3076–3085.

[81] C. Cortés-Mendez et J.-B. Hayet, "Exploring the usage of diffusion models for thermal image super-resolution : A generic uncertainty-aware approach for guided and non-guided schemes," dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, p. 3123–3130.

[82] K. Bayoudh, R. Knani, F. Hamdaoui et A. Mtibaa, "A survey on deep multimodal learning for computer vision : advances, trends, applications, and datasets," *The Visual Computer*, vol. 38, n°. 8, p. 2939–2970, août 2022. [En ligne]. Disponible : https://link.springer.com/10.1007/s00371-021-02166-7

[83] S. Woo, S. Lee, Y. Park, M. A. Nugroho et C. Kim, "Towards good practices for missing modality robust action recognition," dans *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. [En ligne]. Disponible : https://doi.org/10.1609/aaai.v37i3.25378

[84] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu et X. Peng, "Smil : Multimodal learning with severely missing modality," *ArXiv*, vol. abs/2103.05677, 2021. [En ligne].

Disponible : https://api.semanticscholar.org/CorpusID:232170317

[85] Z. Wang, E. Simoncelli et A. Bovik, "Multiscale structural similarity for image quality assessment," dans *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.* Pacific Grove, CA, USA : IEEE, 2003, p. 1398–1402. [En ligne]. Disponible : http://ieeexplore.ieee.org/document/1292216/

[86] K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition," dans *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio et Y. LeCun, édit., 2015.

[87] A. Mittal, R. Soundararajan et A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, n°. 3, p. 209–212, 2013.

[88] "FREE - FLIR Thermal Dataset for Algorithm Training | Teledyne FLIR," accessed : 2024-10-03. [En ligne]. Disponible : https://www.flir.com/oem/adas/adas-dataset-form/

[89] K. He, J. Sun et X. Tang, "Guided image filtering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, p. 1397–1409, 06 2013.

[90] T. Hastie, R. Tibshirani et J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY : Springer, 2009. [En ligne]. Disponible : http://link.springer.com/10.1007/978-0-387-84858-7

[91] N. R. Draper et H. Smith, "Ridge Regression," dans *Applied Regression Analysis*. John Wiley & Sons, Ltd, 1998, p. 387–400, section : 17 _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118625590.ch17. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118625590.ch17

[92] A. Ignatov, R. Timofte, M. Denna, A. Younes, G. Gankhuyag, J. Huh, M. K. Kim, K. Yoon, H.-C. Moon, S. Lee, Y. Choe, J. Jeong, S. Kim, M. Smyl, T. Latkowski, P. Kubik, M. Sokolski, Y. Ma, J. Chao, Z. Zhou, H. Gao, Z. Yang, Z. Zeng, Z. Zhuge, C. Li, D. Zhu, M. Sun, R. Duan, Y. Gao, L. Kong, L. Sun, X. Li, X. Zhang, J. Zhang, Y. Wu, J. Pan, G. Yu, J. Zhang, F. Zhang, Z. Ma, H. Wang, H. Cho, S. Kim, H. Li, Y. Ma, Z. Luo, Y. Li, L. Yu, Z. Wen, Q. Wu, H. Fan, S. Liu, L. Zhang, Z. Zong, J. Kwon, J. Zhang, M. Li, N. Fu, G. Ding, H. Zhu, Z. Chen, G. Li, Y. Zhang, L. Sun, D. Zhang, N. Yang, F. Liu, J. Zhao, M. Ayazoglu, B. B. Bilecen, S. Hirose, K. Arunruangsirilert, L. Ao, H. C. Leung, A. Wei, J. Liu, Q. Liu, D. Yu, A. Li, L. Luo, C. Zhu, S. Hong, D. Park, J. Lee, B. H. Lee, S. Lee, S. Y. Chun, R. He, X. Jiang, H. Ruan, X. Zhang, J. Liu, G. Gendy, N. Sabor, J. Hou et G. He, "Efficient and accurate quantized image super-resolution on mobile npus, mobile ai & aim 2022 challenge : Report," dans *Com-*

puter Vision – ECCV 2022 Workshops, L. Karlinsky, T. Michaeli et K. Nishino, édit. Cham : Springer Nature Switzerland, 2023, p. 92–129.

[93] C. Peng, S. K. Zhou et R. Chellappa, "Da-vsr : Domain adaptable volumetric super-resolution for medical images," dans *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng et C. Essert, édit. Cham : Springer International Publishing, 2021, p. 75–85.

[94] D. Qiu, Y. Cheng et X. Wang, "Medical image super-resolution reconstruction algorithms based on deep learning : A survey," *Computer Methods and Programs in Biomedicine*, vol. 238, p. 107590, août 2023. [En ligne]. Disponible : https://linkinghub.elsevier.com/retrieve/pii/S0169260723002559

[95] N. Yakob, S. Laliberte, P. Doyon-Poulin, P. Jouvet et R. Noumeir, "Data representation structure to support clinical decision-making in the pediatric intensive care unit : Interview study and preliminary decision support interface design," *JMIR Formative Research*, vol. 8, 02 2024.

[96] V. Boivin, M. Shahriari, G. Faure, S. Mellul, E. D. Tiassou, P. Jouvet et R. Noumeir, "Multimodality video acquisition system for the assessment of vital distress in children," *Sensors*, vol. 23, n°. 11, p. 5293, 2023. [En ligne]. Disponible : https://doi.org/10.3390/s23115293

[97] J. Torra, F. Viela, D. Megías, B. Sot et C. Flors, "Versatile near-infrared super-resolution imaging of amyloid fibrils with the fluorogenic probe cranad-2," *Chemistry - A European Journal*, vol. 28, 02 2022.

[98] Y. Cao, G. L. Li, Y. K. Luo, Q. Pan et S. Y. Zhang, "Monitoring of sugar beet growth indicators using wide-dynamic-range vegetation index (wdrvi) derived from uav multispectral images," *Computers and Electronics in Agriculture*, vol. 171, p. 105331, 2020. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0168169920301861

[99] H. Qi, B. Zhu, Z. Wu, Y. Liang, J. Li, L. Wang, T. Chen, Y. Lan et L. Zhang, "Estimation of peanut leaf area index from unmanned aerial vehicle multispectral images," *Sensors*, vol. 20, n°. 23, 2020. [En ligne]. Disponible : https://www.mdpi.com/1424-8220/20/23/6732

[100] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao et C. Change Loy, "Esrgan : Enhanced super-resolution generative adversarial networks," dans *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, p. 0–0.

[101] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser et I. Polosukhin, "Attention is all you need," *Advances in neural information processing*

systems, vol. 30, 2017.

[102] H. Zhang, Y. Hu et M. Yan, “Thermal Image Super-Resolution Based on Lightweight Dynamic Attention Network for Infrared Sensors,” *Sensors*, vol. 23, n°. 21, p. 8717, oct. 2023. [En ligne]. Disponible : https://www.mdpi.com/1424-8220/23/21/8717

[103] L. Chen, X. Chu, X. Zhang et J. Sun, “Simple baselines for image restoration,” dans *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella et T. Hassner, édit. Cham : Springer Nature Switzerland, 2022, p. 17–33.

[104] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong et Z. Luo, “Target-aware Dual Adversarial Learning and a Multi-scenario Multi-Modality Benchmark to Fuse Infrared and Visible for Object Detection,” dans *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA : IEEE, juin 2022, p. 5792–5801. [En ligne]. Disponible : https://ieeexplore.ieee.org/document/9879642/

[105] L. Zhou et S. Feng, “A review of deep learning for single image super-resolution,” dans *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, 2019, p. 139–142.

[106] N. Genser, J. Seiler et A. Kaup, “Camera Array for Multi-Spectral Imaging,” *IEEE Transactions on Image Processing*, vol. 29, p. 9234–9249, 2020, conference Name : IEEE Transactions on Image Processing.

[107] S. Klein, M. Staring, K. Murphy, M. Viergever et J. Pluim, “Elastix : A Toolbox for Intensity-Based Medical Image Registration,” *IEEE Transactions on Medical Imaging*, vol. 29, n°. 1, p. 196–205, janv. 2010. [En ligne]. Disponible : http://ieeexplore.ieee.org/document/5338015/

[108] A. H. Ornek, M. Ceylan et S. Ervural, “Health status detection of neonates using infrared thermography and deep convolutional neural networks,” *Infrared Physics & Technology*, vol. 103, p. 103044, déc. 2019. [En ligne]. Disponible : https://linkinghub.elsevier.com/retrieve/pii/S1350449519303123

ANNEXE A EXEMPLE DE SUPER-RÉSOLUTION DANS UN CONTEXTE MÉDICAL [2]

En abscisse, différentes méthodes de super-résolution et en ordonnée, les échelles de super-résolution.

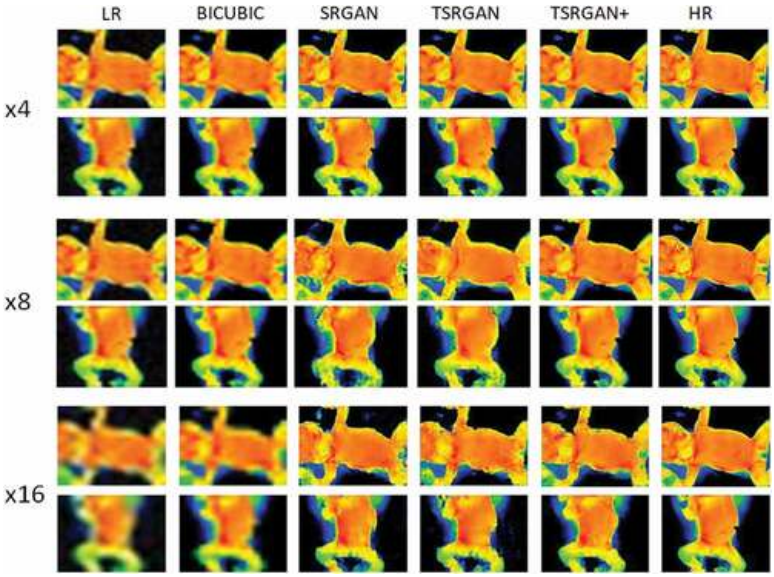


FIGURE A.1 Résultat de super résolution d'images IR de nouveau-nés

ANNEXE B EXEMPLES D'ALGORITHMES DE BLIND SR POUR L'INFÉRENCE À PARTIR D'UNE IMAGE BASSE RÉSOLUTION

Algorithm 1 The overall inference procedure

Input: The LR image to be reconstructed: $\mathbf{y}_{test}$
The external HR image dataset: $\{\mathbf{x}_i\}_{i=1}^{N}$
SR model finetuning step: T
Output: The SR image: $\mathbf{x}_{test}$

// Lines 1-3: Estimate the degradation parameters
1: $\sigma_x, d_x, \sigma_y, d_y = \mathcal{E}_b(\mathbf{y}_{test})$ *// Blur estimation*
2: $\sigma_g, d_g, \sigma_p, d_p = \mathcal{E}_n(\mathbf{y}_{test})$ *// Noise estimation*
3: $q, d_q = \mathcal{E}_j(\mathbf{y}_{test})$ *// JPEG compression estimation*

// Lines 4-11: Synthesize LR images via the degradation model
4: **for** $i = 1 : N$ **do** *// Degradation parameter sampling*
5: $\sigma_{x_i} \sim \mathcal{U}(\sigma_x - d_x, \sigma_x + d_x)$
6: $\sigma_{y_i} \sim \mathcal{U}(\sigma_y - d_y, \sigma_y + d_y)$
7: $\sigma_{g_i} \sim \mathcal{U}(\sigma_g - d_g, \sigma_g + d_g)$
8: $\sigma_{p_i} \sim \mathcal{U}(\sigma_p - d_p, \sigma_p + d_p)$
9: $q_i \sim \mathcal{U}(q - d_q, q + d_q)$
10: $\mathbf{y}_i = [(\mathbf{x}_i \otimes \mathbf{k}(\sigma_{x_i}, \sigma_{y_i})) \downarrow_s + \mathbf{n}(\sigma_{g_i}, \sigma_{p_i})]_{\text{JPEG}(q_i)}$
11: **end for**

// Lines 12-17: Finetune the SR model
12: $j = 0$
13: **while** $j \leq T$ **do**
14: $j \leftarrow j + 1$
15: $\mathcal{L}_{SR} = \frac{1}{n}\sum_{m=1}^{n}|\mathbf{x}_m - \mathcal{G}(\mathbf{y}_m; \theta_G)|$
16: $\theta_G \leftarrow \theta_G - \nabla_{\theta_G}\mathcal{L}_{SR}$
17: **end while**
18: **return** $\mathbf{x}_{test} = \mathcal{G}(\mathbf{y}_{test})$

FIGURE B.1 Algorithme d'inférence proposée par [21]

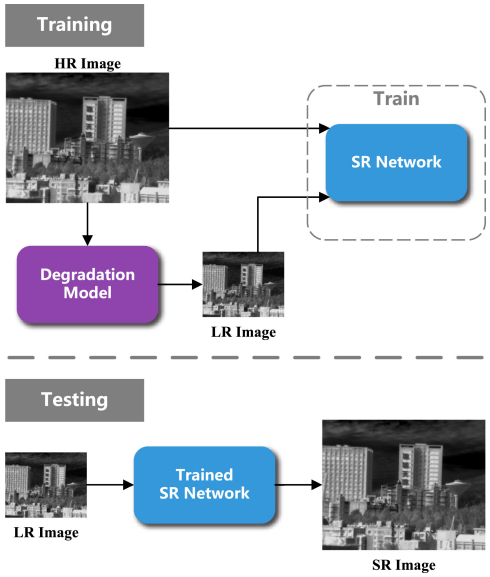


FIGURE B.2 Algorithme d'inférence proposée par [10]

ANNEXE C EXEMPLE DE SUPER RÉSOLUTION GUIDÉE ×8 ET ×16 [17]

La ligne du haut correspond à une SR de ×8 et celle du bas à une SR de ×16.
La première colonne correspond aux images RGB haute résolution (image guide). La deuxième colonne correspond aux images IR basse résolution. La 3ème correspond à l'image haute résolution après application de l'algorithme de SR. La 4ème colonne correspond à la vraie image IR haute résolution.

FIGURE C.1 Résultats de super résolution d'images IR issues de l'architecture gagnante de la compétition [17]

ANNEXE D PROTOCOLE ÉTHIQUE ACCEPTÉ À SAINTE JUSTINE

Les pages suivantes décrivent le protocole éthique proposé et validé pour l'acquisition des images présentées au chapitre 6, un processus qui a débuté en juin 2024.

PROTOCOLE DE RECHERCHE

Création d'une base de données vidéos d'enfants hospitalisés pour l'estimation de pose et la super résolution d'images thermiques

Investigateur principal :

Philippe JOUVET, MD PhD
Directeur de l'unité de recherche en soins critiques respiratoires du Centre de Recherche du CHU Sainte-Justine.
3175 Chemin de Côte Sainte Catherine
Montréal (Québec) H3T 1C5
Tel : 514 345 4927
Fax : 514 345 7731
Courriel : philippe.jouvet@umontreal.ca

Co-investigateurs :

Lama SEOUD, PhD
Chercheure, Centre de recherche du CHU Sainte-Justine
Prof. adjointe, Département de génie informatique et génie logiciel, Polytechnique Montréal
2500 chemin de Polytechnique
Montréal, QC, H3T 1J4
Tél : 514-340-4711 poste 3699
Courriel : lama.seoud@polymtl.ca

Collaborateurs :

- Hugo Rodet,
  Étudiant au doctorat, Département de génie informatique et logiciel, Polytechnique Montréal
- Cyprien Arnold,
  Étudiant à la maitrise, Département de génie informatique et logiciel, Polytechnique Montréal

- Aya Chetto,
  Stagiaire de recherche à Polytechnique Montreal

RÉSUMÉ DU PROTOCOLE

L'unité des Soins Intensifs Pédiatriques (USIP) fournit des soins aux patients en état grave, présentant des maladies mettant en jeu leur pronostic vital. Cette prise en charge nécessite notamment une surveillance étroite de leurs paramètres vitaux, en continu, 24 heures sur 24. Dans le cadre de la surveillance des signes cliniques d'ordre neurologique, une attention particulière est portée sur l'estimation en temps réel de la pose des patients, c'est-à-dire l'analyse des mouvements en 3 dimensions (3D) de la tête et des membres.

Les algorithmes d'estimation de pose humaine les plus performants à ce jour nécessitent de larges bases de données annotées pour entrainer un réseau de neurones profond. Cependant, les bases de données existantes présentent pour la plupart des adultes réalisant des tâches de la vie quotidienne en position debout ou assise. Par des expérimentations préliminaires de notre équipe, nous avons relevé la faible généralisation de ces algorithmes aux vidéos de jeunes patients alités à l'USIP. Ceci s'explique notamment par la variabilité de l'âge des patients, allant de 0 à 18 ans pour la population pédiatrique, entrainant une grande variabilité de la morphologie des enfants en fonction de leur taille et de leur poids. L'utilisation de matériel médical (couverture, équipement…) peut aussi entrainer des difficultés quant à l'analyse de ces vidéos.

Pour permettre l'estimation de pose, notamment de personnes sous des couvertures, il est nécessaire d'utiliser l'information des images thermiques. Cependant, l'utilisation de caméra thermique haute résolution comme celle que nous envisageons d'utiliser (FLIR T1020) coute très chèr, de l'ordre de 50000$ CAD. Pour une utilisation à grande échelle dans l'hôpital, il semble peu réaliste d'utiliser ce type de caméra très dispendieuse. Pour remédier à ce problème, il est possible d'utiliser des caméras thermiques basse résolution puis d'utiliser un algorithme de super-résolution pour augmenter la qualité de l'image. En effet, les caméras thermiques basse résolution sont beaucoup plus économiques, de l'ordre de quelques centaines de dollars. Donc, en plus de pouvoir aider à entrainer l'algorithme d'estimation de pose, la base de données sera aussi utilisée pour entrainer l'algorithme de super résolution à reconstruire une image thermique haute résolution.

Face à ces limitations, et devant le manque d'études sur la gamme de mouvements des patients pédiatriques alités, nous jugeons nécessaire et pertinente la création d'une base de données vidéos annotées en 3D étudiant les patients de l'USIP du CHU. Cette étude est en lien étroit avec la création de la « basede données vidéo des détresses vitales de l'enfant » (appelée MEDEVAC) menée actuellement par les Professeurs Noumeir et Jouvet, le système d'acquisition de vidéos et images infrarouges étant le même, mis à part le capteur infrarouge qui est remplacé par une caméra infrarouge haute résolution. Elle se distingue de l'étude précédente par le fait qu'on n'utilise plus de capteurs inertiels pour l'annotation de pose. Les patients ciblés dans la présente étude sonten surveillance de fin de séjour à l'USIP donc ne sont pas dans un état critique.

Objectif principal : L'objectif principal de notre travail est de développer une base de données vidéo de patients de moins de 18 ans, alités, pour bâtir et évaluer des algorithmes d'estimation de pose humaine et de suivi du mouvement en 3D ainsi que des algorithmes de super-résolution d'images thermiques. En se concentrant, dans cette étude, sur des patients qui sont dans les 24 dernières heures de surveillance de leur séjour à l'USIP, nous ciblons des mouvements « non-critiques» de patients en état stable dans leur lit d'hôpital.

Critère d'éligibilité : tous les patients de moins de 18 ans admis aux soins intensifs du CHU Sainte Justine du 1er novembre 2023 au 30 Juin 2026.

Critères d'inclusion :

1) Patients avec sortie des soins intensifs prévue dans les 24 heures selon le médecin traitant (en surveillance) étant soignés au CHU Sainte Justine
2) Consentement signé

Critères d'exclusions :

1) Matériel d'enregistrement vidéo non disponibles.

Durée de participation d'un patient : 10 minutes d'enregistrement et 10 minutes de préparation
Matériel :

1)Montage similaire au protocole MEDEVAC existant, intégrant une caméra de profondeur Kinect Azure (utile pour le calcul des volumes pulmonaires, l'estimation de pose, etc.) et une caméra infrarouge FLIR T1020 (utile pour le calcul de thermographie, pour l'estimation de pose, etc. en haute résolution) relié à un ordinateur portatif pour la capture des vidéos.

Centre participant : CHU Sainte-Justine

Promoteur	CHU Sainte-Justine
Titre	Création d'une base de données vidéos d'enfants hospitalisés pour l'estimation de pose et le suivi du mouvement 3D
Investigateur principal	Dr. Philippe JOUVET. Co-investigatrice : Lama SEOUD
Collaborateurs	Hugo RODET Cyprien ARNOLD
Comité de la direction de la base des données	
Version du protocole	2 novembre 2023
Objectif	Développer une base de données vidéo de patients de moins de 18 ans, alités, pour bâtir et évaluer des algorithmes d'estimation de pose humaine et de suivi du mouvement 3D et la super-résolution d'images thermiques.
Méthodologie	Création d'une base de données vidéos (RGB-D et infrarouge) pour l'estimation de pose et le suivi du mouvement chez les jeunes patients aux soins intensifs, ainsi que la super-résolution d'images thermiques. Pour l'acquisition des images et vidéos, nous utiliserons un montage existant, similaire à celui déjà utilisé dans le protocole MEDEVAC.

1. Justification de la recherche

L'Unité des Soins Intensifs Pédiatriques (USIP) fournit des soins aux patients en état grave, présentant des maladies mettant en jeu leur pronostic vital. Cette prise en charge nécessite notamment une surveillance étroite de leurs paramètres vitaux, en continu, 24 heures sur 24. Dans le cadre de la surveillance des signes cliniques d'ordre neurologique, une attention particulière est portée sur l'estimation en temps réel de la pose des patients, c'est-à-dire l'analyse des mouvements en 3 dimensions (3D) de la tête et des membres.

L'estimation de pose humaine est un sujet très actif en vision par ordinateur et de nombreux algorithmes ont été proposés au cours des 20 dernières années. Les algorithmes les plus performants à ce jour reposent sur de larges bases de données annotées utilisées pour entrainer un réseau de neurones profond [1] à estimer la pose humaine dans de nouvelles images/vidéos, en maximisant le pouvoir de généralisation à de nouvelles données. Les bases de données en question, comme dans [2] et [3], portent en grande majorité sur la population adulte, et incluent des positions et mouvements relatifs aux activités de la vie quotidienne, comme la marche, la course, la danse, etc. Lors de ces activités, les patients sont debout.

À l'USIP, les patients sont souvent allongés dans un lit d'hôpital, une couverture les recouvrant au moins partiellement. De plus, on retrouve chez ces patients une grande variabilité en termes d'âge (0 à 18 ans) et donc en termes de taille et de proportions morphologiques plus généralement. Notre équipe a conduit des tests moyennant des algorithmes d'estimation de pose 3D offerts par des caméras RGB-D commerciales, comme la Kinect Azure, sur des images de patients à l'USIP. Les résultats ont démontré une très mauvaise estimation de pose 3D sur les enfants alités. Notre explication est que l'algorithme intégré à la Kinect offre un pouvoir de généralisation très faible dans notre contexte particulier.

Il existe à notre connaissance uniquement deux bases de données publiques qui se rapprochent de notre contexte. La première (intitulée SLP) comporte des images RGB-D et en infrarouge de patients alités [4]. Toutefois les patients sont uniquement des adultes. La deuxième (intitulée MINI-RGBD) comporte des images RGB-D de nouveaux nés [5], cependant les images ont été obtenues de manière synthétique, ce qui constitue un frein à la performance des algorithmes entrainés sur cette base [6] lorsqu'appliqués à des images réelles. De plus, les annotations offertes dans ces bases de données sont en 2D, c'est-à-dire que la troisième dimension des articulations et extrémités est inconnue. La 3^{e} dimension permet de lever l'ambiguïté de position dans le cas d'auto-occlusion, comme un patient couché sur le côté par exemple.

Dans cet article [7], les auteurs utilisent des images thermiques néonatales de bébés pour détecter d'éventuelles maladies. Ils montrent qu'en utilisant un algorithme de super-résolution, ils arrivent à passer de 90 à 99% de bonne classification de maladies. Alors que cet article montre l'utilité de la super résolution dans un contexte médical, ce projet proposera une méthode inédite puisqu'elle porte sur de l'estimation de pose d'images thermiques d'enfants. Ils ont acquis une base de données d'images thermiques bébés d'une manière analogue à celle qu'on souhaiterait procéder. Malheureusement, leur base de données n'est pas publique pour des raisons de confidentialités.

Pour pouvoir entrainer les modèles utiles à l'estimation de pose ainsi qu'à la super résolution, nous jugeons nécessaire et pertinente la création d'une base de données vidéos en 3D, propre au contexte de l'USIP.

Cette étude est en lien étroit avec la création de la « base de données vidéo des détresses vitales de l'enfant » (appelée MEDEVAC) menée actuellement par Dr Philippe Jouvet et Prof. Rita Noumeir,

puisque le système de caméras 3D est le même. Toutefois, dans la présente étude, nous remplacerons le capteur infrarouge Lepton par une caméra infrarouge haute résolution.

2. Objectif

L'objectif principal de ce projet de recherche est de développer une base de données d'images et de vidéos de patients de moins de 18 ans hospitalisés dans le service des soins intensifs pédiatriques du à CHU Sainte Justine, dans le but d'analyser les mouvements des jeunes patients alités.

Cette base de données permettra l'entrainement d'algorithmes de super-résolution et d'estimation de pose d'enfants alités.

En se concentrant dans cette étude sur des patients en état stable en fin d'hospitalisation en USIP, nous pourrons caractériser les mouvements dits « normaux » (par opposition à ceux retrouvés lors de détresses vitales) de cette population dans le contexte de l'USIP.

3.Méthodologie

1.1. Éligibilité

Tous les patients de moins de à 18 ans, admis consécutivement dans le service de soins intensifs du CHU Sainte Justine du 1er novembre 2023 au 30 Juin 2026.

1.2. Critères d'inclusion :

1) Patients soignés à l'USIP avec sortie des soins intensifs dans les 24 heures selon le médecin traitant
2) Consentement signé

1.3. Critères d'exclusion :

1) Matériel d'enregistrement vidéo non disponible.

1.4. Lieu de réalisation de la recherche

L'étude se déroulera dans les services de soins intensifs pédiatriques du CHU Sainte Justine de Montréal.

1.5. Consentement

Concernant l'information des patients et de leurs parents, une affiche présentant l'étude sera placée dans le salon des parents. Une assistante de recherche approchera les parents pour recueillir le consentement dès l'admission de l'enfant dans l'unité.

2. Déroulement pratique de la recherche

2.1. Environnement d'acquisition

L'acquisition de données sera menée dans une chambre du service des soins intensifs de l'Hôpital Saint-Justine dont les dimensions sont approximativement égales à 6 x 5 mètres. Chaque acquisition sera effectuée dans les conditions d'éclairage normal (lumière ambiante). Un unique montage contenant une caméra RGB et de profondeur (utile pour le calcul des volumes pulmonaires, l'estimation de pose, etc.) et une caméra infrarouge (utile pour le calcul de thermographie, pour l'estimation de pose,etc.) sera utilisé. Il pourra être placé soit en tête de lit avec un angle de 45°, soit directement au plafond au-dessus du lit grâce à un montage fixe. Les caméras seront coordonnées par un ordinateurportable.

2.2. Matériel d'acquisition

4.2.1. Caméra Kinect Azure

La caméra de profondeur utilisée est la caméra Azure Kinect de la société Microsoft. Un aperçu global de ses caractéristiques est donné ci-dessous. Dans notre cas d'application, nous utiliserons la configuration Near Field Of View (NFOV) unbinned.

Dimensions	103x39x126 (mm) – 440g
Résolution	512 x 512 (px)
Champ de Vision	75° x 65°
Profondeur de vue	0.25 – 6.86 (m)
Vitesse d'acquisition	30 FPS

Tableau *1* - Propriétés Hardware de la Caméra Kinect Azure

4.2.2. Caméra infrarouge FLIR T1020

Dimensions	167x204x188 (mm) – 2100 g
Résolution	1024 x 768 (px)
Champ de Vision	45° x 34°
Plage de température	-40 – 2000 (°C)
Vitesse d'acquisition	30 FPS

2.3. Scénario d'acquisition

4.3.1. Personnes présentes

Toute acquisition sera effectuée en présence :

- Du patient
- D'au moins un membre du corps médical. Cette personne devra être présente au début et à la fin de l'acquisition, mais pourra quitter la salle entre temps si elle le souhaite.
- D'au moins un des membres inscrits en tête de ce rapport (Investigateur Principal, Co-Investigateur ou Collaborateurs). Cette personne restera dans la salle durant toute l'acquisition. Dans la suite, nous réfèrerons à cette personne comme un « membre du corps de recherche ».

4.3.2. Mise en place du matériel

Le support contenant les caméras pourra être placé soit en tête de lit avec un angle de 45°, soit directement au plafond au-dessus du lit grâce à un montage fixe. Afin d'assurer une plus grande variabilité dans notre base de données, nous alternerons les vues. N'ayant aucun contact avec le patient, il pourra être mis en place par le membre du corps de recherche.

4.3.3. Calibration du matériel

Les caméras étant fixes les unes par rapport aux autres, la calibration de leurs paramètres intrinsèques et extrinsèques sera effectuée une unique fois, en amont de toutes les acquisitions. Afin de pallier les très légers déplacements des caméras dans le support au cours du temps, la calibration pourra être refaite toutes les 100 acquisitions.

4.3.4. Durée d'acquisition

Chaque acquisition aura une durée continue entre 5 et 10 minutes avec une fréquence d'échantillonnage de 30 images par seconde dans chacune des 2 modalités (RGB et IR).

3. Taille de l'échantillon

L'objectif principal de ce travail est la constitution d'une base de données vidéos des patients

avec une très haute précision des coordonnées 3D des points clés des patients. Du fait du nombre important de travaux en estimation de pose chez les adultes, nos acquisitions se focaliseront donc sur les enfants en bas âge. Le tableau ci-dessous inclut le nombre d'acquisitions que nous souhaitons réaliser en fonction de la tranched'âge.

Tranche d'Age	0 – 2 ans	2 ans – 6 ans	6 ans – 12 ans	12 ans +
Nombre d'acquisitions	100	100	100	100

Tableau *4* - Nombre de séquences enregistrées et de capteurs placés en fonction de la tranche d'âge

4. Modalité de recueil des données

L'ensemble des données cliniques sera recueilli grâce au dossier clinique informatisé (ICCA) disponible dans l'unité des soins intensifs. Les séquences vidéo seront classifiées et stockées dans une base de données sécurisée par un mot de passe.

4.1. Données recueillies

Nous recueillerons l'ensemble des caractéristiques cliniques du patient, à savoir :

- L'âge de l'enfant
- Le poids et la taille
- Le motif de l'hospitalisation
- - Le diagnostic principal
- En vue d'une reconstruction précise du squelette du patient, nous mesurerons et stockerons la taille des différents membres, à savoir :
- - Membres inférieurs : Cuisses et Jambes (Droite et Gauche)
- - Membres supérieurs : Bras et Avant-Bras (Droite et Gauche)
- - Tronc
- - Tête
- En anticipation des travaux futurs qui viseront à détecter une détresse neurologique chez le patient, nous relèverons les informations neurologiques suivantes :
- - Le score de Glasgow
- - La présence ou non d'un déficit focal (ex : hémiplégie)
- - La présence ou non de convulsions (locales ou généralisées).

5. Considérations éthiques

5.1. Bénéfices, risques et contraintes de l'étude pour le patient

Cette étude observationnelle ne comporte aucun risque ni contrainte pour le patient. Il n'y aura aucune modification de la prise en charge du patient, aucun examen complémentaire. Les vidéos recueillies seront codées. Les vidéos ont pour but la création de la base de données à des fins de recherche et d'enseignement seulement.

5.2. Bénéfices, risques et contraintes de l'étude pour l'établissement

Il n'y a pas de coût supplémentaire pour l'établissement.

5.3. Mesures de protection de la base de données

Dans un souci de respect et de confidentialité des données médicales recueillies et de fiabilité des résultats des études réalisées, un comité de direction dirigé par le chercheur principal sera responsable de la base de données.

La base de données sera équipée d'un système de sécurité visant à :

- ✓ Empêcher toute personne non autorisée d'accéder aux installations,
- ✓ Empêcher l'utilisation des données par des personnes ou organismes non autorisés par le CÉR,
- ✓ Empêcher que des supports de données puissent être lus, copiés, modifiés ou déplacés par une personne non autorisée,
- ✓ Empêcher l'introduction non autorisée de données dans le système informatique.

Le CER sera informé de toute modification des conditions d'utilisation de la base ou de réévaluation des critères de sélection des sujets. Ces modifications ne seront applicables par l'équipe de recherche qu'après obtention de l'approbation du CÉR.

Chaque étude faite à partir de cette base de données devra préalablement être approuvée par le Comité de direction de la base des données. L'étude fera ensuite l'objet d'une soumission au comité d'éthique et justifiera l'élaboration d'une convention d'utilisation de la base. Par cette convention, les chercheurs s'engageront à conserver la confidentialité des données dans les conditions définies préalablement à la constitution de la base.

Les résultats de recherche pourront être diffusées, publiées ou communiquées dans un congrès scientifique sous une forme telle qu'il est impossible de les attribuer à un individu en particulier.

6. Faisabilité

6.1. Recrutement et expérience de l'équipe

L'unité des soins intensifs admet plus de 1000 patients par an âgés de 0 à 18 ans. La majeure partie (> 80%) des patients présentent une ou plusieurs défaillances vitales. L'équipe médicale du service de soins intensifs pédiatriques du CHU Sainte Justine est une équipe internationalement reconnue pour son expérience et la qualité de la prise en charge des enfants aux soins intensifs et pour la conduite de recherches cliniques dans ce domaine. Elle dispose de l'ensemble de l'infrastructure pour la conduite de la recherche, le CCEsAm du CHU Sainte-Justine, qui est une équipe multidisciplinaire, constituée de médecins, chargés de projets, assistants de recherche, coordonnateurs, informaticiens et personnel de soutien. Elle possède une expertise reconnue dans le domaine du développement de systèmes d'aide à la décision, ainsi qu'une expérience dans notre contexte spécifique via le protocole MEDEVAC.

P. Jouvet exerce au sein de l'équipe médicale du service de soins intensifs du CHU Sainte Justine. Son travail sur ce projet s'inscrit dans le cadre de son activité quotidienne dans le service et dans l'équipe de recherche des soins intensifs. Son implication dans ce travail ne justifie aucun salaire complémentaire.

L. Séoud est professeure adjointe au département de génie informatique et génie logiciel à Polytechnique Montréal et chercheure au Centre de recherche du CHU Sainte-Justine. Son expertise porte sur l'apprentissage profond et la visionique appliquée au domaine biomédical. Son implication dans ce travail ne justifie aucun salaire complémentaire.

Perspectives

Une fois le modèle d'estimation de pose humaine 3D entrainé sur la base de données ainsi collectée, nous pourrons déployer le modèle en temps réel à l'USIP, avec uniquement le montage des caméras 3D et infrarouge.

Le modèle sera entrainé à reconnaître des poses 3D et des mouvements relatifs à des enfants alités mais qui ne sont pas en situation de détresse vitale au moment de l'acquisition. Cela nous permettra d'étudier et de caractériser les mouvements "non-critiques", pour pouvoir alimenter un système d'aide à la décision qui utilisera les données des caméras 3D et infrarouge pour détecter des mouvements atypiques, critiques, indiquant possiblement une détresse neurologique.

7. Budget

Le projet nécessitera de payer les assistants de recherche pour le temps de recueil des consentements. Ce coût est estimé à 2,200$/an et sera assuré par les fonds obtenus par le Dr Jouvet dans le cadre de la chaire en intelligence artificielle et santé numérique.

RÉFÉRENCES

[1] J. Martinez, R. Hossain, J. Romero, et J. J. Little, « A Simple yet Effective Baseline for 3D Human Pose Estimation », 2017, p. 2640-2649. Consulté le: juin 01, 2021. [En ligne]. Disponible sur: https://openaccess.thecvf.com/content_iccv_2017/html/Martinez_A_Simple_yet_ICCV_2017_paper.html

[2] C. Ionescu, D. Papava, V. Olaru, et C. Sminchisescu, « Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, nº 7, p. 1325-1339, juill. 2014, doi: 10.1109/TPAMI.2013.248.

[3] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, et G. Pons-Moll, « Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera », in *Computer Vision – ECCV 2018*, vol. 11214, V. Ferrari, M. Hebert, C. Sminchisescu, et Y. Weiss, Éd. Cham: Springer International Publishing, 2018, p. 614-631. doi: 10.1007/978-3-030-01249-6_37.

[4] S. Liu, X. Huang, N. Fu, C. Li, Z. Su, et S. Ostadabbas, « Simultaneously-Collected Multimodal Lying Pose Dataset: Towards In-Bed Human Pose Monitoring under Adverse Vision Conditions », *ArXiv200808735 Cs*, août 2020, Consulté le: juin 01, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2008.08735

[5] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, et A. S. Schroeder, « Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D sequences », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, nº 10, p. 2540-2551, oct. 2020, doi: 10.1109/TPAMI.2019.2917908.

[6] X. Huang, N. Fu, S. Liu, K. Vyas, A. Farnoosh, et S. Ostadabbas, « Invariant Representation Learning for Infant Pose Estimation with Small Data », *ArXiv201006100 Cs*, déc. 2020, Consulté le: avr. 26, 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/2010.06100

[7] F. M. Senalp et M. Ceylan, « A new approach for super-resolution and classification applications on neonatal thermal images », Quant. InfraRed Thermogr. J., p. 1-18, févr.2023, doi:10.1080/17686733.2023.2179282.

ANNEXE E EXEMPLE DE SUPER-RÉSOLUTION POUR L'ESTIMATION DE POSE

L'idée derrière cette méthode (voir Figure E.1) n'est pas seulement d'entraîner l'architecture de super-résolution pour améliorer la qualité de l'image, mais également d'orienter l'entraînement afin d'optimiser l'estimation de pose à partir de l'image thermique. Pour cela, il serait possible de geler les poids de l'algorithme d'estimation de pose et d'ajouter un terme lié à cette estimation dans la fonction de perte utilisée lors de l'entraînement. Ainsi, le modèle serait également optimisé pour améliorer l'estimation de pose humaine, ce qui est pertinent si cette tâche est notre objectif final.

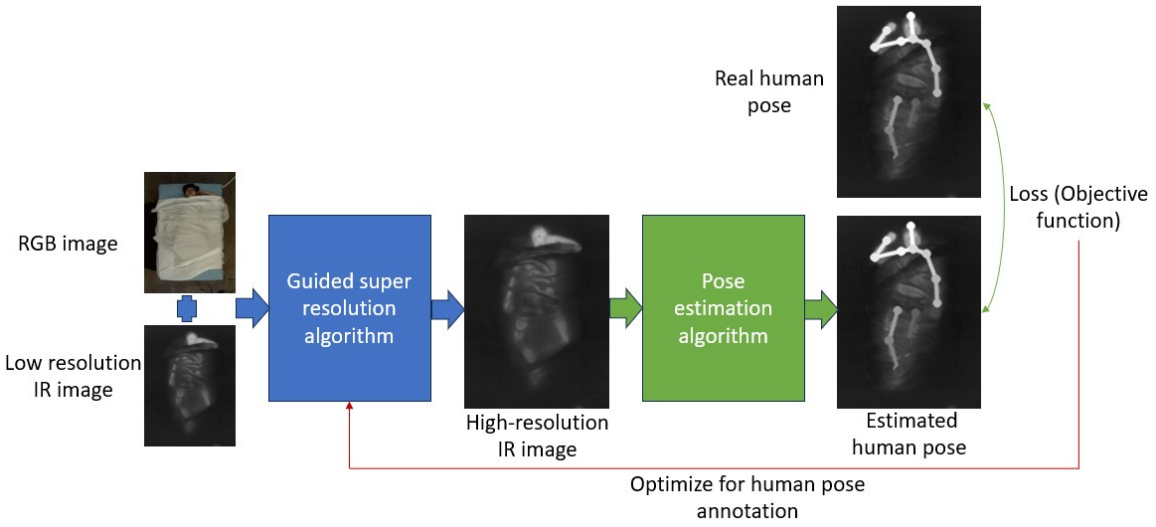


FIGURE E.1 Schéma de l'entraînement d'une super-résolution pour l'estimation de pose

ANNEXE F PROJET UNDERCOVER : UTILISER L'IMAGERIE INFRAROUGE POUR CRÉER UNE IMAGE RGB SANS COUVERTURE

A l'unité de soins intensifs, les patients nécessitent une surveillance continue via des moniteurs au chevet, complétée par des observations humaines intermittentes pour détecter des signes d'aggravation tels que la douleur, l'agitation, ou les changements de conscience. Une première problématique se pose pour l'estimation de pose humaine. Comment estimer la pose d'un enfant en présence d'occlusions visuelles, comme par exemple lorsqu'un patient est couvert par une couverture ?

Aussi, les systèmes de vision par ordinateur peuvent capturer des données sensibles (comme les visages), ce qui pose des problèmes de confidentialité. L'enjeu est de développer des systèmes de vision capables d'éviter la collecte d'informations visuelles détaillées et privées tout en extrayant des données utiles, comme la pose ou les mouvements, pour comprendre les événements en cours (exemple : détection de motifs de mouvement critiques).

De ces deux problématiques est né le projet Undercover : produire une méthode qui permet de résoudre les problèmes d'occlusion, tout en préservant l'anonymat des patients.

Le schéma décrivant l'objectif du projet avec des images issues de SLP est présenté à la Figure F.1.

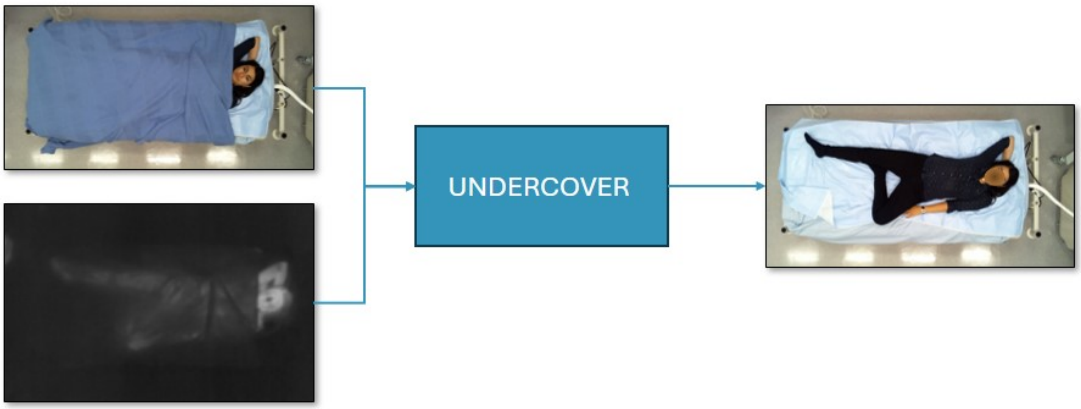


FIGURE F.1 Schéma de l'objectif du projet Undercover avec les images de SLP

Pour atteindre cet objectif, une première approche a été de s'inspirer de SwinFusion [12]. Les entrées de l'algorithme ainsi que la fonction de perte (loss function) ont été modifiées pour s'adapter à cet objectif. Le jeu de données SLP a été utilisé, car il fournit des images

RGB et IR d'une personne sous couverture (entrées de l'algorithme) ainsi que l'image RGB correspondante sans la couverture (image cible).
La Figure F.2 présente l'architecture de la solution proposée.

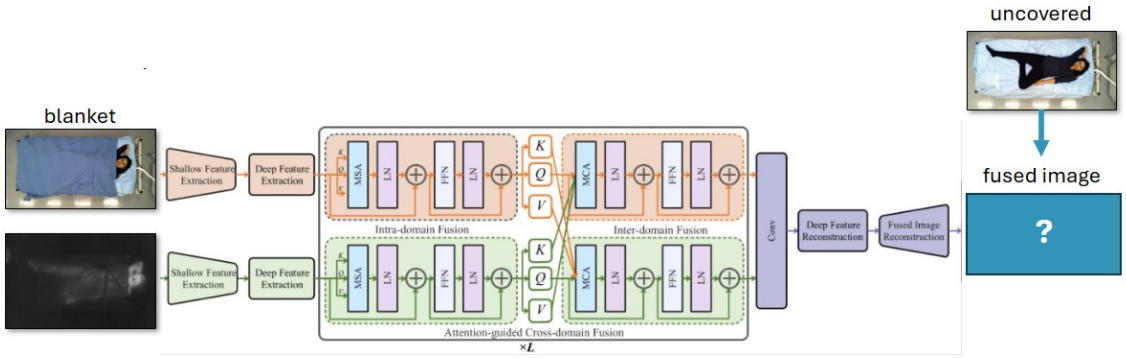


FIGURE F.2 Architecture proposée pour le projet Undercover

L'ensemble de données est composé de 102 patients, chacun réalisant 45 poses différentes, ce qui constitue un total de $102 \times 45 = 4590$ échantillons. Pour des raisons de consommation mémoire de la GPU, toutes les images sont redimensionnées à la taille 64×128. Le jeu de données a été divisé en 70% pour l'entraînement, et 15% pour la validation et le test. Les résultats pour une couverture fine sont présentés en Figure F.3.

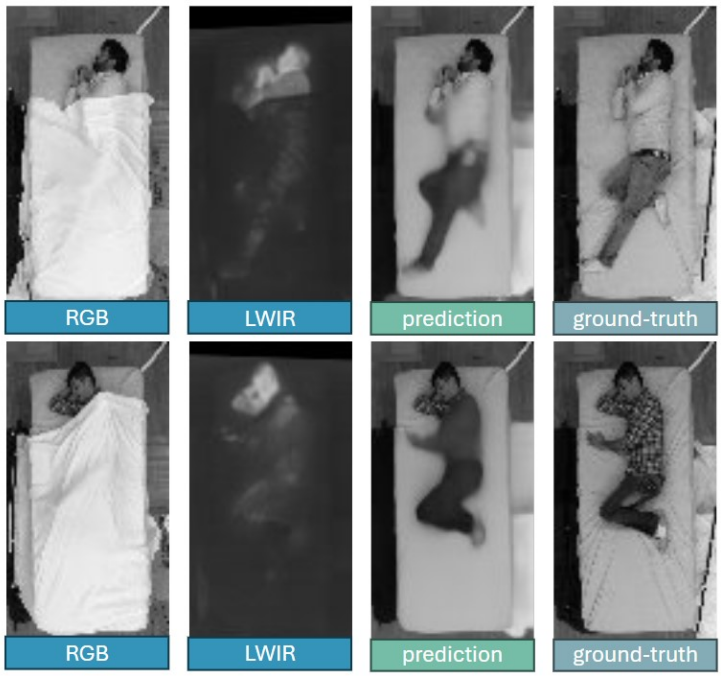


FIGURE F.3 Résulats Undercover, couverture fine

On remarque que le modèle a appris de lui-même à ne pas modifier la partie de l'image située au-dessus de la couverture. La pose est reconstruite de manière grossière, mais reste reconnaissable malgré le flou, présent notamment au niveau des détails fins, comme les mains.

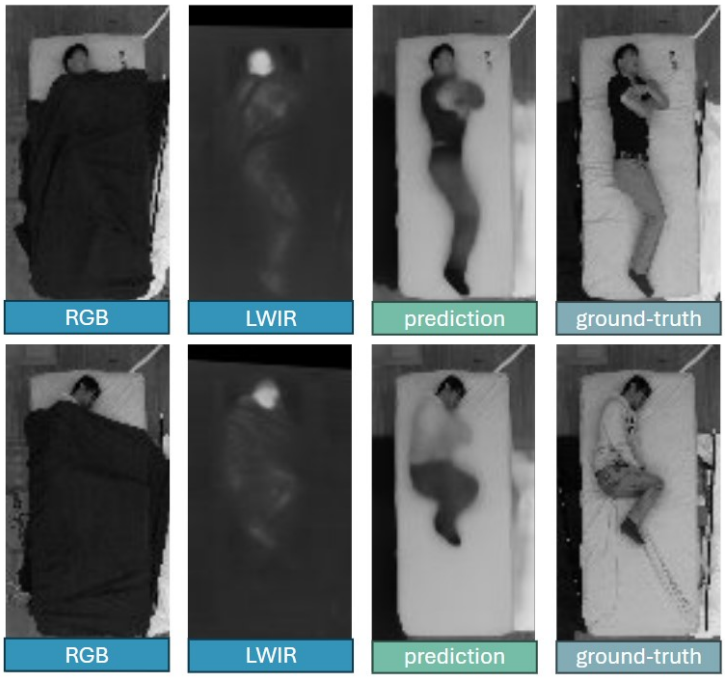


FIGURE F.4 Résulats Undercover, couverture épaisse

Pour une couverture plus épaisse, les résultats sont très similaires (voir Figure F.4). L'infrarouge semble apporter autant d'informations qu'avec une couverture plus fine, ce qui pourrait expliquer la faible différence observée.

Cette première méthode constitue une solution initiale pour répondre à l'objectif d'Undercover. Cependant, l'image générée reste floue, et certaines poses, en particulier celles des bras, demeurent ambiguës. Il serait pertinent d'ajouter un terme supplémentaire dans la fonction de perte afin de favoriser une meilleure conservation de la pose.
De plus, le visage reste encore reconnaissable, ce qui ne correspond pas à l'objectif d'anonymisation des patients. Une solution envisageable serait d'intégrer un algorithme de détection de visage, appliqué à l'image générée, pour flouter cette zone de manière automatique.
Une autre étape importante serait de procéder à un "fine-tuning" du système en utilisant des images provenant du CHUSJ. Cela permettrait d'adapter davantage le modèle aux spécificités des images issues de cet environnement hospitalier.