Amélioration des résolutions spatiale et spectrale d'images satellitaires par réseaux antagonistes.

Anaïs Gastineau

université de BORDEAUX

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE
L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE : MATHÉMATIQUES ET INFORMATIQUE

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES ET CALCUL SCIENTIFIQUE

Par **Anaïs GASTINEAU**

Amélioration des résolutions spatiale et spectrale d'images satellitaires par réseaux antagonistes

Soutenue le 06 décembre 2021

Membres du jury :

Jocelyn CHANUSSOT	Professeur	Grenoble INP	Rapporteur
Ronan FABLET	Professeur	IMT Atlantique	Rapporteur
Marie CHABERT	Professeure	Toulouse INP	Présidente du jury
Andrés ALMANSA	Directeur de Recherche	Université de Paris	Examinateur
Jean-François AUJOL	Professeur	Université de Bordeaux	Directeur
Yannick BERTHOUMIEU	Professeur	Bordeaux INP	Co-directeur
Christian GERMAIN	Professeur	Bordeaux Sciences Agro	Encadrant

Remerciements

Titre — Amélioration des résolutions spatiale et spectrale d'images satellitaires par réseaux antagonistes

Résumé — De plus en plus d'applications, telles que la cartographie ou la classification de l'occupation des sols, nécessitent des images hautes résolutions de la surface de la Terre, mais ces données ne sont pas toujours disponibles. Ainsi, cette thèse porte sur le problème de fusion d'images panchromatiques et multispectrales dans le but d'exploiter au mieux les richesses spatiale et spectrale de chacune de ces données. Pour atteindre cet objectif, cette thèse explore plusieurs aspects liés à l'optimisation du problème ou bien aux architectures considérées.

De manière générale, la paramétrisation des réseaux convolutifs est souvent suffisante pour supporter la diversité des problèmes rencontrés. La base de données d'apprentissage est alors considérée comme le vecteur principal de conditionnement au problème traité. Ainsi, dans un contexte de réseaux antagonistes génératifs, nous proposons d'intégrer une modélisation plus fine du problème de "pansharpening" quant à la conception même du réseau. Nous avons également évalué l'impact sur les performances de reconstruction de différentes formulations de la fonctionnelle globale à minimiser tenant compte des spécificités de l'application.

Dans un premier temps, nous étudions les différents types de régularisation existant dans un cadre variationnel pour ensuite utiliser cette connaissance afin d'ajouter ce type de contraintes géométrique et spectrale dans la fonction de perte du générateur.

Dans un second temps, nous étudions des solutions liées aux architectures considérées pour le générateur et le discriminateur. En effet, nous proposons l'utilisation de plusieurs discriminateurs, chacun répondant à une tâche différente mais complémentaire. Le premier discriminateur se concentre sur la préservation de la résolution spatiale en prenant en compte la luminance et la composante infra-rouge, très informative d'un point de vue de la texture pour la végétation, des images satellites. Le second discriminateur préserve la résolution spectrale en comparant les composantes chromatiques Cb et Cr. Nous étudions également l'ajout de mécanismes d'attention dans le générateur. Nous considérons des mécanismes d'attention spatiale et spectrale pour améliorer la précision de reconstruction du générateur. En effet, ces mécanismes ont pour objectif d'attirer l'attention du générateur sur les parties de l'image les plus pertinentes pour améliorer le résultat.

L'ensemble des pistes que nous avons explorées a conduit à des résultats convaincants, à la fois quantitatifs et visuels. En effet, nous avons pu observer une amélioration notable de la précision des reconstructions spatiale et spectrale, contribuant ainsi à résoudre le problème de fusion d'images panchromatique et multispectrale.

Mots clés — Fusion d'images, télédétection, méthodes variationnelles, réseaux antagonistes génératifs, termes de régularisation, multi-discriminateurs, mécanismes d'attention.

Laboratoires d'accueils — Institut de Mathématiques de Bordeaux, UMR 5251, Université de Bordeaux, 351 Cours de la Libération, F-33405 Talence, France.

Laboratoire de l'Intégration du Matériau au Système, UMR 5218, Université de Bordeaux, 351 Cours de la Libération, F-33405 Talence, France.

Title — Improvement of the spatial and spectral resolutions of satellite images by generative adversarial networks

Abstract — More and more applications, such as mapping or soil classification, need high-resolution images of the surface of the Earth, but this high-resolution data are not always available. Thus, this thesis deals with the problem of fusing panchromatic and multispectral images in order to make the best use of spatial and spectral content of each of these data. To reach this goal, this thesis explores several solutions linked with the optimization of the problem or with the architectures considered.

In general, the parameterization of convolutional networks is often sufficient to support the diversity of the problems considered. The training database is then considered as the main vector for conditioning the addressed problem. Thus, in a context of generative adversarial networks, we propose to integrate a finer modeling of the pansharpening problem with regard to the network design. We also compared the impact on the reconstruction performance of different formulations of the global functional to minimize by taking into account the specificities of the application.

In the first place, we study the different types of existing terms of regularization in a variational framework in order to use this study to add this type of geometric and spectral constraints in the loss function of the generator.

Then, we study solutions related to the architectures considered for the generator and the discriminator. Indeed, we propose the use of several discriminators, each one responding to a different but complementary task. The first discriminator focuses on preserving spatial resolution by taking into account the luminance and the infrared component, highly informative for the texture in vegetation area, of satellite images. The second discriminator preserves the spectral resolution by comparing the Cb and Cr components. We also study the addition of attention mechanism in the generator. We consider spatial and spectral attention mechanisms to improve the reconstruction accuracy of the generator. Indeed, these mechanisms aim to focus the attention of the generator to the most relevant parts of the image to improve the result.

All the proposed methods lead to convincing results, both quantitative and visual. Indeed, we can see an improvement in the precision of the spatial and spectral reconstructions, thus helping to resolve the pansharpening problem.

Keywords — Pansharpening, remote sensing, variational methods, generative adversarial networks, regularization terms, multi-discriminator, attention mechanisms.

Institutes — Institut de Mathématiques de Bordeaux, UMR 5251, Université de Bordeaux, 351 Cours de la Libération, F-33405 Talence, France.

Laboratoire de l'Intégration du Matériau au Système, UMR 5218, Université de Bordeaux, 351 Cours de la Libération, F-33405 Talence, France.

Table des matières

CHAPITRE 1

Introduction

1.1 Contexte

1.1.1 La télédétection

La télédétection est l'ensemble des techniques qui permettent, par l'acquisition d'images, de fournir de l'information sur la surface de la Terre, sans contact direct avec celle-ci. La télédétection englobe tout le processus consistant à capturer et enregistrer l'énergie d'une radiation électromagnétique émise ou réfléchie, traiter et analyser l'information qu'elle représente pour ensuite appliquer cette information [20].

Les satellites sont donc essentiels pour l'observation de la Terre. Ils embarquent divers capteurs permettant l'acquisition d'images avec une précision géométrique qui dépend de l'objectif visé par le satellite.

Ainsi, en télédétection, la résolution spatiale est exprimée en fonction de la taille du terrain capturée par un pixel. Elle affecte donc directement la reproduction de détails spatiaux présents dans la scène observée. En effet, quand la taille d'un pixel est réduite, plus de détails sont capturés et ainsi préservés dans l'image. La résolution spectrale est donnée à la fois par le nombre de canaux ou de capteurs et par la bande passante du signal associée à chacun des canaux capturés par les capteurs produisant l'image. Plus la bande passante est étroite, plus haute est la résolution spectrale. De même, plus le nombre de canaux est élevé, plus la résolution spectrale est élevée. Pour des questions de coût et de technologie, en règle générale, les capteurs embarqués résultent d'un compromis entre résolutions spatiales et spectrales. En effet, certains capteurs sont intègrent l'énergie à l'entrée du capteur sur une large bande en longueur d'onde, et donc ils offrent peu de richesse spectrale mais proposent une résolution spatiale élevée. Lorsque la bande spectrale recouvre l'ensemble du spectre visible, on appelle ce type d'images, les images panchromatiques. Au contraire des images dites multispectrales ou hyperspectrales, acquises par des capteurs sur des bandes beaucoup plus étroites du spectre, offrent une résolution spectrale élevée mais au prix d'une plus faible résolution spatiale. La résolution spatiale est imposée par le facteur de résolution entre la basse et la haute résolution dépendant du satellite considéré.

La télédétection est très utilisée pour la classification, la cartographie, le suivi de la végétation ou encore de l'expansion urbaine. Idéalement, ces applications bénéficieraient utilement d'images à hautes résolutions spatiale et spectrale.

En pratique, comme mentionné plus haut, les images panchromatiques offrent une bonne résolution spatiale mais ne sont souvent pas suffisantes pour beaucoup d'applications, notamment à cause de leur pauvreté spectrale. A contrario, les images multispectrales, voire hyperspectrales, offrent une richesse spectrale d'intérêt au détriment des détails spatiaux. Une première possibilité

pour obtenir des images plus riches serait d'améliorer les performances des capteurs embarqués à bord des satellites mais les progrès sont lents. En effet, il y a des contraintes imposées par le coût de fabrication de ces capteurs mais également d'autres contraintes liées au satellite lui-même, c'est-à-dire son poids, sa puissance ou encore son rayonnement électromagnétique. De plus, les données à haute résolution acquises par le satellite doivent être stockées et transmises au sol, ce qui nécessite beaucoup de mémoire et une large bande passante de communication.

Une autre solution pour répondre à ce problème est d'utiliser un modèle mathématique pour améliorer la résolution spatiale de certaines bandes spectrales. Ainsi, cette thèse s'intéresse à ce problème à travers la fusion d'images satellites. Cette fusion a pour objectif de combiner des images panchromatiques et des images multispectrales dans le but d'améliorer la résolution spatiale des images multispectrales à l'aide des images panchromatiques, conservant ainsi le meilleur de chaque donnée source.

1.1.2 Satellites d'intérêts

Dans la suite de la thèse, nous travaillons avec des images satellites Pléiades et WorldView 3 ce qui nous impose un facteur de résolution spatiale assez classique égal à 4 entre les images panchromatiques et multispectrales. La résolution spectrale, imposée par les capteurs du satellite, peut légèrement varier d'un satellite à l'autre.

Pour les images du satellite Pléiades, les images panchromatiques se caractérisent par une large bande spectrale de longueurs d'ondes comprises entre 480 nm et 800 nm, pour une résolution spatiale au sol de 0.7 m. Les images multispectrales sont composées de 4 bandes spectrales beaucoup plus fines, centrées sur le bleu, le vert, le rouge (RGB) et le proche infra-rouge (NIR). La gamme de longueurs d'ondes pour la bande bleue est comprise entre 430 nm et 550 nm, 490 nm et 610 pour la verte, 600 nm et 720 nm pour la rouge et 750 nm et 950 nm pour le proche infra-rouge, pour une résolution spatiale au sol de 2.8 m.

Pour le satellite WorldView 3, l'image panchromatique est composée d'une bande dont les longueurs d'ondes sont comprises entre 450 nm et 800 nm pour une taille au sol de 0.31 m. Les bandes multispectrales ont une résolution spatiale de 1.24 m avec des longueurs d'ondes comprises entre 450 nm et 510 nm pour le bleu, 510 nm et 580 nm pour le vert, 630 nm et 690 nm pour le rouge et 770 nm et 895 nm pour le proche infra-rouge.

En terme de contenu, il est important de noter les points suivants :

i) Les gammes des longueurs d'ondes de l'image multispectrale peuvent se chevaucher. Cela signifie donc que celles-ci partagent de l'information et qu'elles sont donc par conséquent corrélées.

ii) Le proche infra-rouge reflète mieux la végétation. La plus grande énergie reçue par le capteur nous permet donc de disposer de plus d'informations spatiales, d'informations de textures et de géométrie, dans les zones de végétation.
En effet, la réflectance spectrale est la signature spécifique de la végétation. On peut voir en Figure 1.1 que la réflectance de la végétation dans le visible est basse parce que la majorité de la lumière visible est absorbée pendant la photosynthèse lors de la création de la chlorophylle. En outre, la réflectance est très élevée dans le proche infra-rouge car ces radiations ne sont pas utilisées dans le processus de photosynthèse. On peut donc en conclure que la végétation en bonne santé peut être facilement identifiée grâce à la bande proche infra-rouge.

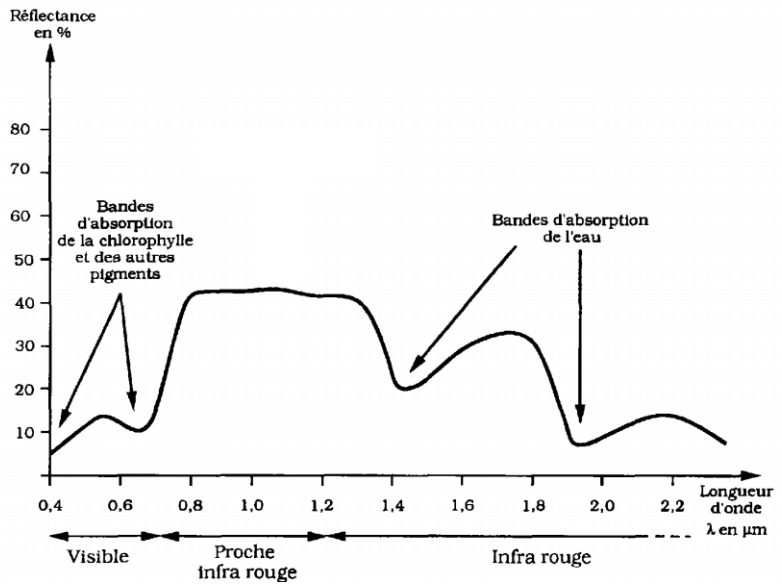
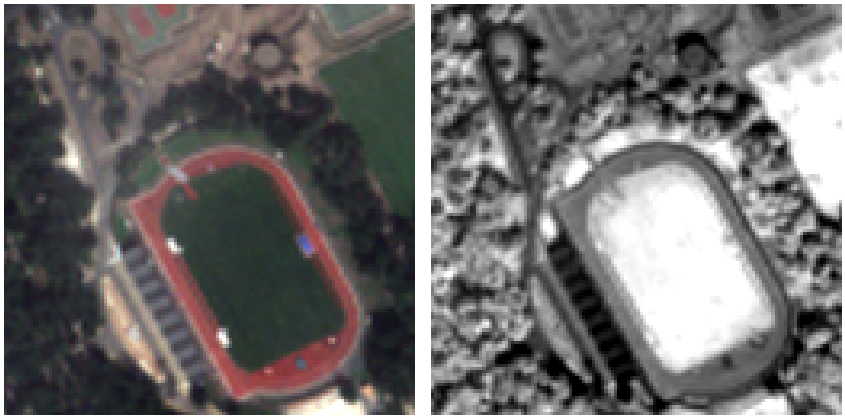


FIGURE 1.1 – Spectre de la végétation, *Télédétection et cartographie, ED. AUP ELF - UREF. Les presses de l'Université du Québec, 1993, p.251*

La Figure 1.2 montre un exemple de ce dernier point : l'intensité du pixel est élevée là où la végétation est dense ou en bonne santé. Lorsque les applications visées sont liées à l'observation de la végétation, il faudra donc être particulièrement attentif à la bande recouvrant le proche infra-rouge.



(a) (b)

FIGURE 1.2 – Illustration de la réflectance de la végétation sur une image satellite Pléiades. Une zone de végétation sur les canaux RGB (a) et dans le proche infra-rouge (NIR) (b). La bande NIR apparaît plus texturée que l'image RGB dans les zones de végétation.

1.2 Problème de pansharpening

1.2.1 Modélisation

Le "pansharpening" est une technique de fusion d'images consistant à mélanger les caractéristiques d'une image panchromatique à haute-résolution spatiale et d'une image multispectrale basse-résolution spatiale afin d'obtenir une image multispectrale haute-résolution spatiale. L'image panchromatique apporte en effet la résolution spatiale, avec en particulier les hautes fréquences de

l'image caractérisant la texture et la géométrie alors que l'image multispectrale apporte la diversité spectrale.

Le problème de pansharpening peut être formulé comme la reconstruction d'une image haute résolution multispectrale u à partir de P l'image panchromatique haute résolution et de l'image u_S multispectrale de basse résolution. Cela revient alors à considérer le modèle suivant :

$$u_S^k = SH^k u^k + B^k \quad \forall k \leq N \tag{1.1}$$

où S est un opérateur de sous échantillonnage, H le noyau d'un opérateur de convolution (ex : filtre), B un bruit gaussien de moyenne nulle et N le nombre de bandes de l'image multispectrale.

Exposé ainsi, le modèle (1.1) est le modèle de super-résolution, il ne prend pas en compte l'image panchromatique. C'est pour cela que beaucoup de modèles supposent que l'image panchromatique puisse être approchée par une combinaison linéaire des différentes bandes spectrales de l'image multispectrale haute résolution que l'on cherche. Cela revient à introduire la contrainte suivante :

$$P = \sum_{k \leq N} \alpha_k u^k \tag{1.2}$$

Avec la contrainte (1.2) le problème s'écrit alors de la façon suivante :

$$\begin{cases} u_S^k = SH^k u^k + B^k \quad \forall\ k \leq N \\ P = \sum\limits_{k \leq N} \alpha_k u^k \end{cases} \tag{1.3}$$

1.2.2 Motivations

Pour aborder le problème de pansharpening, cette thèse propose et développe des méthodes d'apprentissage par réseaux de neurones convolutifs. En effet, depuis quelques années, les réseaux de neurones convolutifs ont montré leur efficacité dans beaucoup de domaines et particulièrement en traitement d'images en général.

En traitement d'images, ces méthodes constituent aujourd'hui l'état-de-l'art pour la majorité des applications. Elles ont pour avantage de modéliser très efficacement la relation entre les variables d'entrée et de sortie par l'apprentissage des poids de différents filtres convolutifs. Cependant, au début de la thèse, nous avons remarqué que beaucoup de ces méthodes innovent sur les architectures des réseaux sans vraiment prendre en compte la modélisation du problème en lui-même. Ainsi, beaucoup de ces réseaux, proposés dans l'état-de-l'art, peuvent s'adapter relativement facilement à différents problèmes tout en donnant de bons résultats, ce qui est à la fois un avantage et un inconvénient de ce type d'approches. La question est alors de savoir si la modélisation du problème devient "inutile", la paramétrisation du réseau étant suffisante pour encaisser la pluralité des tâches considérées ou s'il est possible d'améliorer les résultats en intégrant plus spécifiquement certains modèles en liens avec la physique du problème traité.

Dans cette perspective, cette thèse a pour objectif d'intégrer la modélisation du problème de pansharpening dans un cadre de réseau de neurones convolutifs pour en améliorer la précision de reconstruction des images multispectrales. Pour cela, nous cherchons à conserver les résolutions spatiales et spectrales des images panchromatiques et multispectrales respectivement, par l'ajout de contraintes spatiales et spectrales performantes pour le problème de fusion mais également par l'adaptation aux entrées des différents sous-réseaux composant l'architecture complète. En effet, toutes les bandes spectrales utilisées pour le problème de pansharpening n'apportent pas la même information et nous cherchons donc à utiliser l'information disponible le plus judicieusement possible.

Parmi tous les types de réseaux de neurones convolutifs qui existent, nous avons choisi de nous placer dans un cadre de réseaux génératifs antagonistes (GAN). En effet, depuis quelques années, les GANs commencent à émerger pour le problème de super-résolution, notamment à la suite de la méthode proposée par Ledig *et al.* [44] mais ils étaient encore très peu utilisés pour le problème de pansharpening il y a quelques années.

1.3 Plan du manuscrit

Cette thèse est ainsi orientée sur la résolution du problème de pansharpening à l'aide de méthodes basées apprentissage, en particulier sur l'utilisation des réseaux antagonistes génératifs. Nos contributions ont pour objectif de prendre en compte la modélisation physique du problème de pansharpening et de l'intégrer au mieux dans la conception de l'architecture d'un GAN.

Le Chapitre 2 est consacré à un état-de-l'art général sur les méthodes existantes pour aborder et résoudre le problème de pansharpening. Il existe une large variété de méthodes répondant au problème de pansharpening. En effet, il existe des méthodes par substitution de composantes, des méthodes injectant les hautes fréquences de l'image haute résolution, des méthodes d'analyse multi-résolutions, des méthodes bayésiennes et variationnelles ou encore des méthodes axées sur les réseaux de neurones convolutifs. Ce chapitre a donc pour objectif d'exposer les points forts et les points faibles de chaque type de méthodes.

Le Chapitre 3 présente un peu plus en détails les méthodes variationnelles utilisées pour répondre au problème de pansharpening. Il a principalement pour objectif d'étudier l'influence de différents termes de régularisation modélisant le problème et que nous pourrions utiliser dans la suite de cette thèse. Dans un premier temps, nous présentons un état-de-l'art détaillé sur les termes de régularisation couramment utilisés dans la littérature. Ensuite, nous proposons une méthode PXSNL admettant deux termes de régularisation non-locaux. Le premier est un terme de régularisation géométrique forçant l'alignement des gradients et le second travaille à la reconstruction des intensités de l'image. Ce travail a fait l'objet d'une publication lors de la conférence GRETSI [24].

Les Chapitres 4 et 5 s'intéressent à la résolution problème de pansharpening à l'aide de réseaux antagonistes génératifs (GANs). Plus particulièrement, nous nous intéressons à l'intégration de la modélisation du problème de pansharpening dans des méthodes basées réseaux. En effet, beaucoup de méthodes présentent des architectures de plus en plus développées et complexes tout en gardant une norme l_1 ou l_2 de la différence entre l'image reconstruite et l'image cible dans la fonction de perte à minimiser [86, 49, 79].

Ainsi, le Chapitre 4 se concentre sur les méthodes basées sur les réseaux antagonistes génératifs (GANs) et en particulier sur une de nos contributions, la méthode RDGAN-Geom. La méthode proposée adapte une architecture du type dense résiduel, en ré-injectant l'information des couches précédentes, afin d'éviter le problème de l'évanescence des gradients, qui gêne fortement l'entraînement d'un réseau convolutif. De plus, la reconstruction spatiale étant l'un des points importants pour le problème de pansharpening, nous intégrons un terme de régularisation dans la fonction de perte du générateur. Ce terme, inspiré des méthodes variationnelles, permet de préserver la géométrie de l'image cible en minimisant le produit scalaire du gradient de l'image reconstruite avec l'orthogonal de celui de l'image de référence. Cette contribution a fait l'objet d'une publication à la conférence IEEE ICIP [25].

Le Chapitre 5 est focalisé d'une part sur une considération multi-discriminateurs et d'autre part sur l'introduction d'un terme spectral pour guider de façon plus efficace le générateur. En effet, la reconstruction spectrale étant tout aussi importante que la reconstruction spatiale, notre troisième

contribution est un prolongement de la précédente. D'une part, la méthode s'appuie sur plusieurs discriminateurs. Le premier discriminateur préserve la géométrie (i.e. la résolution spatiale) et le second discriminateur préserve la couleur (i.e. la résolution spectrale). Ainsi, l'objectif est d'entraîner deux discriminateurs, chacun avec une tâche différente mais complémentaire. D'autre part, un terme de régularisation spectral, basé sur la métrique SAM qui minimise la distorsion spectrale, est ajouté à la fonction de perte du générateur. Cette méthode permet donc d'équilibrer la balance entre les reconstructions spatiale et spectrale qui sont essentielles pour notre problème. Cette contribution a été publiée dans le journal IEEE TGRS [26].

Le Chapitre 6 est orienté sur les mécanismes d'attention. Ces mécanismes sont surtout utilisés pour aider le générateur à se concentrer sur les parties de l'image pertinentes à la reconstruction et ainsi améliorer les résultats. Ainsi, nous proposons une méthode permettant d'étudier l'influence des mécanismes d'attention sur notre précédente méthode basée multi-discriminateurs. Pour cela, nous considérons deux mécanismes d'attention spatiale et spectrale dans l'architecture du générateur dans le but d'améliorer à nouveau la précision de notre reconstruction. Dans ce chapitre, nous restons dans le prolongement de notre précédente contribution [26], en travaillant toujours dans un cadre multi-discriminateurs avec deux termes de régularisations spatial et spectral dans la fonction de perte du générateur.

Un dernier chapitre conclut cette thèse et propose des pistes d'amélioration et des perspectives pour le problème de fusion d'images.

CHAPITRE 2

État-de-l'art général

Ce chapitre vise à présenter un panorama des méthodes permettant d'améliorer la précision spatiale d'une image multispectrale en s'appuyant sur une image panchromatique mieux résolue spatialement (pansharpening). Ces méthodes peuvent être divisées en deux grandes familles d'approches : les approches sans apprentissage et les méthodes avec apprentissage.

Les approches sans apprentissage peuvent être répertoriées en plusieurs catégories. Nous retrouvons les méthodes par substitution de composantes, les méthodes de contribution spectrale relative, les méthodes par injection des hautes fréquences, les méthodes d'analyse multi-résolution ou encore les méthodes basées *a priori*. Les méthodes par substitution de composantes sont très utilisées pour leur faible temps de calcul et pour leur bonne capacité à retenir les détails spatiaux de l'image panchromatique. Cette catégorie regroupe des méthodes telles que l'ACP (Analyse par Composantes Principales), la transformation de Gram-Schmidt ou l'IHS (Intensity Hue Saturation) par exemple. Les méthodes de contribution spectrale relative telles que la transformée de Brovey ou l'IM (Intensity Modulation), incluent les méthodes où une combinaison linéaire des bandes spectrales est appliquée, au lieu d'une substitution. Les méthodes basées sur l'injection des hautes fréquences incluent les techniques HPF (High-Pass Filtering) et HPM (High Pass Modulation). Ces deux méthodes injectent les hautes fréquences de l'image, obtenues en soustrayant un filtrage passe bas de l'image panchromatique à celle d'origine. La catégorie des méthodes d'analyse multi-résolution regroupe celles décomposant l'image, par des applications itératives ou opérateurs, en séquence de signaux ou pyramides avec un contenu d'information décroissant. On peut y trouver des méthodes basées sur les ondelettes, la pyramide de Laplace, etc. Les méthodes basées *a priori* regroupent les approches bayésiennes et variationnelles. Les méthodes bayésiennes qui reposent sur l'utilisation d'une distribution *a posteriori* de l'image que l'on souhaite obtenir. Les méthodes variationnelles regroupent les approches par optimisation déterministes, c'est-à-dire que l'on considère un modèle que l'on régularise en donnant un *a priori* sur la solution recherchée. Cela peut être un *a priori* de texture, de géométrie ou d'intensité par exemple.

La seconde grande famille d'approches concerne les méthodes d'apprentissage. Ce type de méthodes permet de modéliser efficacement la relation entre les variables d'entrée et de sortie par la composition de plusieurs niveaux (convolution, activation, etc.). Le principal avantage de ce type de méthodes repose sur la performance des réseaux de neurones, les résultats obtenus donnant souvent l'état-de-l'art. De plus, une fois l'apprentissage aboutit, les résultats sur les images tests sont rapides à acquérir et généralement très bons. Cependant, le temps d'apprentissage peut être long et fastidieux, notamment lors du réglage des hyper-paramètres. Le second inconvénient provient

des bases de données disponibles. De plus en plus de bases de données d'images sont disponibles, en particulier pour le problème de super-résolution par exemple. En revanche, pour le problème de pansharpening, l'acquisition de données satellites haute-résolution n'est pas aisée car celles-ci sont coûteuses et souvent non libre de droits. La diffusion de ces bases de données est donc difficile ou impossible.

Cet état-de-l'art donnera un aperçu général et représentatif des méthodes existantes en se basant notamment sur les états-de-l'art de Loncan *et al.* [53] et de Vivone *et al.* [69].

2.1 Méthodes sans apprentissage

2.1.1 Méthodes basées décomposition

2.1.1.1 Méthodes par substitution de composantes

Ces méthodes commencent par sur-échantillonner l'image multispectrale basse résolution à la même résolution que l'image panchromatique. Ensuite, l'image multispectrale est transformée en un ensemble de composantes en utilisant une transformation linéaire des bandes spectrales.

- ACP (Analyse par Composantes Principales)

L'ACP [17] est très utilisée en analyse du signal. Elle transforme les bandes de l'image multispectrale en un nouvel ensemble de combinaisons de composantes spectrales non corrélées. En général, la première composante principale collecte les informations qui sont communes aux différentes bandes spectrales utilisées en données d'entrée à l'ACP, c'est-à-dire l'information spatiale, c'est donc la composante idéale pour être remplacée par l'image panchromatique dans le processus de pansharpening. L'information spectrale, qui est spécifique à chaque bande, est collectée dans les autres composantes principales.

L'ACP pour le pansharpening peut se résumer de la manière suivante :

1. Augmenter par interpolation bicubique la résolution de l'image multispectrale pour atteindre celle de l'image panchromatique,
2. Calculer l'ACP des bandes de l'image multispectrale,
3. Remplacer la première composante de l'ACP par l'image panchromatique,
4. L'image finale est obtenue en inversant la projection initiale effectuée par l'ACP.

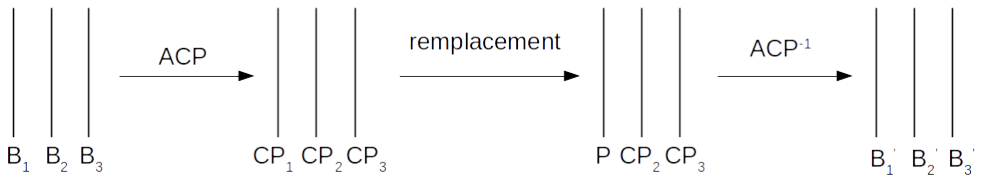


FIGURE 2.1 – Schéma représentatif de la méthode utilisant l'ACP [17].

Cette méthode, représentée en Figure 2.1, amène toutefois à des résultats présentant des distorsions au niveau du spectre, notamment pour les images de végétation. En effet, lors de l'ACP, les composantes principales sont déterminées à partir de la matrice de covariance des données multispectrales et lors de la projection inverse, les bandes spectrales n'ont plus la même importance car la première composante principale a été remplacée par l'image panchromatique. Cela peut donc amener à de fortes distorsions car l'importance des bandes est déterminée par la variance des composantes principales.

Afin d'amoindrir les distorsions spectrales, Gonzalez-Audicana*et al.* [29] ont proposé une méthode mélangeant les ondelettes et l'ACP. Dans cette méthode, seuls les détails de la première composante sont remplacés par les détails de l'image panchromatique.

Cependant, l'approche initiale [29] consistant à remplacer la première composante principale, n'est basée sur aucune statistique entre l'image panchromatique et la première composante principale. Une grande variance de la première composante principale ne signifie pas forcément qu'il y a une forte corrélation avec l'image panchromatique. Dans ce cas, le remplacement de la première composante principale n'est pas un choix judicieux. Shah *et al.* [65] ont proposé une méthode combinant l'ACP et les contourlettes [15] afin d'adapter la sélection des composantes pour la substitution ou l'injection de détails spatiaux.

• IHS (Intensity-Hue-Saturation)

La méthode IHS [16] est une technique en pansharpening connue pour son efficacité. Cependant, elle amène également une forte distorsion spectrale. Cette technique utilise l'espace couleur IHS qui est souvent choisi à cause de la capacité du système visuel humain à détecter l'intensité (I), les teintes (H) et la saturation (S).

La capacité de la transformation en IHS de séparer efficacement les informations spatiales (bande I) des informations spectrales (bandes H et S) rend cette méthode applicable pour le pansharpening.

La technique IHS comporte plusieurs étapes :

1. Sur-échantillonner l'image multispectrale pour obtenir une résolution égale à celle de l'image panchromatique et normaliser chaque bande,
2. Convertir l'image RGB en IHS,
3. Remplacer la composante I par l'image panchromatique,
4. L'image multispectrale finale est obtenue en appliquant la transformée inverse, c'est-à-dire en convertissant l'image IHS en RGB.

Les principaux inconvénients de cette technique sont que l'on ne peut l'appliquer qu'à trois bandes spectrales et que l'on peut observer des distorsion spectrales sur les bandes obtenues. Cependant, plusieurs généralisations ou adaptations ont été proposées. Par exemple, [61] propose d'adapter cette méthode afin d'obtenir une meilleure résolution en adaptant les coefficients, ou bien d'augmenter la qualité spectrale par l'extraction des contours de l'image panchromatique et les combinant avec l'image multispectrale.

• Orthogonalisation de Gram-Schmidt

Une autre méthode de la catégorie substitution de composantes est la méthode de Gram-Schmidt, utilisée en algèbre linéaire et en statistiques multivariées. Cette technique [42] permet d'orthogonaliser les matrices de données ou dans ce cas, les bandes de l'image multispectrale en supprimant la corrélation entre celles-ci.

La procédure est la suivante :

1. Calculer une version basse résolution de l'image panchromatique à partir de l'image multispectrale. Pour cela, il existe plusieurs méthodes : une moyenne pondérée des bandes multispectrales où les poids dépendent de la réponse spectrale des bandes spectrales, de l'image panchromatique et de la transmittance optique de la bande panchromatique, ou encore en sous échantillonnant l'image panchromatique haute résolution.
2. Traiter chaque bande en tant que vecteur de grande dimension en commençant par l'image panchromatique basse résolution générée précédemment comme premier vecteur. Rendre les bandes orthogonales en utilisant l'orthogonalisation de Gram-Schmidt.
3. Remplacer la première bande obtenue après la procédure d'orthogonalisation par l'image

panchromatique haute résolution.

4. L'image finale est obtenue en faisant la transformation inverse en utilisant les mêmes coefficients.

Dans la première étape, le choix de la méthode permettant de simuler l'image panchromatique basse résolution est déterminant, les distorsions spatiales peuvent être plus importantes avec la première méthode. En effet, ces distorsions sont dues au fait que l'image obtenue en faisant la moyenne des bandes spectrales n'a pas la même radiométrie que l'image panchromatique. La deuxième méthode n'est donc pas affectée par les distorsions spectrales mais généralement, on peut observer une moins bonne qualité spatiale.

Aiazzi *et al.* [4] ont proposé une méthode permettant d'éviter cela en générant l'image panchromatique en faisant une moyenne pondérée où les poids sont estimés pour minimiser l'erreur moyenne au carrée des résidus avec l'image panchromatique sous échantillonnée.

De manière générale, l'ensemble de ces méthodes amènent une meilleure qualité spatiale mais introduisent également une forte distorsion spectrale.

2.1.1.2 Méthodes à contribution spectrale relative

Ces méthodes peuvent être considérées comme une variante des méthodes par substitution de composantes car à la place d'une substitution, on applique une combinaison linéaire des bandes spectrales.

• Transformée de Brovey

La transformée de Brovey [27] permet de fusionner des données provenant de différents capteurs. Elle est basée sur une transformée chromatique.

Dans cette méthode, l'image multispectrale est normalisée et chaque bande spectrale de l'image multispectrale haute résolution est obtenue en multipliant les bandes normalisées avec l'image panchromatique de la manière suivante :

$$u^k = \frac{\widetilde{u}_S^k}{\sum_{j \leq N} \widetilde{u}_S^j} P \tag{2.1}$$

où N désigne le nombre de bandes de l'image multispectrale, u^k la k-ème bande de l'image fusionnée, $\widetilde{u}_S^k$ la k-ème bande de l'image multispectrale ré-échantillonnée à la même résolution que l'image panchromatique P.

Une variante de cette transformée consiste à soustraire l'intensité de l'image multispectrale à l'image panchromatique avant d'utiliser l'équation (2.1). La première méthode donne de bons contrastes sur l'image finale mais apporte également une distorsion spectrale alors que la variante préserve mieux les détails spectraux.

• Méthode IM (Intensity Modulation)

Pour cette méthode [48], on peut reprendre l'équation de la transformée de Brovey (2.1) en remplaçant la somme des bandes spectrales de l'image multispectrale par la composante I représentant l'intensité dans la transformation IHS. On a alors :

$$u^k = \frac{\widetilde{u}_S^k}{I} P \tag{2.2}$$

où u^k correspond à l'image fusionnée finale, P, l'image panchromatique et $\widetilde{u}_S^k$, la bande spectrale k de l'image multispectrale u_S agrandie à la taille de l'image panchromatique.

Cette méthode est assez similaire à la transformée de Brovey mais peut amener à des distorsions de couleur si la composante I n'a pas le même ordre de grandeur que les bandes de l'image multispectrale.

2.1.1.3 Méthodes par injection de hautes fréquences

• Algorithme HPF (High-Pass Frequency)

La méthode de pansharpening HPF [17] extrait les hautes fréquence de l'image panchromatique pour après les injecter ou les ajouter dans l'image multispectrale agrandie à la taille de l'image panchromatique. L'extraction de ces informations spatiales se fait par un filtrage passe bas de l'image panchromatique :

$$filtered_P = h_0 * P \tag{2.3}$$

où P désigne l'image panchromatique et h_0 le filtre passe bas. Plusieurs filtres sont possibles : Gaussien, Box filter, Laplacian, etc.

L'information est ensuite injectée en additionnant, pixel par pixel, l'image obtenue en soustrayant $filtered_P$ et P avec l'image multispectrale, c'est-à-dire :

$$u^k = \widetilde{u}_S^k + (P - filtered_P) \tag{2.4}$$

où $\widetilde{u}_S$ est l'image multispectrale agrandie à la taille de l'image panchromatique, P l'image panchromatique et u l'image fusionnée.

Cette méthode présente une faible distorsion spectrale. Cependant, l'ondulation dans la réponse en fréquence est forte.

• Algorithme HPM (High-Pass Modulation)

Avec l'algorithme HPM [64], l'image panchromatique mieux résolue P est multipliée par chaque bande de l'image multispectrale agrandie $\widetilde{u}_S^k$ et normalisée par l'image panchromatique filtrée $filtered_P$, c'est à dire :

$$u^k = \widetilde{u}_S^k \frac{P}{filtered_P} \tag{2.5}$$

Cet algorithme suppose que chaque bande de l'image multispectrale obtenue est proportionnelle à l'image P mieux résolue à chaque pixel. Dans cette méthode, le filtre passe bas peut être choisi de la même façon que dans la sous-section précédente.

Vivone *et al.* [68] ont proposé une méthode utilisant un modèle de régression linéaire, basée sur la méthode des moindres carrés, servant à faire correspondre spectralement chaque bande de l'image multispectrale avec l'image panchromatique, en l'appliquant au modèle d'injection HPM, afin d'améliorer la qualité des résultats. Cela revient à réécrire (2.5) de la façon suivante :

$$u^k = \widetilde{u}_S^k \frac{P + c_k}{\widetilde{P^{LR}} + c_k} \qquad avec \;\; c_k = \frac{\mu_{u_S^k}}{g_k} - \mu_P \;\; et \;\; g_k = r_{u_S^k, P^{LR}} \frac{\sigma_{u_S^k}}{\sigma_{P^{LR}}} \tag{2.6}$$

où $\widetilde{P^{LR}}$ est obtenue par un filtre passe bas et sur-échantillonnée à partir de l'image panchromatique P, r le coefficient de corrélation, σ la covariance et μ la moyenne.

• Amélioration Spectrale

Zhou *et al.* [89] proposent une méthode permettant d'améliorer la diversité spectrale de l'image multispectrale sur-échantillonnée avant l'injection des détails (Figure 2.2). Beaucoup de méthodes de pansharpening (ACP, Brovey, IHS, etc.) utilisent ce procédé afin de pouvoir améliorer la résolution de l'image multispectrale en injectant par la suite des détails obtenus grâce à l'image panchromatique. De manière générale, une interpolation bicubique est utilisée afin d'agrandir l'image multispectrale mais ce type d'interpolation peut causer de fausses informations spectrales car elles ignorent les caractéristiques spectrales mélangées dans l'image multispectrale, donc ne prennent pas en compte les pixels mélangés.

En effet, les images satellites sont composées de pixels purs et de pixels mélangés [18]. On appelle pixels purs les pixels recouvrant un sol composé d'un seul et unique matériau. Au contraire, on désigne par pixels mélangés les pixels qui représentent plusieurs matériaux. En effet, lorsque le sol est constitué de deux ou trois types de matériaux différents, le pixel représentant cette région résulte d'une combinaison de la réflectance de ces différents matériaux.

Cependant, ces auteurs proposent une méthode d'interpolation basée sur l'information spatiale présente dans l'image panchromatique.

La méthode proposée se décompose en trois étapes :

1. Un algorithme de sur-segmentation est appliqué à l'image panchromatique afin d'en obtenir une carte de superpixels,
2. Catégorisation des pixels de l'image multispectrale
3. L'image multispectrale mieux résolue est ensuite obtenue en estimant les pixels mélangés et en étendant les pixels purs.

De manière un peu plus détaillée, pour l'étape de création de la carte de superpixels, les auteurs proposent d'utiliser un algorithme de sur-segmentation afin d'obtenir une carte de superpixels dans le but d'extraire l'information spatiale de l'image panchromatique. L'image panchromatique est alors entièrement sur-segmentée en différentes régions. Ces superpixels sont ensuite labellisés afin de pouvoir catégoriser les pixels de l'image multispectrale. Au final, les pixels purs sont étendus directement et les pixels mélangés sont estimés. Ces pixels sont estimés de la façon suivante : on suppose que chaque pixel mélangé peut s'écrire comme une combinaison linéaire parcimonieuse de pixels purs des k superpixels. Cela revient alors à minimiser une fonctionnelle du type :

$$\alpha^\star = \underset{\alpha}{argmin}\ \frac{1}{2}||x - D\alpha||_2^2 + \lambda||W\alpha||_1 \tag{2.7}$$

où x est le pixel mélangé, D le dictionnaire joint à x avec un vecteur parcimonieux α que l'on cherche et W est l'adaptateur local qui donne une liberté différente à chaque vecteur de base.

Considérer la représentation parcimonieuse des pixels mélangés pour sur-échantillonner l'image multispectrale semble donner de meilleurs résultats quantitatifs lorsque la méthode d'injection est appliquée par la suite.

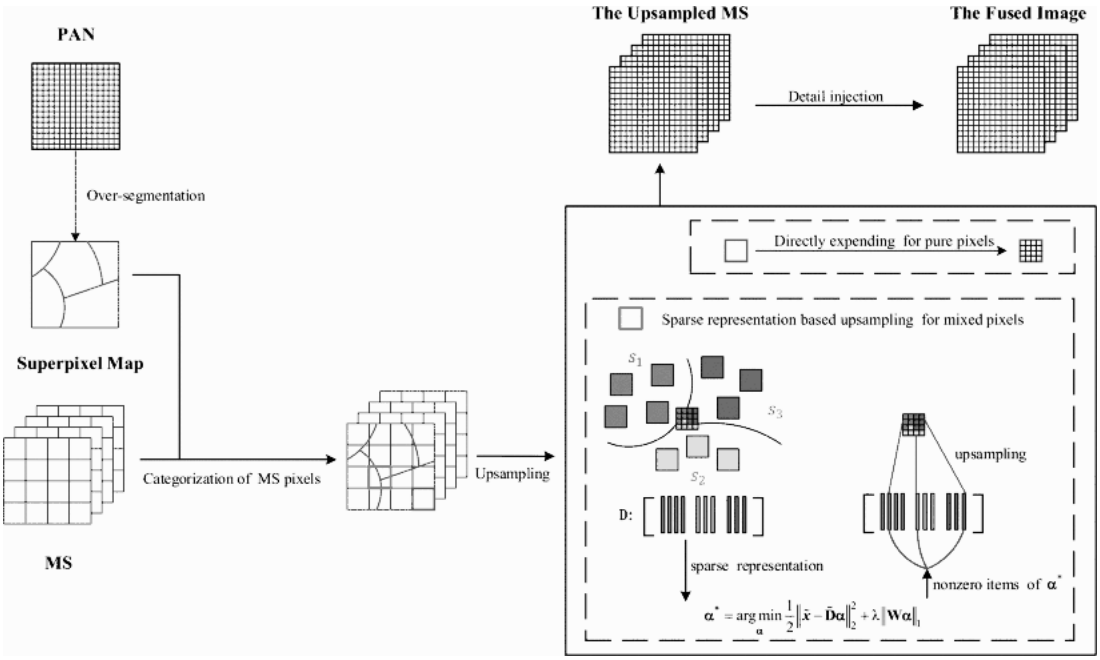


FIGURE 2.2 – Schéma récapitulatif de la méthode proposée par Zhou *et al.* [89].

2.1.1.4 Méthode d'analyse multi-résolution

Les méthodes d'analyse multi-résolutions consistent à décomposer les images panchromatiques et multispectrales en séquences de signaux ou pyramides avec un contenu d'information décroissant. Les hautes fréquences de l'image panchromatique sont alors ajoutées à l'image multispectrale afin d'obtenir l'image fusionnée. Généralement, l'algorithme est appliqué pour chaque bande de l'image multispectrale séparément.

• Pyramide de Laplace

La pyramide de Laplace, originellement proposée par Burt *et al.* [13], est une décomposition d'images passe-bande dérivée de la pyramide gaussienne, qui est obtenue par une réduction récursive (filtrage passe-bas et décimation) de l'image (Figure 2.3).

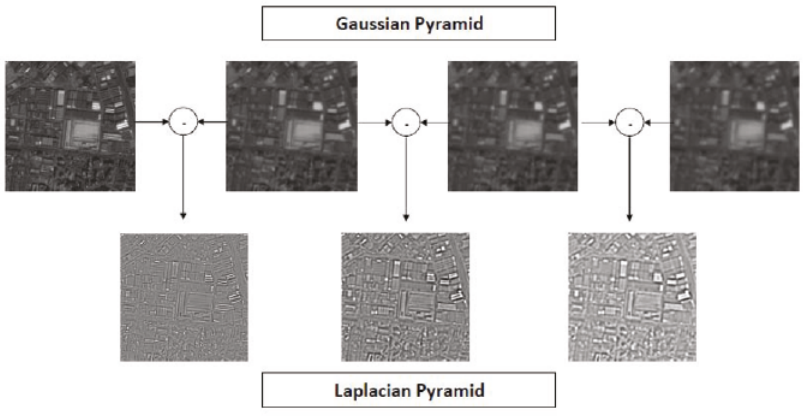


FIGURE 2.3 – Construction d'une pyramide de Laplace [13].

En pansharpening, on utilise une version généralisée de la pyramide de Laplace (GLP). La GLP peut être formulée de la façon suivante :

$$u^k = \widetilde{u}_S^k - g_k(P - \widetilde{P_L}) \quad \forall\ i = 1...N \tag{2.8}$$

où u^k est la bande k de l'image multispectrale fusionnée, $\widetilde{u}_S^k$ l'image multispectrale agrandie à la taille de P l'image panchromatique et $\widetilde{P_L}$ est l'image panchromatique sur-échantillonnée à partir de l'image panchromatique de la pyramide. Les coefficients g_i représentent les poids pour les détails injectés à chaque bande spectrale. Pour la fusion avec la pyramide de Laplace généralisée, plusieurs modèles existent pour obtenir ces coefficients. Le plus courant est le modèle minimisant la distorsion spectrale :

$$g_k = \beta \frac{\widetilde{u}_S^k}{\widetilde{P_L}} \quad \forall\ k \leq N \tag{2.9}$$

où β est défini comme le ratio entre l'écart type standard moyen local de l'image multispectrale sur-échantillonnée et l'écart type de $\widetilde{P_L}$, c'est à dire :

$$\beta = \sqrt{\frac{\frac{1}{N}\sum_{k=1}^{N} var(\widetilde{u}_S^k)}{var(\widetilde{P_L})}} \tag{2.10}$$

Pour résumer, cette méthode peut être utilisée en pansharpening de la manière suivante :

1. Sur-échantillonner chaque bande de l'image multispectrale à la taille de l'image panchromatique,
2. Appliquer la pyramide de Laplace à l'image panchromatique,
3. Sélectionner les poids g_k pour les détails à partir de la pyramide à chaque niveau (2.9) et (2.10),
4. L'image finale est obtenue en ajoutant les détails de la pyramide à chaque bande spectrale pondérée par les coefficients obtenus à l'étape précédente (2.8).

• Algorithme "à trous"

La transformée en ondelettes "à trous" est une technique couramment utilisée en pansharpening qui extrait les détails spatiaux en utilisant les ondelettes "à trous". Ce schéma utilise un modèle d'injection additif :

$$u^k = \widetilde{u}_S^k + P - \widetilde{P_L} \quad \forall\ i \leq N \tag{2.11}$$

où $\widetilde{P_L}$ est la composante basse fréquence de l'image panchromatique et généré par les ondelettes "à trous".

Les ondelettes "à trous" sont des ondelettes non-orthogonales qui n'utilisent pas la décimation lors de la décomposition et qui sont invariantes par translation, une caractéristique qui les rend appréciables pour la fusion d'images. En effet, les ondelettes non-invariantes par translation ont pour conséquence que les coefficients peuvent changer drastiquement pour une petite perturbation du signal. Ici, une translation du signal initial va simplement translater les coefficients en ondelettes. A cause du changement par translation, le niveau du pixel dans l'image finale fusionnée dépendra du décalage et donc introduira une distorsion spectrale.

Pour le pansharpening, on peut résumer cette méthode :

1. Décomposer de l'image panchromatique en utilisant la transformée en ondelettes "à trous" en n niveaux. En général, $n = 2$ ou 3,
2. Ajouter les détails spatiaux de l'image panchromatique décomposée à chaque bande de l'image multispectrale.
3. L'image finale est obtenue en appliquant la transformée inverse.

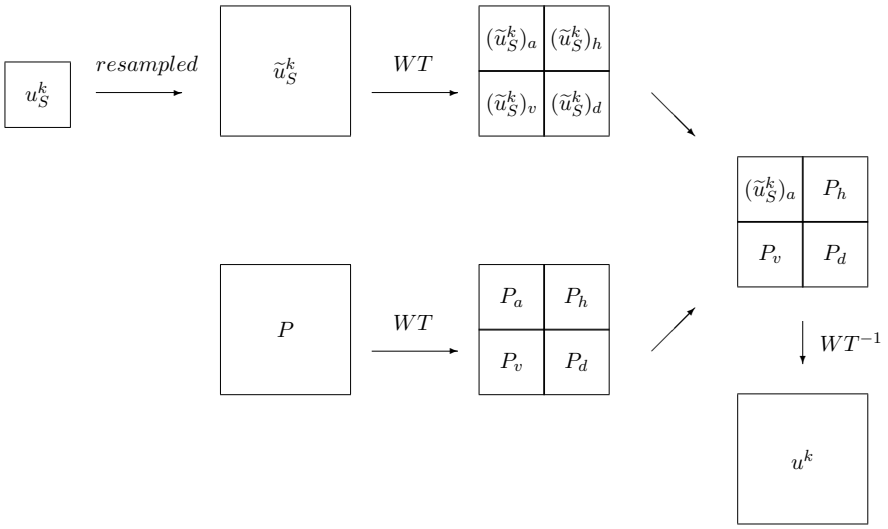


FIGURE 2.4 – Schéma représentatif de la décomposition en ondelettes pour le pansharpening qui s'applique pour chaque bande k de l'image multispectrale u_S avec P représentant l'image panchromatique, P_a (resp. $(\widetilde{u}_S^k)_a$) les coefficients d'approximation de P (resp. MS^k) et P_d, P_v et P_h (resp. $(\widetilde{u}_S^k)_d$, $(\widetilde{u}_S^k)_v$ et $(\widetilde{u}_S^k)_h$) les détails diagonaux, verticaux et horizontaux de P (resp. $\widetilde{u}_S^k$).

Cette méthode, qui peut s'appliquer également avec d'autres types d'ondelettes (décimées, orthogonales, biorthogonales, etc.) [52, 81] est schématisée en Figure 2.4. L'avantage majeur des méthodes multi-résolutions avec les ondelettes est la distorsion spectrale minimale des données. En effet, avec ces méthodes, seule la structure/géométrie contenue dans l'image panchromatique est ajoutée. En revanche, la géométrie est ajoutée à chaque bande, indépendamment. Cependant, cette méthode n'est adaptée pour la fusion d'images que dans le cas où le ratio de taille entre l'image panchromatique haute résolution et l'image multispectrale basse résolution est une puissance de deux.

• Méthode basée sur les opérateurs morphologiques

Restaino *et al.* [63] ont proposé une méthode d'analyse multi-résolutions basée sur les opérateurs morphologiques. Cette approche s'appuie sur une pyramide morphologique qui est un schéma de décomposition non linéaire fondée sur les opérateurs morphologiques. Le choix d'utiliser ces opérateurs repose sur le fait que ceux-ci affectent directement l'extraction des détails de l'image panchromatique. Cette approche revient à considérer le modèle général

$$u^k = \widetilde{u}_S^k - g_k(P - \widetilde{P_L}) \quad \forall\ k = 1...N \tag{2.12}$$

mais en changeant la construction de la pyramide. Ici, les auteurs considèrent pour chaque niveau de la pyramide $l \leq L$:

$$P_l = S\psi_l \tag{2.13}$$

où S correspond à un opérateur de sous-échantillonnage et ψ_l un opérateur morphologique. L'utilisation de ces opérateurs permet d'extraire beaucoup de détails de l'image panchromatique et les résultats montrent une très bonne qualité spatiale.

Cette méthode a pour but de mieux détecter la structure et la texture de l'image panchromatique et ainsi d'incorporer plus de détails dans l'image multispectrale. Cela amène donc à des résultats ayant une bonne résolution spatiale. De manière générale, les approches morphologiques sont utilisées dans l'analyse de texture, la segmentation et la classification. Ici, les opérateurs

morphologiques sont utilisés dans le processus de fusion afin d'extraire la structure et la texture de l'image panchromatique et ainsi d'incorporer plus de détails dans l'image multispectrale. Cela amène à des résultats présentant une bonne qualité spatiale.

2.1.2 Méthodes basées *a priori*

2.1.2.1 Méthodes bayésiennes

Les méthodes bayésiennes [53, 75] modélisent la dégradation de l'image d'origine u, que l'on cherche, en calculant la probabilité conditionnelle que les images multispectrale u_S et panchromatique P soient observées, sachant u, c'est-à-dire en calculant $p(u_S, P|u)$. Elles prennent en compte la connaissance des caractéristiques a priori de l'image fusionnée attendue pour ainsi déterminer la probabilité *a posteriori* $p(u|u_S, P)$.

Les modèles les plus courants sont de mettre un *a priori* gaussien ou un *a priori* de parcimonie.

-*A priori* gaussien

On note u^k les colonnes de la matrice u que l'on suppose mutuellement indépendantes et on leur donne un *a priori* gaussien [53], c'est à dire qu'on suppose

$$p(u^k|\mu_k, \Sigma_k) = \mathcal{N}(\mu_k, \Sigma_k) \tag{2.14}$$

où

- μ_i est une image fixé définie par l'interpolation de l'image u_S,
- Σ_i est une matrice d'hyperparamètres inconnus.

Pour réduire le nombre de paramètres à estimer, on suppose que les Σ_i sont identiques. Et pour résoudre ce type de problème, on peut utiliser l'algorithme MCMC.

La méthode MCMC génère une collection de N échantillons qui sont asymptotiquement distribués selon la loi a posteriori. Et l'estimateur bayésien associé est construit à partir de ces échantillons.

-*A priori* de parcimonie

Dans ce cas, au lieu de mettre simplement un *a priori* gaussien, une représentation parcimonieuse est utilisée afin de régulariser le problème de fusion. Plus précisément, les patchs de l'image cible (projeté sur un sous espace défini par H) sont représentés comme une combinaison linéaire creuse des éléments d'un dictionnaire approprié.

Par exemple, Wei *et al.* [75] introduisent le terme de régularisation suivant :

$$\phi(u) = \frac{1}{2} \sum_{k=1}^{N} ||u^k - \mathcal{P}(\bar{D}_k \bar{A}_k)||_2^2 \tag{2.15}$$

où

- $u^k \in \mathbb{R}^n$ est la k-ième bande de $U \in \mathbb{R}^{\tilde{m}_\lambda \times n}$, avec $k = 1...\tilde{m}_\lambda$,
- $\mathcal{P}(.) : \mathbb{R}^{n_p \times n_{pat}} \rightarrow \mathbb{R}^{n \times 1}$ un opérateur linéaire qui fait la moyenne du chevauchement des patchs de chaque bande, n_{pat} étant le nombre de patchs associé à la k-ième bande,
- $\bar{D}_k \in \mathbb{R}^{n_p \times n_{at}}$ le dictionnaire de la k-ième bande, n_{at} nombre d'atomes du dictionnaire,
- $\bar{A}_k \in \mathbb{R}^{n_{at} \times n_{pat}}$ est le code de la k-ième bande.

Dans ce modèle, la représentation parcimonieuse assure que l'image que l'on cherche est bien représentée par les atomes du dictionnaire, défini par les observations. Comparée à d'autres méthodes, celle-ci donne de moins bons résultats en terme de qualité spatiale mais moins de distorsions spectrales.

2.1.2.2 Méthodes variationnelles

Les méthodes variationnelles regroupent les approches par optimisation déterministes, c'est-à-dire que l'on considère un modèle que l'on régularise en donnant un *a priori* sur la solution recherchée. Le problème de pansharpening (1.3) est un problème mal posé : l'existence, l'unicité ou la stabilité de la solution n'est pas garantie. Afin d'obtenir un problème bien posé, on ajoute donc un terme de régularisation.

Selon l'Équation (1.3), cela revient alors à considérer une fonctionnelle contenant un ou plusieurs termes d'attaches aux données. Le premier terme d'attache aux données décrit la préservation des couleurs de l'image basse résolution et le second décrit la relation entre l'image panchromatique et l'image haute résolution multispectrale.

• Le modèle VWP (Variational Wavelet Pansharpening) proposé par Moeller *et al.* est une méthode combinant les ondelettes et la géométrie de l'image panchromatique comme fonctionnelle d'énergie à minimiser [56, 57].

Cela consiste à minimiser la fonctionnelle

$$E(u) = E_g + E_s + E_w + E_c \tag{2.16}$$

où

$$E_g = \sum_{k \leq N} \int_\Omega |\theta^\perp . \nabla u^k|^2$$

est un terme, repris du modèle $P + XS$, permettant d'améliorer la qualité spatiale de l'image multispectrale en préservant la géométrie de l'image panchromatique,

$$E_w = \sum_n c_0 (a_L^i(n) - \alpha_L^i(n))^2 \phi_{j,n}^2 + \sum_n \sum_{j=1}^{L} \sum_{k=1}^{N} c_j (d_{k,j}(n) - \beta_{k,j}^i(n))^2 \psi_{j,n}^k$$

avec ϕ la fonction d'échelle correspondant à l'ondelette ψ, $a_j^i(n) = <\uparrow u^i, \psi_{j,n}^2>$, $d_{k,j}(n) = < P, \phi_{j,n}^k >$, α_j^i les coefficients pour la bande i et $\beta_{k,j}^i(n)$ les coefficients des détails. Ce terme représente l'attache aux données avec une approche par ondelettes dans le domaine fréquentiel. Pour faire correspondre l'intensité de l'image basse résolution avec les contours plus nets, il faut décomposer l'image panchromatique et chaque bande de l'image multispectrale en ondelette. Puis les coefficients du plus haut niveau de décomposition de l'image panchromatique sont fusionnés avec les coefficients du plus bas niveau de décomposition en ondelette de l'image multispectrale. Dans le domaine spatial, ce terme s'écrit $E_w = \sum_n (Z - u)^2$, où Z correspond à l'image fusionnée obtenue avec la méthode en ondelettes schématisée en figure 2.4,

$$E_c = \nu \sum_{i=1}^{N} \int_{\Omega - \Gamma} (u^i - \uparrow u_S^i)^2 dx$$

représentant le terme préservant les informations spectrales dans lequel chaque bande préserve la couleur de l'image multispectrale sur-échantillonnée là où il n'y a pas de contours ni de texture. C'est à dire en dehors de $\Gamma = exp(-\frac{cst}{|\nabla P|^2})$ et

$$E_s = \mu \sum_{i,j=1, i<j}^{N} \int_\Omega (u^i \uparrow u_S^j - u_j \uparrow u_S^i)^2 dx$$

le terme représentant la préservation des informations fréquentielles de l'image basse résolution en préservant tout les ratios possible entre les bandes spectrales différentes de l'image que l'on souhaite u et l'image u_S sur-échantillonnée.

• Le modèle introduit par Palsson *et al.* [59] reprend le modèle (1.3) mais le régularisant avec un terme de variation totale. La variation totale favorise des solutions qui sont lisses par morceaux avec des discontinuités nettes, induisant des contours d'objets nets dans l'image.

Cela revient à minimiser :

$$\mu \sum_{k \leq N} \int_\Omega |SH^k u^k - u_S^k|^2 + \lambda \int_\Omega \left(\sum_{k \leq N} \alpha_k u^k - P \right)^2 + \gamma \sum_{k \leq N} \int_\Omega |\nabla u_k(x)| \tag{2.17}$$

Ce modèle est basé sur l'hypothèse (1.2) mais cela n'est pas forcément avantageux car ne correspond pas nécessairement à la réalité. En effet, les données satellites peuvent avoir un défaut de repérage spectral qui est un décalage de la longueur d'onde dans le domaine spectral.

2.2 Méthodes avec apprentissage

Ces dernières années, beaucoup de méthodes basées apprentissage ont été proposées dans la littérature. En effet, l'utilisation des réseaux de neurones pour le traitement d'images a démontré leur efficacité à produire des résultats de l'état-de-l'art. Les précédentes recherches ont montré qu'un réseau de neurones peut modéliser efficacement la relation entre les variables d'entrées et de sortie via la composition de plusieurs niveaux (convolution, fonction d'activation, etc.). Ce paragraphe est volontairement peu développée car les approches existantes de l'état-de-l'art sont présentées en détails tout au long de cette thèse dans les chapitres appropriés.

• Huang *et al.* [37] ont proposé un algorithme permettant d'entraîner la relation entre la haute et la basse résolution des différents images. Cette méthode peut se décomposer en plusieurs étapes. Étant donné une image panchromatique P haute résolution et une image multispectrale basse résolution, u_S, l'algorithme proposé est le suivant :

1. Sur-échantillonner u_S à la taille de P, on obtient alors une image multispectrale $\uparrow u_S$,
2. Normaliser P et chaque bande de $\uparrow u_S$ entre $[0, 1]$,
3. Calculer $\hat{P}$ l'image panchromatique obtenue par une combinaison linéaire des bandes de $\uparrow u_S$,
4. Le réseau est entraîné avec les patchs des images $\hat{P}$ et P,
5. L'image finale est obtenue en utilisant le réseau avec chaque bande spectrale de $\uparrow u_S$ et P.

• Masi *et al.* [55] proposent une méthode de pansharpening basée sur un CNN (Convolutional Neural Network). Ils adaptent un réseau déjà utilisé en super résolution pour le problème de pansharpening.

Les auteurs s'inspirent du papier de Dong *et al.* [21] dans lequel les auteurs considèrent un CNN s'inspirant du comportement d'une représentation parcimonieuse pour le problème de super résolution. La représentation parcimonieuse pour le problème de super résolution est une méthode de machine learning qui comprend trois étapes principales :

— la projection de chaque patch de l'image sur un dictionnaire basse résolution,
— la cartographie entre les patchs basse résolution et leurs correspondants dans le dictionnaire haute résolution,
— la reconstruction grâce à la combinaison des patchs du dictionnaires haute résolution

Ce CNN imite ce comportement en étant composé de trois couches qui correspondent approximativement aux trois étapes citées précédemment.

Pour adapter ce CNN au pansharpening, les auteurs proposent d'agrandir l'image multispectrale basse résolution à la taille de l'image panchromatique par interpolation bicubique puis d'utiliser le réseau pour la super résolution avec l'image multispectrale interpolée concaténée avec l'image panchromatique (figure 2.5).

On considère un patch de l'image d'entrée x_p centré en p de taille 9×9. Ce patch est projeté ($y_p = w_1 * x_p$) sur 64 différents patchs de tailles $9 \times 9 \times 3$, constituant ainsi une projection sur un dictionnaire basse résolution. Ensuite, comme w_2 est défini sur un filtre de taille 1×1, y_p (vecteur de taille 64×64) subit une transformation non linéaire $z_p = f_2(y_p)$, où z_p est un vecteur de taille 32, par analogie avec la translation avec la base haute résolution (dictionnaire haute résolution). Pour terminer, grâce à une convolution 5×5 dans la dernière couche, z_p contribue à la reconstruction en p et à son voisinage de taille 5×5, par analogie avec la moyenne pondérée pour les patchs haute résolution.

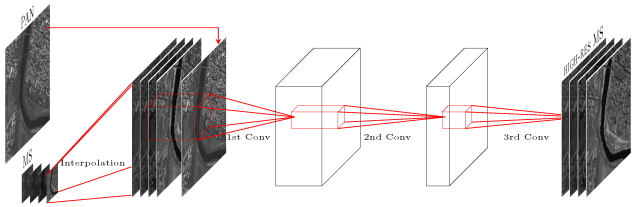


FIGURE 2.5 – Schéma représentatif du réseau de neurones de la méthode pour le pansharpening proposé par Masi *et al.* [55].

Enfin, pour optimiser ce réseau, les auteurs utilisent l'erreur moyenne quadratique entre l'image reconstruite $\tilde{u}$ et l'image de référence u, i.e. :

$$\mathcal{L} = \frac{1}{|\Omega|} ||\tilde{u} - u||_2^2, \tag{2.18}$$

où Ω est le domaine de l'image.

• Yang *et al.* [79] ont proposé une nouvelle architecture (PanNet) pour le pansharpening dans le but de préserver la qualité spatiale et spectrale des images. Ce réseau peut être utilisé avec une grande variété d'images de différents satellites sans avoir besoin d'être ré-entraîné.

Ce réseau à les caractéristiques suivantes :

— Il incorpore les connaissances spécifiques sur le pansharpening, c'est-à-dire qu'il propage l'information spectrale, à travers le réseau en utilisant des images multispectrales sur-échantillonnées, afin de la préserver. Pour la structure spatiale de l'image panchromatique, le réseau est entraîné dans le domaine de filtrage passe-haut plutôt que dans le domaine de l'image.

— Contrairement aux méthodes traditionnelles, les convolutions permettent de capturer les corrélations intra-bandes.

— Beaucoup des méthodes traditionnelles nécessitent un ajustement des paramètres en fonction des satellites fournissant les données. Cependant, ce réseau travaillant dans le domaine des hautes fréquences, les auteurs indiquent qu'il n'y aurait alors pas besoin d'ajuster les paramètres, ce que nos tests n'ont pas permis de confirmer.

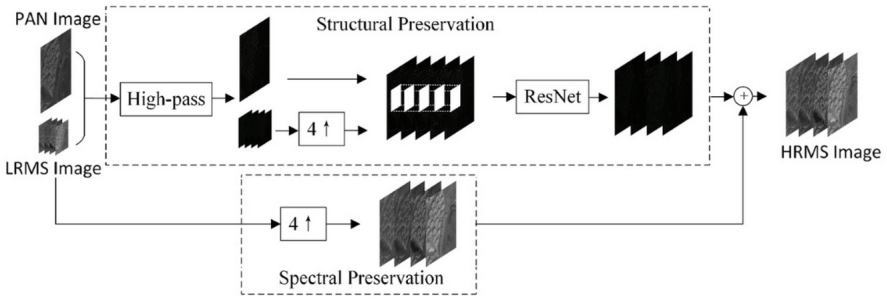


FIGURE 2.6 – Schéma représentatif du réseau PanNet proposé par Yang *et al.* [79].

Dans cette méthode, la norme l_2 de la différence entre l'image reconstruite et l'image de référence est utilisée pour optimiser les poids du réseau.

• Guo *et al.* [32] proposent un réseau constitué de 4 couches d'inférence pour le problème de pansharpening tout en étant robuste aux différents satellites. Ils utilisent une structure multi-niveaux afin d'exploiter pleinement les détails extraits par les couches convolutives. L'architecture est la suivante :

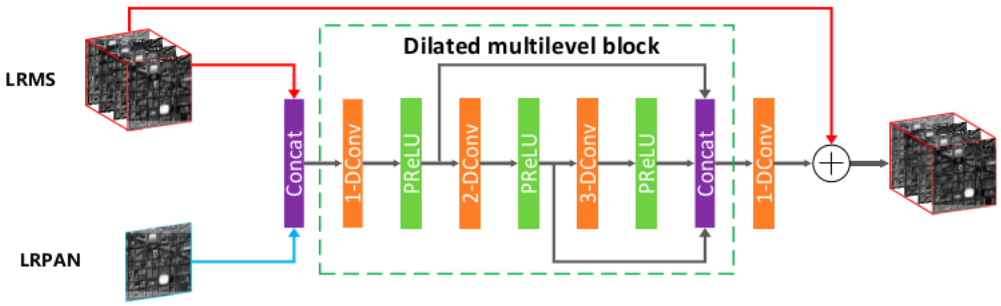


FIGURE 2.7 – Architecture du réseau proposé par Guo *et al.* [32].

Cette architecture a pour objectif de faciliter la reconstruction de l'image en élargissant le champ récepteur après chaque couche de convolution. Ainsi, la considération de convolution dilatée (i.e. une déconvolution) permet de mieux capturer l'information contextuelle pour mieux la reconstruire.

De plus, les auteurs choisissent également d'ajouter un terme de régularisation sur les poids du réseau dans la fonctionnelle à minimiser :

$$\underset{\omega,b}{argmin}\ \frac{1}{N}\sum_{n\leq N}||\tilde{u}^l - u^l||_2^2 + \frac{\lambda}{2}||\omega||_2^2, \tag{2.19}$$

où ω et b sont respectivement les poids et les biais à optimiser, $\tilde{u}$ est la sortie du réseau donc l'image fusionnée et u l'image cible c'est-à-dire l'image multispectrale haute résolution de référence.

Le premier terme permet de minimiser l'erreur entre l'image reconstruite et l'image de référence en utilisant la norme l_2 et le second terme est une régularisation sur les poids du réseau. De manière générale, une régularisation sur les poids du réseau permet d'éviter le sur-apprentissage. En effet, une régularisation l_2 sur les poids du réseau à optimiser permet de pénaliser les poids élevés, mais ne favorise pas la parcimonie de la solution contrairement à une norme l_1. De plus, un réseau ayant des poids élevés peut souvent être instable, i.e. une petite perturbation dans les données initiales peut amener à des résultats totalement différents, et ce n'est pas ce que l'on souhaite. C'est pourquoi il est possible d'utiliser ce type de régularisation dans la fonction de perte à minimiser.

2.3 Comparaison des méthodes

Dans cette section, nous abordons les métriques utilisées pour comparer les images fusionnées. En effet, le problème de pansharpening repose sur deux points clés : la reconstruction spatiale et la reconstruction spectrale. Ainsi, il existe plusieurs mesures utilisées ou créées pour le problème de pansharpening permettant d'évaluer ces deux aspects de manière distincte ou conjointe.

Pour la suite, on note X et Y des images multispectrales, L le nombre de bandes spectrales et N le nombre de pixels dans chacune des images.

• SAM

La mesure Spectral Angle Mapper (SAM) mesure la distorsion spectrale entre l'image obtenue et l'image de référence. En effet, cette mesure calcule la valeur absolue de l'angle entre deux vecteurs qui ont pour éléments les valeurs des pixels pour les différentes bandes de l'image multispectrale haute résolution et de l'image de référence à chaque position. Une valeur de SAM égale à zéro indique une absence de distorsion spectrale mais des distorsions radiométriques peuvent être présentes, dans ce cas, les deux vecteurs sont parallèles mais ont des longueurs différentes. Pour obtenir une mesure globale de l'image, la moyenne de la valeur SAM à chaque position est calculée.

$$SAM(X,Y) = \frac{1}{N}\sum_{i=1}^{N} SAM(x_i, y_i) \qquad avec\ SAM(x_i, y_i) = arccos\left(\frac{< x_i, y_i >}{||x_i||_2||y_i||_2}\right), \tag{2.20}$$

où x_i et y_i représente des vecteurs de $\mathbb{R}^L$.

• CC

La mesure Cross Correlation (CC) repose quant à elle sur un critère spatial calculant la corrélation intra et inter bandes,

$$CC(X,Y) = \frac{1}{L}\sum_{i=1}^{L} CCS(X^i, Y^i) \qquad avec\ CCS(X^i, Y^i) = \frac{\sum_{j=1}^{N}(X_j^i - \mu_X)(Y_j^i - \mu_Y)}{\sqrt{\sum_{j=1}^{N}(X_j^i - \mu_X)^2 \sum_{j=1}^{N}(Y_j^i - \mu_Y)^2}}, \tag{2.21}$$

où μ_X et μ_Y représentent les moyennes de X et Y. La valeur idéale de ce critère est 1. En effet, le coefficient de corrélation varie entre -1 et 1. Un coefficient de valeur 1 indique que les deux images sont corrélées positivement, c'est-à-dire qu'elles sont très proches l'une de l'autre, dans le sens où elles partagent beaucoup d'informations spatiales. Alors qu'un coefficient de valeur -1 indique que les deux images sont corrélées négativement, donc opposées.

• RMSE

La Root Mean Square Error (RMSE) est un critère global qui mesure l'erreur l_2 entre deux images. Le calcul du RMSE s'effectue entre chaque bande spectrale de l'image de référence et de l'image fusionnée et mesure les changements de radiance des valeurs des pixels :

$$RMSE(X,Y) = \frac{||X - Y||_2}{\sqrt{N \times L}} \tag{2.22}$$

• ERGAS

L'Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [71] est également un critère global proposé spécifiquement pour le problème de fusion d'images :

$$ERGAS(X,Y) = 100d\sqrt{\frac{1}{L}\sum_{i=1}^{L}\left(\frac{RMSE_i}{\mu_i}\right)^2}, \tag{2.23}$$

où $RMSE_i = \frac{||X^i - Y^i||_2}{\sqrt{N}}$ et d le ratio entre la haute et la basse résolution. Cette mesure a été proposée dans le but d'être indépendante du ratio entre les images, du nombre de modalités traitées, de la résolution spatiale, de la dynamique, du gain, des coefficients d'étalonnage et des unités.

2.4 Bilan

Cet état-de-l'art montre la très grande variété des méthodes permettant de résoudre le problème de pansharpening. Les méthodes les plus récentes et les plus prometteuses reposent essentiellement sur l'apprentissage, par l'utilisation de réseaux convolutifs. Cette thèse est donc naturellement tournée vers l'utilisation de réseaux de neurones convolutifs pour résoudre le problème de pansharpening.

Cependant, beaucoup de méthodes basées réseaux mettent en place une architecture dans l'objectif d'obtenir les meilleurs résultats quantitatifs et visuels sans vraiment prendre en compte le problème. Dans cette thèse, nous avons décidé d'intégrer la modélisation du problème de pansharpening dans un réseau. Ainsi, le prochain chapitre de cette thèse est consacré aux méthodes variationnelles pour étudier les différents termes de régularisation existant et performant pour le problème. Ensuite, le chapitre 4 est centré sur les réseaux convolutifs et plus particulièrement sur les GANs (réseaux génératifs antagonistes) dans lesquels nous intégrons la modélisation des deux enjeux principaux : la reconstruction spectrale et la reconstruction spatiale des images multispectrales haute-résolution.

CHAPITRE 3

Fusion d'images non locale préservant la géométrie basée sur les méthodes variationnelles

Ce chapitre a pour objectif de présenter la modélisation du problème de pansharpening d'un point de vue variationnel. En effet, le choix d'un terme de régularisation est très important car il va influencer la solution. Le but est alors de se faire une idée sur les régularisations permettant de préserver les résolutions spatiale et spectrale à l'aide des images panchromatique et multispectrale.

On rappelle que le modèle direct associé au problème de pansharpening est le suivant :

$$\begin{cases} u_S^k = SH^k u^k + B^k \quad \forall\ k \leq N \\ P = \sum_{k \leq N} \alpha_k u^k \end{cases} \tag{3.1}$$

où P est l'image panchromatique, u_S l'image multispectrale basse résolution, u l'image que l'on souhaite reconstruire, N le nombre de bandes des images multispectrales, S un opérateur de sous-échantillonnage, H du flou et B du bruit. Ce problème est mal posé et les méthodes variationnelles permettent de résoudre ce problème en ajoutant un terme de régularisation donnant un a priori sur la solution recherchée.

Pour commencer, un bref aperçu des différentes régularisations proposées par la littérature est présenté. Ensuite, une des contributions de cette thèse, visant à résoudre le problème de pansharpening par l'utilisation de termes non locaux est développée.

3.1 État-de-l'art basé méthodes variationnelles

Les termes de régularisation proposés dans la littérature pour résoudre le problème de pansharpening ont généralement le même objectif : reconstruire l'image multispectrale à la résolution spatiale de l'image panchromatique. En effet, la majorité des méthodes se basent sur les termes d'attaches aux données, conformément au modèle (3.1), pour la reconstruction spectrale.

• Régularisation géométrique :

Ce type de régularisation a pour but de préserver la géométrie présente dans l'image panchromatique.

- Le modéle P+XS est une approche introduite par Ballester *et al.* [10] qui revient à minimiser les deux termes d'attaches aux données du modèle (1.3) en ajoutant un terme de régularisation forçant

l'alignement des courbes de niveau de chaque canal multispectral de l'image que l'on souhaite améliorer avec celles de l'image panchromatique. Cela revient à transférer la géométrie de l'image panchromatique à l'image multispectrale. Cela se traduit par le terme de régularisation suivant :

$$\sum_{k\leq N}\int_{\Omega}|\theta^{\perp}.\nabla u^k|^2 \tag{3.2}$$

où $\theta^{\perp}$ est le champs de vecteurs normaux unitaires des lignes de niveau de l'image panchromatique.

Cette régularisation se fonde sur l'hypothèse que pour "les images multispectrales satellitaires, la géométrie des canaux spectraux est, dans une large mesure, contenue dans la carte topographique de son image panchromatique" [10]. Cela signifie que dans le cas des images multispectrales satellitaires, les images d'une même scène prises à des longueurs d'ondes différentes partagent des informations géométriques.

Au final, on minimise :

$$\inf_{u}\ \mu\sum_{k\leq N}\int_{\Omega}|SH^k u^k-u_S^k|^2+\lambda\int_{\Omega}\left(\sum_{k\leq N}\alpha_k u^k-P\right)^2+\sum_{k\leq N}\gamma_k\int_{\Omega}|\theta^{\perp}.\nabla u^k|^2 \tag{3.3}$$

Le principal inconvénient de ce modèle est qu'il suppose que les différentes bandes spectrales aient exactement les mêmes informations géométriques, ce qui n'est pas toujours le cas. En effet, les bandes dans le visible partagent la plupart des informations géométriques, mais elles en partagent moins avec les bandes spectrales dans l'infra rouge, ce qui peut influencer les résultats. Par exemple, des variations subtiles dans la végétation peuvent créer des structures invisibles dans les canaux R, V et B mais visibles dans le canal IR.

- Le modèle VWP proposé par Moeller *et al.* est une méthode combinant les ondelettes et la géométrie de l'image panchromatique comme fonctionnelle d'énergie à minimiser [56, 57] présentée plus en détails dans l'état-de-l'art général (Cf. Chap.2).

• Régularisation du type variation totale :

- He *et al.* [34] sont partis du modèle (1.1), en exploitant un terme de régularisation correspondant à une variante de la variation totale. Le problème de pansharpening est, dans ce cas, considéré comme un problème de colorisation de chaque pixel de l'image panchromatique. L'ajout du gradient de l'image panchromatique dans la fonction de variation totale a pour but de préserver les contours de l'image panchromatique. Cela revient alors à considérer la fonctionnelle suivante :

$$\mu\sum_{k\leq N}\int_{\Omega}|SH^k u^k-u_S^k|^2+\int_{\Omega}\sqrt{\sum_{k\leq N}|\nabla u_k(x)|^2+\gamma|\nabla P(x)|^2} \tag{3.4}$$

où γ correspond au poids que l'on souhaite donner pour la contribution de l'image panchromatique dans la régularisation.

- Bungert *et al.* [12] proposent une approche considérant simultanément le problème de pansharpening et la déconvolution aveugle. Cette méthode consiste à chercher $u^\star$, l'image fusionnée, en considérant le problème d'optimisation suivant :

$$(u^\star,k^\star)=\underset{(u,k)\in U\times K}{argmin}\ \frac{1}{2}||A_k u-y||^2+\mathcal{R}_u(u)+\mathcal{R}_k(k) \tag{3.5}$$

où $A_k = SC_k$, avec S l'opérateur de sous-échantillonnage et C_k opérateur de convolution avec le noyau k, y l'image multispectrale basse résolution (les observations) et $\mathcal{R}_u$ et $\mathcal{R}_k$ sont les termes de régularisation.

Cela revient à considérer le modèle "classique" de super-résolution puis de régulariser en prenant en compte l'image panchromatique. En effet, on a :

$$\mathcal{R}_u(u) = \lambda_u dTV(u) + i_{[0,\infty[^m} \tag{3.6}$$

et

$$dTV(u) = \sum_i ||P\nabla u_i|| \quad et \quad P\nabla u_i = \nabla u_i - < \frac{\nabla v}{|\nabla v|}, \nabla u_i > \frac{\nabla v}{|\nabla v|} \tag{3.7}$$

où v est l'image panchromatique. Le second terme de régularisation concerne le noyau k :

$$\mathcal{R}_k(k) = \lambda_k TV(k) + i_{\mathbb{S}} \quad o\grave{u} \;\; \mathbb{S} = \left\{ k \in K, k_i \geq 0, \sum_i k_i = 1 \right\}. \tag{3.8}$$

Comme la variation totale ne prend pas en compte l'image haute résolution qui est une importante source d'informations, les auteurs proposent alors d'utiliser la variation totale directionnelle. Ce terme lisse donc les régions où le gradient de P et de u sont orthogonaux et favorise les gradients qui sont alignés. Ce terme de régularisation ne force pas le gradient de la solution dans le sens de celui de l'image panchromatique car $\nabla u = 0$ minimise la fonctionnelle mais prend avantage de l'alignement des gradients des images.

- Adesso *et al.* [1, 2] proposent une méthode de pansharpening basée sur une méthode de super résolution qui suppose que l'image Z que l'on cherche puisse se décomposer dans un sous espace de dimension inférieure. En effet, ils considèrent le modèle suivant :

$$\begin{cases} H &= ZBM + N_k \\ P &= RZ + N_p \end{cases} \tag{3.9}$$

où H est l'image multispectrale, P l'image panchromatique, B l'opérateur de flou, M l'opérateur de sous échantillonnage, R le poids associé à chaque bande, N_h et N_p les bruits. En supposant que Z appartient à un sous espace de dimension inférieure, on a alors $Z = EX$ avec E correspondant à la base de ce sous espace. Au final, cela revient à minimiser :

$$\frac{1}{2}||H - EXBM||_F^2 + \frac{\lambda_m}{2}||P - REX||_F^2 + \lambda_\varphi \varphi(X) \tag{3.10}$$

Ici, $\varphi(X)$ est le terme de régularisation correspondant à la variation totale collaborative et il est possible de considérer deux normes différentes :

1. $\varphi(X) = ||A(X)||_{p,q,r} = \left(\sum_{j=1}^{N} \left(\sum_{i=1}^{L} \left(\sum_{k=1}^{M} |A_{i,j,k}|^p \right)^{\frac{q}{p}} \right)^{\frac{r}{q}} \right)^{\frac{1}{r}}$
2. $\varphi(X) = (\mathbb{S}(bd), l^q(x))(A(X)) = \left(\sum_{j=1}^{N} \left\| \begin{matrix} A_{1,j,1} & ... & A_{1,j,M} \\ \vdots & & \vdots \\ A_{L,j,1} & ... & A_{L,j,M} \end{matrix} \right\|_{\mathbb{S}^p}^q \right)^{\frac{1}{q}}$

où A correspond à la concaténation des matrices des gradients en x et y et $||.||_{\mathbb{S}^p}$ correspond à la norme de Schatten [1, 2].

Ce terme de régularisation permet de considérer le gradient d'une image multispectrale en tant que matrice de trois dimensions, où ces dimensions correspondent au nombre de pixels, au nombre

de bandes et au nombre de dérivées directionnelles calculées à chaque pixel. De plus, en choisissant différents types de normes, le type de régularisation change et amène à des résultats pouvant varier.

- Le modèle introduit par Palsson *et al.* [59], présenté en détail dans l'état-de-l'art général (cf. Chap.2).

• Régularisation non locale :

La régularisation non locale fait interagir chaque point avec tous les autres du domaine. La relation de proximité (présente dans la régularisation locale) est remplacée par une mesure de similarité mettant en relation les points qui ont le plus de similarités géométriques ou texturales.

Dans ce cas, le terme de régularisation prend avantage du principe d'auto-similarité des images naturelles, appliqué à l'image panchromatique.

- Le modèle NLV introduit par Duran *et al.* [22] a été inspiré par l'algorithme de débruitage moyen non local. Il revient à minimiser la fonctionnelle :

$$\inf_u \mu \sum_{k\leq N} \int_\Omega |SH^k u^k - u_S^k|^2 + \lambda \int_\Omega \left(\sum_{k\leq N} \alpha_k u^k - P \right)^2 + \sum_{k\leq N} \int\int_{\Omega\times\Omega} (u^k(x) - u^k(y))^2 \omega_P(x,y) \quad (3.11)$$

où

$$\omega_P(x,y) = \frac{1}{\gamma(x)} exp\left(-\frac{d_\rho(P(x),P(y))}{h^2} \right)$$

représente la mesure de dissimilarité,

$$\gamma(x) = \int_\Omega exp\left(-\frac{d_\rho(P(x),P(y))}{h^2} \right)$$

le facteur de normalisation et

$$d_\rho(P(x),P(y)) = \int_\Omega G_\rho(t) |P(x+t) - P(y-t)|^2 dt$$

la distance entre les deux patchs centrés en x et y avec G_ρ est un noyau gaussien.

Le poids ω_P décroît quand la dissimilarité augmente, donc la moyenne est faite entre des patchs de région de l'image très similaire et préserve l'image mais réduit les petites fluctuations (en particulier le bruit).

De manière générale, beaucoup de méthodes donnent des résultats avec une bonne résolution spatiale mais avec une perte d'informations spectrales. Cependant, ce modèle permet de réduire fortement cette perte d'informations spectrales tout en gardant une bonne résolution spatiale.

- Duran *et al.* [23] ont également proposé le modèle NLVD qui considère le problème de pan-sharpening comme un problème d'optimisation minimisant une fonctionnelle avec un terme de régularisation non local. La minimisation de cette fonctionnelle est découplée sur chaque bande spectrale, qui permet alors de considérer des composantes spectrales non recalées.

Ce modèle considère une nouvelle contrainte qui impose la préservation du ratio radiométrique entre l'image panchromatique et chaque composante spectrale afin de conserver la géométrie de l'image.

Cela revient à minimiser la fonctionnelle suivante :

$$\inf_u \sum_{k\leq N} \int\int_{\Omega\times\Omega} (u^k(y)-u^k(x))^2 \omega_{P_k}(x,y) + \mu \sum_{k\leq N} \int_\Omega |SH^k u^k - u_S^k|^2 + \sum_{k\leq N} \frac{\delta}{||P_k||^2} \int_\Omega \left(u^k(x)\tilde{P}_k(x) - \tilde{u}^k(x)P_k(x)\right)^2 \quad (3.12)$$

avec $\tilde{P}_k$ est l'interpolation bicubique de l'image panchromatique P exprimée dans la même référence que u^k et $\tilde{u}^k$ l'interpolation bicubique de u_S^k.

Les principaux avantages de cette méthode sont de ne pas utiliser l'hypothèse de combinaison linéaire entre l'image panchromatique et l'image multispectrale et de ne pas avoir besoin de recalage des données spectrales initiales.

3.2 Contribution : modèle non local PXSNL+NLV

3.2.1 Modèle

Le modèle P+XS [10], présenté plus haut, minimise les termes d'attaches aux données en forçant l'alignement des lignes de niveau de chaque canal multispectral de l'image que l'on cherche avec celles de l'image panchromatique. Cela revient à transférer la géométrie de l'image panchromatique à l'image multispectrale recherchée de la manière suivante :

$$\sum_{k\leq N} \int_\Omega |\theta^\perp . \nabla u^k|^2 dx, \quad (3.13)$$

où $\theta = \frac{\nabla P}{|\nabla P|}$, Ω correspond au domaine de la solution u. Ici et dans la suite de cet article, le gradient est défini de la façon suivante :

$$\nabla u = (\partial_x u, \partial_y u).$$

Ce modèle est sensible au bruit contrairement au modèle NLV proposé par Duran *et al.* [22] qui considère un terme de régularisation non local inspiré de l'algorithme de débruitage NL-means :

$$\sum_{k\leq N} \int_\Omega \int_{\mathcal{N}_x} (u^k(x)-u^k(z))^2 \omega(x,z) dz dx, \quad (3.14)$$

où $\mathcal{N}_x = \{z \in \Omega \ t.q. \ |x-z| \leq r\}$ représente le voisinage de x. Pour ce modèle, les poids ω au point x sont calculés de la façon suivante :

$$\omega(x,z) = \frac{1}{C(x)} exp\left(-\frac{d(P(x),P(z))}{h^2}\right), \quad (3.15)$$

où $C(x)$ est la constante de normalisation et d la distance entre les patchs $P(x)$ et $P(z)$ de l'image panchromatique, $z \in \mathcal{N}_x$ et h est un paramètre de filtrage. Les poids décroissent quand la dissimilarité augmente : la moyenne est calculée entre des patchs très similaires et préserve ainsi l'image tout en réduisant les petites fluctuations, en particulier celles causées par le bruit.

D'après les différents modèles présentés précédemment, nous proposons ici un terme non local (PXSNL) combinant les idées présentes dans les deux termes précédents (3.13) et (3.14). C'est-à-dire que nous proposons un terme préservant la géométrie tel que celui proposé par Ballester *et al.* tout en gardant le caractère non local de Duran *et al.* :

$$\sum_{k\leq N} \int_\Omega \int_{\mathcal{N}_x} |\theta^\perp(z) . \nabla u^k(x)|^2 \omega(x,z) dz dx. \quad (3.16)$$

Ce terme permet d'aligner le gradient de u au point x en prenant en compte les vecteurs au voisinage de ce point. Ainsi, comme le gradient est assez sensible au bruit, utiliser les vecteurs au voisinage permet d'atténuer ce bruit.

Nous considérons deux termes d'attaches aux données pour le modèle (3.1) ainsi, cela revient à minimiser :

$$\begin{aligned} \underset{u}{argmin} \sum_{k\leq N}\int_\Omega |SH^k u^k - y^k|^2 + \lambda \int_\Omega \left(\sum_{k\leq N} \alpha_k u^k - P\right)^2 \\ + \mu \sum_{k\leq N}\int_\Omega \int_{\mathcal{N}_x} |\theta^\perp(z).\nabla u^k(x)|^2 \omega(x,z)dzdx. \end{aligned} \tag{3.17}$$

Pour cela, nous utilisons l'algorithme de descente du gradient à pas fixe. En effet, si nous considérons

$$F(u) = \int_\Omega \int_{\mathcal{N}_x} |\theta^\perp(z).\nabla u(x)|^2 \omega(x,z)dzdx, \tag{3.18}$$

le gradient est alors :

$$\nabla F(u) = -2div\left(\int_{\mathcal{N}_X} (\theta^\perp(z).\nabla u(x))\theta^\perp(z)\omega(x,z)dz\right), \tag{3.19}$$

avec $div(v) = \partial_x v_1 + \partial_y v_2$, où $v = (v_1, v_2)$.

3.2.2 Résultats

Dans ce paragraphe, nous comparons les méthodes mentionnées en simulant une image panchromatique et une image multispectrale à partir des images de la Figure 3.1, en suivant le modèle (3.1).

FIGURE 3.1 – Images de références utilisées pour les tests.

Les résultats quantitatifs obtenus sur ces données simulées sont les suivants :

Image	Modèle	CC	SAM	RMSE	PSNR
	Valeurs idéales	1	0	0	max
Cactus	P+XS	0.9917	0.0656	6.07	32.53
	NLV	0.992	0.054	5.96	32.66
	PXSNL	0.9923	0.0515	5.85	32.83
	PXSNL + NLV	**0.9925**	**0.047**	**5.77**	**32.95**
Burano	P+XS	0.997	0.086	5.99	32.7
	NLV	0.9977	0.0585	5.33	33.73
	PXSNL	0.9979	0.0616	5.07	34.12
	PXSNL + NLV	**0.9980**	**0.0515**	**4.87**	**34.46**
Maisons	P+XS	0.996	0.1257	5.37	33.59
	NLV	0.9973	0.0769	4.41	35.26
	PXSNL	0.9974	0.083	4.32	35.42
	PXSNL + NLV	**0.9977**	**0.0645**	**4.05**	**35.99**
Ballons	P+XS	0.995	0.0299	5.66	33.07
	NLV	0.9977	0.0191	3.85	36.84
	PXSNL	0.998	0.0201	3.54	37.12
	PXSNL + NLV	**0.9985**	**0.0157**	**3.05**	**38.41**
Feuilles	P+XS	0.9966	0.0750	6.06	32.81
	NLV	0.9978	0.0460	4.83	34.85
	PXSNL	**0.9979**	0.0479	**4.71**	35.01
	PXSNL + NLV	0.9975	**0.0391**	5.19	**35.20**

Image	Modèle	CC	SAM	RMSE	PSNR
	Valeurs idéales	1	0	0	max
Phare	P+XS	0.9936	0.0281	5.66	33.09
	NLV	0.9953	0.0257	4.87	34.39
	PXSNL	0.9956	0.0239	4.69	34.73
	PXSNL + NLV	**0.996**	**0.0215**	**4.49**	**35.11**
Satellite4	P+XS	0.9896	0.0520	8.09	29.99
	NLV	0.98898	0.0534	8.08	30.00
	PXSNL	0.9892	0.0517	7.98	30.10
	PXSNL + NLV	**0.9894**	**0.0516**	**7.94**	**30.16**
Satellite3	P+XS	0.986	0.0910	8.44	29.61
	NLV	0.986	0.0905	8.40	29.65
	PXSNL	0.9866	0.0881	8.26	29.80
	PXSNL + NLV	**0.9867**	**0.0871**	**8.21**	**29.85**
Satellite2	P+XS	0.9717	0.1096	4.89	34.34
	NLV	0.9869	0.0692	3.34	37.63
	PXSNL	0.9854	0.0829	3.48	37.28
	PXSNL + NLV	**0.9890**	**0.0583**	**3.07**	**38.35**
Satellite1	P+XS	0.9733	0.0980	5.59	33.18
	NLV	0.9876	0.0603	4.06	35.94
	PXSNL	0.9861	0.0698	4.14	35.76
	PXSNL + NLV	**0.9898**	**0.0511**	**3.79**	**36.53**

TABLE 3.1 – Résultats quantitatifs obtenus avec les images de la Figure 3.1, les moins bons résultats sont soulignés et les meilleurs sont en gras. Nous pouvons voir que le modèle P+XS donne les moins bons résultats sur chaque image testée et que le terme PXSNL couplé avec le terme NLV donne les meilleurs résultats.

Nous avons utilisé les mesures de qualité globale PSNR et RMSE, ainsi que les mesures SAM et CC qui calculent respectivement la qualité spectrale et spatiale des images fusionnées.

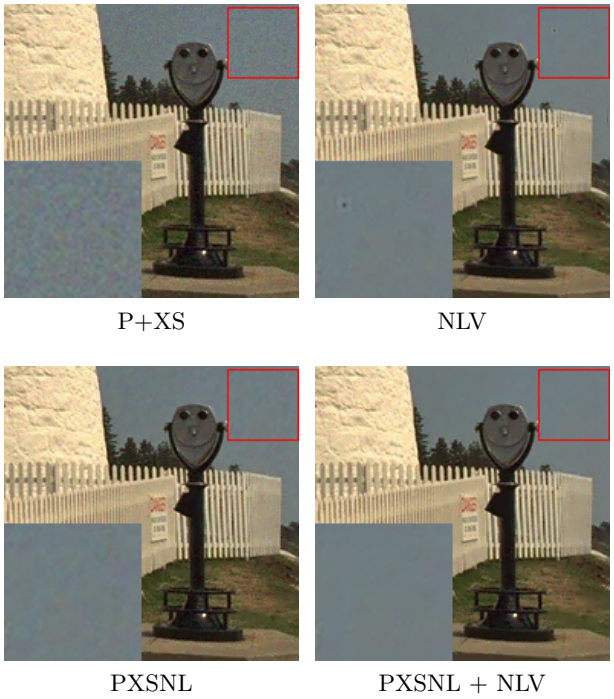


FIGURE 3.2 – Zooms sur les résultats obtenus avec l'image "Phare". Les termes PXSNL et NLV débruitent la solution contrairement au modèle P+XS.

Nous pouvons voir qu'avec les données simulées, le terme PXSNL proposé comparé aux modèles P+XS et NLV, donne de meilleurs résultats quantitatifs (Table 3.1) pour la majorité des images testées. Pour les images où les résultats quantitatifs sont inférieurs au modèle NLV, considérer les deux termes (NLV et PXSNL) permet d'obtenir un gain visuel et quantitatif. De plus, nous pouvons observer que contrairement au modèle P+XS le modèle que nous proposons débruite la solution

(Figure 3.2).

Nous avons également testé ces mêmes méthodes sur des images du satellite Pléiades. En utilisant notre modèle sur ces images, nous pouvons voir (Figure 3.3) qu'il est plus difficile d'observer de grandes différences entre les modèles comparés. Cependant, le modèle que nous proposons conserve très bien la structure de l'image panchromatique. Nous avons comparé ce résultat avec ceux donnés par deux méthodes basées réseaux convolutifs. Ces deux méthodes considèrent chacune, soit un Convolutional Neural Network [55] (PNN : Pansharpening Neural Network) ou un Deep Residual Neural Network [76] (DRPNN : Deep Residual Pansharpening Neural Network) adaptés pour le pansharpening. Ces réseaux donnent une bonne résolution spatiale mais une résolution spectrale plus faible que les méthodes variationnelles considérées. En effet, on peut voir que les résultats obtenus avec les images Pléiades présentent des distorsions au niveau du spectre (Figure 3.3). Cela peut être expliqué par les données utilisées lors de l'apprentissage de ces réseaux qui ne correspondent pas à celles utilisées pour tester ces réseaux.

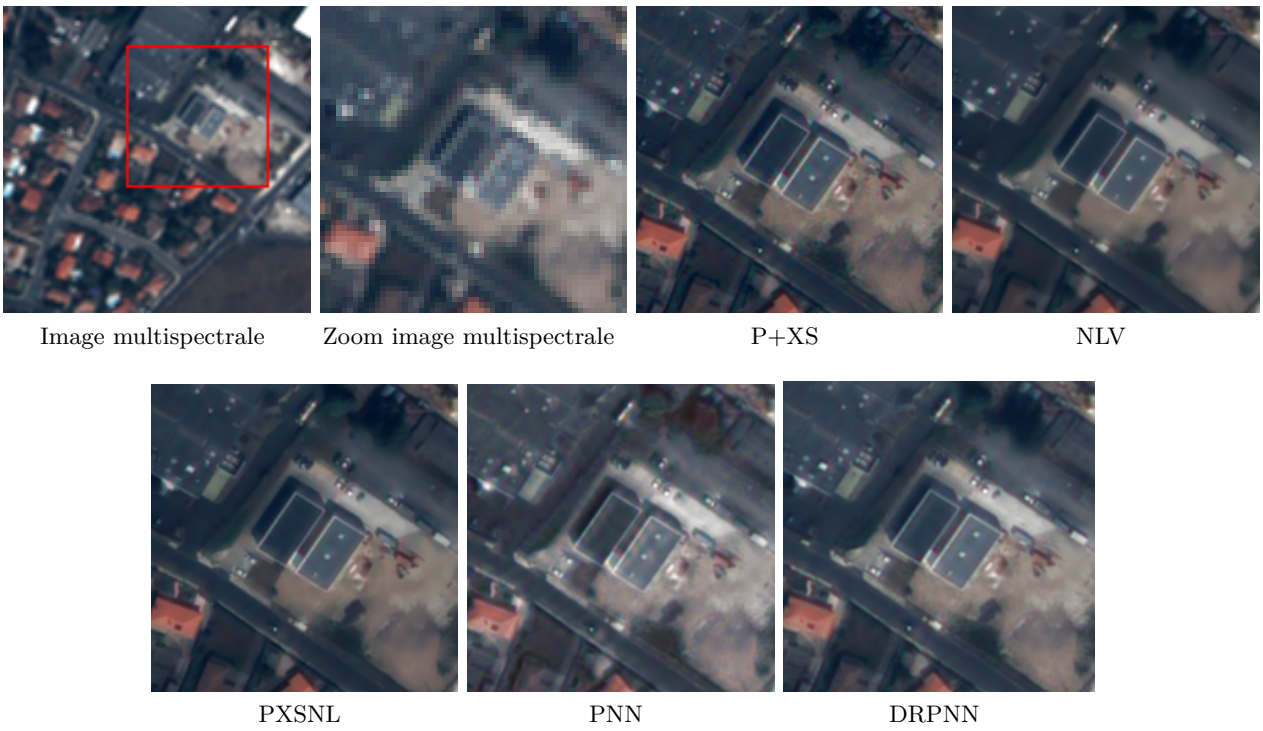


FIGURE 3.3 – Zooms sur des résultats obtenus avec une image Pléiades. Nous pouvons voir que le modèle que nous proposons (PXSNL) préserve mieux la géométrie que le modèle NLV, les contours sont plus nets. Les résultats obtenus avec les méthodes basées réseaux donnent une bonne résolution spatiale mais apporte un changement au niveau du spectre.

Ces résultats visuels restent néanmoins à prendre avec précaution car aucun des réseaux utilisés pour la comparaison n'a été ré-entraîné sur des données Pléiades. En effet, chaque satellite a des capteurs qui lui sont propres, ainsi, entraîner un réseau sur un seul satellite peut donner de mauvaises reconstructions spectrales les images d'autres satellites.

3.3 Conclusion

Nous avons proposé un terme de régularisation aux données non local permettant de transférer la géométrie de l'image panchromatique à l'image multispectrale haute résolution recherchée [24]. De plus, le terme que nous proposons permet de débruiter la solution en considérant les gradients

donnés par le voisinage en chaque point.

Les résultats obtenus sur des données simulées et satellites montrent une amélioration visuelle ou quantitative par rapport aux méthodes comparées. Cependant les résultats visuels sur les données satellites sont à considérer avec prudence pour l'instant.

Ainsi, l'utilisation de méthodes d'apprentissage reste encourageante car la résolution spatiale est bien reconstruite malgré l'absence de ré-entraînement pour les deux méthodes testées. C'est pourquoi la suite de cette thèse repose sur l'utilisation des réseaux convolutifs tout en ajoutant une dimension de modélisation du problème de pansharpening, que ce soit dans la fonction de perte ou l'architecture par exemple.

CHAPITRE 4

Reconstruction de la géométrie par l'utilisation de GANs

Ce chapitre est consacré aux méthodes basées sur les réseaux antagonistes génératifs (GANs) et présente notre deuxième contribution.

Dans un premier temps, un bref aperçu des outils utilisés est proposé. Une présentation des réseaux convolutifs et particulièrement des GAN est suivi d'un état-de-l'art sur les GANs afin de décrire plus précisément le cadre de travail de cette thèse.

Ensuite, une méthode RDGAN-Geom est proposée considérant un réseau générateur dense résiduel dans un contexte GANs tout en ajoutant une contrainte géométrique dans la fonction de perte inspirée par les termes de régularisation étudiés au chapitre précédent (Cf. Chap.3). Cette partie permet de mesurer l'influence de cette contrainte dans les résultats en fonction de l'architecture proposée.

4.1 Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs sont une catégorie particulière de réseau de neurones conçus pour le traitement d'images en particulier. Le premier réseau convolutif LeNet a été proposé par LeCun *et al.* en 1998 [43] mais l'engouement autour des réseaux convolutifs est survenu à la suite du réseau AlexNet [41], proposé par Krizhevsky *et al.* et fortement inspiré du réseau LeNet, pour le problème de classification.

4.1.1 Composition d'un réseau de neurones convolutif

Les réseaux de neurones convolutifs (ou CNN) permettent de modéliser efficacement la relation entre les variables d'entrée et de sortie par la composition de plusieurs niveaux. Généralement, ce sont des couches convolutives ou denses, des couches de pooling et des fonctions d'activations détaillées ci-dessous.

• Couches convolutives :

Les couches convolutives permettent d'extraire des caractéristiques de complexités variables. L'extraction de ces caractéristiques s'adapte au problème considéré par l'apprentissage des poids de chaque couche convolutive. C'est l'apprentissage automatique des poids qui rend les réseaux convolutifs très performants. Ces couches ont pour objectif de rechercher l'ensemble des caractéristiques des images d'entrée par filtrage convolutionnel. Ainsi, la carte de caractéristiques résultante peut être vue comme un filtre qui indique où se situent les caractéristiques d'intérêt dans l'image.

Cependant, pour un problème de reconstruction, la dernière couche convolutives est un peu particulière car elle a un noyau de taille 1×1. Dans ce contexte, son objectif n'est plus d'extraire des caractéristiques mais d'écraser la dimension profondeur du volume d'entrée afin de la faire correspondre avec la dimension de la sortie souhaitée.

• Couches de pooling :

La couche de pooling réalise une opération permettant de sous-échantillonner une carte de caractéristiques. Cette opération suit généralement une couche convolutive dans un réseau de classification. Les deux types de pooling utilisés sont :

▷ Le pooling maximal qui sélectionne la valeur maximale pour chaque voisinage, permettant ainsi de garder les caractéristiques les plus fortes,

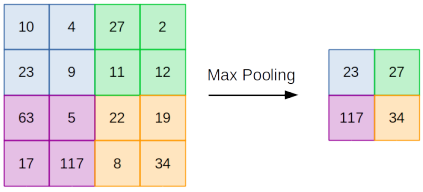


FIGURE 4.1 – Schéma représentatif d'une couche de pooling maximale.

▷ Le pooling moyennant qui calcule la valeur moyenne des caractéristiques dans chaque voisinage.

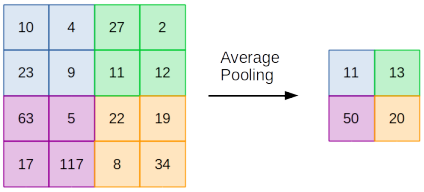


FIGURE 4.2 – Schéma représentatif d'une couche de pooling moyennant.

• Les fonctions d'activation :

Les fonctions d'activation ont pour objectif de reproduire le potentiel d'activation du cerveau humain et permettent donc le passage ou non de l'information. Par conséquent, ces fonctions vont décider si la réponse d'un neurone doit être activée ou non. Elles permettent d'introduire des complexités non-linéaires au réseau. En effet, considérer des fonctions non linéaires permet d'augmenter la capacité du réseau à modéliser des données plus complexes.

Les deux principales fonctions d'activations sont les suivantes :

▷ la fonction sigmoïde $g(x) = \frac{1}{1+e^{-x}}$ qui est la plus ancienne.

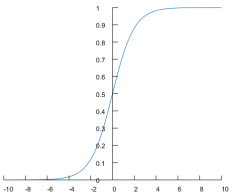


FIGURE 4.3 – Fonction sigmoïde définie par $g(x) = \frac{1}{1+e^{-x}}$.

Cette fonction écrase les valeurs d'entrée en condensant la sortie entre $[0, 1]$. De plus, cette fonction est différentiable, ce qui la rend appréciable lors de l'entraînement d'un réseau.

Cependant, elle est de moins en moins utilisée car elle peut être la cause du problème de la dégénérescence du gradient vers 0 si le réseau est trop profond. En effet, elle perd l'information due à la saturation et donc le gradient a de très forte chance d'arriver à 0 lorsque les valeurs sont très grandes ou très petites.

▷ la fonction ReLU (Rectified Linear Unit) $g(x) = max(0, x)$.

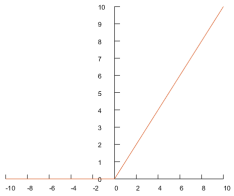


FIGURE 4.4 – Fonction ReLU définie par $g(x) = max(0, x)$.

Cette fonction laissent toutes les valeurs supérieures à 0 inchangées et attribue 0 aux valeurs négatives. La plupart des réseaux utilisent cette fonction ou l'une de ses variantes telles que leaky ReLU ou ELU. Les variantes ont généralement pour objectif de nuancer les valeurs négatives afin que le gradient ne soit pas toujours égal à 0 dans cet intervalle.

Par exemple, la fonction leaky ReLU $g(x) \begin{cases} ax & si\ x < 0\ avec\ a > 0 \\ x & sinon \end{cases}$, ou la fonction ELU (Exponential Linear Unit) $g(x) \begin{cases} a(e^x - 1) & si\ x < 0\ avec\ a > 0 \\ x & sinon \end{cases}$ attribuent une fonction linéaire ou exponentielle aux valeurs négatives.

En effet, ces fonctions d'activations sont différentiables ou sous-différentiables [58] donc très bien adaptées pour les algorithmes d'apprentissage avec rétro-propagation du gradient.

4.1.2 Entraînement d'un réseau de neurones convolutif

L'entraînement d'un CNN revient à optimiser les poids à chacune de ces couches. L'idée générale est de calculer l'erreur entre l'image prédiction et l'image de référence puis de mettre à jour ces poids à l'aide de l'algorithme de rétro-propagation du gradient de l'erreur.

Dans un cadre idéal, une fois arrivé à convergence, le réseau est quasiment aussi performant sur les données tests que sur les données d'entraînement. Cependant, même si le réseau a convergé, plusieurs problèmes peuvent apparaître. C'est ce qui rend l'apprentissage parfois difficile à maîtriser :

— Le problème de sur-apprentissage : le modèle appris n'est pas aussi performant sur les données tests que sur les données d'apprentissage. Une solution peut être d'élargir la base de données d'apprentissage.
— Le problème de sous-apprentissage : le modèle n'est pas assez performant que ce soit sur les données d'entraînement ou les données tests. De manière générale, cela provient d'un problème d'architecture. En effet, le modèle n'est pas assez complexe pour capturer toute la richesse de la distribution des données d'entraînement. Choisir une architecture plus complexe, plus profonde ou plus adaptée au problème, peut réduire ce risque.

4.2 Réseaux antagonistes génératifs (GANs)

4.2.1 Généralités

Les GANs (Generative Adversarial Networks ou réseaux antagonistes génératifs en français) sont une classe d'algorithmes d'apprentissage non-supervisé, introduit par Goodfellow *et al.* [30] en 2014.

Ce type de réseau convolutif cherche à imiter n'importe quelle distribution de données. De manière générale, un GAN est un modèle génératif où deux réseaux sont placés en compétition. Le premier réseau est le générateur G. Il génère un échantillon, tandis que son adversaire, le réseau discriminant D essaie de détecter si un échantillon est réel ou bien s'il s'agit du résultat du générateur.

— Le discriminateur D est un réseau de classification basique, qui retourne la probabilité que l'échantillon en entrée appartienne à la base de données (qu'il soit réel). Idéalement, on souhaite que, pour un échantillon y, $D(y) = 1$ si y appartient à la base de données et $D(y) = 0$ si y est un échantillon généré.

— Le générateur G est situé avant le discriminateur, dont l'entrée est un échantillon aléatoire et dont la sortie est un échantillon généré. L'objectif de G est alors de tromper le réseau D afin de faire croire que l'échantillon généré est réel.

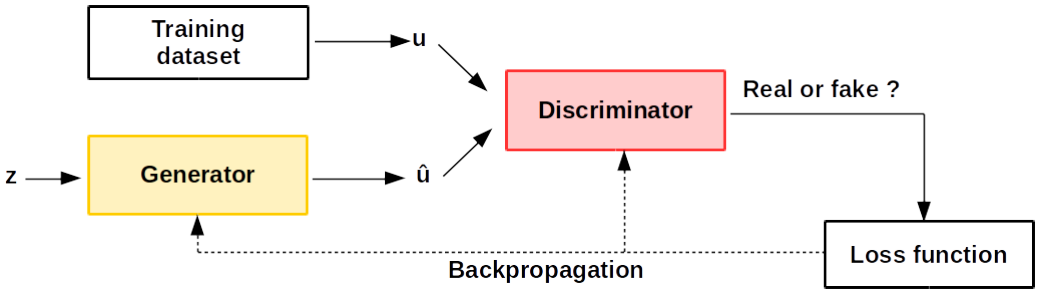


FIGURE 4.5 – Schéma représentatif d'un GAN.

Comme l'objectif est de générer un échantillon selon la distribution des données que l'on possède, la fonction de perte considérée repose sur l'entropie croisée qui est utilisée pour évaluer la précision de prévisions probabilistes [30] :

$$L(\hat{y}, y) = ylog(\hat{y}) + (1 - y)log(1 - \hat{y}). \tag{4.1}$$

Cette formule permet de quantifier la différence entre deux distributions, c'est un moyen de savoir à quel point la distribution que l'on cherche est proche de celle que l'on souhaite.

On note p_{data} la distribution représentée par les données, x suit la distribution p_{data} et z un échantillon aléatoire qui suit la distribution p_z. On doit alors considérer le label pour les données venant de la distribution p_{data} et le label venant des données générées suivant la distribution p_z.

— D'un côté, nous avons les données venant du jeu de données. On veut qu'elles soient labellisées 1 car on veut que $D(x) = 1$. Alors en prenant $y = 1$ et $\hat{y} = D(x)$ dans (4.1), on obtient :

$$L(D(x), 1) = log(D(x)). \tag{4.2}$$

En effet, si l'on trace la courbe représentative de (4.2) (Figure 4.6 (a)), la fonction est croissante de $]-\infty, 0]$ sur $[0, 1]$ et le maximum est atteint en 1, on en déduit que maximiser (4.2) va forcer les données réelles à avoir une probabilité de 1.

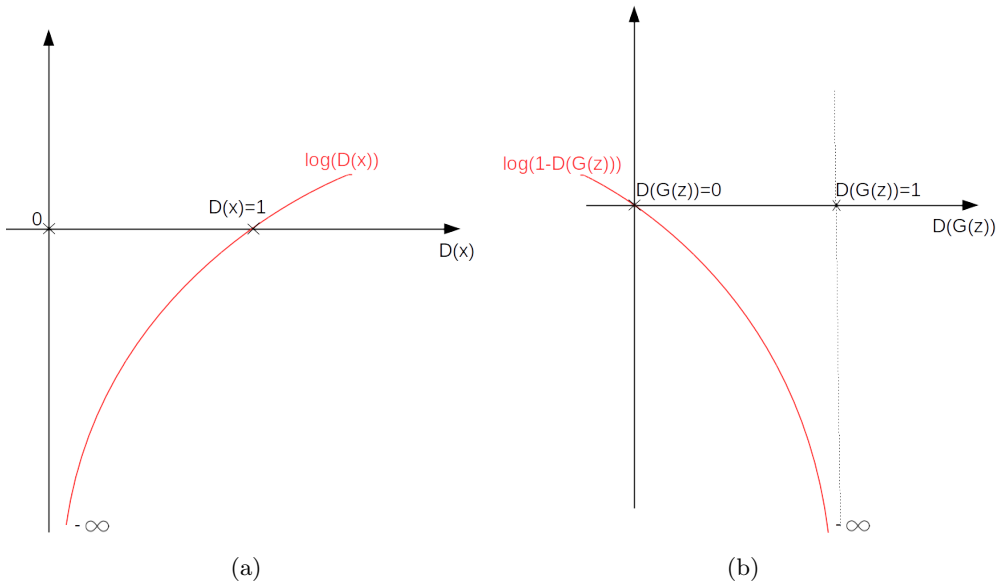
— De l'autre côté, nous avons les données venant du générateur. Dans ce cas, on veut que le label soit 0 car on veut $D(G(z)) = 0$. Alors en prenant $y = 0$ et $\hat{y} = D(G(z))$ dans (4.1), on obtient :

$$L(D(G(z)), 0) = log(1 - D(G(z))). \tag{4.3}$$

Et si l'on trace la courbe associée à (4.3) (Figure 4.6 (b)), la fonction est décroissante de $[0, -\infty[$ sur $[0, 1]$ et le maximum est atteint en 0. Maximiser (4.3), revient donc à forcer les données générées à avoir une probabilité de 0.

L'objectif de D étant de bien classer les données réelles et les données générées, cela revient à maximiser les équations (4.2) et (4.3) :

$$\max_D \; log(D(x)) + log(1 - D(G(z))). \tag{4.4}$$



(a) (b)

FIGURE 4.6 – Fonctions présentes dans le terme (4.4).

Parallèlement, l'objectif de G est de tromper le discriminateur, c'est-à-dire que G est entraîné à minimiser la probabilité que le discriminateur fasse la bonne prédiction pour une donnée générée, donc l'on veut $D(G(z)) = 1$. En regardant sur la figure 4.6 (b), on voit que cela revient à minimiser $log(1 - D(G(z)))$ par rapport aux paramètres de G.

Au final, la fonction de perte est alors

$$\min_G \max_D \; \mathbb{E}_x \left[log(D(x))\right] + \mathbb{E}_z \left[log(1 - D(G(z)))\right]. \tag{4.5}$$

L'espérance est ajoutée car on utilise plusieurs échantillons (selon x ou z).

4.2.2 Limitations

Les GAN présentent deux principaux inconvénients. Le premier est le problème du *vanishing gradient*. En effet, ce phénomène peut surgir lors de la phase d'apprentissage des paramètres du réseau G. Le générateur G n'étant pas encore assez entraîné, il va donc produire de faux résultats facilement reconnaissables par le discriminateur D. Lors de la minimisation de l'entropie croisée

$$\min_G \; \mathbb{E}_z \left[log(1 - D(G(z)))\right], \tag{4.6}$$

la dérivée va être très proche de 0 et ainsi la descente de gradient sera très lente ou n'aura pas lieu (Figure 4.7(a)). Par conséquent, G peut produire des résultats très similaires (et non réalistes) pendant de nombreuses itérations.

Le second inconvénient est le fait que la fonction de perte basique (Figure 4.6(b)) est strictement décroissante et non bornée inférieurement, ainsi la fonction va vers $-\infty$ pendant le processus de minimisation.

C'est pour cela qu'en pratique, il est préférable de changer la fonction (4.6) par

$$\max_G \mathbb{E}_z \left[log(D(G(z))) \right]. \tag{4.7}$$

Cela signifie qu'au lieu de minimiser la probabilité que D fasse une bonne prédiction, on cherche à maximiser la probabilité que D fasse une mauvaise prédiction (Figure 4.7(b)).

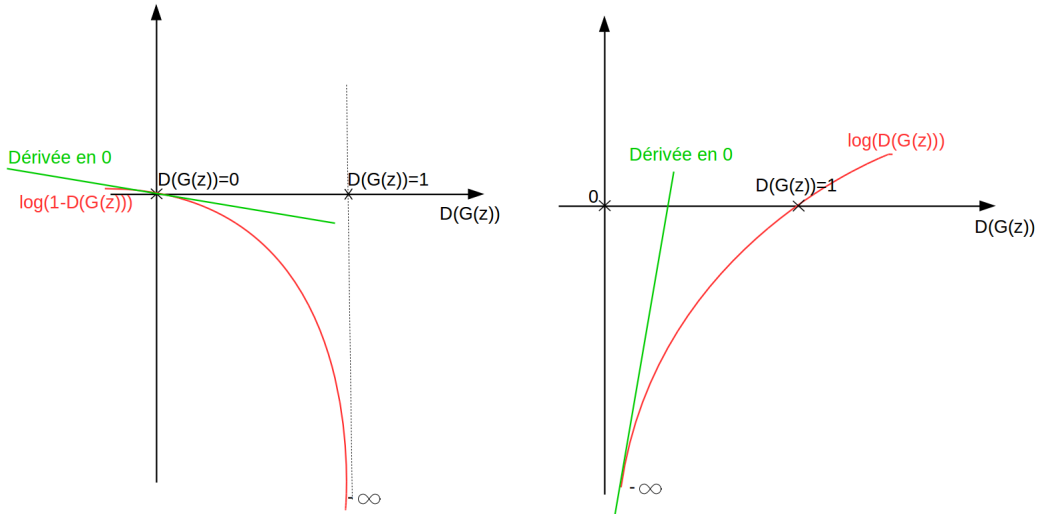


FIGURE 4.7 – Comparaison des fonctions utilisées pour l'optimisation des paramètres de G.

4.2.3 Optimisation des poids des réseaux

L'optimisation des poids du générateur et du discriminateur se fait de manière alternée, en optimisant sur D puis sur G :

- ▶ Faire k pas de montée de gradient pour le discriminateur :
  - ▷ Générer m échantillons aléatoires $\{z_1, ..., z_m\}$ à partir d'une loi $p(z)$.
  - ▷ Prendre m échantillons $\{x_1, ..., x_m\}$ à partir de $p_{data}(x)$.
  - ▷ Étape de SGD (Descente de Gradient Stochastique) :

$$\theta_D \leftarrow \theta_D + \varepsilon \nabla_{\theta_D} \frac{1}{m} \sum_{i \leq m} \left[log(D(x_i)) + log(1 - D(G(z_i))) \right] \tag{4.8}$$

- ▶ Faire un pas de descente de gradient pour le générateur :
  - ▷ Générer m échantillons $\{z_1, ..., z_m\}$ à partir de $p(z)$.
  - ▷ Étape de SGD :

$$\theta_G = \theta_G - \varepsilon \nabla_{\theta_G} \frac{1}{m} \sum_{i \leq m} -log(D(G(z_i))) \tag{4.9}$$

où

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i \leq m} \left[log(D(x_i)) + log(1 - D(G(z_i))) \right] = \frac{1}{m} \sum_{i \leq m} \left[\frac{1}{D(x_i)} \nabla D(x_i) + \frac{1}{1 - D(G(x_i))} \nabla D(G(z_i)) \right], \tag{4.10}$$

et

$$\nabla_{\theta_G} \frac{1}{m} \sum_{i \leq m} -log(D(G(z_i))) = \frac{1}{m} \sum_{i \leq m} -\frac{1}{D(G(z_i))} D'(G(z_i)) \nabla G(z) \tag{4.11}$$

En pratique, l'hyperparamètre k est choisi égal à 1 et l'étape SGD (Stochastic Gradient Descent) se fait en utilisant l'algorithme ADAM (ADAptive Moment estimation). ADAM [38] est un algorithme d'optimisation de descente de gradient stochastique adaptatif spécialement conçu pour la formation de réseaux de neurones profonds. C'est un algorithme dans lequel le gradient utilisé dans chaque itération est mis à jour à partir du précédent en utilisant une technique basée sur les moments.

En plus de l'hyperparamètre k lié à l'algorithme de descente, le paramètre m, qui correspond à la taille du batch, est très important. Il permet d'optimiser au mieux les poids du réseau pour l'ensemble des données. Une taille de batch trop petite ne permettra pas au réseau de capter pleinement la distribution des données d'entraînement. En effet, lors de l'entraînement, les poids sont optimisés en faisant la moyenne des gradients de chaque image du batch. Ainsi, il faut qu'il y ait suffisamment d'images pour avoir une estimation correcte. Si il y a trop peu d'images par batch, les poids du réseau peuvent alors changer drastiquement d'une itération à l'autre et donc impacter les performances. D'autre part, une taille de batch trop grande peut ralentir fortement l'entraînement car beaucoup d'images doivent être prises en compte à chaque passage.

4.3 État-de-l'art basé GAN

• Ledig *et al.* [44] sont les pionniers à proposer une méthode de super-résolution utilisant un GAN. Afin d'obtenir un bon générateur G_{θ_G}, les auteurs utilisent une architecture GAN du type résiduelle (Figure 4.8) pour le générateur et un réseau de classification combinant des couches convolutives et denses pour le discriminateur.

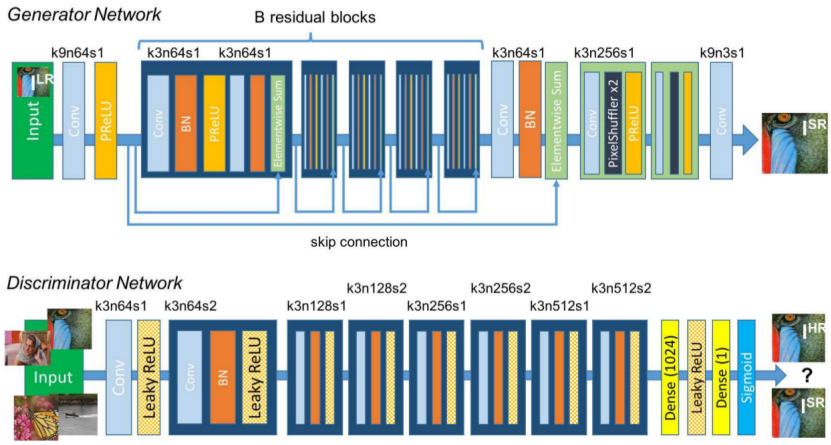


FIGURE 4.8 – Architecture utilisée par les auteurs pour le problème de super-résolution où k correspond à la taille du noyau, n à la taille de la carte de caractéristiques et s le décalage de la fenêtre à chaque couche.

Cependant, les auteurs proposent d'utiliser une autre fonction de perte pour le générateur que celle classiquement utilisée dans un cadre GAN. Un terme de perte perceptuelle, en plus de l'entropie croisée, est ajouté. La perte "adversarial" pousse la solution à avoir l'air réaliste en utilisant le réseau discriminant entraîné pour différentier les images réalistes des images non réalistes. Le terme perceptuel est motivée par le fait qu'une fonction de perte perceptuelle donne des résultats visuellement meilleurs qu'une fonction de perte dans l'espace des pixels. Les auteurs proposent alors la fonction suivante :

$$l^{SR} = l^{SR}_{VGG/ij} + 10^{-3} l^{SR}_{Gen}, \tag{4.12}$$

où

$$l^{SR}_{VGG/ij} = \frac{1}{W_{ij}H_{ij}} \sum_{x \leq W_{ij}} \sum_{y \leq H_{ij}} (\phi_{ij}(I^{HR})_{x,y} - \phi_{ij}(G_{\theta_G}(I^{LR}))_{x,y})^2, \tag{4.13}$$

W_{ij} et H_{ij} correspondent à la taille de l'image à la sortie de la couche ϕ_{ij}, I^{HR} l'image haute résolution de référence et I^{LR} l'image en basse résolution qui sert en entrée du réseau générateur G_{θ_G}. Et

$$l^{SR}_{Gen} = \sum_{n \leq N} -log(D_{\theta_D}(G_{\theta_G}(I^{LR}))), \tag{4.14}$$

où N correspond à la taille du batch.

• Liu *et al.* ont été les précurseurs de l'utilisation des GANs pour le problème de pansharpening. Dans un premier temps [50], ils ont considéré une simple architecture composée d'un empilement de couches convolutives pour les réseaux générateur et discriminateur. Plus récemment [49], ils ont proposé une extension de cette méthode visant à améliorer les performances de leur algorithme en changeant l'architecture du générateur. Pour ce faire, ils considèrent une architecture du type résiduel encodeur-décodeur comprenant deux sous-réseaux.

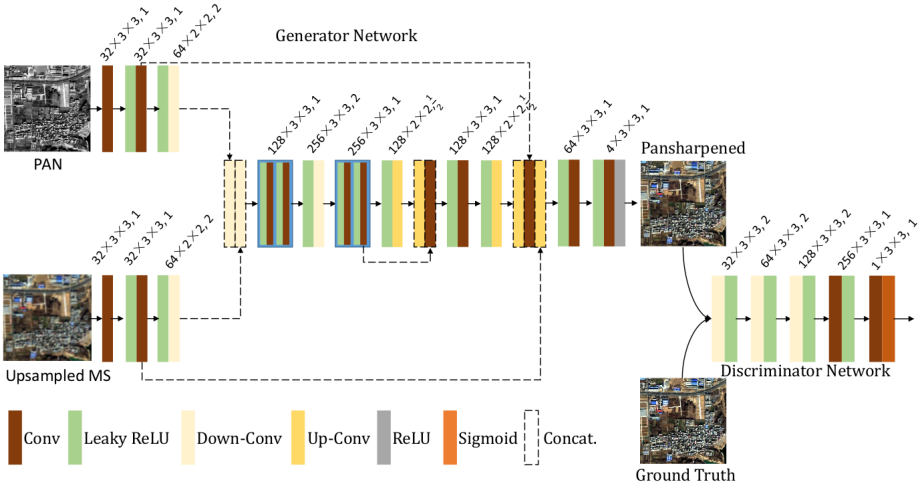


FIGURE 4.9 – Schéma de l'architecture proposée par Liu *et al.* pour la méthode PSGAN.

Ces deux sous-réseaux (Figure 4.9) permettent d'extraire les caractéristiques de chacune des images panchromatique et multispectrale avant de les concaténer pour les utiliser en entrée du réseau principal.

Finalement, les poids de ces réseaux sont optimisés en utilisant les termes d'entropie croisée et en ajoutant un terme l_1 dans la fonction de perte du générateur :

$$\mathcal{L}(G) = \sum_{n \leq N} -\alpha log(D(u_S, \tilde{u}))) + \beta ||u - \tilde{u}||_1 \tag{4.15}$$

et

$$\mathcal{L}(D) = \sum_{n \leq N} 1 - log(D(u_S, \tilde{u})) + log(D(u_S, u)), \tag{4.16}$$

où $\tilde{u} = G(u_S, P)$ est l'image reconstruite par le générateur G qui prend en entrée l'image panchromatique P et l'image multispectrale basse résolution u_S. D est le discriminateur, N la taille du batch et α et β sont des paramètres permettant de mettre plus ou moins de poids sur les différents

termes de la fonctionnelle.

• Zhang *et al.* ont proposé un réseau SFTGAN [86] utilisant un réseau convolutif profond considérant un module SFT (Spatial Features Transfer) permettant de transférer les caractéristiques spatiales. Cette structure a pour objectif de reproduire les caractéristiques spatiales de l'image panchromatique dans l'image multispectrale.

Introduites par Wang *et al.* [74] pour le problème de super-résolution, les couches SFT sont initialement conditionnées par des cartes de segmentation pour lesquelles ces couches génèrent une paire de paramètres ensuite utilisés pour appliquer par transformation affine sur l'entrée de la couche.

Ainsi, *Zhang et al.* ont choisi d'adapter l'utilisation de ces couches SFT pour le problème de pansharpening. Dans ce contexte, ces couches génèrent deux paires de paramètres, une pour l'image panchromatique et la seconde pour l'image multispectrale, afin de prendre en compte par la suite les caractéristiques de l'image panchromatique et celles de l'image multispectrale.

Les auteurs considèrent alors l'architecture suivante pour le générateur :

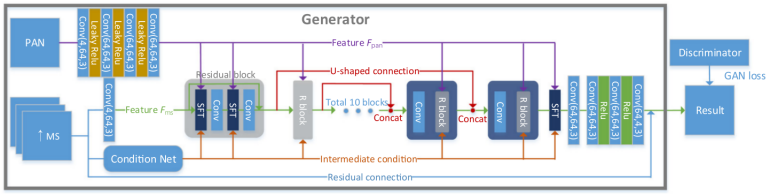


FIGURE 4.10 – Architecture du générateur pour la méthode SFTGAN proposée par Zhang *et al.*.

Ensuite, les poids des réseaux sont optimisés à l'aide des fonctions de perte suivantes :

$$L(G_\theta) = \sum_{i=1}^{N} [\alpha log(D_\eta(G_\theta(\uparrow u_S, P))) + \delta ||u - G_\theta(\uparrow u_S, P)||_1] \tag{4.17}$$

et

$$L(D_\eta) = \sum_{i=1}^{N} [log(1 - D_\eta(G_\theta(\uparrow u_S, P))) + log(D_\eta(u))] \,, \tag{4.18}$$

où u est l'image de cible, P l'image panchromatique et $\uparrow u_S$ l'image multispectrale basse résolution agrandie à la taille de P par une interpolation bicubique.

4.4 Contribution : méthode RDGAN

La préservation spatiale étant un des points importants du problème de pansharpening, plusieurs modèles ont été mis en place pour répondre à cette problématique. Par exemple, Yang *et al.* ont proposé un réseau PanNet [79] entraîné dans les hautes fréquences de l'image.

De notre côté, nous proposons une méthode fondée sur les GANs considérant une architecture dense résiduelle tout en ajoutant une contrainte géométrique dans la fonction de perte. Ce terme a été choisi en fonction de l'étude des termes de régularisation présentée au Chap.3.

4.4.1 Architecture

Dans la littérature sur les GANs, beaucoup de méthodes, par exemple les méthodes PSGAN [50, 49] et SFTGAN [86], considèrent les fonctions de perte suivantes :

$$L(G_\theta) = \sum_{i \leq N} \alpha log(D_\eta(G_\theta(z))) + \delta ||u - G_\theta(z)||_1 \tag{4.19}$$

pour le générateur et

$$L(D_\eta) = \sum_{i \leq N} log(1 - D_\eta(G_\theta(z))) + log(D_\eta(u)) \tag{4.20}$$

pour le discriminateur, où z est l'entrée du réseau, D_η le discriminateur, G_θ le générateur et N la taille du batch. Ces fonctions utilisent les termes d'entropie croisée classiques dans un cadre GAN tout en ajoutant une norme l_1 de la différence entre l'image de référence u et l'image reconstruite $G_\theta(z)$, permettant ainsi de guider la reconstruction en se basant sur l'image de référence.

De plus, une grande partie des méthodes proposées prennent en entrée les images panchromatique et multispectrale pour obtenir en sortie une image multispectrale haute-résolution.

Dans ce contexte de pansharpening où la reconstruction des hautes fréquences est primordiale, nous décidons de travailler dans un cadre résiduel. Ainsi, le réseau est dédié à l'apprentissage des détails spatiaux et spectraux hautes fréquences manquant dans les images initiales. Cela signifie que notre réseau prend en entrée les images panchromatique et multispectrale et la sortie du réseau correspond à une image résiduelle. Ce résidu contient alors toutes les informations spatiales et spectrales nécessaires à la haute résolution. L'image multispectrale finale est obtenue en ajoutant ce résidu à l'image multispectrale sur-échantillonnée à l'aide d'une interpolation bicubique. Cela revient à reformuler les fonctions de perte (4.20) et (4.19) dans un cadre résiduel :

$$\mathcal{L}(G_\theta) = \sum_{i \leq N_b} \alpha log(D_\eta(G_\theta(\uparrow y, P) + \uparrow y)) + \delta ||u - G_\theta(\uparrow y, P) - \uparrow y||_1 \tag{4.21}$$

et

$$\mathcal{L}(D_\eta) = \sum_{i \leq N_b} log(1 - D_\eta(G_\theta(\uparrow y, P) + \uparrow y)) + log(D_\eta(u)). \tag{4.22}$$

Ensuite, nous considérons une architecture dense résiduelle, présentée en Figure 4.11, pour le générateur. Cette architecture permet de combiner les avantages des architectures denses [33] et résiduelles [36].

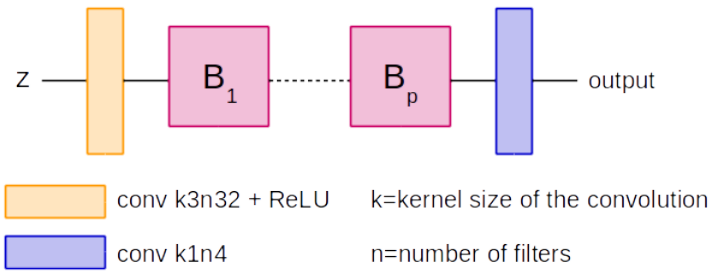


FIGURE 4.11 – Architecture générale utilisée pour le générateur, où l'entrée du réseau $z = [P, \uparrow y]$ est la concaténation [.] de l'image panchromatique P et de l'image multispectrale $\uparrow y$, $y = (y^1, ..., y^N)$, ré-échantillonnée à la taille de P. Les blocs B_i, $i \leq p$, sont les blocs denses résiduels présentés en Figure 4.14.

Elle prend en entrée $z = [P, \uparrow y]$ qui est la concaténation de l'image panchromatique P et de l'image multispectrale basse-résolution $\uparrow y$ interpolée à la taille de P et renvoie $G_\theta(z)$ correspondant

à une image résiduelle. Pour simplifier les notations, on appelle $\hat{u}$ l'image fusionnée finale , obtenue de la façon suivante :

$$\hat{u} = G_\theta(\uparrow y, P) + \uparrow y. \tag{4.23}$$

Les types d'architectures résiduelles et denses ont été introduites pour résoudre les problèmes de minimisation du gradient (i.e. *the vanishing gradient problem*), souvent rencontrés pendant la phase d'entraînement lors de l'utilisation de réseaux convolutifs profonds [28]. De manière générale, dans un réseau, les poids de celui-ci reçoivent une mise à jour proportionnelle à la dérivée partielle de la fonction de perte calculée en fonction du poids actuel à chaque itération pendant l'entraînement. Il est alors possible que le gradient devienne très petit et ainsi que les poids ne changent pas ou très peu d'une itération à l'autre. Par conséquent, le réseau n'apprend plus alors qu'il n'est pas encore arrivé à convergence.

Pour contrer ces problèmes, ces deux types d'architecture ont été proposés, les réseaux ResNet en 2016 et ensuite DenseNet en 2018. Ces architectures permettent d'entraîner des réseaux plus profonds. La principale différence entre ces deux architectures réside dans la façon d'injecter l'information au fil des couches convolutives. L'architecture ResNet additionne deux sorties de couches non consécutives alors que l'architecture DenseNet les concatène. Cependant, l'objectif reste le même : ré-injecter l'information obtenue aux niveaux des couches précédentes afin de garder un gradient non nul ou non proche de zéro.

• ResNet :

Cette architecture a été proposée par *He et al.* [33] pour la reconnaissance d'images en 2016 et est depuis utilisée pour divers problèmes comme la classification, la détection, la segmentation, etc. L'apprentissage résiduel signifie que chaque couche du réseau de neurones est uniquement responsable du réglage précis de la sortie d'une couche précédente en ajoutant simplement le résidu appris à la couche précédente à l'entrée de cette couche.

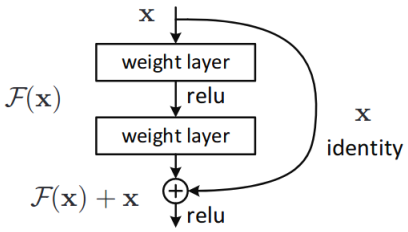


FIGURE 4.12 – Illustration d'une connexion résiduelle.

De manière très générale, un réseau résiduel peut se formuler à l'aide de l'équation suivante :

$$y = f(x) + x, \tag{4.24}$$

où x est l'entrée du bloc résiduel et $f(x)$ les couches convolutives constituant ce bloc. Alors à la couche l, on peut réécrire (4.24),

$$\begin{aligned} y_l &= f(x_l) + x_l \\ x_{l+1} &= ReLU(y_l). \end{aligned} \tag{4.25}$$

Dans cet exemple, l'entrée x_l est ajoutée à la sortie de la couche convolutive sans modification. Cependant, l'entrée peut être modifiée par une fonction H avant d'être ajoutée. Cela donne alors

$$\begin{aligned} y_l &= f(x_l) + H(x_l) \\ x_{l+1} &= ReLU(y_l). \end{aligned} \tag{4.26}$$

De même, la fonction ReLU appliquée peut être n'importe quelle autre fonction g :

$$\begin{aligned} y_l &= f(x_l) + H(x_l) \\ x_{l+1} &= g(y_l). \end{aligned} \tag{4.27}$$

Pour être plus précis, dans le cas d'un ResNet, $H = id$ car l'entrée d'origine n'est pas modifiée avant d'être additionnée. La fonction g, est également l'identité afin d'assurer que l'information soit bien transmise sans modification. En effet, si l'on considère $g \neq id$, l'entrée de la couche suivante ne contiendra plus exactement la même donnée qu'en entrée, elle aura été modifiée par g.

Pour finir, après N couches, on obtient alors

$$x_{l+N} = x_l + f(x_l) + f(x_{l+1}) + ... + f(x_{l+N-1}). \tag{4.28}$$

• DenseNet :

DenseNet est une architecture de réseaux de neurones convolutifs introduite par *Huang et al.* [36]. Afin d'améliorer le flux d'informations entre les couches du réseau, les auteurs proposent un nouveau schéma de connexion entre les couches du réseau : une connexion directe de chaque couche avec les suivantes.

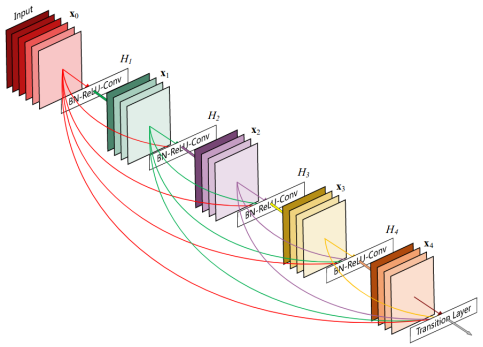


FIGURE 4.13 – Architecture d'un bloc dense dans un réseau de neurones convolutifs.

En effet, si on note H_l la transformation non linéaire donnée par la couche l du réseau et x_l la sortie de cette couche, la l^{ieme} couche reçoit les cartes de caractéristiques de toutes les couches précédentes, c'est à dire $x_0, ..., x_{l-1}$. Ainsi, chaque couche a accès au gradient à partir de la fonction de perte et du signal d'entrée, amenant à une supervision profonde implicite. Par conséquent, la sortie de la couche l s'écrit de la façon suivante :

$$x_l = H_l([x_0, ..., x_{l-1}]), \tag{4.29}$$

où [.] représente la concaténation.

Dans cette architecture, chaque bloc dense est séparé par une couche de transition consistant en une convolution avec un noyau de taille 1 suivie d'un pooling moyennant, dont le but est de réduire la taille de certaines dimensions.

Architecture dense résiduelle

Dans l'architecture proposée en Figure 4.11, chaque bloc dense résiduel est composé de quatre couches convolutives. Et la $l^{ième}$ couche d'un bloc prend en entrée les cartes de caractéristiques de toutes les couches précédentes (4.29). Cela signifie que chaque couche a accès à toute l'information antérieure dans un bloc. De plus, chacun de ces blocs est composé d'une connexion résiduelle représentée par l'addition de l'entrée de chaque bloc avec la sortie de la dernière couches. De cette façon, l'information en entrée de chaque bloc est transmise sans aucune modification.

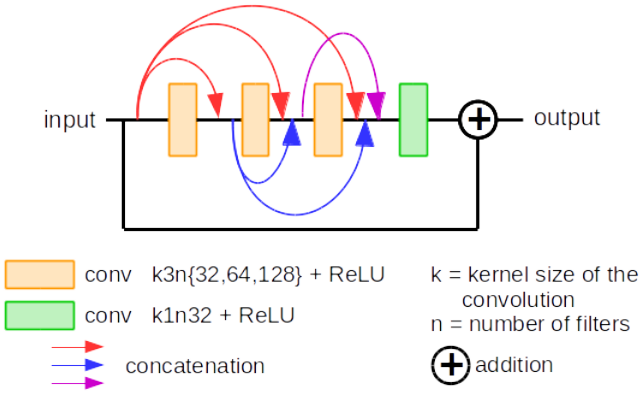


FIGURE 4.14 – Architecture de chaque bloc dense résiduel B_i pour $i = 1, ..., p$. Les connexions denses sont représentées par la concaténation de chaque sortie de couches précédentes et les connexions résiduelles par l'addition de deux sorties de couches différentes.

4.4.2 Fonction de perte

Dans un second temps, pour renforcer la reconstruction géométrique des images, nous proposons de considérer un terme de régularisation dans la fonction de perte du générateur. Ce terme est initialement proposé par Ballester *et al.* [10] dans un cadre variationnel pour le problème de pansharpening. Nous proposons alors de minimiser la fonctionnelle suivante :

$$L(G_\theta) = \sum_{i \leq N} \alpha log(D_\eta(\hat{u})) + \delta ||\hat{u} - u||_1 + \beta \sum_{x \in \Omega} |\nabla u(x)^\perp \cdot \nabla \hat{u}|, \tag{4.30}$$

où $\nabla(.)$ est le gradient, $\perp$ le vecteur orthogonal et Ω le domaine de l'image. Ce troisième terme force l'alignement des gradients de chaque bande spectrale de l'image reconstruite avec ceux de l'image cible. Cela permet ainsi de transférer la géométrie de l'image de référence à celle que l'on cherche à reconstruire. Cela se traduit alors par le produit scalaire entre le vecteur gradient orthogonal de l'image cible et le vecteur gradient de l'image reconstruite en tout point du domaine. En effet, un produit scalaire nul indique que les vecteurs sont colinéaires et par conséquent que la direction des gradients est préservée.

4.4.3 Résultats

- Base de données :

Des images satellites Pléiades et World View 3 sont utilisées pour entraîner et tester les différentes méthodes. Ces deux images ont été acquises dans la zone de Bordeaux, l'image Pléiades en août 2012 et l'image World View 3 en juillet 2018.

La base de données World View 3 est composée d'une image panchromatique et d'une image multispectrale de 4 canaux (bleu, vert, rouge et infra-rouge). La bande spectrale pour le bleu couvre les longueurs d'ondes comprises entre 450 nm et 510 nm, entre 510 nm et 580 nm pour le vert, entre 630 nm et 690 nm pour le rouge et entre 750 nm et 895 nm pour l'infra-rouge. Concernant l'image panchromatique, la bande spectrale couvre les longueurs d'ondes comprises entre 450 nm et 800 nm. La résolution spatiale est donnée par la distance d'échantillonnage au sol qui est de 0.31 m pour l'image panchromatique et 1.24 m pour l'image multispectrale. Cela impose un facteur 4 entre la basse et la haute résolution.

Pour les données Pléiades, la distance d'échantillonnage au sol est de 0.7 m pour l'image panchromatique avec la bande spectrale comprise entre 480 nm et 830 nm. L'image multispectrale quant à elle à une distance d'échantillonnage au sol de 2.8 m avec une bande spectrale comprise

entre 430 nm et 550 nm pour la bande bleue, 490 nm et 610 nm pour la verte, 600 nm et 720 nm pour la rouge et 750 nm et 950 nm pour l'infra-rouge.

Ces différentes caractéristiques indiquent que ces deux bases de données ont des caractéristiques spectrales et spatiales différentes mais le facteur de résolution spatiale entre la basse et la haute résolution est le même (égale à 4). La distance d'échantillonnage au sol est plus petite sur la base de données World View 3, ce qui signifie que l'on peut observer plus de détails sur ces images.

Les satellites fournissent des images représentant plusieurs km^2, et donc de plusieurs millions de pixels. Comme nous n'avons pas d'images de références pour entraîner et tester un réseau de neurones, nous utilisons le protocole de Wald [72], couramment utilisé pour le problème de pansharpening. Ainsi, la création des bases de données suit plusieurs étapes :

— Étape 1 : Les images panchromatiques I_p et multispectrales I_{ms} sont sous-échantillonnées d'un facteur 4 de la façon suivante :

$$\begin{cases} \hat{I}_p = SHI_p \\ \hat{I}_{ms} = SHI_{ms} \end{cases} \tag{4.31}$$

où S est un opérateur de sous-échantillonnage et H un opérateur de flou. L'image $\hat{I}_p$ joue alors le rôle de l'image panchromatique haute-résolution, $\hat{I}_{ms}$ celui de l'image multispectrale basse-résolution et I_{ms} est ainsi utilisée comme image de référence pour l'entraînement mais également pour l'évaluation des métriques telle que le PSNR par exemple.

— Étape 2 : Comme notre réseau prend en entrée l'image multispectrale sur-échantillonnée à la taille de la panchromatique, l'image $\hat{I}_{ms}$ est ensuite sur-échantillonnée à la taille de $\hat{I}_p$ à l'aide d'une interpolation bicubique :

$$\tilde{I}_{ms} = \uparrow_{bic} \hat{I}_{ms}, \tag{4.32}$$

où $\uparrow_{bic}$ représente l'interpolation bicubique.

— Étape 3 : Finalement, les images panchromatiques et multispectrales sont découpées en patchs de taille 128×128 et ces patchs sont triés afin d'enlever les zones inutilisables, en particulier les zones nuageuses où aucun détail spatial n'est visible ou alors les zones "sensibles", comme l'aéroport, qui ont été floutées par le fournisseur de données et qui ne sont donc pas exploitables.

• Métriques utilisées :

Pour évaluer et comparer les performances de notre méthode et des méthodes de l'état-de-l'art, nous utilisons les métriques PSNR, RMSE, CC, SAM et ERGAS présentées en détail au Chap. 2 2. Les mesures PSNR, RMSE et ERGAS sont des mesures globales, mesurant la qualité de la reconstruction dans sa globalité. Les mesures CC et SAM donnent une indication de qualité des reconstructions spatiale et spectrale respectivement.

• Détails d'implémentation :

La méthode proposée a été implémentée en Tensorflow avec ADAM pour minimiser la fonction de perte. Les tailles de batchs ont été fixées à 19 pour la base de données Pléiades et 17 pour la base de données World View 3. Les tailles de batchs ont été choisies de façon à diviser parfaitement le nombre d'images d'entraînement pour que chacune des images passe exactement une fois par epoch. De plus, nos bases de données sont relativement petites donc choisir une taille de batch trop grande n'est pas pertinent.

Les poids de la fonction de perte à l'Eq. 4.30 ont été paramétrés afin de donner les meilleures mesures quantitatives, i.e. $\alpha = 1$, $\gamma = 100$ et $\beta = 0.5$.

• Comparaison des méthodes :

La méthode proposée a été comparée à plusieurs méthodes de l'état-de-l'art. Notamment avec les méthodes de réseaux PSGAN [50] et PanNet [79] qui ont été ré-entraînés sur nos bases de données. Mais également avec la méthode variationnelle P+XS [10] dont notre terme de régularisation s'inspire.

Cependant, afin de valider l'architecture et l'apport du terme géométrique dans la fonction de perte du générateur, nous avons ajouté ce terme dans la fonction de perte de plusieurs architectures. Les résultats quantitatifs sont présentés en Tab.4.1

Méthode	PSNR	CC	SAM	RMSE	ERGAS
Valeur idéale	max	1	0	0	0
P+XS [10]	19.37	0.860	0.317	28.48	11.29
PanNet [79]	28.36	<u>0.950</u>	<u>0.157</u>	10.30	4.69
PSGAN [50]	<u>26.59</u>	0.952	0.155	<u>10.93</u>	4.23
PSGAN-Geom	27.18	0.955	0.145	10.14	3.99
GAN ResNet	28.38	0.960	0.153	10.07	4.57
GAN ResNet-Geom	28.42	0.961	0.152	10.02	4.51
GAN ResNet HF	28.24	0.9595	0.154	10.22	<u>4.81</u>
GAN ResNet HF-Geom	28.26	0.960	0.152	10.18	4.76
Residual dense l_2	29.29	0.965	0.143	9.04	4.00
RDGAN	29.37	**0.969**	0.141	8.94	**3.94**
RDGAN-Geom	**29.38**	**0.969**	**0.138**	**8.93**	**3.94**
RDGAN HF	29.17	0.967	0.145	9.15	4.03
RDGAN-Geom HF	29.20	0.968	0.144	9.11	4.00

TABLE 4.1 – Résultats quantitatifs obtenus sur les images tests Pléiades. Les meilleurs résultats sont en gras et les pires soulignés.

D'après les résultats en Tab.4.1, lorsque le terme géométrique est ajouté à la méthode PSGAN proposée par Liu *et al.*, les résultats quantitatifs sont améliorés mais l'architecture proposée par Liu *et al.* est très simple car composée d'une succession de couches convolutives. En changeant cette architecture pour une architecture de type ResNet, ligne 'GAN ResNet' du tableau, on voit que les résultats sont déjà améliorés et l'ajout du terme géométrique, 'GAN ResNet-Geom', améliore les résultats de façon moins marquée qu'avec l'architecture PSGAN. Ce test permet de mettre en évidence l'importance de l'architecture.

Nous avons également testé l'entraînement dans les hautes fréquences comme proposé par Yang *et al.* pour la méthode PanNet. Ces tests sont nommés dans le tableau avec l'abréviation HF. Les résultats ne semblent pas concluants dans le sens où les performances quantitatives sont inférieures quelles que soient les architectures et les fonctions de perte considérées.

Finalement, motivés par les performances de l'architecture de type ResNet et par l'ajout du terme de régularisation géométrique dans la fonction de perte du générateur, la méthode proposée 'RDGAN-Geom' donne les meilleurs résultats quantitatifs. Avec cette architecture, l'ajout du terme géométrique donne des résultats similaires ou légèrement meilleurs.

Un exemple sur les bandes RGB est affiché en Fig.4.15.

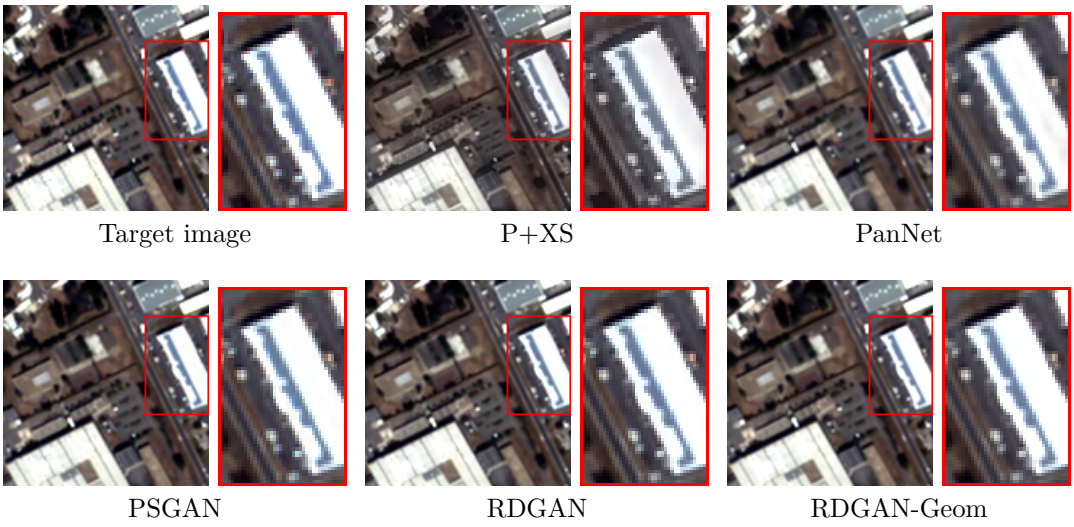


FIGURE 4.15 – Résultats obtenus avec les différentes méthodes sur un échantillon de la base de données Pléiades. Les résultats RGB ne permettent pas de conclure visuellement car on ne peut pas observer de différence majeure entre les méthodes.

Les résultats visuels obtenus sur les bandes RGB sont difficilement différentiables car il n'y a pas de différences perceptibles entre les méthodes comparées. Une représentation de la différence entre l'image de référence et l'image reconstruite permet toutefois de mieux percevoir les performances de ces méthodes en Fig.4.16.

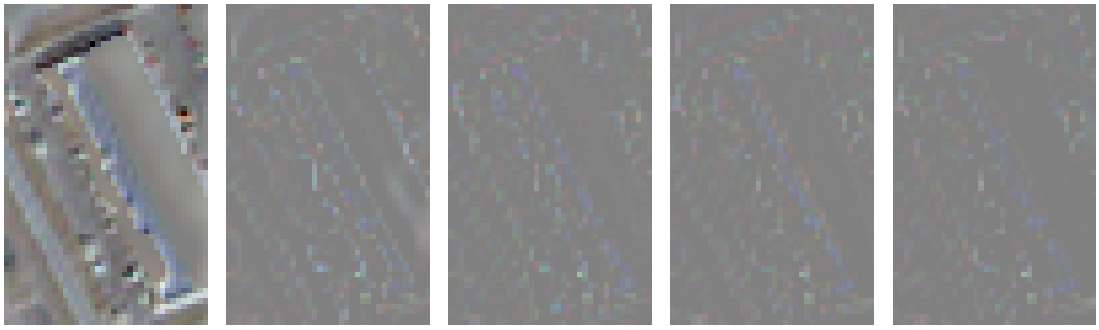


FIGURE 4.16 – Différence entre l'image de référence et l'image reconstruite obtenue la partie zoomée de la Figure 4.15. On peut voir que la méthode proposé reconstruit mieux la géométrie.

Sur cet exemple, présenté en Fig.4.16, on peut voir que la méthode proposée préserve mieux la géométrie car moins de différences sont visibles.

Afin de quantifier cette amélioration, nous avons proposé d'utiliser des mesures basées sur le gradient et le Laplacien définies de la façon suivante :

$$M_A(X,Y) = \frac{1}{|\Omega|}\sum_{\Omega} ||AX - AY||_2, \tag{4.33}$$

où $|\Omega|$ est le nombre de pixels et $A = \nabla$ ou $A = \Delta$. La mesure M_∇ permet de comparer géométrie dans l'image et M_Δ de comparer la détection des points [8]. On obtient alors les résultats suivants :

Modèle	M_∇	M_Δ
Valeur idéale	0	0
P+XS	<u>17.22</u>	<u>23.96</u>
PanNet	9.07	13.52
PSGAN	8.66	13.18
RDGAN	8.20	12.55
RDGAN-Geom	**8.10**	**11.87**

TABLE 4.2 – Résultats quantitatifs obtenus sur les images tests de la base de données Pléiades. Ces résultats montrent une meilleure préservation de la géométrie avec la méthode proposée RDGAN-Geom. Les meilleurs résultats sont en gras et les pires sont soulignés.

Ces résultats sont en adéquation avec les résultats visuels obtenus en Fig.4.16. On peut donc en conclure que ces mesures jugent bien la qualité de la reconstruction spatiale.

Nous avons également comparé notre méthode avec les méthodes de l'état-de-l'art sur la base de données World View 3. Les résultats quantitatifs sont présentés en Tab.4.3.

Modèle	CC	SAM	PSNR	ERGAS	RMSE	M_∇	M_Δ
Valeur idéale	1	0	max	0	0	0	0
P+XS	<u>0.887</u>	<u>0.166</u>	23.59	<u>10.35</u>	18.89	<u>16.07</u>	<u>25.26</u>
PanNet	0.930	0.135	<u>23.22</u>	8.01	<u>19.17</u>	14.81	23.61
PSGAN	**0.966**	0.110	29.30	**6.33**	9.61	9.99	17.16
RDGAN-Geom	**0.966**	**0.107**	**29.39**	6.58	**9.50**	**9.73**	**16.58**

TABLE 4.3 – Résultats quantitatifs obtenus sur les images tests de la base de données World View 3. Les meilleurs résultats sont en gras et les pires sont soulignés. La méthode proposée donnent des résultats similaires ou meilleurs pour l'ensemble des mesures utilisées.

De manière générale, la comparaison des résultats quantitatifs sur la base de données WorldView 3 est assez différente de ceux obtenus sur la base Pléiades. En effet, on remarque que la méthode PSGAN donne de meilleurs résultats sur cette base que sur la base Pléiades, au contraire de la méthode PanNet qui donne de beaucoup moins bons résultats sur la base WorldView 3. Cependant, pour l'ensemble des mesures, excepté ERGAS, la méthode proposée donne en moyenne de meilleurs résultats sur les données tests.

L'exemple visuel présenté en Fig.4.17 montre les résultats obtenus sur une image de la base de données WorldView 3. Sur cet exemple, on peut voir une différence sur les bandes RGB. Pour commencer, on remarque que la méthode PanNet reconstruit une image plus sombre. Le contraste ne semble pas bien reconstruit avec ce réseau. D'un point de vue des détails et des textures, on voit, dans la partie zoomée, que les méthodes P+XS et PanNet donnent des résultats plus lissés que les autres méthodes.

Ensuite, lorsque l'on regarde l'image des différences entre les hautes fréquences de l'image de référence et celle de l'image fusionnée (troisième et quatrième ligne en Fig.4.17), les observations faites sur les canaux RGB semblent confirmées concernant les détails géométriques. En effet, les méthodes P+XS et PanNet donnent les moins bons résultats et la meilleure reconstruction est obtenue avec la méthode proposée. Concernant la méthode P+XS, cela peut s'expliquer par la fait qu'il s'agit d'une méthode variationnelle qui n'est pas aussi performante qu'une méthode basée apprentissage. Il est également intéressant de relever que même si la méthode PSGAN donne le meilleur PSNR sur cette image, les détails géométriques sont mieux reconstruits avec la méthode RDGAN-Geom. Ce fait est mis en évidence par la mesure M_∇ qui est bien meilleure pour la méthode RDGAN.

Cet exemple visuel est donc représentatif des mesures quantitatives obtenues en Tab.4.3.

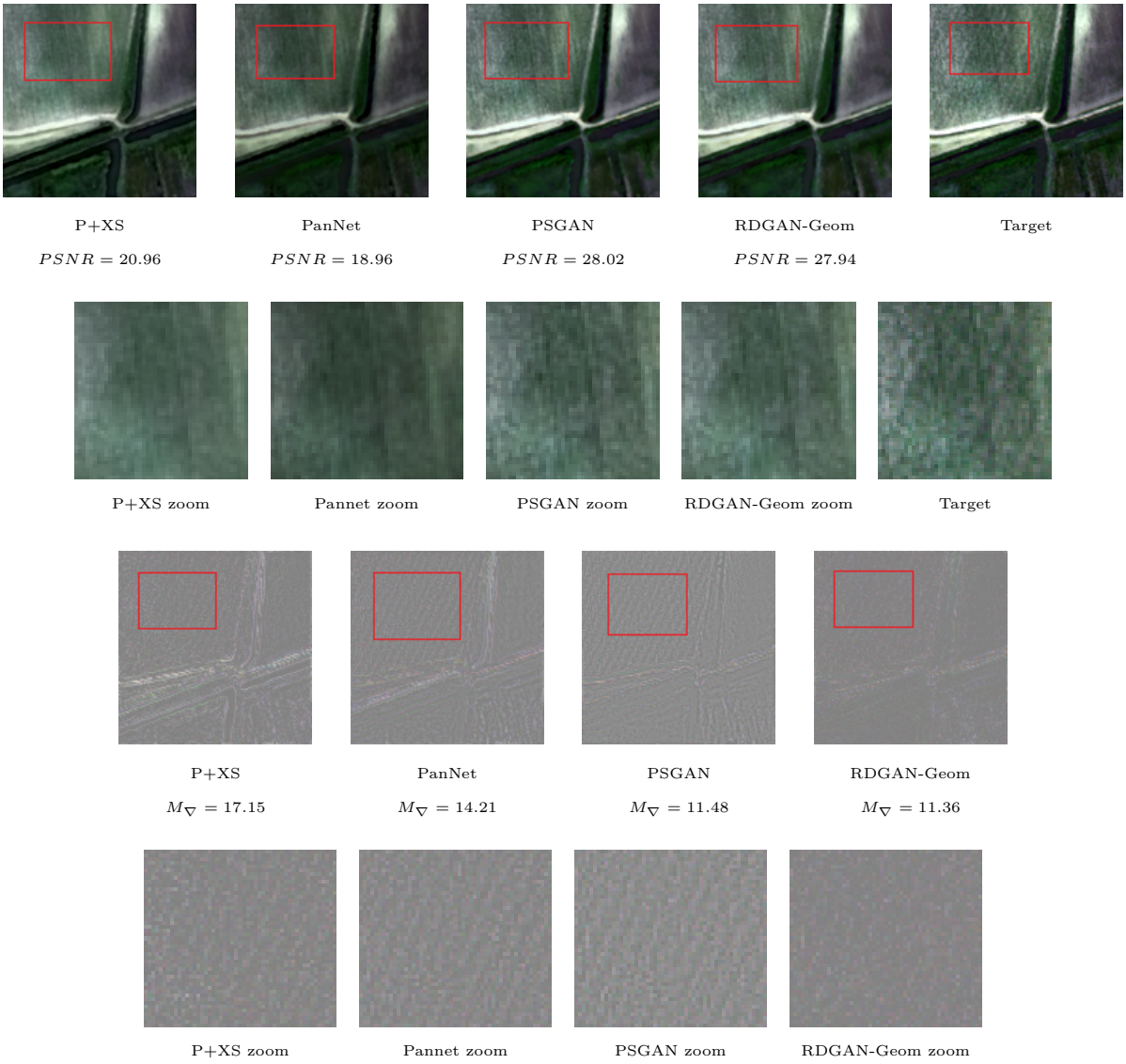


FIGURE 4.17 – Exemple avec la base de données WorldView 3. La première ligne affiche les canaux RGB et la seconde ligne affiche le zoom associé. La troisième et la quatrième lignes affichent la différence entre les hautes fréquences de l'image de référence et les hautes fréquences de l'image reconstruite.

4.5 Conclusion

Dans ce Chapitre, nous avons étudié l'influence d'un terme de régularisation dans la fonction de perte d'un réseau convolutif. Dans l'objectif d'améliorer la reconstruction géométrique, nous avons proposé une méthode basée GAN qui considère un terme de régularisation dans la fonction de perte du générateur [25]. Ce terme aligne le gradient de l'image à reconstruire avec celui de l'image de référence.

Les résultats montrent une amélioration quantitative et visuelle surtout lorsque l'on compare la reconstruction des hautes-fréquences des images reconstruites.

Cependant, la reconstruction spectrale est un point tout aussi important pour le problème de pansharpening et doit être prise en compte lors de la fusion des images. Une perspective d'amélio-

ration de cette méthode serait donc d'ajouter l'aspect spectral, que cela soit par l'architecture ou la fonction de perte.

Dans cet esprit, le chapitre suivant prend en compte non seulement sur la reconstruction spatiale mais également la reconstruction spectrale des images satellites.

CHAPITRE 5

Préservation des résolutions spatiale et spectrale dans un cadre GAN basé multi-discriminateur

Dans une perspective de pansharpening, la reconstruction spectrale est tout aussi importante que la reconstruction spatiale. Dans ce chapitre, nous proposons une nouvelle méthode modélisant ces deux aspects. Ainsi, nous présentons un second modèle, basé sur le précédent, considérant deux discriminateurs. De plus, afin d'équilibrer la reconstruction spectrale et la reconstruction spatiale, une contrainte spectrale a été ajoutée à la fonction de perte du générateur.

5.1 État-de-l'art basé multi-discriminateurs

• Zhu *et al.* [90] proposent une méthode de super-résolution basée sur une approche multi-discriminateurs. Cette méthode combine plusieurs discriminateurs dans le but d'améliorer la qualité visuelle et la précision de l'image générée de la façon suivante :

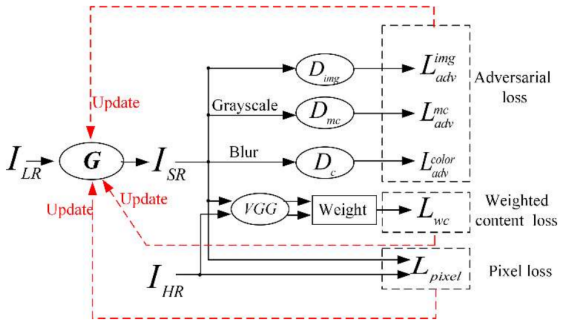


FIGURE 5.1 – Schéma de la méthode proposée par *Zhu et al.*

Plus précisément, ils proposent d'utiliser trois discriminateurs, chacun apportant une caractéristique à la solution. Dans la suite, on note I_{HR} l'image haute résolution, I_{LR} l'image basse résolution et I_{SR} l'image super-résolue.

— Le premier discriminateur permet de comparer les images entre elles. Ce réseau prend en entrée l'image générée et l'image haute résolution afin de les comparer sans modifications. La fonction de perte à minimiser est donc la même que dans un contexte GAN "traditionnel" :

$$\mathcal{L}_{D_{img}} = -log(D_{img}(I_{HR})) - log(1 - D_{img}(I_{SR})) \tag{5.1}$$

— Le second réseau compare les couleurs des images en entrée. Ce réseau prend en entrée les images $I^{\downarrow}_{HR}$ et $I^{\downarrow}_{SR}$ obtenues en sous échantillonnant les images I_{HR} et I_{SR} respectivement :

$$I^{\downarrow}_x = I_x * B, \quad B(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right). \tag{5.2}$$

La fonction de perte est alors :

$$\mathcal{L}_{D_c} = -log(D_c(I^{\downarrow}_{HR})) - log(1 - D_c(I^{\downarrow}_{SR})) \tag{5.3}$$

Ce choix de discriminateur est motivé par le fait que la vision humaine est sensible aux couleurs principales, à la luminosité et au contraste des objets présents dans l'image. Ainsi, les auteurs proposent de mettre en entrée du discriminateur les images sous-échantillonnées afin de conserver uniquement les couleurs principales sans trop de détails.

— Le troisième compare la géométrie et les textures. Il prend en entrée les images I^g_{HR} et I^g_{SR} obtenues en faisant une combinaison linéaire des bandes des images I_{HR} et I_{SR} respectivement :

$$I^g_x = 0.299 I_{x,red} + 0.578 I_{x,green} + 0.114 I_{x,blue}. \tag{5.4}$$

La fonction de perte est donc (toujours la même) :

$$\mathcal{L}_{D_{mc}} = -log(D_{mc}(I^g_{HR})) - log(1 - D_{mc}(I^g_{SR})) \tag{5.5}$$

Les auteurs proposent d'utiliser les images en niveaux de gris (sans teinte ni saturation) en entrée de ce réseau car ces images permettent de mieux souligner les détails géométriques et les textures.

La fonction de perte combinant ces trois réseaux discriminateurs est une combinaison linéaire des équations (5.3), (5.1) et (5.5) :

$$\mathcal{L}_D = \mathcal{L}_{D_{img}} + 10^{-1}\mathcal{L}_{D_{mc}} + 4.10^{-3}\mathcal{L}_{D_c} \tag{5.6}$$

Pour le générateur, les auteurs proposent de considérer deux termes dans la fonction de perte, souvent utilisés pour le problème de super-résolution :

$$\mathcal{L}_G = \mathcal{L}_{pixel} + \mathcal{L}_{adv} + \mathcal{L}_{low-level} + 10^{-5}\mathcal{L}_{high-level}, \tag{5.7}$$

où $\mathcal{L}_{adv}$ correspond au terme d'entropie croisée où intervient le générateur dans les équations (5.3), (5.1) et (5.5), $\mathcal{L}_{pixel}$ correspond à la norme l_2 entre la I_{SR} et I_{HR},

$$\mathcal{L}_{low-level} = ||\alpha^{SR}_{i,j}\Phi_{2,2}(I_{SR}) - \alpha^{HR}_{i,j}\Phi_{2,2}(I_{HR})||^2, \tag{5.8}$$

où α représente le poids spatial affecté à chacun des canaux (moyenne pondérée sur la troisième dimension pour obtenir une carte de caractéristique en 2D) et enfin

$$\mathcal{L}_{high-level} = ||\Phi_{5,4}(I_{SR}) - \Phi_{5,4}(I_{HR})||^2. \tag{5.9}$$

Les "fonctions" Φ représentent des sorties de couches VGG pré-entraînées [66], c'est-à-dire des cartes de caractéristiques. Plus Φ représente une sortie de couches profondes, plus le niveau de détails est élevé. Cela signifie que ces deux termes font la différence l_2 entre deux cartes de caractéristiques obtenues par sortie de couches VGG.

Pour chacune des deux fonctions de perte, les combinaisons linéaires des termes sont estimées par expérimentation.

Les résultats obtenus avec cette méthode ne sont pas les meilleurs d'un point de vue quantitatifs. En effet, le PSNR ne s'améliore pas mais on peut distinguer une amélioration visuelle.

• *Lee et al.* [45] proposent de considérer une architecture du type résiduel pour le générateur et des réseaux discriminateurs constitués de couches convolutives. Les réseaux discriminateurs sont entraînés afin de comparer les pixels, les couleurs et la géométrie (contours et textures) des images en entrée. La méthode proposée se présente de la façon suivante :

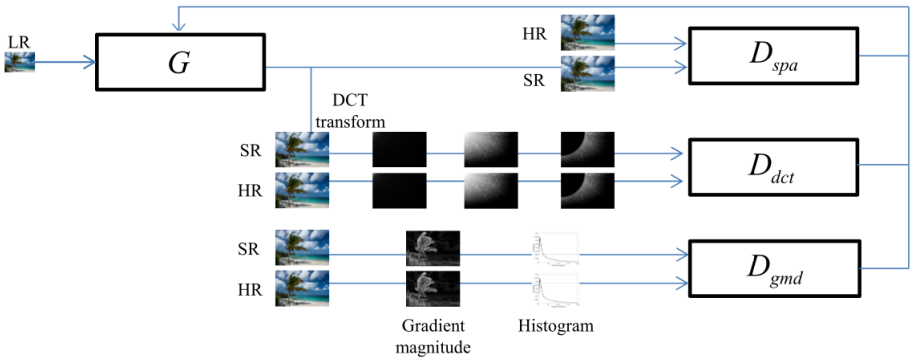


FIGURE 5.2 – Schéma de la méthode proposée par *Lee et al.*

— La premier réseau discriminateur D_{spa} est le discriminateur "pixel" usuellement utilisé pour les GANs. Il prend en entrée l'image haute résolution I_{HR} et l'image générée super-résolue I_{SR} et détermine si l'image générée est vraie ou fausse.

— Le second réseau D_{dct} prend en entrée la transformée en cosinus discret des images I_{HR} et I_{SR}, notée I_{HR}^{DCT} et I_{SR}^{DCT} respectivement. Ce réseau a pour objectif de supprimer les artefacts en damier souvent rencontrés lorsque l'on augmente la résolution d'une image. La transformée en cosinus discrète est obtenue de la façon suivante :

$$I^{DCT} = W_r(B_{th}(DCT(I))), \tag{5.10}$$

où

$$W_r(I) = \begin{cases} 0 & si \quad x^2 + y^2 \leq r \\ 1 & sinon \end{cases}$$

et

$$B_{th}(I) = \begin{cases} 0 & si \quad |DCT(I(x,y))| < th \\ 1 & sinon \end{cases}$$

— Le troisième réseau D_{gmd} prend en entrée les histogrammes des magnitudes des gradients des images I_{HR} et I_{SR}, notée I_{HR}^{hist} et I_{SR}^{hist} respectivement. Cela permet de restaurer les hautes fréquences des images super-résolues, c'est-à-dire de conserver la géométrie des images. On remarque que ce réseau prend en entrée une donnée 1D donc les filtres des convolutions sont en 1D également (contrairement aux deux réseaux précédents pour lesquels les entrées et les filtres sont en 2D). L'entrée de ce réseau est calculée de la façon suivante :

$$I^{hist} = hist(GM(I)), \quad \text{où} \quad GM(I) = \sqrt{\nabla_x I + \nabla_y I}. \tag{5.11}$$

Ainsi, la fonction de perte globale pour les discriminateurs est la suivante :

$$\begin{aligned} \mathcal{L}_D = & - log(D_{gmd}(I_{SR}^{hist})) - log(1 - D_{gmd}(I_{HR}^{hist})) \\ & + \beta \left[-log(D_{spa}(I_{SR})) - log(1 - D_{spa}(I_{HR})) \right] \\ & + \gamma \left[-log(D_{dct}(I_{SR}^{DCT})) - log(1 - D_{dct}(I_{HR}^{DCT})) \right], \end{aligned} \tag{5.12}$$

où les paramètres β et γ sont obtenus en effectuant plusieurs tests (les valeurs ne sont pas données dans l'article).

Pour le générateur, les auteurs pré-entraînent celui-ci (dans un contexte hors GAN, donc sans discriminateur) avec une fonction de perte de type MSE, c'est-à-dire

$$\mathcal{L}_G^{pre-trained} = ||I_{HR} - I_{SR}||_2. \tag{5.13}$$

Une fois que le générateur est pré-entraîné, les auteurs utilisent ce réseau dans un contexte GAN, face aux trois discriminateurs présentés précédemment. Ils considèrent une fonction de perte comparant la sortie des couches d'un réseau VGG :

$$\mathcal{L}_G = ||\Phi(I_{HR}) - \Phi(I_{SR})||_2 - \lambda_1 log(D_{spa}(I_{SR})) - \lambda_2 log(D_{dct}(I_{SR}^{DCT})) - \lambda_3 log(D_{gmd}(I_{SR}^{hist})), \tag{5.14}$$

où les paramètres λ_i, $i = 1, ..., 3$ sont les poids attribués à chacun des termes.

• Dans un contexte un peu différent, Park *et al.* [60] proposent une méthode basée CycleGAN considérant plusieurs discriminateurs appliqués aux images sous-marines. L'objectif est de supprimer le trouble présent sur les images sous-marines. Pour cela, les auteurs proposent une architecture du type CycleGAN.

Ce type de réseau permet d'entraîner un réseau pour une tâche précise sans avoir besoin d'un jeu de données par paires (c'est-à-dire que l'image de référence n'est pas nécessaire). Pour cela, deux jeux de données sont nécessaires : le premier contenant des images du domaine I (dans cet article, des images troubles) et le deuxième contenant des images du domaine J (ici, des images claires).

Un CycleGAN est donc un Generative Adversarial Network utilisant deux générateurs G et F et deux discriminateurs D_i et D_j. Le générateur G va apprendre la transformation de I dans J et le générateur F la transformation inverse :

$$\begin{aligned} G &: I \to J \\ F &: J \to I \end{aligned} \tag{5.15}$$

Du coup, chaque réseau a son discriminateur associé, D_j pour G et D_i pour F. Ainsi, D_j va distinguer les images réelles $j \in J$ des images $G(i) \in J$ générées par G alors que D_i va distinguer les images réelles $i \in I$ des images $F(j) \in I$ générées par F.

La fonction de perte associée à ces réseaux est :

$$\mathcal{L} = \mathcal{L}_a dv + \lambda \mathcal{L}_{cyc}, \tag{5.16}$$

où

$$\mathcal{L}_{adv}(G, D_j, I) = \frac{1}{m} \sum_{i \leq m} (1 - D_j(G(i)))^2, \tag{5.17}$$

est obtenue en utilisant la méthodes des moindres carrées et

$$\mathcal{L}_{cyc} = \frac{1}{m} \sum_{i \leq m} (F(G(i)) - i) + (G(F(j)) - j). \tag{5.18}$$

qui permet de garder une image similaire lorsqu'elle passe dans les deux réseaux de manière successive. En effet, en passant une image en entrée de G puis sa sortie dans F, on doit obtenir une

image similaire à celle qui est entrée dans G initialement. Sans ce terme, la fonction de perte oblige à ce que l'image générée appartienne à l'autre domaine mais n'oblige pas les images en entrée et en sortie à se ressembler. Pour résumer, ce second terme ajoute une contrainte à la sortie de chaque réseau.

Les CycleGANs présentent donc un très bon avantage qui est un entraînement sans jeu de données par paires. Cependant, comme ce type de réseau est entraîné à trouver la transformation entre des images sans paires, l'information contenue dans l'image en entrée comme les formes, les contours ou les couleurs ne sont pas nécessairement bien reconstruits. Des artefacts ou des couleurs non réalistes peuvent donc apparaître.

Pour revenir à l'article, les auteurs choisissent donc de se placer dans ce contexte en ajoutant un discriminateur pour chaque générateur. Le premier discriminateur D_{cnt} va discriminer le contenu de l'image et le second D_{stl} le style de l'image :

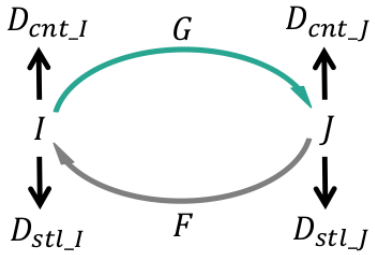


FIGURE 5.3 – Schéma représentatif de la méthode proposée par *Park et al.*

Les fonctions de perte associées à chaque discriminateur sont les suivantes :

$$D_{stl} \ : \ \mathcal{L}_{stl} = \mathbb{E}_{j\sim p_{data}(j)}\left[(D_{stl}(j)-1)^2\right] + \mathbb{E}_{i\sim p_{data}(i)}\left[D_{stl}(G(i))^2\right]. \tag{5.19}$$

Cette fonction de perte est équivalente à celle considérée dans un GAN traditionnel. La seconde fonction de perte est :

$$D_{cnt} \ : \ \mathcal{L}_{cnt} = \mathbb{E}_{j\sim p_{data}(j)}\left[(D_{cnt}(\Phi(j))-1)^2\right] + \mathbb{E}_{i\sim p_{data}(i)}\left[D_{cnt}(G(\Phi(i)))^2\right], \tag{5.20}$$

où $\Phi(.)$ est une carte de caractéristiques correspondant à la sortie d'une couche d'un réseau VGG. La fonction de perte finale est alors une moyenne pondérée des précédents termes :

$$\mathcal{L}_{\mathcal{D}} = \lambda_{cnt}\mathcal{L}_{cnt} + \lambda_{stl}\mathcal{L}_{stl} + \lambda_{cyc}\mathcal{L}_{cyc}, \tag{5.21}$$

où $\mathcal{L}_{cyc}$ est le terme inhérent au CycleGAN. Il garantit la transformation "individuelle" des différents domaines. Cela évite que certains types d'échantillons soient transférés d'un ensemble à l'autre. Le poids λ_{cyc} est fixé alors que les poids λ_{cnt} et λ_{stl} sont calculés de façon à balancer les effets des deux types de générateurs :

$$\begin{aligned}
\lambda_{cnt} &= \alpha * \frac{\sum_{x\in\Omega} \pi(j(x)) + \pi(i(x))}{\sum_{x\in\Omega} \pi(j(x)) + \pi(i(x)) + \pi(F(j(x))) + \pi(G(i(x)))}, \\
\lambda_{stl} &= \alpha * \frac{\sum_{x\in\Omega} \pi(F(j(x))) + \pi(G(i(x)))}{\sum_{x\in\Omega} \pi(j(x)) + \pi(i(x)) + \pi(F(j(x))) + \pi(G(i(x)))},
\end{aligned} \tag{5.22}$$

où α et β sont des constantes et $\pi(.)$ est le détecteur de contours de Sobel. Ces poids sont complémentaires : quand un des deux augmente, l'autre décroît de manière proportionnelle. Cela permet de mettre plus ou moins de poids sur un terme en fonction de la sortie des générateurs.

• Ma *et al.* [54] ont proposé une méthode Pan-GAN basée multi-discriminateurs pour le problème de pansharpening représentée en Figure 5.4. Un premier discriminateur pour la reconstruction spatiale et le second pour la reconstruction spectrale :

— Le discriminateur spatial prend en entrée l'image panchromatique et la luminance de l'image multispectrale, permettant au réseau de se focaliser seulement sur la résolution spatiale des images.

— Le discriminateur spectral prend en entrée l'ensemble des bandes multispectrales basse résolution ainsi que l'image de cible sous-échantillonnée. Ainsi, le réseau prend en compte seulement les intensités de l'image, sans prendre en compte les hautes fréquences.

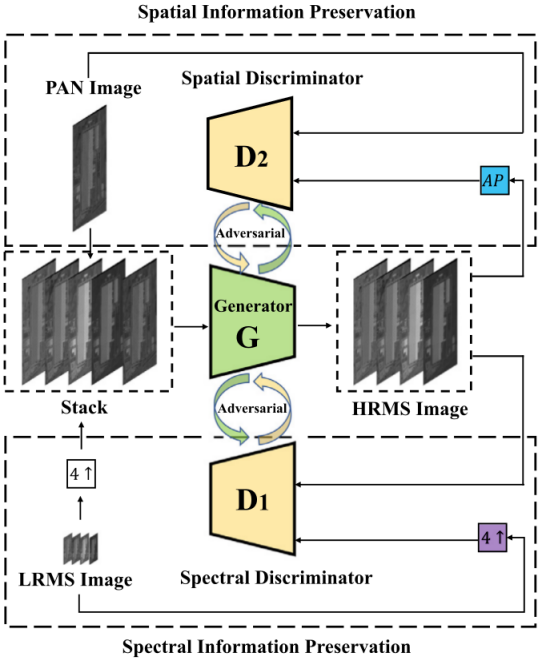


FIGURE 5.4 – Schéma représentatif de la méthode Pan-GAN proposée par Ma *et al.*.

Du point de vue des architectures, les auteurs proposent seulement des réseaux générateur et discriminateur considérant des couches convolutives avec un ajout de connexions résiduelles dans le générateur.

En plus des termes antagonistes classiques dans un cadre de GAN, les auteurs ajoutent une contrainte spatiale

$$\frac{1}{N}\sum_{n\leq N}||\nabla AP(\tilde{u}) - \nabla P||_F^2 \tag{5.23}$$

et une contrainte spectrale dans la fonction de perte

$$\frac{1}{N}\sum_{n\leq N}||\downarrow \tilde{u} - u_S||_2^2, \tag{5.24}$$

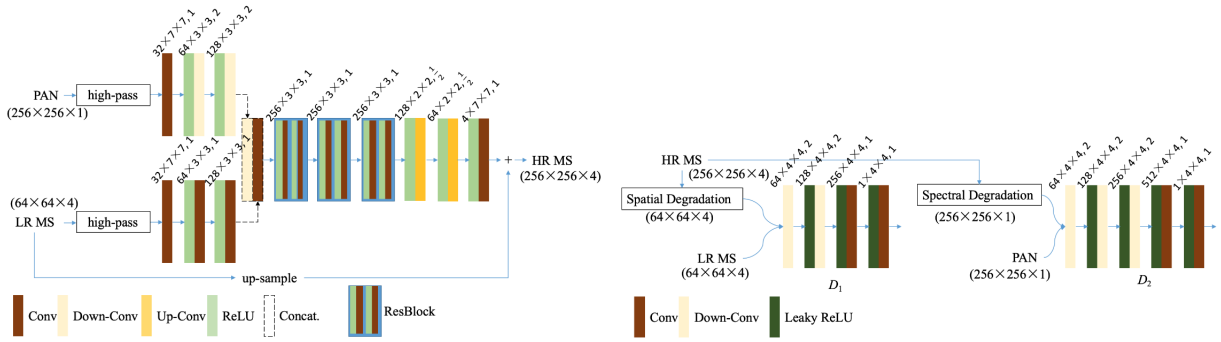
où P est l'image panchromatique, $\tilde{u}$ l'image reconstruire, u_S l'image multispectrale basse-résolution, N la taille du batch, ∇ l'opérateur gradient, $\downarrow$ un opérateur de sous-échantillonnage et $AP(\tilde{u})$ est la luminance de $\tilde{u}$ obtenue à l'aide d'une fonction de moyennage "Average Pooling".

On remarque que chaque contrainte suit la tâche du discriminateur associé. En effet, comme le discriminateur spatial prend en entrée l'image panchromatique et la luminance de l'image multispectrale : les auteurs choisissent de minimiser la norme l_2 de la différence. L'idée est la même pour la contrainte spatiale, les auteurs ont choisi d'ajouter la norme de Frobenius de la différence entre les gradients des images en entrée du second discriminateur.

• Zhou *et al.* [88] ont également proposé une méthode PGMAN basée multi-discriminateurs dans un contexte GAN. Les discriminateurs proposés sont très similaires à ceux considérés par Ma *et al.* [54] :

— le discriminateur spatial prend en entrée l'image panchromatique et la luminance de l'image reconstruite,

— le discriminateur spectral prend en entrée l'image multispectrale basse-résolution et l'image reconstruite spatialement dégradée, i.e. sous-échantillonnée à la taille de la basse résolution.



(a) (b)

FIGURE 5.5 – Architecture du générateur (a) et des discriminateurs (b) de la méthode PGMAN proposée par Zhou *et al.*.

Dans un premier temps, on remarque que l'architecture du générateur, présentée en Figure 5.5 (a), est similaire à celle proposée par Liu *et al.* pour la méthode PSGAN [49]. En effet, deux sous-réseaux sont utilisés afin d'extraire les caractéristiques des images panchromatiques et multispectrales. De plus, inspiré par la méthode PanNet proposée par Yang *et al.* [79] le réseau est entraîné dans les hautes fréquences.

Les auteurs proposent d'ajouter une nouvelle fonction de perte pour optimiser les poids des réseaux. Cette fonctionnelle ne considère pas les termes d'entropie croisée usuels mais considère un modèle WGAN-GP (Wasserstein GAN with Gradient Penalty), utilisant la distance de Wasserstein et ajoutant une pénalité de gradient dans le but de stabiliser l'entraînement. Cette pénalité, correspondant à la norme l_2 du gradient $(||\nabla_x D(x)||_2 - 1)^2$, force le gradient de la solution à être de norme 1 et donc stabilise la solution [31].

De plus, les auteurs choisissent d'utiliser la métrique QNR [6] en tant que contrainte dans la fonction de perte du générateur. Cette mesure permet de quantifier la qualité de fusion des images panchromatique et multispectrale sans image de référence.

Cela revient alors à considérer les fonctionnelles suivantes :

$$\mathcal{L}(G) = \frac{1}{N} \sum_{n \leq N} -\alpha D_1(\tilde{u}, P, u_S) - \beta D_2(\tilde{u}, P, u_S) + (1 - QNR(\tilde{u}, P, u_S)), \tag{5.25}$$

$$\mathcal{L}(D_1) = \frac{1}{N} \sum_{n \leq N} -D_1(P) + D_1(\tilde{u}_L) + \lambda GP(D_1, P, \tilde{u}_L), \tag{5.26}$$

et

$$\mathcal{L}(D_2) = \frac{1}{N}\sum_{n\leq N} -D_2(u_S) + D_2(\downarrow \tilde{u}) + \lambda GP(D_2, u_S, \downarrow \tilde{u}), \tag{5.27}$$

où G est le générateur, $D_{1,2}$ les discriminateurs, N la taille du batch, $\downarrow$ l'opérateur de sous-échantillonnage, $\tilde{u}_L$ la luminance de l'image reconstruite $\tilde{u}$, P l'image panchromatique et u_S l'image multispectrale basse-résolution.

5.2 Contribution : méthode MDSSCGAN-SAM

Comme on a pu le voir précédemment, la modélisation du problème est importante dans un réseau convolutif et permet d'améliorer les résultats. Nous proposons ici, une nouvelle méthode basée multi-discriminateurs extension de RDGAN-Geom. Elle a pour objectif de renforcer la reconstruction spatiale tout en considérant la reconstruction spectrale qui est tout aussi importante pour le problème de pansharpening.

Cette méthode exploite conjointement les sources d'informations spatiales et spectrales. Pour cela, nous proposons de séparer ces deux tâches en considérant deux discriminateurs "orthogonaux" schématisés en Fig. 5.6. Le premier est optimisé pour préserver la texture et la géométrie en prenant en entrée la luminance et la bande proche infrarouge (NIR) des images. Le second préserve la couleur et la résolution spectrale en prenant en entrée les composantes chromatiques Cb et Cr.

Ainsi, cette méthode nous permet d'entraîner deux discriminateurs, chacun d'eux associé à des tâches différentes mais complémentaires. De plus, pour renforcer cet aspect, une contrainte spatiale et une contrainte spectrale sont ajoutées dans la fonction de perte du générateur.

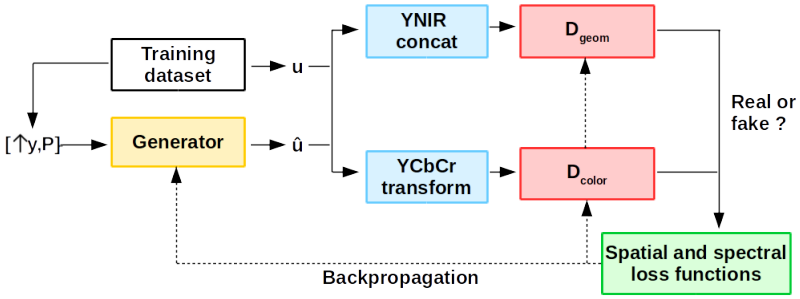


FIGURE 5.6 – Schéma général de la méthode proposée, où P est l'image panchromatique, $\uparrow y$ est l'image multispectrale sur-échantillonnée à la taille de P, u l'image de référence et $\hat{u}$ l'image fusionnée. D_{geom} désigne le discriminateur spatial et D_{color} le discriminateur spectral.

5.2.1 Discriminateur spatial

Motivés par la reconstruction spatiale et la préservation de la texture, le premier discriminateur D_{geom} prend en entrée la luminance et la bande proche infra-rouge des images. En effet, les images en niveau de gris permettent de mieux mettre en évidence les textures et la géométrie des images. De plus, comme la végétation reflète fortement dans le proche infra-rouge (Cf. Chap.1), cette bande est donc très importante pour obtenir l'information de texture dans ces zones. Ainsi, la bande infra-rouge est aussi sélectionnée en entrée de ce discriminateur.

La luminance Y est obtenue en faisant une combinaison linéaire des bandes rouge R, verte V et bleue B :

$$Y = 0.299R + 0.587V + 0.114B. \tag{5.28}$$

Ensuite, la bande Y et la bande NIR sont concaténées et mises en entrée du discriminateur.
Pour finir, la fonction de perte optimisée pour ce discriminateur est la suivante :

$$\mathcal{L}_{D_{geom}} = \sum_{i \leq N_b} log(1 - D_{\eta_g}(\hat{u}_{YIR})) + log(D_{\eta_g}(u_{YIR})), \tag{5.29}$$

où $\hat{u}_{YIR}$ et u_{YIR} sont la concaténation de la luminance Y et la bande NIR de l'image reconstruite $\hat{u}$ et de l'image de référence u respectivement. N_b est la taille du batch et η_g sont les paramètres du réseau D_{geom} à optimiser.

5.2.2 Discriminateur spectral

Dans un second temps, pour préserver les couleurs de l'image et ainsi la résolution spectrale, nous avons considéré un second discriminateur comparant les couleurs de l'image cible avec celles de l'image reconstruite à l'aide des composantes chromatiques Cb et Cr de la transformation YCbCr.

Le choix de ces composantes repose sur le fait que l'interdépendance des bandes R, G et B est généralement plus élevée que celle des composantes Y, Cb et Cr. Par conséquent, pour préserver la couleur, nous choisissons d'utiliser les bandes Cb et Cr. Cela revient alors à minimiser :

$$\mathcal{L}_{D_{color}} = \sum_{i \leq N_b} log(1 - D_{\eta_c}(\hat{u}_{CbCr})) + log(D_{\eta_c}(u_{CbCr})), \tag{5.30}$$

où u_{CbCr} et $\hat{u}_{CbCr}$ correspondent à la concaténation des bandes Cb et Cr de l'image cible u et de l'image fusionnée $\hat{u}$ respectivement. La variable η_c correspond quant à elle aux poids du réseau discriminateur D_{color}.

5.2.3 Architecture des discriminateurs

L'architecture de chaque discriminateur, présentée en Figure 5.7, est composée de sept couches convolutives avec un nombre de cartes de caractéristiques croissant de 32 à 1024. Ces couches convolutives sont utilisées pour extraire suffisamment de caractéristiques afin de capturer la représentation des données.

Ensuite, deux couches denses sont ajoutées pour la classification. Ce type de couches permet d'apprendre une fonction dans l'espace des données détectant si l'image générée est fausse ou vraie. En effet, contrairement aux couches convolutives qui extraient de l'information locale, les couches denses apprennent l'information à partir de l'ensemble des caractéristiques de la couche précédente, c'est-à-dire en combinant toute l'information donnée en entrée de cette couche.

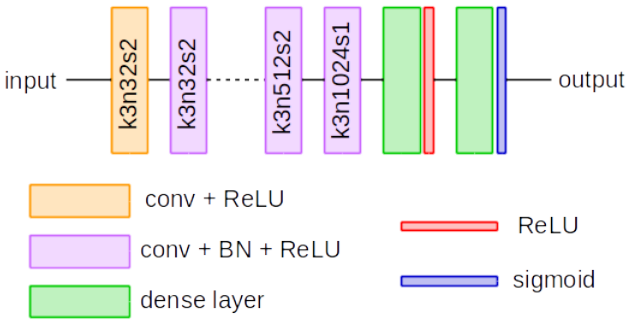


FIGURE 5.7 – Architectures des discriminateurs, où k est la taille du noyau, n le nombre de filtres and s la taille du pas pour chacune des couches convolutives.

5.2.4 Générateur

Pour le générateur, la même architecture que dans la précédente contribution est considérée. C'est-à-dire une architecture de type dense résiduelle, combinant les avantages des architectures

denses [36] et résiduelles [33], permettant ainsi d'éviter le problème du *vanishing gradient* souvent rencontré lors de l'apprentissage.

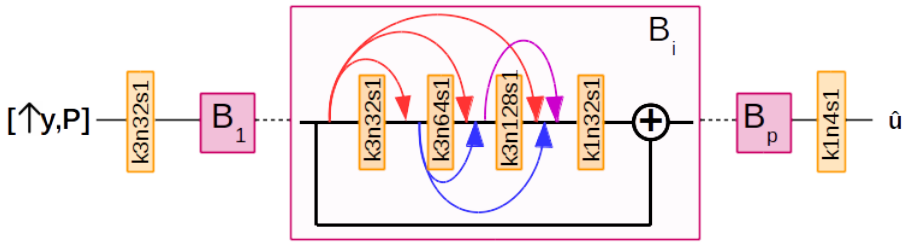


FIGURE 5.8 – Architecture générale pour le générateur. $[P, \uparrow y]$ est la concaténation [.] de l'image panchromatique P et de $\uparrow y$ l'image multispectrale $y = (y^1, ..., y^N)$ ré-échantillonnée à la taille de P, les blocs B_i, $i \leq p$, sont denses résiduels, k est la taille du noyau de convolution, n le nombre de filtres et s le pas. Chaque couche convolutive est suivie d'une couche ReLU, exceptée la dernière. Les flèches représentent les connexions denses et les $+$ les connexions résiduelles.

Ce réseau, présenté en Figure 5.8 prend en entrée l'image panchromatique P et l'image multispectrale ré-échantillonnée $\uparrow y$ à la taille de P à l'aide d'une interpolation bicubique. Elle est composée de plusieurs blocs denses résiduels et renvoie une image résiduelle car nous travaillons toujours dans un cadre résiduel, comme mentionnée au début de ce chapitre (4.21) et (4.22).

En revanche, la fonction de perte considérée est différente. En plus de la contrainte spatiale, nous proposons d'ajouter une contrainte spectrale. Cela revient alors à minimiser la fonctionnelle suivante :

$$\mathcal{L}(G_\theta) = \sum_{i \leq N_b} \alpha_g log(D_{\eta_g}(\hat{u}_{YIR})) + \alpha_c log(D_{\eta_c}(\hat{u}_{CbCr}) + \alpha_{l1}||\hat{u} - u||_1 + \alpha_t \sum_{x \in \Omega} |\langle \nabla u(x)^\perp, \nabla \hat{u}(x) \rangle| + \alpha_{sam} \sum_{x \in \Omega} arccos\left(\frac{\langle u(x), \hat{u}(x) \rangle}{||u(x)||_2 ||\hat{u}(x)||_2}\right), \tag{5.31}$$

où ∇ est le gradient, $\perp$ le vecteur orthogonal, $\{,\}$ est le produit scalaire, Ω le domaine de l'image et $\alpha_{g,c,l_1,t,sam}$ sont les poids attribués à chaque terme. Le premier et le second terme sont les termes d'entropie croisée associés à chaque discriminateur. Le troisième terme est la norme l_1 de la différence entre l'image cible et l'image reconstruite. Le quatrième terme est la contrainte géométrique forçant l'alignement des gradients de l'image reconstruite avec ceux de l'image de référence. Ce terme, utilisé dans notre approche RDGAN-Geom, permet d'améliorer la reconstruction géométrique des images. Enfin, le dernier terme est la contrainte spectrale. Cette contrainte est basée sur la mesure SAM. Cette mesure calcule la valeur absolue des angles entre deux vecteurs composés de la valeur des pixels de l'image cible et de l'image reconstruite à chaque point du domaine. Une mesure SAM égale à zéro, signifie donc une absence de distorsion spectrale. Cependant, des distorsions radiométriques peuvent être présentes. Cela signifie que les vecteurs sont parallèles mais de longueur d'ondes différentes.

Ces deux contraintes ont pour objectif de renforcer les reconstructions spatiale et spectrale d'une autre façon que le discriminateur.

Dans le cas de la reconstruction spectrale, le discriminateur travaille à l'aide des composantes Cb et Cr alors que le discriminateur cherche à minimiser la mesure SAM. Dans ce cas, le problème est le même mais il est abordé d'une manière différente dans le but de se compléter. Il en va de même avec la reconstruction spatiale, le discriminateur utilise la luminance et la bande NIR alors que le générateur veut aligner les gradients des images.

5.3 Résultats

• Métriques de comparaison et bases de données :

Pour comparer nos résultats avec l'état-de-l'art, nous utilisons les mesures présentées au Chap.1, c'est-à-dire les mesures PSNR, SAM, ERGAS, RMSE et CC. Les mesures PSNR et ERGAS sont des mesures globales, CC mesure la distorsion spectrale et SAM la distorsion spectrale. De plus, comme les textures et géométries sont plus visibles dans les hautes fréquences, nous proposons de calculer le PSNR sur des images filtrées à l'aide d'un filtre passe haut. Cela revient alors à considérer la mesure suivante :

$$PSNR_h(X, Y) = PSNR(X * h, Y * h), \tag{5.32}$$

où h est le filtre de Butterworth [14].

Par ailleurs, nous utilisons les mêmes données qu'au Chap.4 : les bases de données Pléiades et WorldView 3, toutes les deux créées en suivant le protocole de Wald [72].

• Détails d'implémentation :

La méthode proposée est implémentée avec TensorFlow 1.2 et utilise l'algorithme ADAM pour optimiser les poids des réseaux. Les tailles de batch utilisées sont 19 pour la base Pléiades et 17 pour la base WorldView 3.

Les paramètres $\alpha_{g,c,l_1,t,sam}$ sont optimisés de façon à obtenir la meilleure balance entre toutes les métriques de comparaison. De plus, le même poids a été attribué à chaque discriminateur pour leur donner la même importance, i.e. $\alpha_g = 0.5$, $\alpha_c = 1$, $\alpha_{l_1} = 100$, $\alpha_t = 1$ and $\alpha_{sam} = 50$.

5.3.1 Étude d'ablation

Dans un premier temps, pour souligner l'avantage de la méthode proposée ainsi que l'influence de chaque discriminateur, il convient de comparer plusieurs méthodes. Les deux premières méthodes ne considèrent qu'un seul des deux discriminateurs individuellement. La troisième méthode considère un discriminateur prenant en entrée la concaténation de la luminance, de la bande infrarouge et des composantes chromatiques Cb et Cr. Et enfin, la dernière méthode considérée est une approche basée multi-discriminateurs mais sans la contrainte spectrale afin d'évaluer l'impact de ce terme de régularisation dans la reconstruction. Donc, cela revient à considérer 5 approches, résumées dans le Tab. 5.1 :

Méthode	Nombre de discriminateurs	Entrée des discriminateurs	Contraintes dans la fonction de perte du générateur
'MDSSC-GAN concat'	1	[Y, NIR, Cb, Cr]	spatial
'MDSSC-GAN texture'	1	[Y, NIR]	spatial
'MDSSC-GAN color'	1	[Cb, Cr]	spatial
'MDSSC-GAN'	2	[Y,NIR] et [Cb, Cr]	spatial
'MDSSC-GAN SAM'	2	[Y,NIR] et [Cb, Cr]	spatial et spectral

TABLE 5.1 – Résumé des méthodes comparées, où [.] représente la concaténation, Y la luminance, NIR la bande infrarouge et Cb et Cr les composantes chromatiques.

Les résultats quantitatifs obtenus sont alors présentés en Tab. 5.2.

Method	CC	SAM	PSNR	$PSNR_h$	ERGAS
ideal value	1	0	max	max	0
MDSSC-GAN concat	0.969	0.141	29.416	27.224	3.888
MDSSC-GAN color	**0.970**	0.139	29.492	27.406	**3.875**
MDSSC-GAN texture	0.969	0.140	29.394	**27.5015**	3.887
MDSSC-GAN	**0.970**	0.138	**29.493**	27.361	3.884
MDSSC-GAN SAM	**0.970**	**0.137**	29.455	27.455	3.881

TABLE 5.2 – Résultats quantitatifs obtenus en comparant les performances de chaque discriminateur individuellement avec la méthode proposée sur la base de donnée Pléiades. 'MDSSC-GAN color' considère seulement le discriminateur spectral, 'MDSSC-GAN texture' considère seulement le discriminateur spatial, 'MDSSC-GAN concat' considère un seul discriminateur prenant en entrée la concaténation de la luminance, de la bande infrarouge et des composantes chromatiques, 'MDSSC-GAN' et 'MDSSC-GAN SAM' correspondent à la méthode proposée sans et avec la contrainte spectrale dans la fonction de perte. Les meilleurs résultats sont en gras et les pires sont soulignés.

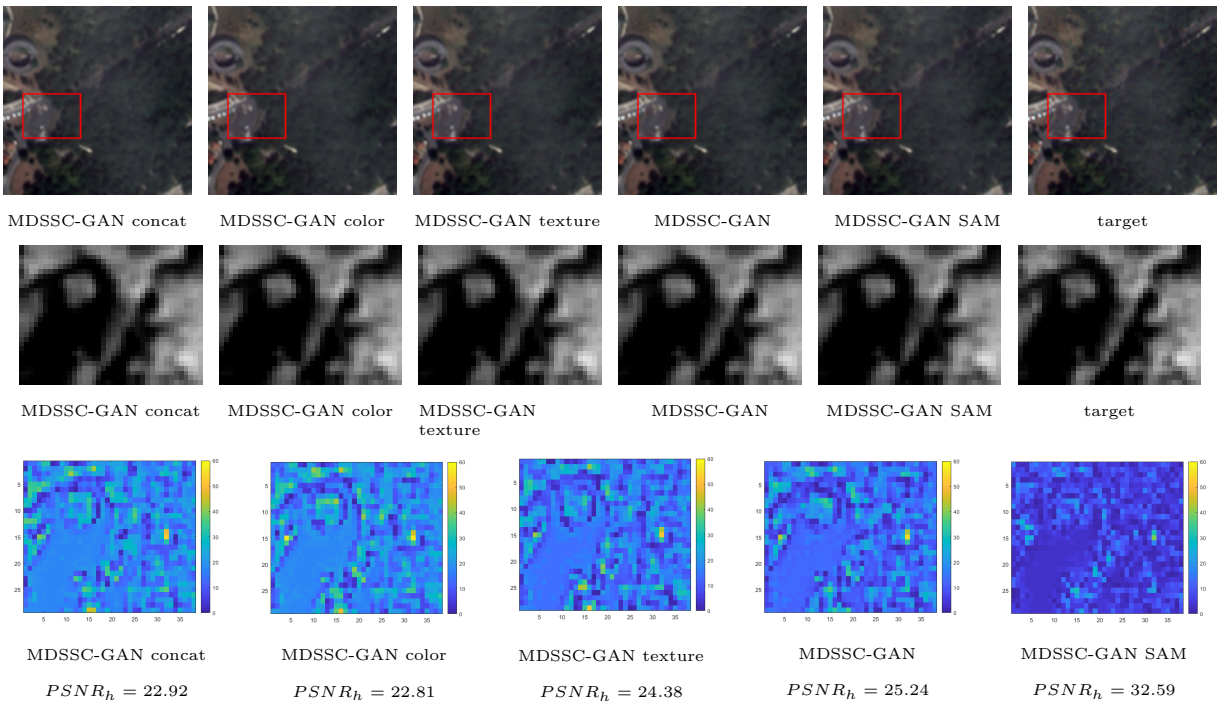


FIGURE 5.9 – Résultats obtenus sur une image Pléiades en zone de végétation lorsque l'on compare les performances de chaque discriminateur individuellement avec la méthode proposée 'MDSSC-GAN SAM'. 'MDSSC-GAN color' considère seulement le discriminateur spectrale, 'MDSSC-GAN texture' considère seulement le discriminateur spatial, 'MDSSC-GAN concat' considère un seul discriminateur prenant en entrée la concaténation de la luminance, de la bande infrarouge et des composantes chromatiques Cb et Cr, 'MDSSC-GAN' considère la méthode multi-discriminateur sans la contrainte spectrale et 'MDSSC-GAN SAM' est la méthode proposée. La première ligne est l'image RGB, la seconde ligne la bande infrarouge et le troisième ligne affiche la différence entre les hautes fréquences de l'image de référence et de l'image reconstruite. On peut voir que le meilleur résultat est donné par la méthode 'MDSSC-GAN SAM'.

Dans un premier temps, on peut noter que l'aspect multi-discriminateurs améliore de manière significative les résultats quantitatifs (Tab. 5.2). En effet, la méthode 'MDSSC-GAN concat' montre que considérer un seul discriminateur concaténant les entrées de deux discriminateurs n'est

pas suffisant pour améliorer les résultats. Ensuite, quand on considère seulement le discriminateur spatial, les résultats quantitatifs ne sont pas convaincants, excepté pour la mesure PSNR_h. Cela signifie que ce discriminateur reconstruit bien les hautes fréquences comme voulues, mais n'est pas assez performant pour obtenir une reconstruction spectrale ou à l'échelle du pixel suffisante. En revanche, lorsque l'on considère seulement le discriminateur spectral, les résultats sont similaires à ceux obtenus avec la méthode multi-discriminateurs proposée, excepté pour la mesure PSNR_h. Ainsi, ce discriminateur seul a du mal à bien reconstruire la géométrie dans l'image. On voit bien que, considérés indépendamment, chacun des discriminateurs améliore la mesure associée à leur rôle. En regardant les résultats obtenus avec la méthode proposée, on peut donc en déduire que la combinaison des deux discriminateurs améliore l'ensemble des métriques. Finalement, lorsque l'on compare la méthode 'MDSSC-GAN' et 'MDSSC-GAN SAM', on peut voir qu'ajouter une contrainte spectrale semble renforcer la reconstruction spectrale en améliorant les mesures SAM et PSNR_h mais en dégradant légèrement le PSNR.

Les résultats en Fig. 5.9 montrent un exemple sur une zone de végétation. Comme la végétation réfléchit mieux le rayonnement proche infrarouge, on a aussi choisit d'afficher cette bande. Cet exemple montre bien que la méthode multi-discriminateurs avec contrainte spectrale reconstruit mieux les hautes fréquences que les autres méthodes.

Cette étude d'ablation permet de conclure sur l'avantage d'utiliser plusieurs discriminateurs et sur l'ajout d'une contrainte spectrale. En effet, les résultats quantitatifs montrent une amélioration de l'ensemble des métriques et les résultats montrent une amélioration de la reconstruction des hautes fréquences.

5.3.2 Comparaison avec les méthodes de l'état-de-l'art

Les résultats suivants comparent notre méthode avec les méthodes de l'état-de-l'art pour le problème de pansharpening sur les deux bases de données Pléiades et WorldView 3. Les méthodes par injection de coefficients GLP [70] et GSA [62] et les méthodes de réseaux de neurones convolutifs PanNet [79], PSGAN [50] et RDGAN-Geom [25] sont utilisées. Pour une comparaison plus pertinente, tous les réseaux ont été ré-entraînés sur les mêmes bases de données. Les temps d'entraînement sont d'environ 24h pour la méthode PanNet et la méthode proposée et 15h pour les méthodes PSGAN et RDGAN-Geom.

Méthode	CC	SAM	PSNR	$PSNR_h$	ERGAS
valeur idéale	1	0	max	max	0
GSA [62]	0.871	0.237	21.94	<u>22.41</u>	10.74
GLP [70]	<u>0.866</u>	<u>0.242</u>	<u>21.74</u>	23.05	<u>10.91</u>
PanNet [79]	0.950	0.157	28.36	26.60	8.77
PSGAN [50]	0.952	0.155	26.59	26.96	4.28
RDGAN-Geom [25]	0.969	0.138	29.38	27.23	3.94
MDSSC-GAN	**0.970**	0.138	**29.49**	27.36	**3.88**
MDSSC-GAN SAM	**0.970**	**0.137**	29.45	**27.45**	**3.88**

TABLE 5.3 – Résultats quantitatifs des différentes méthodes de l'état-de-l'art sur la base de données Pléiades. Les meilleurs résultats sont en gras et les pires sont soulignés.

Tab. 5.3 présente la comparaison des résultats quantitatifs des différentes méthodes de l'état-de-l'art sur la base de données Pléiades. Tout d'abord, on peut voir que les méthodes basées réseaux de neurones convolutifs donnent de meilleurs résultats face aux méthodes par injection de coefficients. Cependant, il est important de noter que l'une de ces méthodes requiert le réglage des coefficients

α_k, $k \leq K$, de l'Eq. (1.3) pour la fusion mais que nous ne les connaissons pas. Donc il est peut-être possible d'obtenir des résultats légèrement meilleurs si l'on utilise les coefficients exacts. Ensuite, la méthode proposée donne de meilleurs résultats quantitatifs pour la plupart des mesures considérées. On peut noter une amélioration des mesures SAM, CC et PSNR_h, ce qui signifie une meilleur reconstruction spatiale et spectrale avec la méthode 'MDSSC-GAN SAM', mais un meilleur PSNR avec la méthode 'MDSSC-GAN'.

Cette première figure, Fig.5.10 est un exemple général de la fusion des images panchromatique et multispectrale pour les différentes méthodes de l'état-de-l'art basées apprentissage. Il est très difficile de voir une différence majeure entre ces méthodes sur les bandes R, G, B et NIR affichées ainsi mais sur cet exemple la méthode proposée améliore le PSNR d'environ 0.3dB.

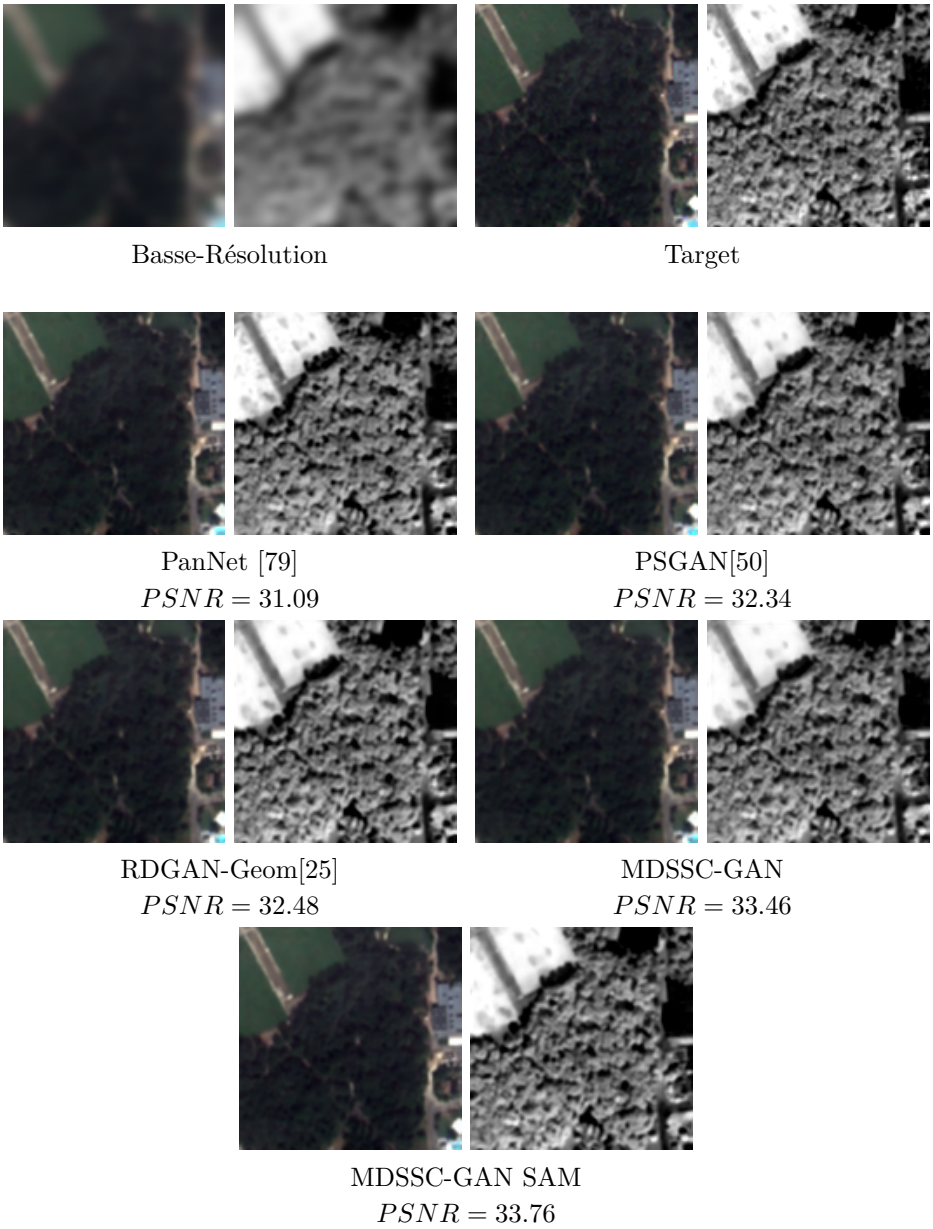


FIGURE 5.10 – Résultats obtenus avec des méthodes de l'état-de-l'art sur une image Pléiades. Pour chaque méthode, est affiché les bandes RGB à gauche et la bande NIR à droite. Une amélioration significative de PSNR peut être observée sur cette image avec la méthode proposée MDSSC-GAN SAM'.

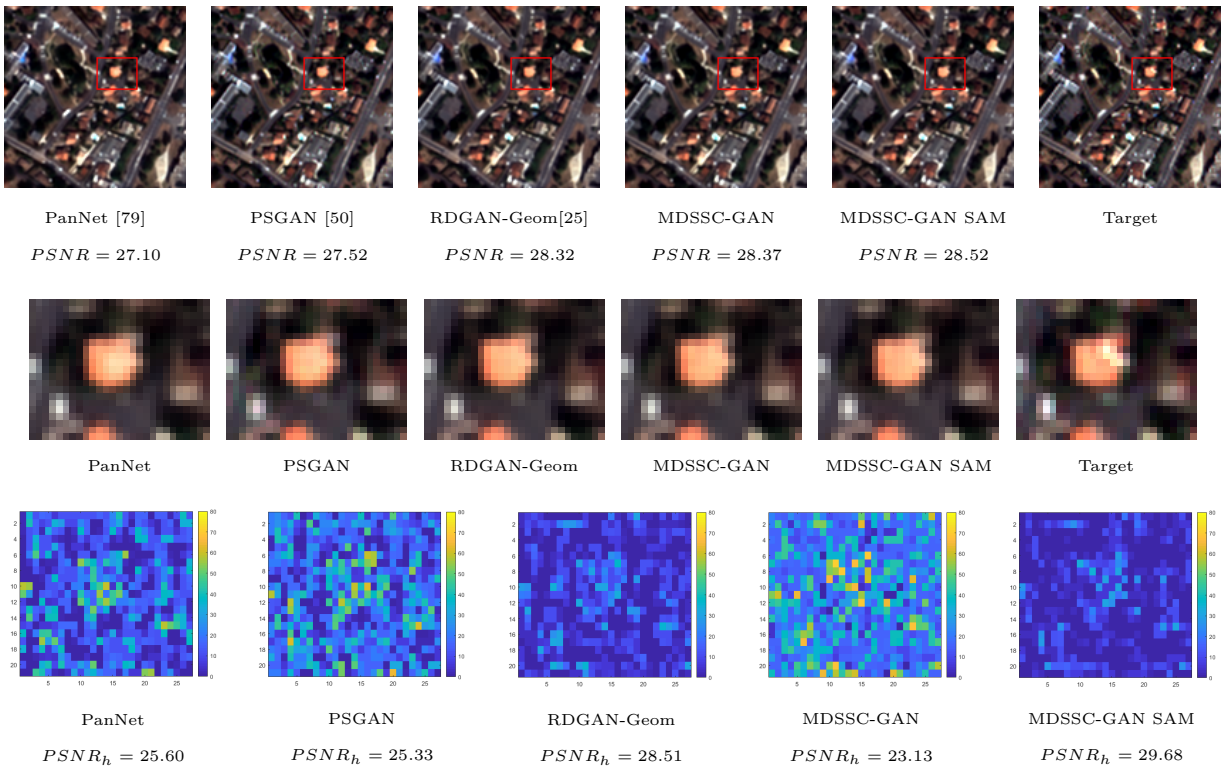


FIGURE 5.11 – Résultats obtenus avec une image Pléiades pour les différentes méthodes de l'état-de-l'art en zone urbaine. La première ligne montre le résultat sur les bandes R, G et B. La seconde ligne en affiche une partie zoomée. La dernière ligne affiche la différence entre les hautes fréquences de l'image de référence et celles de l'image fusionnée. La meilleure reconstruction des hautes fréquences est donnée par la méthode proposée 'MDSSC-GAN SAM'.

Étant donnée la difficulté d'évaluation visuelle des résultats simplement à l'aide des bandes R, G, B et Nir, la Fig. 5.11 a pour objectif de visualiser la reconstruction spatiale. Les composantes R, G et B, ainsi que la différence entre les hautes fréquences de l'image de référence et les hautes fréquences de l'image fusionnée sont présentées.

On peut ainsi voir que la meilleure reconstruction est obtenue à l'aide de la méthode proposée. Dans cet exemple, on peut également remarquer que le meilleur PSNR ne donne pas la meilleure reconstruction des hautes fréquences. En effet, le meilleur PSNR est obtenu avec la méthode 'MDSSC-GAN', mais les hautes fréquences sont moins bien reconstruites avec cette méthode. Au contraire, avec la méthode proposée, le PSNR n'est pas le meilleur mais on peut observer une nette amélioration de la reconstruction spatiale.

Ensuite, pour juger de la reconstruction spectrale, la Fig.5.12 montre un exemple de la préservation spectrale. En effet, cette figure montre la carte SAM, c'est-à-dire l'angle de distorsion entre l'image de référence et l'image fusionnée à chaque pixel. Un pixel sans distorsion est affiché en bleu et un pixel avec une forte distorsion apparaît en jaune. Dans cet exemple, il est plus difficile de voir une différence aussi marquée qu'avec la reconstruction spatiale. Cependant, sur la partie zoomée, on observe que l'ajout du terme de régularisation SAM dans la fonction de perte dans la méthode 'MDSSC-GAN SAM' permet une meilleure reconstruction spectrale.

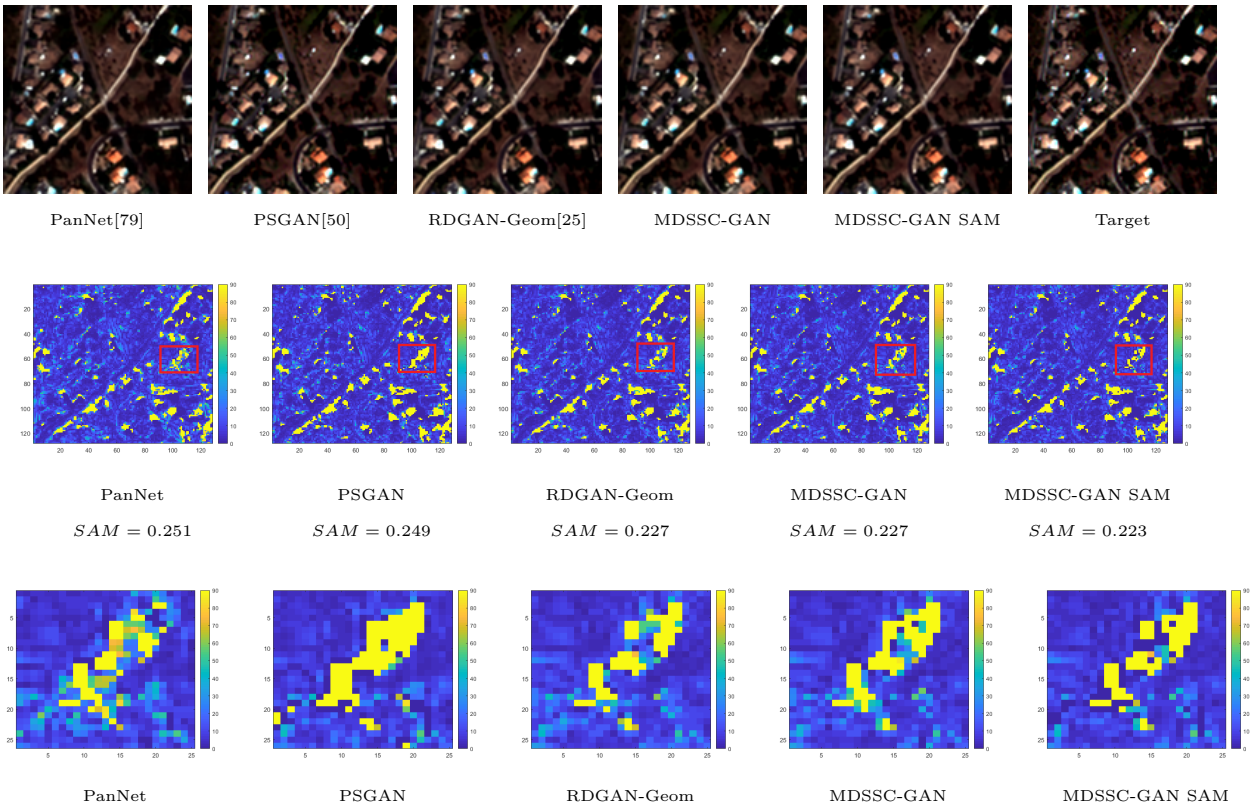


FIGURE 5.12 – Résultats obtenus avec des méthodes de l'état-de-l'art sur une image Pléiades. La première ligne montre les bandes R, G et B. La seconde ligne affiche la carte SAM, i.e. la distorsion de l'angle entre l'image fusionnée et l'image de référence. La troisième ligne montre une partie zoomée de cette carte. La meilleure reconstruction spectrale est obtenue avec la méthode proposée 'MDSSC-GAN SAM'.

Finalement, sur la base de données WorldView 3, les résultats quantitatifs présentés en Tab.5.4 montrent la même amélioration que sur la base de données Pléiades avec la méthode proposée 'MDSSC-GAN SAM'. Cependant, sur cette base de données, on peut noter que la mesure SAM est meilleure pour toutes les méthodes mais les autres métriques sont du même ordre. La meilleure approche en regardant les résultats quantitatifs est encore la méthode 'MDSSC-GAN SAM'.

Méthode	CC	SAM	PSNR	$PSNR_h$	ERGAS
valeur idéale	1	0	max	max	0
GSA [62]	<u>0.879</u>	<u>0.184</u>	<u>22.67</u>	23.14	<u>19.94</u>
GLP [70]	0.880	0.179	<u>22.67</u>	<u>23.07</u>	14.53
PanNet [79]	0.930	0.135	23.22	23.93	8.01
PSGAN [50]	0.966	0.110	29.30	26.82	6.33
RDGAN-Geom [25]	0.966	0.107	29.39	26.20	6.58
MDSSC-GAN	0.967	0.106	29.44	26.59	6.48
MDSSC-GAN SAM	**0.968**	**0.104**	**29.62**	**26.82**	**6.31**

TABLE 5.4 – Résultats quantitatifs pour les différentes méthodes de l'état-de-l'art sur la base de données WorldView 3. Les meilleurs résultats sont en gras et les pires sont soulignés.

Avec la base de données WorldView 3, les résultats quantitatifs présentés en Tab.5.4 comme les résultats visuels présentés en Fig.5.13 montrent que les meilleurs résultats sont obtenus avec la méthode proposée. En effet, sur la Figure 5.13, qui est un exemple en zone de végétation, la texture

de la forêt est mieux reconstruite lorsque l'on considère plusieurs discriminateurs.

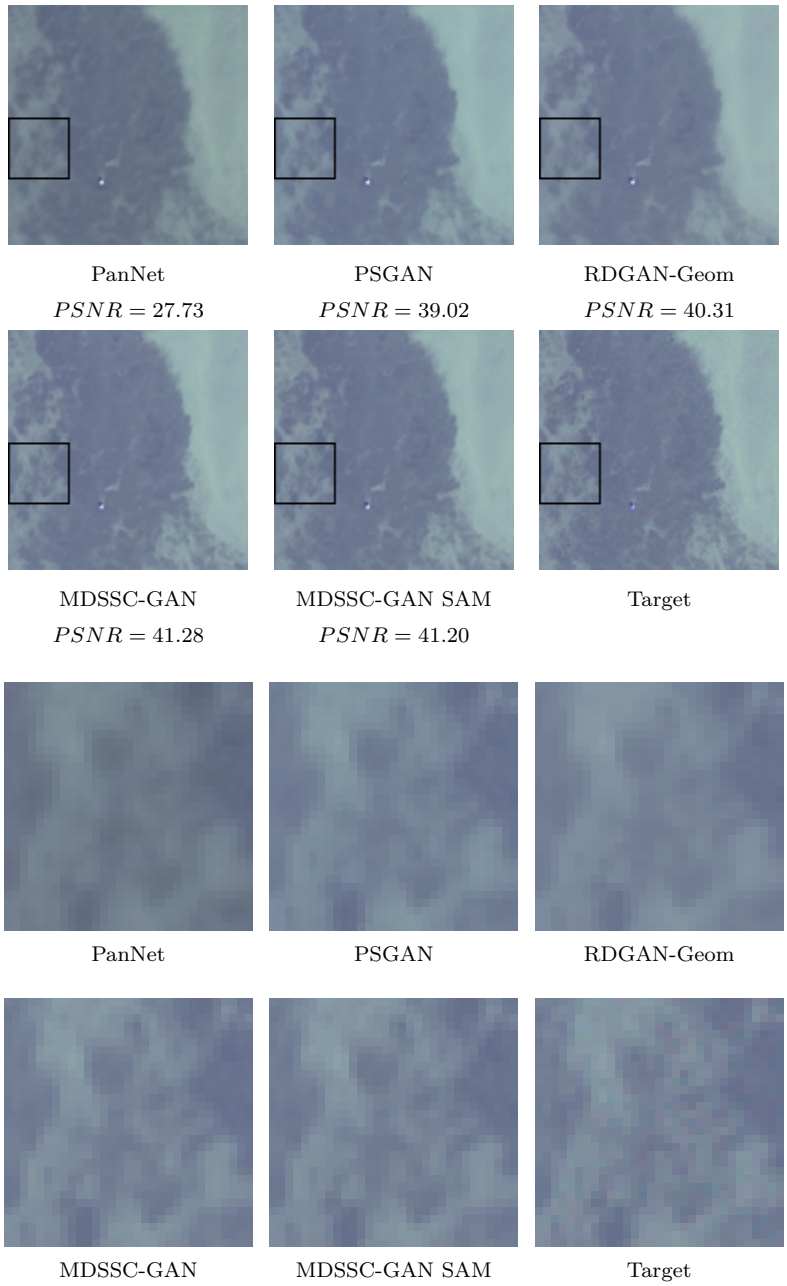


FIGURE 5.13 – Résultats obtenus avec des méthodes de l'état-de-l'art sur une image WorldView 3 dans une zone de végétation. Les deux premières lignes montrent les images RGB et les deux dernières lignes affichent un zoom des canaux RGB. On peut voir une meilleure reconstruction de la texture avec les méthodes considérant plusieurs discriminateurs 'MDSSC-GAN' et 'MDSSC-GAN SAM'.

5.4 Conclusion

Pour conclure ce chapitre, nous avons proposé une méthode basée multi-discriminateurs dans un cadre GAN pour le problème de pansharpening [26]. Cette méthode permet d'entraîner deux discriminateurs complémentaires : le premier améliore la résolution spatiale et le second la résolution spectrale. De plus, pour renforcer ces deux aspects, deux termes de régularisation, spatiale et spectrale, sont ajoutés à la fonction de perte du générateur.

Les expériences menées sur les bases de données satellites Pléiades et WorldView 3 ont montré que la méthode proposée donne de meilleurs résultats, à la fois quantitatifs et visuels. En effet, l'ensemble des métriques spatiales et spectrales sont améliorées et les résultats visuels le confirment.

Une perspective pour améliorer la qualité de reconstruction est l'utilisation de mécanismes d'attention afin de focaliser l'attention du réseau sur les zones de l'image pertinentes pour la reconstruction. Ainsi, le chapitre suivant repose sur l'utilisation des mécanismes d'attention dans une architecture GAN.

CHAPITRE 6

Reconstructions spatiale et spectrale basées sur l'utilisation de mécanismes d'attention

Ce chapitre s'intéresse aux différents mécanismes d'attention et a pour objectif d'étudier leur comportement dans une architecture déjà existante. En effet, nous intégrons un mécanisme d'attention spatiale et un mécanisme d'attention spectrale pour renforcer ces deux aspects très importants pour le problème de pansharpening dans le cadre multi-discriminateurs proposé au Chap.5. Ces mécanismes ont pour objectif de préciser encore plus les reconstructions spatiale et spectrale des images en aidant le générateur à se concentrer sur les zones de l'image pertinentes à l'aide des différentes cartes d'attention.

6.1 Généralités sur les mécanismes d'attention

Initialement, l'attention est un mécanisme développé pour améliorer les performances des architectures du type encodeur-décodeur pour le problème de traduction de textes dans un modèle récurrent. Ce mécanisme est aujourd'hui utilisé pour une très grande variété de problèmes mais de différentes manières.

Ce mécanisme est inspiré de l'attention visuelle humaine. Il vise la capacité d'apprendre à se concentrer sur des zones spécifiques d'une donnée, par exemple un mot dans une phrase ou une zone d'une image. Ces mécanismes d'attention sont ainsi intégrés à des architectures pour aider les réseaux de neurones artificiels à apprendre les zones d'intérêt, de la donnée d'entrée ou des cartes de caractéristiques obtenues au cours du réseau, sur lesquelles se concentrer pour prédire la sortie.

L'attention a donc été initialement proposée par Bahdanau *et al.* en 2015 [9] dans un modèle de traduction automatisé Seq2Seq. Ce modèle est composé d'une architecture encodeur-décodeur, où l'encodeur transforme le séquence d'entrée et compresse cette information dans un vecteur de contexte de taille fixe. Cette représentation est supposée être un bon résumé de la séquence d'entrée. Ensuite, le décodeur produit la traduction à l'aide de ce vecteur.

Cependant, la taille fixe du vecteur de contexte fourni par l'encodeur a montré ses limites. En effet, il a été montré que fixer la taille de ce vecteur est un inconvénient lorsque la phrase considérée en entrée est longue. Très souvent, lorsque le processus de l'encodeur est terminé, les premiers éléments de la phrase ont été oubliés. Le mécanisme d'attention a alors été introduit dans le but de régler ce problème dit "de longues dépendances".

Par la suite, ce mécanisme a été repris par Vaswani *et al.* [67] en 2016, où les mots sont traités en parallèle plutôt que séquentiellement. Cette méthode, appelée *Transformers*, est pionnière

dans l'utilisation de processus parallèles pour le mécanisme d'attention. Les méthodes récentes de l'état-de-l'art vont ensuite reprendre ce procédé pour traiter l'information.

De manière plus détaillée, Bahdanau *et al.* ont proposé une méthode où tous les mots d'entrée peuvent être pris en compte dans le vecteur de contexte, mais où une importance relative est également accordée à chacun d'eux. Ils mettent l'accent sur l'intégration de tous les mots de l'entrée, représentés par des états cachés, lors de la création du vecteur de contexte. Cette intégration se fait en prenant simplement une somme pondérée des états cachés. L'idée est alors d'apprendre les poids de cette somme pondérée pour accorder plus de poids (et ainsi plus d'attention) sur les mots les plus pertinents pour la prédiction. Donc l'attention agit de sorte qu'à chaque fois que le modèle prédit un mot en sortie, le modèle utilise seulement des parties de l'entrée où l'information est la plus pertinente au lieu d'utiliser toute la séquence. En d'autres termes, le modèle prête attention seulement à quelques mots de la phrase d'entrée.

Du point de vue de l'architecture, l'attention assure l'interface entre l'encodeur et le décodeur qui donne au décodeur l'information nécessaire provenant de l'encodeur à chaque état caché. Le modèle est capable de se concentrer de manière sélective sur les différentes parties de la séquence d'entrée et ensuite de faire le lien entre l'entrée et la sortie. Donc cela permet au modèle de traiter des longues phrases d'entrées.

Transposé au traitement d'images, les mécanismes d'attention sont utilisés pour chercher les régions de l'image source d'informations pour la reconstruction ou la classification. De plus en plus de méthodes de l'état-de-l'art basée apprentissage, notamment pour le problème de super-résolution, proposent des architectures plus profondes afin d'améliorer les performances, ce qui rend ces réseaux de plus en plus difficiles à entraîner. Ces mécanismes d'attention sont aujourd'hui une alternative mise en place pour améliorer les performances des architectures proposées sans avoir à ajouter de la profondeur aux réseaux considérés. En effet, en indiquant au réseau où se situent les parties de l'images informatives/pertinentes pour la reconstruction, le réseau serait mieux capable de s'adapter au problème.

6.2 État-de-l'art basé sur les mécanismes d'attention

Dans le domaine du traitement d'images, il existe plusieurs types d'attentions. On peut, par exemple, distinguer l'attention globale de l'attention locale ou encore l'attention spatiale de l'attention spectrale.

Les mécanismes d'attention spatiale ou spectrale sont des mécanismes d'attention locale, c'est à dire qu'ils apprennent les zones pertinentes en utilisant seulement une partie de l'entrée. Dans ce cas, la temps de calcul est réduit mais, dans certains cas, l'apprentissage peut être plus compliqué à cause du manque d'information.

L'attention globale apprend les zones pertinentes en utilisant toute la source. L'inconvénient de ce mécanisme repose sur la taille de la source : plus celle-ci est grande, plus le temps de calcul sera important. C'est ce type d'attention qui a été introduit par Bahdanau *et al.*, et qui a ensuite été repris en traitement d'images par Wang *et al.* [73] sous le nom d'auto-attention. De manière un peu plus détaillée, c'est un mécanisme reliant différentes positions d'une même entrée afin d'en calculer une représentation. Cela permet ainsi d'apprendre la corrélation entre les différentes parties de la séquence d'entrée. La méthode la plus connue est celle proposée par Wang *et al.* [73] schématisée en Fig. 6.1.

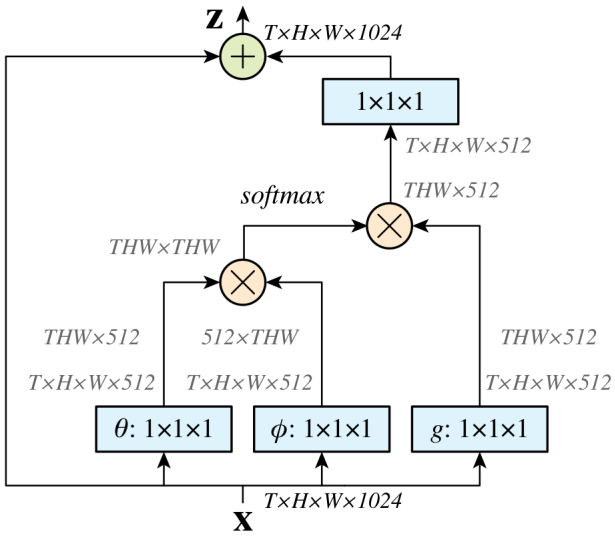


FIGURE 6.1 – Schéma représentatif du module d'attention de la méthode proposée par Wang *et al.*. X représente l'entrée, Z la sortie, $\otimes$ la multiplication matricielle et $\oplus$ la somme matricielle [73].

Cette méthode a pour objectif d'extraire l'information en prenant en compte tous les pixels de l'image à l'aide de l'opération de multiplication matricielle. C'est ce qui fait la force de cette méthode mais aussi sa faiblesse car cette opération est très coûteuse en mémoire et en temps de calcul, la rendant difficilement utilisable en l'état.

6.2.1 Méthodes basées attention spatiale et attention spectrale

• Woo *et al.* [77] ont proposé une méthode basée attention (CBAM) pour le problème de classification. Cette méthode considère deux types d'attention dans le réseau, un module d'attention spatiale et un module d'attention spectrale.

Le module d'attention spatiale (Fig.6.2) a pour objectif de repérer les zones pertinentes pour la classification. La première étape consiste alors à réduire la troisième dimension grâce à des couches de pooling moyennant et de pooling maximal. Les caractéristiques spatiales sont ensuite extraites à l'aide d'une couche convolutive. Finalement, la carte d'attention est obtenue en appliquant une fonction sigmoïde pour obtenir une carte entre $[0, 1]$.

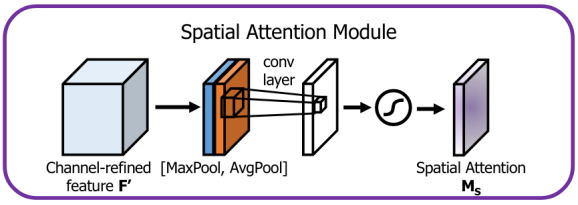


FIGURE 6.2 – Module d'attention spatiale proposé par Woo *et al.* pour la méthode CBAM [77].

Le module d'attention spectrale (Fig.6.3) a pour objectif d'exploiter les relations inter-canal. Pour cela, la dimension spatiale est réduite à l'aide de couches de pooling moyennant et de pooling maximal. Et ensuite, les caractéristiques sont extraites grâce à deux couches denses. Finalement, la carte d'attention est obtenue en appliquant une fonction sigmoïde pour obtenir une carte entre $[0, 1]$.

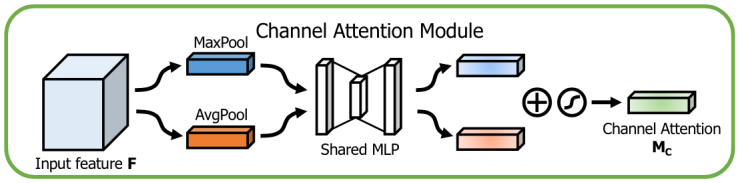


FIGURE 6.3 – Module d'attention spectrale proposé par Woo *et al.* pour la méthode CBAM [77].

Les deux modules d'attention sont intégrés de la façon suivante :

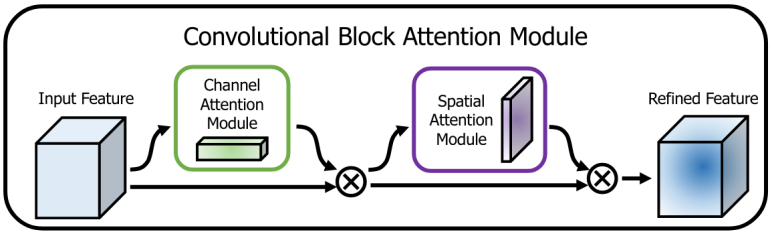


FIGURE 6.4 – Schéma de l'architecture générale proposée par Woo *et al.* [77].

Les modules sont simplement appliqués l'un après l'autre dans chaque bloc dans une architecture de type résiduelle.

• Zhang *et al.* [85] proposent une méthode RCAN basée réseaux pour le problème de super-résolution. Pour ce problème, la profondeur du réseau est cruciale. Cependant, plus un réseau est profond, plus il est difficile à entraîner. De plus, l'abondance d'information basse fréquence dans l'image en entrée et à travers le réseau entrave sa capacité à reconstruire l'image car ces informations basse résolution sont traitées de manière égale tout au long du réseau.

Ainsi, pour faire face à ces problèmes, les auteurs proposent un réseau (Fig. 6.5) résiduel très profond avec un mécanisme d'attention. En effet, ce réseau a une structure RIR (Residual In Residual) composée de plusieurs groupes résiduels avec de longues connexions résiduelles. De plus, chaque groupe est constitué de blocs résiduels avec de courtes connexions. Cette structure autorise l'abondance d'information et permet aussi au réseau de se concentrer sur la reconstruction des hautes fréquences manquantes.

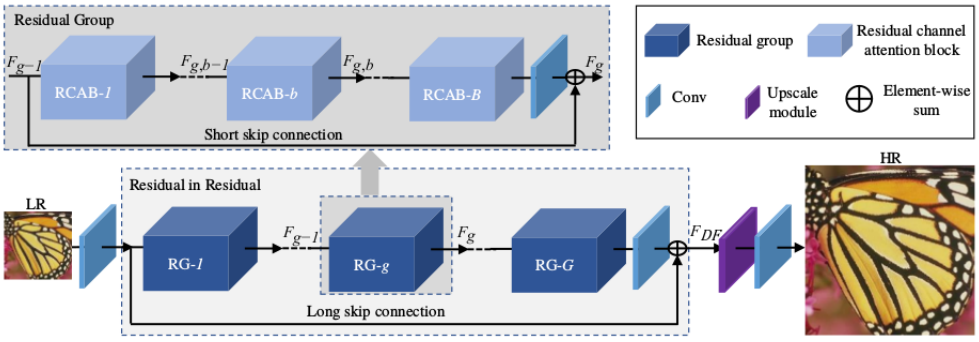


FIGURE 6.5 – Schéma représentatif du réseau proposé par Zhang *et al.* [85].

Le mécanisme d'attention, présenté en Figure 6.7, permet de redimensionner de manière adaptative les caractéristiques de chaque canal et aussi les interdépendances entre canaux.

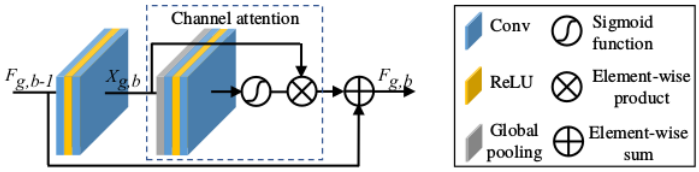


FIGURE 6.6 – Schéma du mécanisme d'attention des canaux proposé par Zhang *et al.* [85].

La couche de 'global pooling' fait la moyenne sur les lignes et les colonnes des éléments de chaque canal. Puis des couches convolutives et des fonctions d'activation sont utilisées pour combiner et mettre en évidence les canaux les plus pertinents pour la reconstruction. Cela se schématise de la façon suivante :

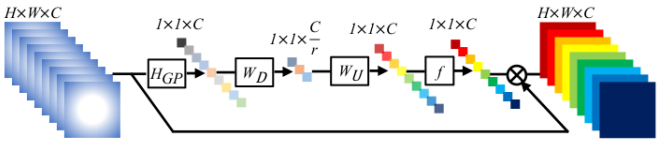


FIGURE 6.7 – Détails du mécanisme d'attention des canaux proposé par Zhang *et al.* [85].

Pour finir, afin d'optimiser les poids du réseau, les auteurs choisissent de minimiser la fonction suivante à l'aide d'une descente de gradient stochastique :

$$L(\Theta) = \frac{1}{N} \sum_{i \leq N} ||H_{RCAN}(I^i_{LR}) - I^i_{HR}||_1, \tag{6.1}$$

où H_{RCAN} est le réseau proposé et N la taille du batch.

• Wu *et al.* [78] proposent une méthode exploitant les avantages de la structure multi-échelle et du mécanisme d'attention pour le problème de super-résolution.

Cette méthode, MGAN, décrite en figures 6.8 et 6.9, se décompose en 2 étapes :
— Une partie extraction des caractéristiques considérant plusieurs blocs d'attention multi-échelle. Le mécanisme permet de capturer l'importance des caractéristiques mais aussi d'exploiter entièrement les indices de contexte spatial. Les connections denses multi-échelle permettent d'exploiter les caractéristiques de différentes couches à plusieurs échelles. Ainsi, des informations contextuelles plus riches sont apprises et la discrimination de la représentation des caractéristiques est aussi améliorée.
— Une partie reconstruction, ayant pour objectif de préparer la sortie du réseau à l'aide d'une couche de déconvolution et d'une couche de convolution renvoyant l'image reconstruite finale.

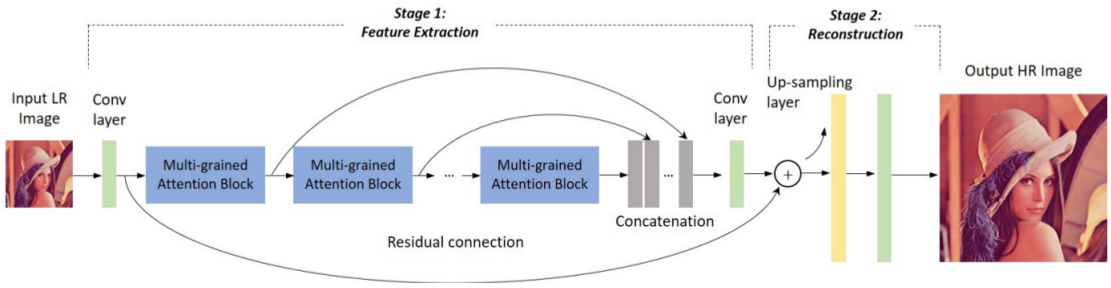


FIGURE 6.8 – Schéma de la méthode MGAN proposée par Wu *et al.* [78].

L'optimisation du réseau se fait via la fonction de perte suivante :

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{N} ||\mathcal{H}_{MGAN}(I_i^{LR}, \theta) - I_i^{HR}||_1, \tag{6.2}$$

où N est la taille du batch, $\mathcal{H}_{MGAN}$ le réseau, θ les paramètres de ce réseau, I^{LR} l'image basse résolution et I^{HR} l'image haute résolution. Cette fonction est simplement la norme L_1 de la différence entre l'image reconstruite et l'image haute résolution de référence.

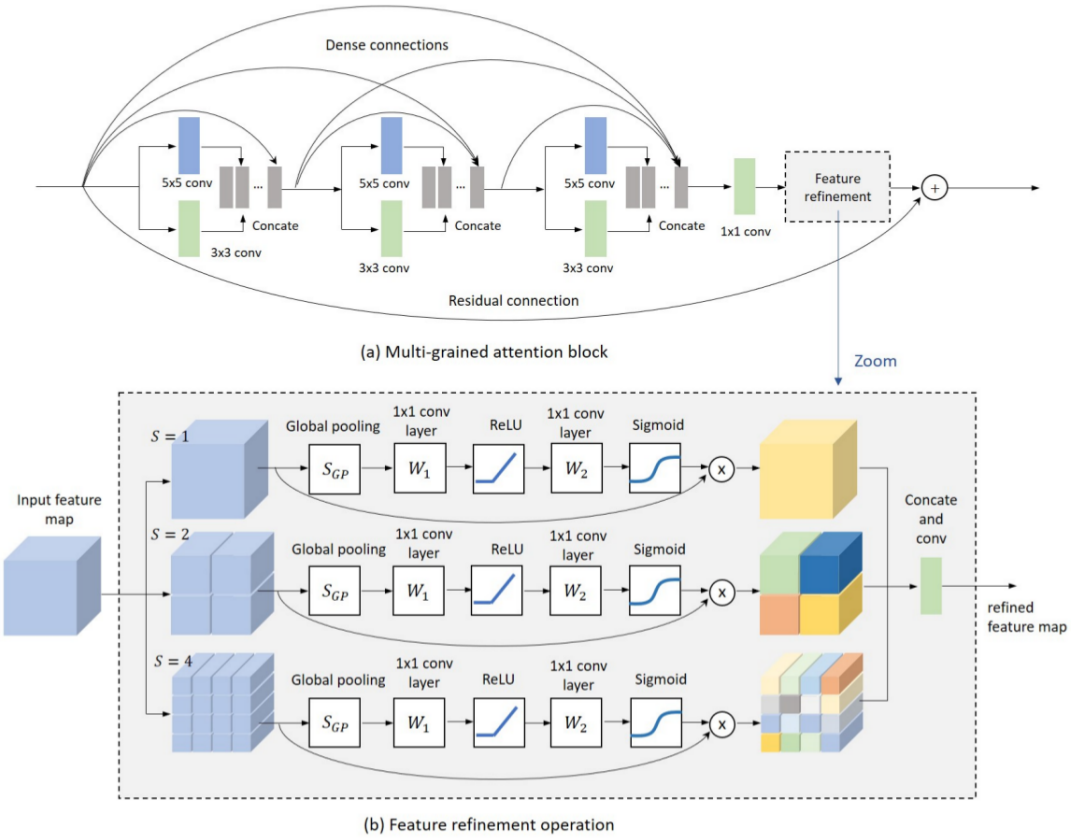


FIGURE 6.9 – Détails d'architecture des blocs d'attention proposés par Wu *et al.* [78].

On remarque que l'attention est spatiale, à la fois globale et locale (Figure 6.9 (b)) mais se caractérise de la même façon que Zhang *et al.* [85] d'un point de vue architecture. C'est à dire que les caractéristiques sont mélangées spatialement à l'aide d'une couche 'global pooling', puis les couches convolutives et les couches d'activation permettent de mettre plus ou moins de poids sur les mélanges les plus intéressants pour la reconstruction.

• Yang *et al.* [80] proposent une méthode basée attention pour le problème de super-résolution. Cette méthode est motivée par le fait que les méthodes de l'état de l'art proposent des réseaux de plus en plus profonds qui deviennent pas conséquent de plus en plus difficiles à entraîner ou à faire converger.

Ainsi, le modèle proposé MAMSR, présenté en figure 6.10, considère une architecture moins profonde mais plus performante par l'utilisation de plusieurs modules d'attention. Un module d'attention pour les canaux et un module d'attention spatiale. Cela permet alors d'apprendre la relation entre les canaux des caractéristiques mais aussi entre les pixels à chaque position.

L'attention des canaux permet de ré-allouer le poids de chaque canal des cartes de caractéristiques pour renforcer l'information haute fréquence dans l'image basse résolution.

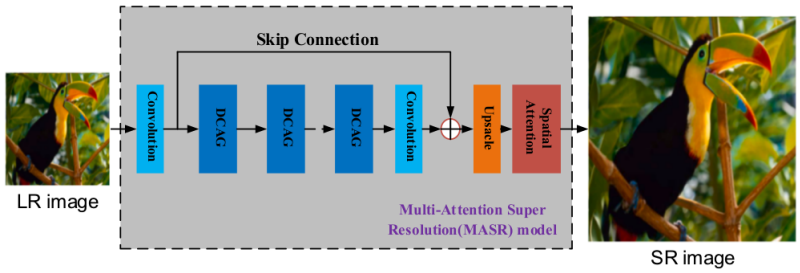


FIGURE 6.10 – Schéma représentatif de la méthode proposée MAMSR [80].

L'attention spatiale permet quant à elle d'entraîner un masque à travers différentes couches convolutives. La valeur du masque à chaque position représente le poids du pixel à la position correspondante dans l'image d'origine. En continuant d'entraîner les poids de ce masque durant l'apprentissage, l'information haute fréquence est renforcée de façon adaptative. Cela a pour objectif de reconstruire une image plus proche de l'image cible.

La combinaison de ces deux modules a alors pour objectif d'améliorer les hautes fréquences de l'image de façon adaptative dans l'image basse résolution.

L'architecture proposée se décompose en 4 parties :

i) Extraction des caractéristiques à l'aide de convolutions.

ii) Plusieurs modules DCAG. Ces modules se décomposent en 3 étapes : la compression des caractéristiques à l'aide d'un max pooling, la sélection des caractéristiques grâce à des couches denses et finalement une opération de multiplication afin d'ajouter l'attention à chaque canal.

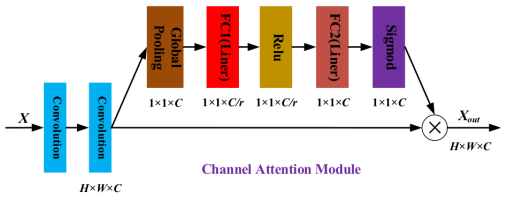


FIGURE 6.11 – Schéma du module d'attention des canaux.

iii) Un module de sur-échantillonnage.

iv) Un module d'attention spatiale.

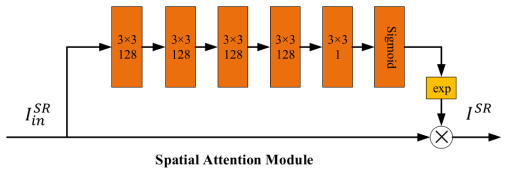


FIGURE 6.12 – Schéma du module d'attention spatiale.

Ce réseau est optimisé avec une norme l_1 de la différence entre l'image cible et l'image reconstruite.

On remarque que cette méthode mélange les deux idées présentes dans les méthodes RCAN [85] et MGAN [78]. En effet, les auteurs associent l'attention des canaux proposée par Zhang *et al.* pour la méthode RCAN et une adaptation un peu plus simple de l'attention spatiale proposée par Wu

et al. pour la méthode MGAN. Les architectures considérées pour les blocs d'attention sont très proches, seulement quelques différences, comme le nombre de couches convolutives par exemple, diffèrent.

• Zhang *et al.* [83] ont proposé une méthode basée attention pour le problème de pansharpening. Cette méthode incorpore l'attention dans une architecture de type U-Net de la façon suivante :

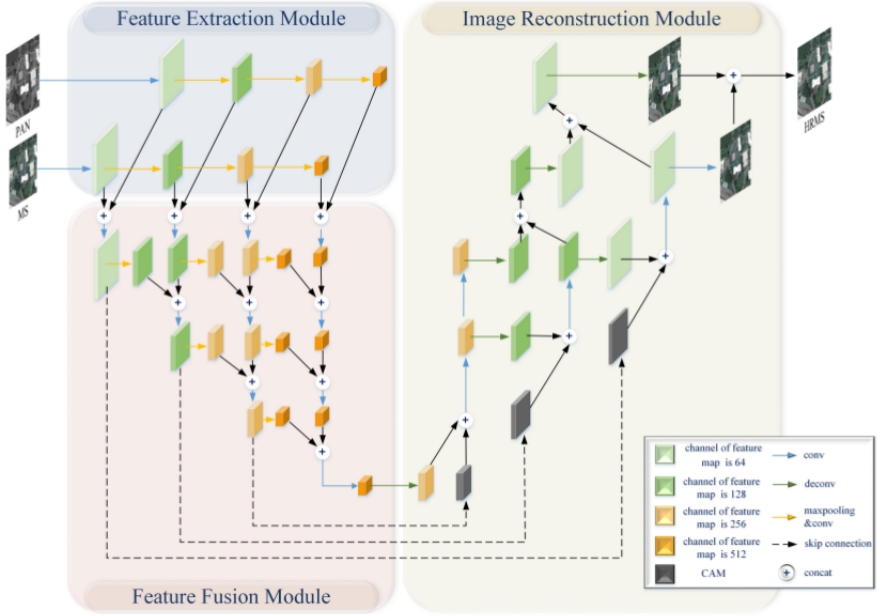


FIGURE 6.13 – Architecture générale de type U-Net proposée par Zhang *et al.* [83].

On peut voir en Fig.6.13 que le module d'attention est appliqué à plusieurs niveaux de l'architecture lors de la fusion des caractéristiques puis ré-injecté au niveau correspondant lors de la reconstruction de l'image.

De manière un peu plus détaillé, les auteurs proposent l'utilisation du module d'attention spectrale suivant :

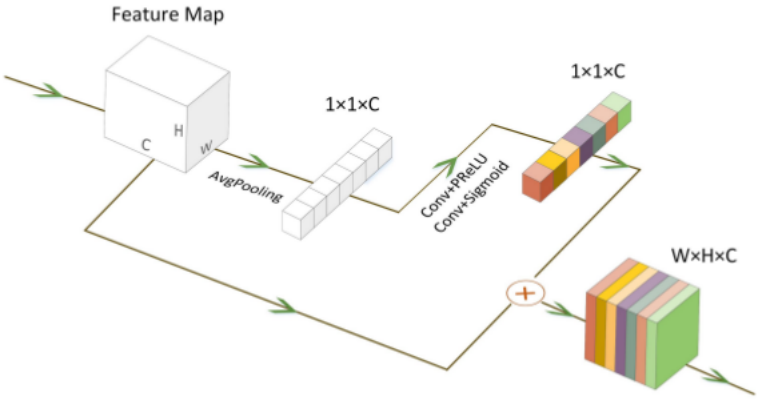
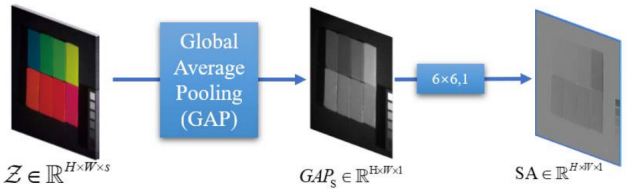


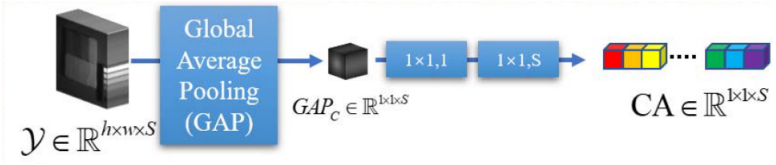
FIGURE 6.14 – Module d'attention utilisé par Zhang *et al.* [83].

Le principe de ce module est le même que présenté précédemment. C'est-à-dire que l'information est compressée à l'aide de couches de pooling puis analysée à l'aides de couches convolutives et de différentes fonctions d'activations.

• Hu *et al.* [35] ont proposé une méthode de fusion d'images hyperspectrales basée sur l'utilisation de mécanismes d'attention. L'attention spectrale est générée à partir de l'image hyperspectrale alors que la carte d'attention spatiale est générée sur l'image RGB haute résolution. Les modules d'attention utilisés sont les suivants :



(a)



(b)

FIGURE 6.15 – Module d'attention spatiale (a) et module d'attention spectral (b) proposés par Hu *et al.* [35].

Ces modules sont ensuite intégrés à une architecture du type ResNet où les cartes d'attention SA et CA sont appliquées à la sortie du dernier bloc.

• Zheng *et al.* [87] proposent une méthode de pansharpening hyperspectrale où des modules d'attention spatiale et spectrale sont intégrés à une architecture de type U-Net. Ces modules sont ajoutés à la sortie de chaque bloc et les cartes générées sont appliquées directement après, comme présentée en Fig. 6.16

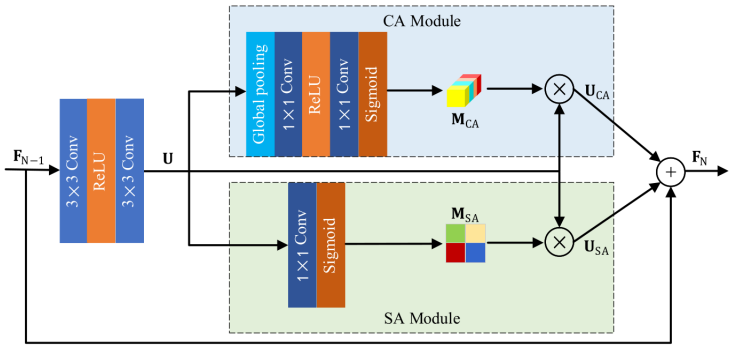


FIGURE 6.16 – Schéma des modules proposés par Zheng *et al.* [87].

Pour résumer l'ensemble des méthodes présentées ci-dessus, on remarque que les mécanismes d'attention proposés suivent un schéma global. Dans un premier temps, une couche de pooling est utilisée pour extraire ou compresser les caractéristiques sur la/les dimension(s) d'intérêt(s) si nécessaire. Ensuite ces caractéristiques sont mélangées à l'aide de couches convolutives ou denses. Finalement, une fonction d'activation sigmoïde est appliquée en sortie de couche pour générer une carte d'attention avec des valeurs comprises entre 0 et 1.

6.2.2 Méthodes utilisant un autre type d'attention

Ce paragraphe décrit quelques méthodes proposant des mécanismes d'attention dans un cadre un peu différent des méthodes présentées précédemment. Par exemple, Liu *et al.* [51] utilisent un réseau du type U-Net pour extraire les cartes d'attention, Dai *et al.* [19] utilisent des caractéristiques de second ordre, Lei *et al.* [46] utilisent ce qu'on appelle la *gate attention* ou encore Li *et al.* [47] proposent d'utiliser l'attention dans un cadre multi-échelle.

• Liu *et al.* [51] proposent une méthode basée attention pour le problème de super-résolution. Cette méthode permet de discriminer les zones de textures des zones non texturées pour ensuite compenser les hautes fréquences dans les zones où le besoin existe.

Pour cela, l'architecture en figure 6.17 est composée de plusieurs blocs denses résiduels pour améliorer la résolution de l'image en entrée, tout en considérant un mécanisme d'attention en parallèle, reconstruisant les hautes fréquences dans les zones texturées. L'architecture de ce mécanisme est inspirée d'une architecture de type encodeur-décodeur où plusieurs blocs résiduels sont utilisés.

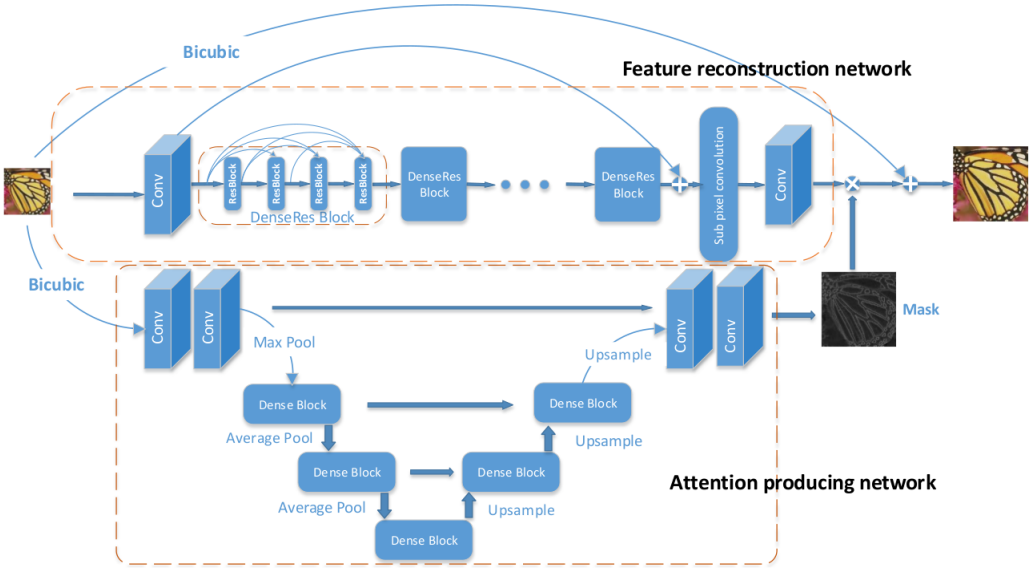


FIGURE 6.17 – Architecture proposée par Liu *et al.* [51].

La fonction de perte utilisée est une norme L_1 de la différence entre l'image reconstruite et l'image de référence. Cependant, ce type d'attention peut être coûteux en terme d'apprentissage car il dépend de la profondeur du réseau U-Net.

• Dai *et al.* [19] proposent une méthode de second ordre basée attention pour le problème de super-résolution. Ces travaux sont motivés par le fait que la plupart des papiers proposent des architecture de plus en plus profondes ou de plus en plus vastes en négligeant d'explorer la corrélation entre les caractéristiques des couches intermédiaires. Cela entrave alors la capacité de représentation des réseaux convolutifs car ils ne sont pas exploités au maximum de leur potentiel.

Ainsi, les auteurs proposent un réseau, SAN présenté en figure 6.18, considérant les statistiques de second ordre dans un mécanisme d'attention.

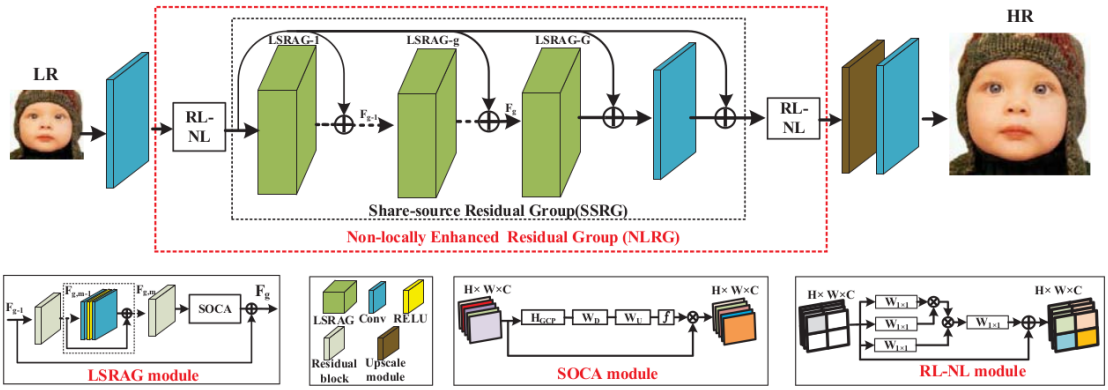


FIGURE 6.18 – Schéma représentatif de la méthode SAN proposée par Dai *et al.* [19].

L'architecture est composée de plusieurs blocs LSRAG puis d'un module RL-NL, tous deux développés par les auteurs. Les blocs LSRAG ont pour objectif de redimensionner de façon adaptative les caractéristiques de chaque canal en utilisant les statistiques du second ordre pour obtenir une représentation de ces caractéristiques plus discriminative. Le module RL-NL a pour but d'incorporer des opérations non locales dans le réseau afin de capturer l'information contextuelle spatiale longue distance mais aussi pour apprendre la représentation de caractéristiques de plus en plus abstraites.

Cette méthode se veut plus performante dans l'apprentissage de la corrélation des caractéristiques et dans l'expression de ces caractéristiques. Cependant, d'un point de vue attention, l'architecture des blocs reste très proche de celles présentées précédemment. La différence majeure se situe dans l'utilisation des statistiques de second ordre en entrée de ces blocs.

Finalement, les poids de ce réseaux sont optimisés à l'aide de la fonction de perte suivante :

$$L(\Theta) = \frac{1}{N}\sum_{N} ||H_{SAN}(I^i_{LR}) - I^i_{HR}||_1, \tag{6.3}$$

où I_{LR} est l'image basse résolution, I_{HR} l'image haute résolution cible, H_{SAN} le réseau proposée et N la taille du batch.

• Lei *et al.* [46] ont proposé une méthode de pansharpening comprenant un module d'attention globale dans une architecture de type ResNet. L'attention est insérée après chaque bloc résiduel tout au long du réseau de la façon suivante :

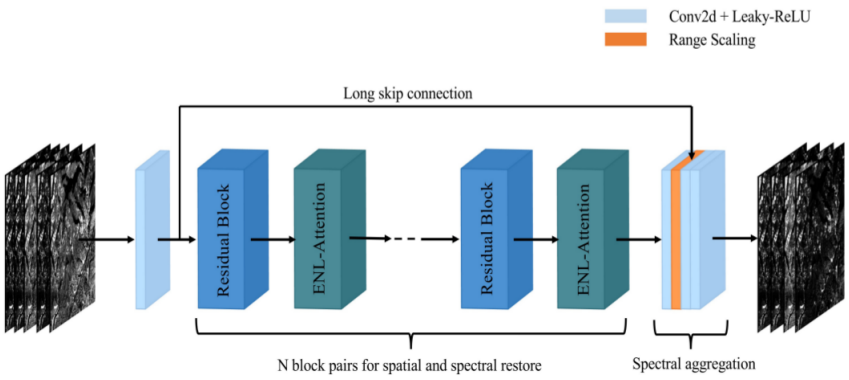


FIGURE 6.19 – Schéma de l'architecture générale proposée par Lei *et al.* [46].

Pour le module d'attention, les auteurs reprennent l'architecture proposée par Wang *et al.*

[73] en améliorant la complexité calculatoire liée au calcul de la matrice d'auto-corrélation. Ainsi, l'architecture proposée pour le module d'attention est la suivante :

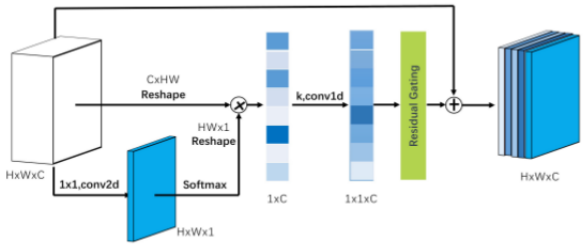


FIGURE 6.20 – Module d'attention proposée par Lei *et al.* [46], inspiré de Wang *et al.*[73].

Comme on peut le voir en Fig.6.20, la matrice de corrélation est obtenue à l'aide d'une couche convolutive et ensuite une opération de soft-max sur la dimension spectrale est utilisée pour obtenir la carte d'attention spatiale.

Dans un second temps, une couche de convolution 1D est utilisée pour modéliser les interactions inter-canaux. Cette information est ensuite intégrée à l'aide d'un "residual gating mechanism". Ce mécanisme a pour objectif de contrôler la sortie afin que suffisamment d'informations soient transmises. Cette dernière couche sert à améliorer la convergence.

• Li *et al.* [47] ont proposé une méthode basée attention pour la problème de pansharpening qui extrait des caractéristiques à plusieurs échelles. Cette architecture se schématise de la façon suivante :

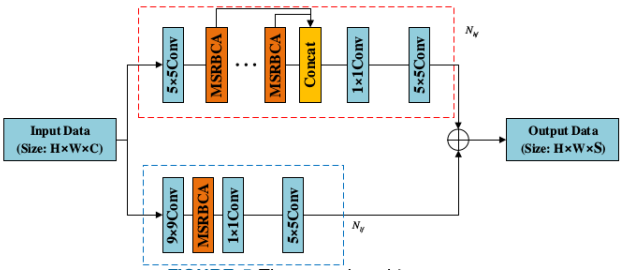
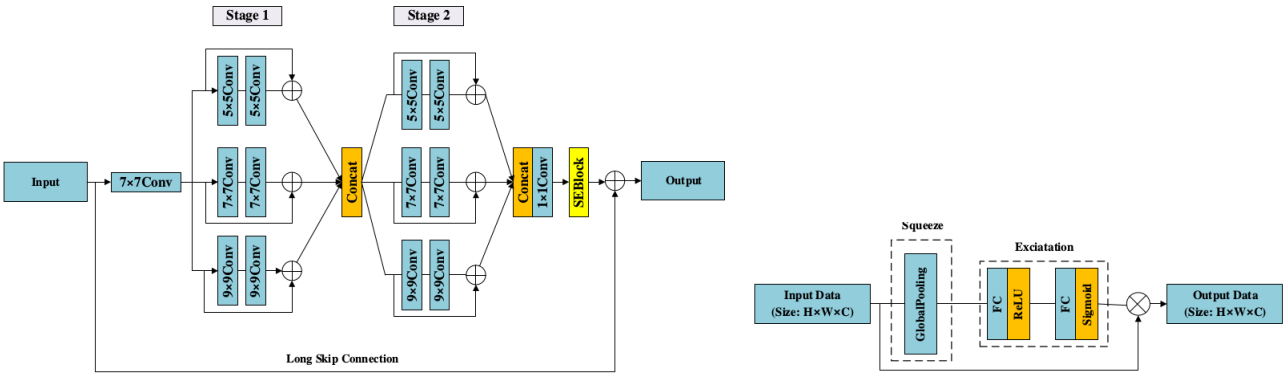


FIGURE 6.21 – Schéma de la méthode proposée par Li *et al.* [47].

L'attention quant à elle est utilisée dans chaque bloc multi-échelle "MSRBCA".



block MSRBCA Attention

FIGURE 6.22 – Schéma des blocs multi-échelles et schéma de l'attention [47].

L'attention est obtenue de la même façon que dans beaucoup de méthodes présentées précédemment. L'information est "compressée" à l'aide de couches de pooling puis analysée à l'aide de couches denses et de fonctions d'activations.

6.3 Contribution : méthode combinant attention spatiale et spectrale

Inspirés par l'état-de-l'art précédent, nous proposons une méthode exploitant deux modules d'attention : un module d'attention spatiale et un module d'attention spectrale. Elle reprend notre précédente méthode basée multi-discriminateurs [26], et se situe donc dans un cadre GAN considérant deux discriminateurs et deux contraintes dans la fonction de perte du générateur.

Cette méthode, illustrée en Figure 6.23, a pour objectif d'exploiter les mécanismes d'attention afin de répondre à notre problème. Chacun des mécanismes se concentre sur l'une des reconstructions à effectuer. En effet, le module d'attention spatiale cherche à focaliser l'attention du réseau sur les zones pertinentes pour améliorer la résolution spatiale, c'est-à-dire la géométrie et les hautes fréquences présentes dans les images. Le module d'attention spectrale, quant à lui, a pour objectif d'améliorer la résolution spectrale en mettant en avant les canaux pertinents à la reconstruction.

Ces deux mécanismes s'inscrivent dans le générateur et chacun prend en entrée l'image la plus adéquate pour accomplir leur tâche. Pour cela, le module d'attention spatiale prend en entrée l'image panchromatique qui contient toute l'information spatiale dont nous avons besoin. Le module d'attention spectrale prend en entrée l'image multispectrale qui possède les informations spectrales nécessaires.

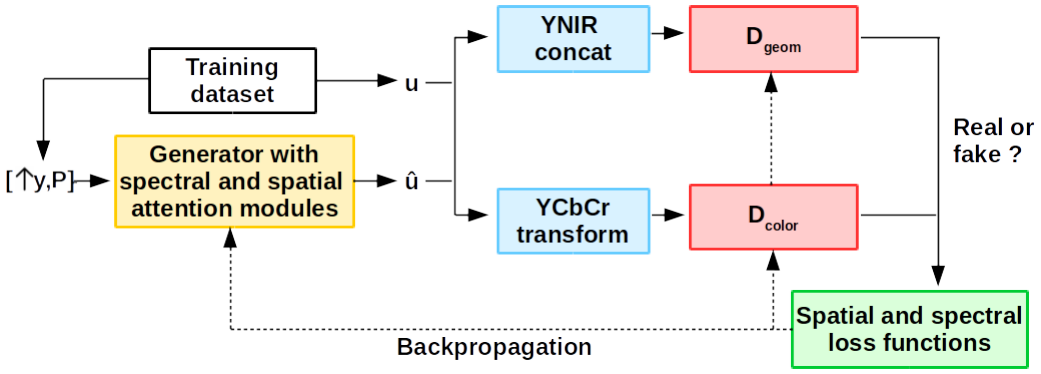


FIGURE 6.23 – Schéma général de la méthode proposée.

6.3.1 Modules d'attention

Nous proposons deux architectures pour les mécanismes d'attention spatiale et spectrale respectivement. Ces architectures sont inspirées de la méthode "générique", c'est-à-dire commune à beaucoup de méthodes proposées dans la littérature, pour construire des cartes d'attention :

- Dans un premier temps, les caractéristiques des images panchromatiques et multispectrales sont extraites à l'aide d'une ou de plusieurs couches convolutives.
- Ces caractéristiques sont ensuite mélangées à l'aide d'une couche de pooling ou avec une couche convolutive de noyau 1×1.
- Finalement, la carte d'attention attendue est obtenue en utilisant une fonction d'activation sigmoïde, qui va réduire l'intervalle des valeurs entre 0 et 1.

Dans le cas de l'attention spectrale, présentée en Fig. 6.24, nous utilisons une couche de pooling moyennant sur les dimensions spatiales pour ainsi réduire les dimensions de l'image et se concentrer sur la dimension spectrale.

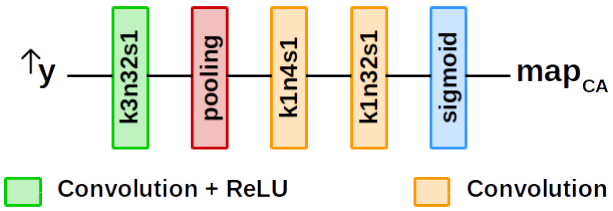


FIGURE 6.24 – Schéma de l'architecture proposée pour le mécanisme d'attention spectrale, où k est la taille du noyau de la convolution, n le nombre de filtres, s la fenêtre de déplacement du noyau et $\uparrow y$ est l'image multispectrale agrandie par interpolation bicubique.

Dans notre mécanisme d'attention spatiale 6.25, nous n'utilisons pas de couche de pooling pour réduire la dimension spectrale car les couches de convolution de noyau 1×1 sont très performantes pour ce genre de tâche.

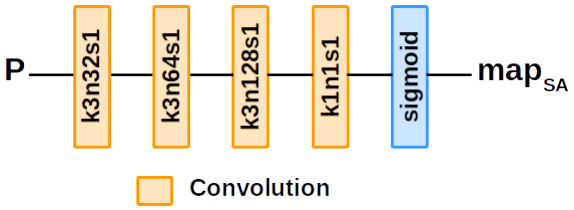


FIGURE 6.25 – Schéma de l'architecture proposée pour le mécanisme d'attention spatiale, où k est la taille du noyau de la convolution, n le nombre de filtres, s la fenêtre de déplacement du noyau et P est l'image panchromatique.

Cependant, quel que soit le mécanisme considéré, nous avons remarqué que la sortie de la dernière couche convolutive a de très grandes valeurs. Par conséquent, les valeurs de sortie de la fonction sigmoïde sont saturées à 1. Cela ne permet donc pas d'obtenir des cartes d'attention exploitables dans le sens où celles-ci n'apportent aucune information et sont constantes et égales à 1.

Pour faire face à ce problème, nous avons choisi de changer légèrement la fonction sigmoïde afin d'obtenir une carte informative. Pour ce faire, nous avons choisi de diviser l'entrée de la sigmoïde par une variable μ. Cela revient alors à utiliser la fonction d'activation suivante :

$$g(x) = \frac{1}{1 + e^{x/\mu}}, \tag{6.4}$$

où $\mu = max(x)$ et x la carte de caractéristiques en entrée de la fonction sigmoïde. La variable μ est donc différente pour chaque image.

La fonction considérée (6.4) a les mêmes propriétés que la fonction sigmoïde d'origine. Elle est donc tout à fait adaptée à notre problème.

6.3.2 Générateur

Pour utiliser au mieux l'information que nous possédons, nous calculons la carte d'attention spatiale sur l'image panchromatique et la carte d'attention spectrale sur le l'image multispectrale, tel que Hu *et al.* [35] le proposent pour le problème de fusion d'images hyperspectrales.

De plus, pour utiliser au maximum l'information fournie par les cartes d'attention, nous les injectons après chaque bloc dense résiduel. Ces deux modules d'attention s'intègrent donc dans l'architecture du générateur, comme illustré la Figure 6.26.

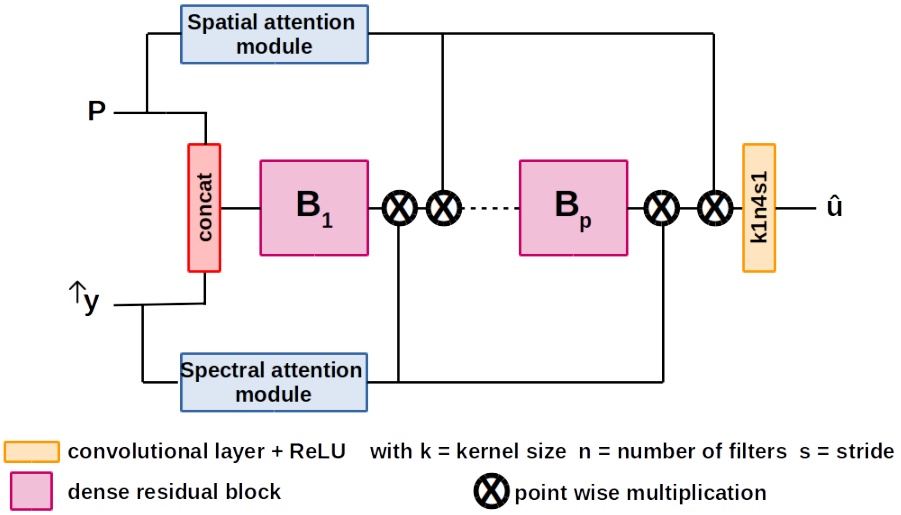


FIGURE 6.26 – Schéma du générateur considéré, où P est l'image panchromatique et $\uparrow y$ est l'image multispectrale agrandie à la taille de P à l'aide d'une interpolation bicubique.

L'attention permet de focaliser le réseau sur les zones pertinentes de l'image. Ainsi, nous avons pu diminuer la profondeur du générateur en considérant un bloc dense résiduel en moins par rapport à nos dernières contributions [25, 26]. Cela signifie donc que le réseau est moins profond que dans nos précédentes méthodes proposées [25, 26] car les mécanismes d'attention sont mis en parallèle.

6.3.3 Discriminateurs

Les discriminateurs considérés sont ceux de notre précédente contribution [26]. Ces deux discriminateurs ont des objectifs distincts. Le premier se concentre sur la résolution spatiale en prenant en entrée la luminance et la bande infra-rouge des images. Le second prend en entrée les composantes chromatiques Cb et Cr pour améliorer la reconstruction de la résolution spectrale.

Nous rappelons que l'architecture des discriminateurs, détaillée au Chap. 5, est la suivante :

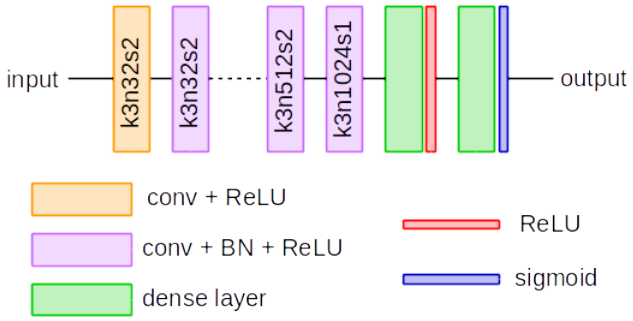


FIGURE 6.27 – Architectures des discriminateurs, où k est la taille du noyau, n le nombre de filtres et s la taille du pas pour chacune des couches convolutives.

6.4 Résultats

6.4.1 Bases de données, détails d'implémentation et métriques de comparaison

- Bases de données et comparaison des méthodes :

Le cadre de comparaison reste le même que précédemment, c'est-à-dire que nous utilisons les métriques PSNR, SAM, ERGAS, RMSE, CC et PSNR_h pour comparer les résultats sur les bases

de données Pléiades et WorldView 3. Rappelons également que ces deux bases de données ont été créées à l'aide du protocole de Wald [72].

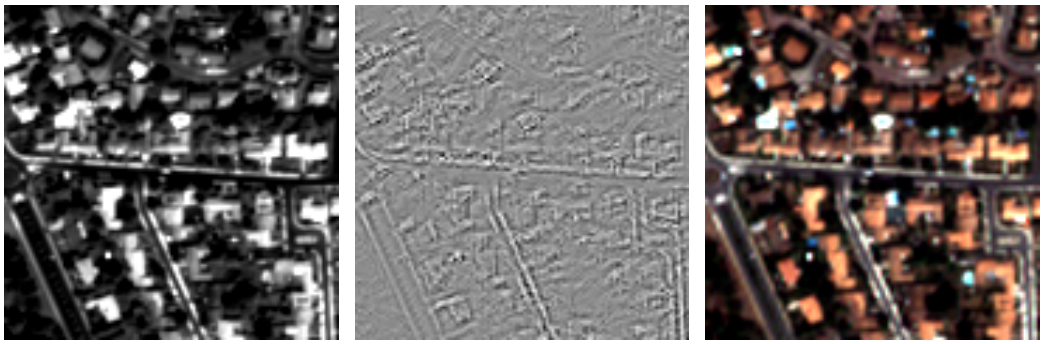
• Détails d'implémentation :

La méthode proposée est implémentée avec TensorFlow 1.2 et utilise l'algorithme ADAM pour minimiser les fonctions de pertes. Les tailles de batchs utilisées sont 17 pour la base Pléiades et 19 pour la base WorldView 3.

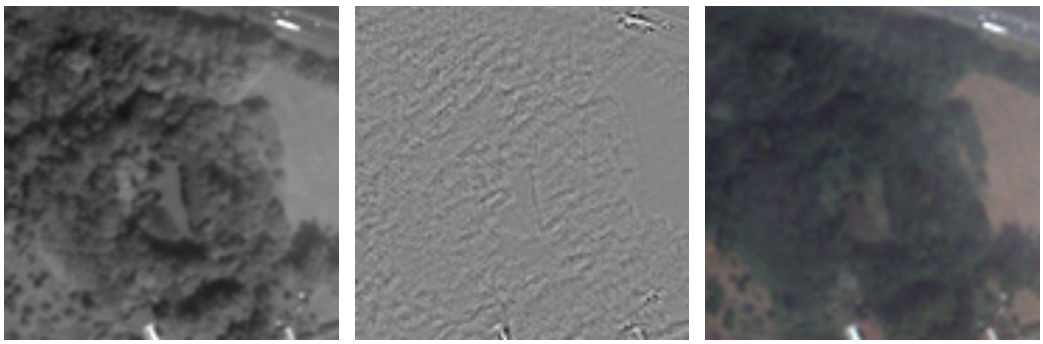
6.4.2 Comparaison avec les méthodes de l'état-de-l'art

• Cartes d'attention :

En premier lieu, il est important d'observer les cartes d'attention spatiales obtenues avec la méthode proposée.



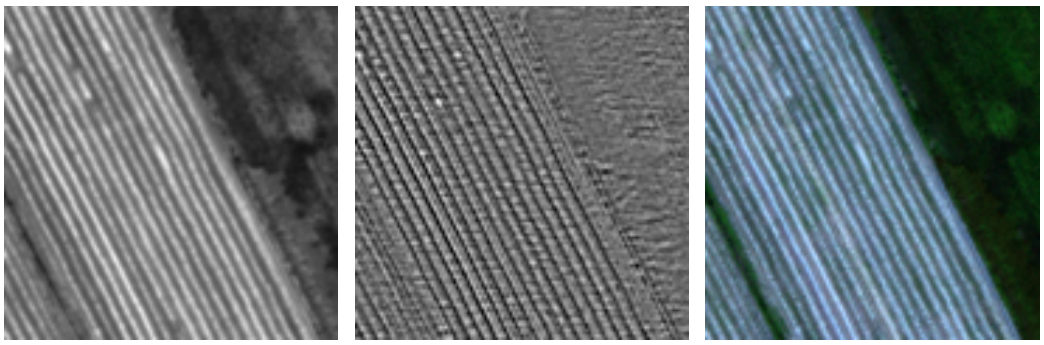
(a)



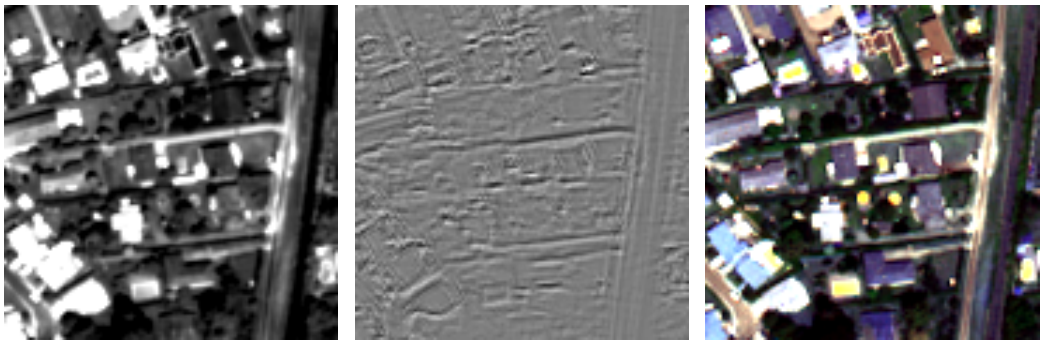
(b)

FIGURE 6.28 – Exemples de cartes d'attention spatiale obtenues sur la base de données Pléiades. La première colonne correspond à l'image panchromatique sur laquelle est calculée la carte d'attention spatiale, la seconde colonne affiche la carte d'attention spatiale obtenue et la troisième colonne l'image RGB correspondante.

Les cartes d'attention spatiale présentées en Figures 6.28 et 6.29 nous permettent de voir que certaines textures sont mieux représentées que d'autres. En effet, les lignes droites présentes en Figure 6.29 (a) ainsi que les zone urbaines en Figures 6.28 (a) et 6.29 (b) sont très bien représentées. En revanche, les zones de végétation comme la forêt en Figure 6.28 (b) est présente mais le contraste est moins marqué. Cela s'explique par le fait que l'image panchromatique associée sur laquelle est calculée la carte d'attention est moins contrastée en zones de végétation qu'en zones urbaines.



(a)



(b)

FIGURE 6.29 – Exemples de cartes d'attention spatiale obtenues sur la base de données WorldView 3. La première colonne correspond à l'image panchromatique sur laquelle est calculée la carte d'attention spatiale, la seconde colonne affiche la carte d'attention spatiale obtenue et la troisième colonne l'image RGB correspondante.

- Comparaison :

Afin de mettre en évidence l'apport des mécanismes d'attention considérés, nous comparons cette méthode avec notre précédente méthode MDSSCGAN-SAM [26]. Nous considérons aussi les méthodes PSGAN [50] et PanNet [79], déjà utilisées en comparaison tout au long de cette thèse. Finalement, nous considérons la méthode HSRNet basée attention proposée par Hu *et al.* [35] pour le problème de fusion d'images hyperspectrales. Ce problème étant légèrement différent du nôtre, nous avons adapté la méthode HSRNet à notre problème pour pouvoir comparer. Toutes les méthodes considérées ont été ré-entraînées sur les deux bases de données.

Méthode	CC	SAM	PSNR	$PSNR_h$	ERGAS
valeur idéale	1	0	max	max	0
PanNet [79]	<u>0.950</u>	<u>0.157</u>	28.36	<u>26.60</u>	<u>8.77</u>
PSGAN [50]	0.952	0.155	<u>26.59</u>	26.96	4.28
HSRNet [35]	**0.971**	0.149	28.81	27.13	4.38
MDSSC-GAN SAM [26]	0.970	**0.137**	**29.45**	27.45	**3.88**
méthode proposée	0.970	0.142	29.32	**27.47**	3.98

TABLE 6.1 – Résultats quantitatifs des différentes méthodes de l'état-de-l'art sur la base de données Pléiades. Les meilleurs résultats sont en gras et les pires sont soulignés.

Dans un premier temps, sur la base Pléiades, on remarque que les meilleurs résultats dépendent à la fois de la méthode considérée, mais également de la métrique utilisée pour la comparaison. La méthode HSRNet [35] donne le meilleur résultat pour la mesure CC. Donc cette méthode reconstruit bien la résolution spatiale des images. En revanche, la méthode proposée semble conserver mieux

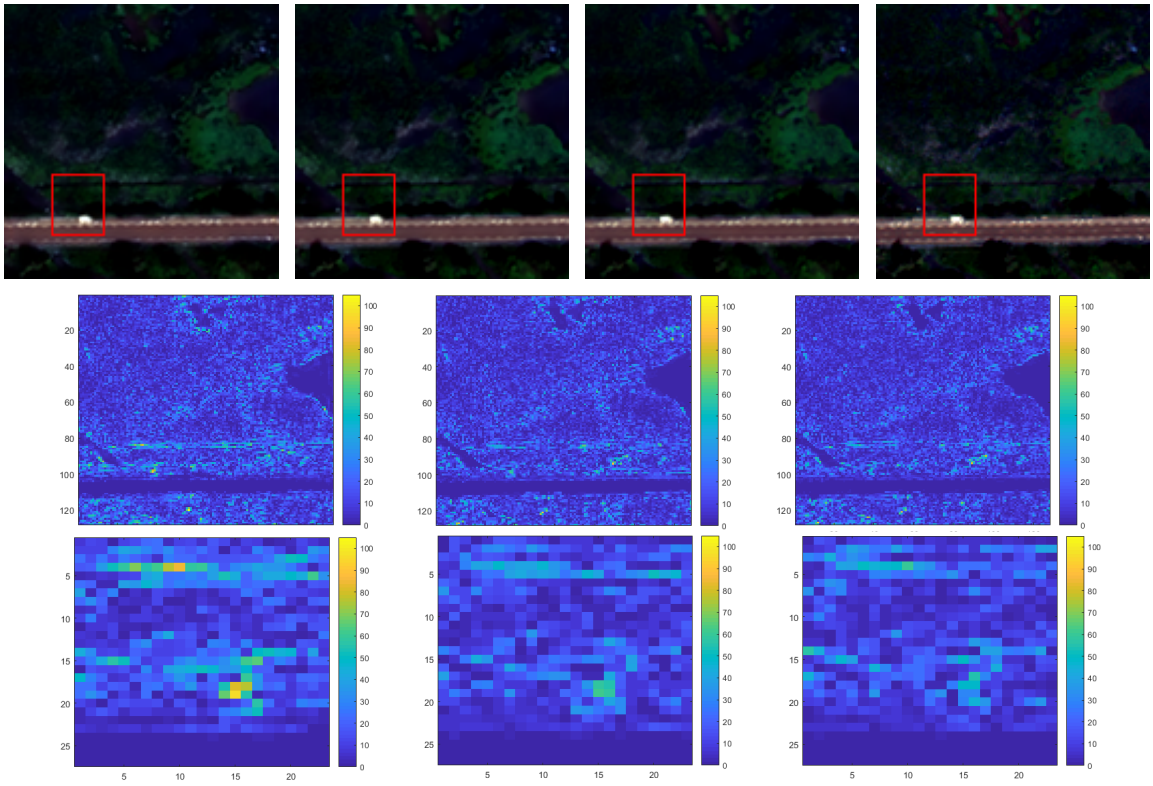
les hautes fréquences de l'image car la métrique PSNR_h est meilleure. La mesure SAM nous indique que la résolution spectrale est mieux conservée avec la méthode MDSSCGAN SAM [26].

On peut dire que, de manière globale, la méthode MDSSCGAN SAM [26] donne de meilleurs résultats quantitatifs.

Méthode	CC	SAM	PSNR	$PSNR_h$	ERGAS
valeur idéale	1	0	max	max	0
PanNet [79]	<u>0.930</u>	<u>0.135</u>	<u>23.22</u>	<u>23.93</u>	<u>8.01</u>
PSGAN [50]	0.966	0.110	29.30	26.82	6.33
HSRNet [35]	**0.960**	0.100	29.44	26.16	6.59
MDSSC-GAN SAM	0.968	0.104	29.62	**26.82**	6.31
méthode proposée	0.968	**0.095**	**29.89**	26.52	**6.17**

TABLE 6.2 – Résultats quantitatifs pour les différentes méthodes de l'état-de-l'art sur la base de données WorldView 3. Les meilleurs résultats sont en gras et les pires sont soulignés.

Les résultats sur la base de données WorldView 3 diffèrent un peu de ceux obtenus avec la base Pléiades. En effet, on peut voir que les meilleurs résultats sont obtenus avec la méthode proposée basée attention. Cependant, on remarque que la méthode HSRNet [35] donne encore le meilleur résultat pour la métrique CC.

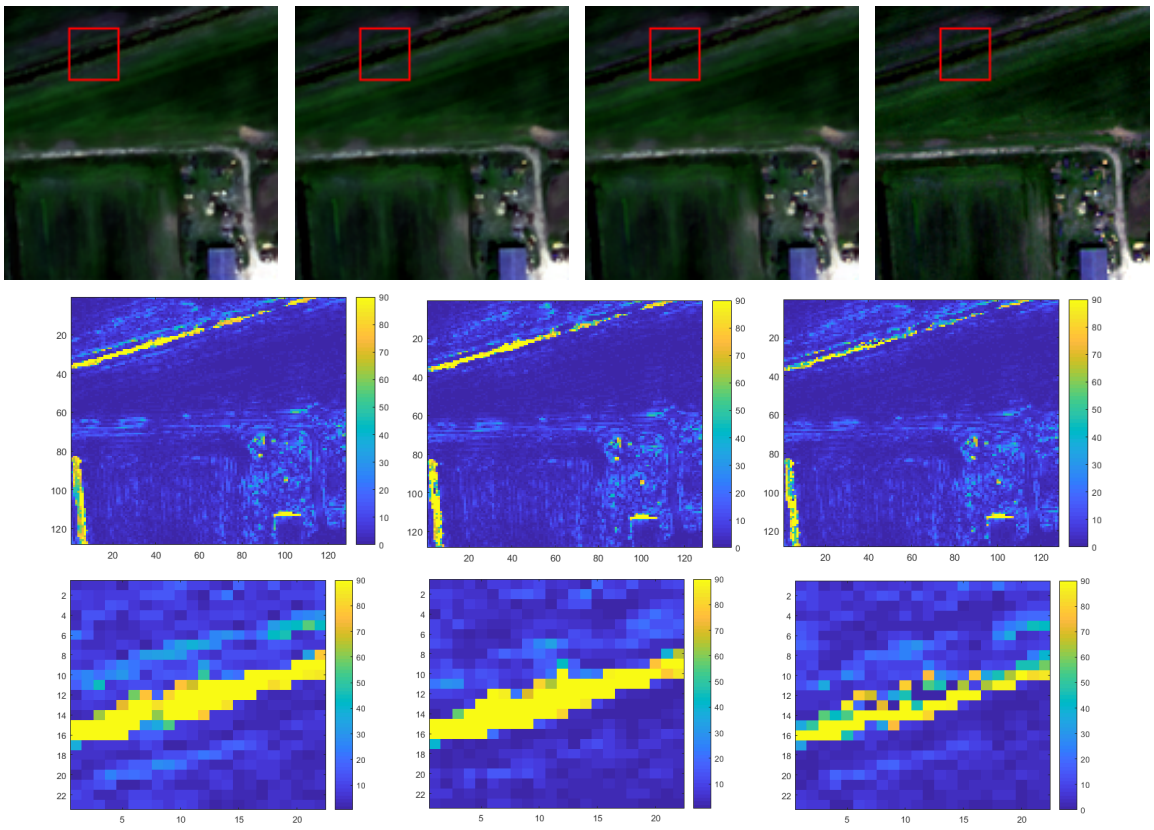


HSRNet [35]
$PSNR = 28.78$

MDSSCGAN SAM [26]
$PSNR = 29.79$

Méthode proposée
$PSNR = 30.01$

FIGURE 6.30 – Résultats obtenus sur une image WorldView 3. De gauche à droite : méthode HSRNet [35], la méthode MDSSCGAN SAM [26], la méthode proposée et l'image de référence. La première ligne correspond à l'image RGB, la seconde ligne à la différence entre l'image de référence et l'image reconstruite et la dernière ligne est une partie zommée. Le bleu indique aucune différence et le jaune une grande différence de valeur entre les pixels. La méthode proposée présente l'image la mieux reconstruite.



HSRNet [35]
$SAM = 0.129$

MDSSCGAN SAM [26]
$SAM = 0.123$

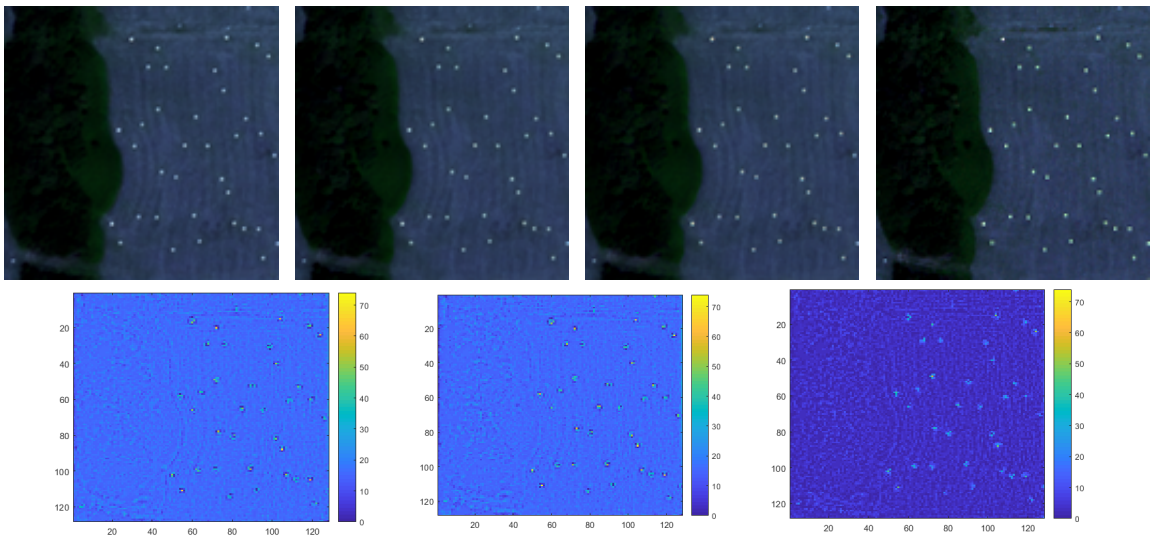
Méthode proposée
$SAM = 0.116$

FIGURE 6.31 – Carte SAM obtenues sur une image WorldView 3. De gauche à droite : méthode HSRNet [35], la méthode MDSSCGAN SAM [26], la méthode proposée et l'image de référence. La première ligne correspond à l'image RGB, la seconde ligne à la carte SAM et la dernière ligne est une partie zommée. Un pixel bleu indique aucune distorsion spectrale, un pixel jaune indique une forte distorsion. On voit que la méthode proposée présente moins de distorsions spectrales.

Les exemples affichés en Figures 6.30, 6.31 et 6.32, pour la base de données WorldView 3, sont bien représentatifs des résultats quantitatifs obtenus en Table 6.2. En effet, on peut voir qu'en affichant la différence entre l'image de référence et l'image reconstruite (Figure 6.30) la méthode proposée reconstruit une image dont les valeurs des pixels correspondent mieux à celles de l'image de référence.

De plus, lorsque l'on regarde la carte de la distorsion spectrale à chaque pixel en Figure 6.31, on voit qu'il y a plus de pixels jaune pour les méthodes HSRNet et MDSSCGAN SAM. Une distorsion plus forte est donc présente sur ces images. Ces méthodes reconstruisent donc moins bien la résolution spectrale des images contrairement à la méthode proposée.

Finalement, lorsque l'on regarde les hautes fréquences des images et notamment la différence entre les hautes fréquences de l'image de référence et celles de l'image reconstruite (Figure 6.32), on remarque très facilement que la méthode proposée donne un meilleur résultat. En effet, on peut voir sur l'image que la majorité des pixels sont bleus foncés pour la méthode proposée et bleus clairs pour les méthodes HSRNet et MDSSCGAN SAM. Cela signifie qu'il y a moins de différence entre les hautes fréquences de l'image reconstruite et celles de l'image de référence avec la méthode proposée. La reconstruction est donc plus précise.

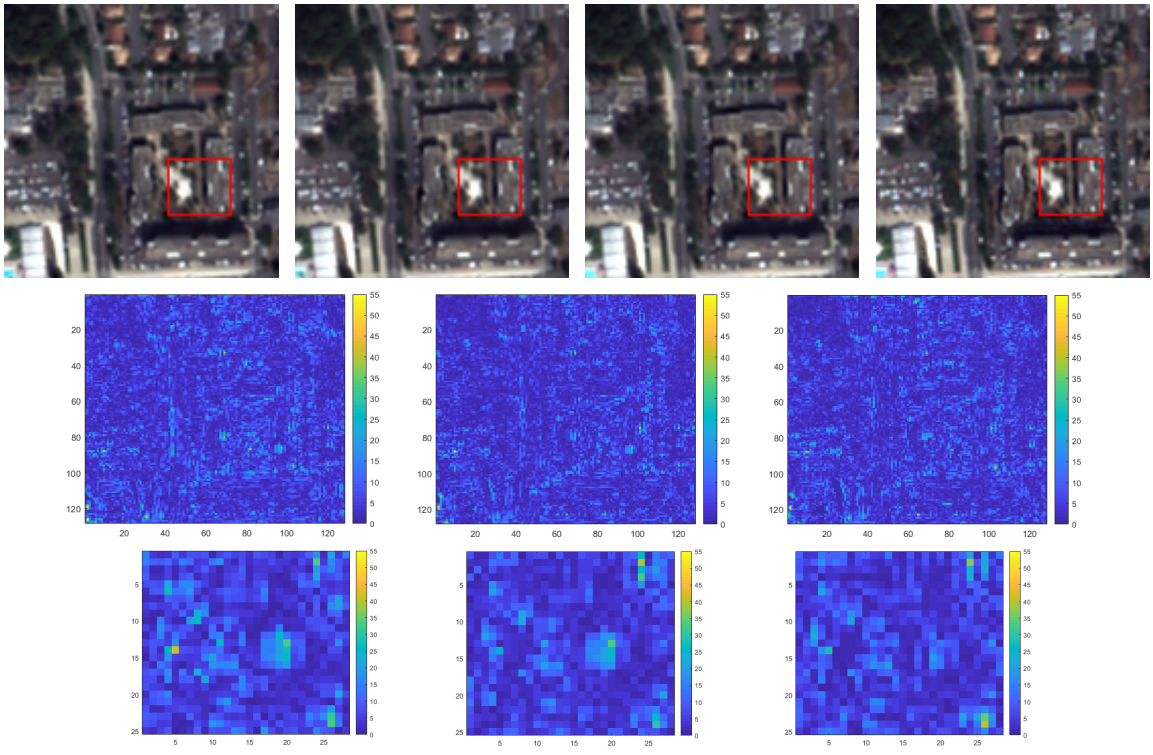


HSRNet [35]
$PSNR_h = 23.67$

MDSSCGAN SAM [26]
$PSNR_h = 23.85$

Méthode proposée
$PSNR_h = 33.49$

FIGURE 6.32 – Résultats obtenus sur une image WorldView 3. De gauche à droite : méthode HSRNet [35], la méthode MDSSCGAN SAM [26], la méthode proposée et l'image de référence. La première ligne correspond à l'image RGB, la seconde ligne à la différence entre les hautes fréquences de l'image de référence et celles de l'image reconstruite et la dernière ligne est une partie zommée. Le bleu indique aucune différence et le jaune une grande différence de valeur entre les pixels.

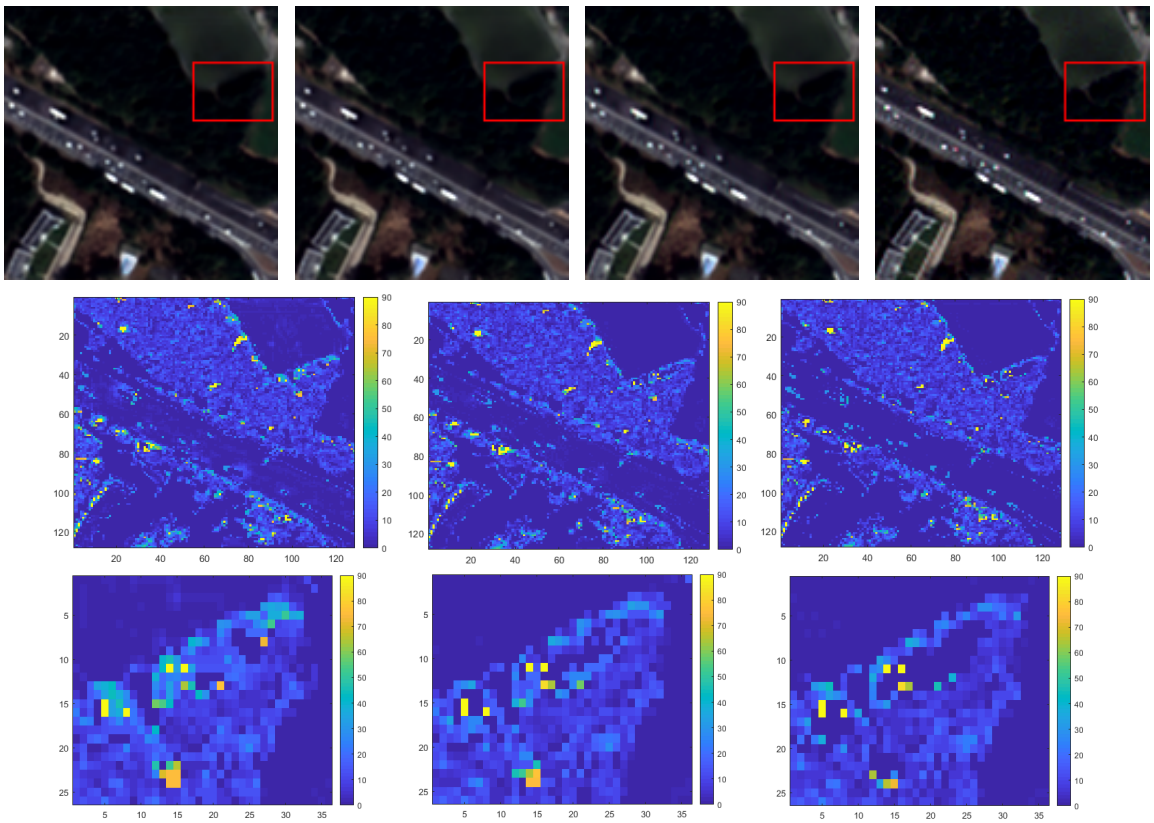


HSRNet [35]
$PSNR = 29.35$

MDSSCGAN SAM [26]
$PSNR = 30.27$

Méthode proposée
$PSNR = 30.15$

FIGURE 6.33 – Résultats obtenus sur une image Pléiades. De gauche à droite : méthode HSRNet [35], la méthode MDSSCGAN SAM [26], la méthode proposée et l'image de référence. La première ligne correspond à l'image RGB, la seconde ligne à la différence entre l'image de référence et l'image reconstruite et la dernière ligne est une partie zommée. Le bleu indique aucune différence et le jaune une grande différence de valeur entre les pixels.



HSRNet [35]
$SAM = 0.113$

MDSSCGAN SAM [26]
$SAM = 0.103$

Méthode proposée
$SAM = 0.101$

FIGURE 6.34 – Carte SAM obtenues sur une image Pléiades. De gauche à droite : méthode HSRNet [35], la méthode MDSSCGAN SAM [26], la méthode proposée et l'image de référence. La première ligne correspond à l'image RGB, la seconde ligne à la carte SAM et la dernière ligne est une partie zommée. Un pixel bleu indique aucune distorsion spectrale, un pixel jaune indique une forte distorsion.

Les résultats visuels obtenus avec la base de données Pléiades sont présentés en Figures 6.33, 6.34 et 6.35. Pour chacune des figures, un pixel bleu foncé indique qu'il y a aucune distorsion ou différence entre l'image reconstruite et l'image de référence alors qu'un pixel jaune indique une forte distorsion ou différence.

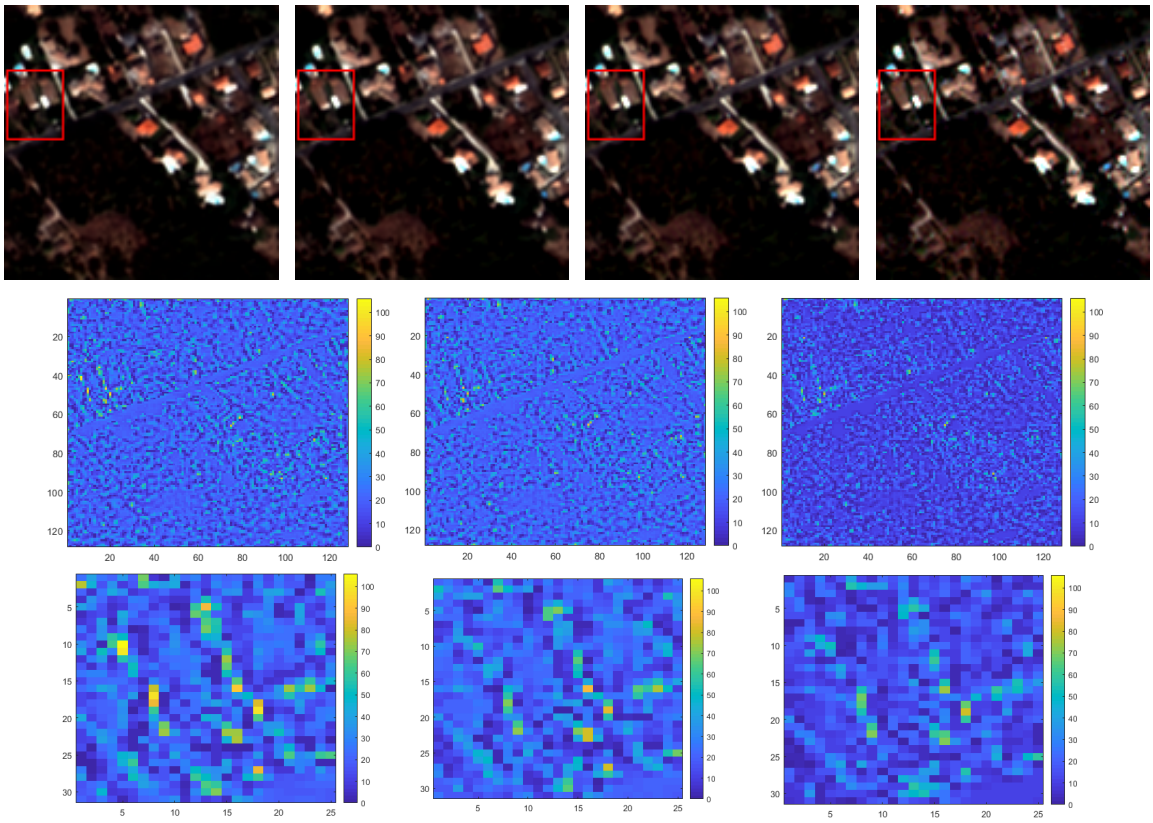
De manière plus détaillée, lorsque l'on regarde la différence pixel à pixel entre l'image reconstruite et l'image de référence (Figure 6.33), on remarque que toutes les méthodes donnent de bons résultats sur l'image globale. Cependant, on peut voir une légère différence dans la partie zoomée, où la méthode proposée présente moins de différence entre les pixels de l'image de référence et ceux de l'image reconstruite.

Dans un second temps, quand on compare les cartes de distorsions spectrales en Figure 6.34, on peut remarquer que la méthode proposée préserve mieux la richesse spectrale.

Finalement, on peut observer la différence entre les hautes fréquences de l'image reconstruite et celles de l'image de référence en Figure 6.35. On peut faire la même remarque que pour l'image WorldView 3 affichée en Figure 6.32. En effet, l'image différence est d'un bleu foncé sur l'ensemble de l'image, ce qui indique une amélioration globale des hautes fréquences.

Sur l'ensemble des deux bases, les résultats sont visuellement assez concordants. En effet, on peut noter une amélioration, plus ou moins marquée, sur l'ensemble des exemples visuels affichés. Les résultats quantitatifs confirment l'amélioration visuelle. En revanche, dans le cas de la base de données Pléiades, les résultats quantitatifs sont légèrement plus favorables à notre méthode

précédente (MDSSCGAN SAM [26]). Néanmoins, notre méthode basée attention est très proche quantitativement et meilleure qualitativement.



HSRNet [35]
$PSNR_h = 21.78$

MDSSCGAN SAM [26]
$PSNR_h = 21.69$

Méthode proposée
$PSNR_h = 25.91$

FIGURE 6.35 – Résultats obtenus sur une image Pléiades. De gauche à droite : méthode HSRNet [35], la méthode MDSSCGAN SAM [26], la méthode proposée et l'image de référence. La première ligne correspond à l'image RGB, la seconde ligne à la différence entre les hautes fréquences de l'image de référence et celles de l'image reconstruite et la dernière ligne est une partie zommée. Le bleu indique aucune différence et le jaune une grande différence de valeur entre les pixels.

6.5 Conclusion

En conclusion de ce chapitre, nous avons étudié l'influence des mécanismes d'attention dans notre architecture multi-discriminateurs. Pour cela, nous avons considéré deux types de mécanismes : un mécanisme d'attention spatiale et un mécanisme d'attention spectrale. La combinaison des deux a pour objectif de focaliser l'attention du générateur sur les parties les plus pertinentes de l'image à reconstruire dans le but d'améliorer la précision de la reconstruction.

Nos expérimentations sur les bases de données Pléiades et WorldView 3 donnent des résultats qui ne sont pas complètement concordants. En effet, le gain quantitatif obtenu sur la base de données WorldView 3 est confirmé visuellement. En revanche, les résultats quantitatifs sur la base Pléiades sont légèrement plus favorable à notre méthode précédente. Cependant, la méthode proposée basée attention est proche quantitativement et semble meilleure visuellement.

CHAPITRE 7

Conclusion : bilan et perspectives

7.1 Bilan

Dans cette thèse, nous nous sommes intéressés à la résolution du problème de pansharpening en utilisant des méthodes d'apprentissage. Les deux principaux enjeux du problème de pansharpening étant d'obtenir une reconstruction permettant de conserver le meilleur des résolutions spatiale et spectrale des données d'entrée, nous avons proposé majoritairement des méthodes basées GAN répondant à ce double enjeu.

La première méthode proposée [25] répond principalement au problème de la reconstruction des hautes fréquences en considérant une fonction de perte basée sur les gradients des images et donc prenant en compte la géométrie présente dans les images. Cette première contribution permet de confirmer le gain de la régularisation dans la fonction de perte du réseau convolutif. Elle prouve également l'intérêt de prendre en compte la modélisation du problème afin d'améliorer les résultats.

Suite à ces premiers résultats, notre seconde contribution [26] est une extension de la première en se concentrant sur la reconstruction conjointe des résolutions spatiale et spectrale. En effet, cette méthode considère deux discriminateurs, chacun répondant à l'un des deux enjeux cités précédemment. Ces discriminateurs utilisent l'information de manière judicieuse en exploitant les bandes adéquates pour la tâche attribuée. Le discriminateur spectral prend en compte les composantes chromatiques Cb et Cr. De son côté, afin d'améliorer la reconstruction des textures et de la géométrie, le discriminateur spatial prend en compte la luminance et la bande infra-rouge. De plus, pour équilibrer la balance entre la reconstruction spatiale et la reconstruction spectrale, nous considérons un terme de régularisation basé sur une métrique spectrale très utilisée pour le problème de pansharpening. Cette méthode répond donc aux deux enjeux du problème de pansharpening, c'est-à-dire une reconstruction disposant d'une composante spatiale aussi détaillée que celle de la bande panchromatique et un contenu spectral aussi fidèle que possible aux bandes multispectrales d'origine. De plus, cette méthode conforte notre idée d'intégrer la modélisation du problème à notre réseau de neurones convolutif.

Enfin, la dernière méthode proposée profite de l'influence des mécanismes d'attention spatiale et spectrale. Dans la littérature, ces mécanismes sont utilisés pour focaliser l'attention du réseau sur les zones d'intérêts de l'image pour la reconstruction. Nous avons donc considéré un mécanisme d'attention spatiale et un mécanisme d'attention spectrale afin de renforcer la précision de la reconstruction des résolutions spatiale et spectrale. Les résultats obtenus sur la base WorldView 3

sont encourageants car l'on peut observer un gain quantitatif confirmé par les résultats visuels. Les résultats sur la seconde base de données sont un peu plus mitigés. En effet, les résultats quantitatifs sont légèrement plus favorables à notre méthode précédente [26] mais notre méthode basée attention est reste proche quantitativement et meilleure qualitativement.

7.2 Perspectives

Bien que cette thèse apporte des réponses au problème de pansharpening selon plusieurs angles d'approche, beaucoup de possibilités n'ont pas été explorées. Plusieurs perspectives peuvent donc être envisagées.

• Une première perspective pourrait être de considérer une méthode basée sur un flux de normalisation. En effet, tout au long de cette thèse, nous nous sommes intéressés aux GANs car ils ont une très bonne capacité à estimer une distribution de données complexe. En revanche, les GANs sont des algorithmes assez instables et peuvent avoir du mal à converger. Cela s'explique par le fait que l'un des réseaux peut prendre le dessus par rapport à l'autre. Ainsi, l'entraînement deviendra instable car le réseau le moins performant ne pourra pas "progresser" et le point d'équilibre que l'on cherche entre les deux réseaux ne sera pas atteint. Le gradient lié à la fonction de perte du réseau qui ne progresse plus va donc petit à petit s'annuler [40]. Dans cette thèse, nous tentons de résoudre ce problème en utilisant une architecture de type dense résiduelle pour le générateur. Cependant, il existe une nouvelle famille d'approches utilisant un flux de normalisation pour faire face à ce problème. En effet, les flux de normalisation sont des modèles génératifs qui produisent des distributions traitables, c'est-à-dire où l'échantillonnage et l'évaluation de la densité peuvent être efficaces et exacts [39].

• Une seconde perspective serait de s'affranchir du protocole de Wald [72] pour l'entraînement d'un réseau de neurones. Bien que le protocole de Wald offre une façon pratique d'utiliser l'image multispectrale en tant qu'image de référence et de "créer" les images panchromatiques et multispectrales en sous-échantillonnant les données d'origines, cela peut conduire à une mauvaise fusion. En effet, le principal problème de ce protocole est que la fusion est évaluée à une échelle différente de celle de l'application. Et selon la résolution d'origine et des détails du paysage, la reconstruction de détails fins peut être problématique à l'échelle d'origine [3, 84].

De plus, plusieurs métriques sans image de référence ont été proposées dans la littérature pour le problème de pansharpening. Par exemple, Alparone *et al.* [5] ont proposé une mesure qui repose sur le concept d'information mutuelle qui représente la mesure d'entropie entre deux sources d'informations A et B en mesurant la redondance d'information entre ces sources. Une autre métrique sans référence a été proposée par Alparone *et al.* [7]. Celle-ci a pour objectif de satisfaire deux propriétés :

— la qualité spectrale des données fusionnées signifie que les relations de similarité entre n'importe quel couple de bandes sont inchangées après la fusion,
— la qualité spatiale indique dans quelle mesure les relations entre chaque bande MS et l'image PAN sont préservées après la fusion.

Ce type de métriques pourrait être exploité dans un cadre d'apprentissage, notamment dans la fonction de perte afin d'éviter la dégradation des données d'origines. De même, des contraintes géométriques dépendantes de l'image panchromatique et des contraintes spectrales basées sur l'image multispectrale basse résolution peuvent être envisagées.

• D'un point de vue applicatif, nous avons choisi la fusion d'images panchromatiques et multispectrales. Hors, beaucoup d'applications nécessitent des données hyperspectrales haute résolution. Une autre perspective serait alors d'utiliser le travail que nous avons effectué afin de fusionner des données hyperspectrales avec des données panchromatiques [52]. Cela imposerait un facteur de

résolution plus grand. Pour gérer de grand facteur de résolution, quelques méthodes basées sur les réseaux de neurones multi-échelles se sont montrées performantes [78, 82].

Une autre éventualité serait une fusion d'images multi-modales, multi-échelles et multi-temporelles comme proposé par Benedetti *et al.* [11] par exemple. Cependant, ce problème est plus complexe car en plus de la fusion des images, un recalage, préalable ou intégré au réseau, est nécessaire.

Dans ces deux cas, la modélisation du problème est différente du problème que l'on a traité mais les enjeux sont similaires : la reconstruction des résolutions spatiale et spectrale. Nous pourrions donc utiliser les connaissances acquises dans cette thèse pour répondre à ces problèmes.

Les perspectives citées ci-dessus ne sont que des exemples. En effet, les possibilités liées à la problématique sont nombreuses et variées.

Bibliographie

[1] P. Addesso, M. Dalla Mura, L. Condat, R. Restaino, G. Vivone, D. Picone, and J. Chanussot. Hyperspectral Pansharpening using Convex Optimization and Collaborative TV. *Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, 2016.

[2] P. Addesso, M. Dalla Mura, L. Condat, R. Restaino, G. Vivone, D. Picone, and J. Chanussot. Collaborative Total Variation for Hyperspetral Pansharpening. *IEEE International Geoscience and Remote Sensing Symposium*, 2017.

[3] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva. MTF-tailored multiscale fusion of high-resolution MS and PAN imagery. *Photogrammetric Engineering & Remote Sensing, vol. 72, no. 5, pp. 591-596*, 2004.

[4] B. Aiazzi, S. Baronti, M. Selva, and L. Alparone. Enhanced Gram-Schmidt spectral sharpening based on multivariate regression of MS and PAN data. *IEEE Inter. Conf. on Geosci. and Remote Sens. Symposium*, 2006.

[5] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, and F. Nencini. A New Method for MS + Pan Image Fusion Assessment Without Reference. *IEEE International Symposium on Geoscience and Remote Sensing, pp. 3802-3805*, 2006.

[6] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva. Multispectral and Panchromatic Data Fusion Assessment Without Reference. *ASPRS Journal of Photogrammetric Engineering and Remote Sensing, vol. 74, no. 2, pp. 193-200*, 2008.

[7] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva. Multispectral and Panchromatic Data Fusion Assessment Without Reference. *Photogrammetric Engineering & Remote Sensing, vol. 8, no. 2 pp. 193-200*, 2008.

[8] G. Aubert, J.-F. Aujol, and L. Blanc-Feraud. Detecting codimension-two objects in an image with Ginzburg-Landau models. *International Journal of Computer Vision*, 2005.

[9] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR*, 2015.

[10] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé. A Variational Model for P+XS Image Fusion. *IJCV, vol. 69, no. 1, pp 43-59*, 2006.

[11] P. Benedetti, D. Ienco, R. Gaetano, K. Ose, R. G. Pensa, and S. Dupuy. <inline-formula><tex-math notation="LaTeX">M^3Fusion</tex-math></inline-formula> : A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 12, pp. 4939-4949*, 2018.

[12] L. Bungert, D. Coomes, M. Ehrhardt, J. Rasch, R. Reisenhofer, and Schönlieb ; C.-B. Blind Image Fusion for Hyperspectral imaging with directional TV. *Inverse Problems, vol. 34, no. 4*, 2018.

[13] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on communications, vol. 31, No. 4, pp.532-540*, 1983.

[14] S. Butterworth. On the Theory of Filter Amplifiers. *Experimental Wireless and the Wireless Engineer, vol. 7, pp. 536–541*, 1930.

[15] E.J. Candès and D. L. Donoho. Curvelets - A Surprisingly Effective Nonadaptive Representation For Objects with Edges. *Curves and Surfaces*, 2000.

[16] W. Carper, T. Lillesand, and R. Kiefer. The use of Intensity-Hue-Saturation transformations for merging SPOT panchromatic and multispectral image data. *Photogramm Eng Remote Sensing, vol. 56, No. 4, pp. 459-467*, 1990.

[17] P.S. Chavez, S.C. Sides, and J.A. Anderson. Comparison of three different methods to merge multiresolution and multispectral data : Landsat TM and SPOT Panchromatic. *Photogrammetric Engineering and Remote Sensing, vol. 57, no. 3 pp. 265-303*, 1991.

[18] A.L. Choodarathnakara, T. Ashok Kumar, S. Koliwad, and C.G. Patil. Mixed Pixels : A Challenge in Remote Sensing Data Classification for Improving Performance. *International Journal of Advanced Research in Computer Engineering and Technology*, 2012.

[19] T Dai, J. Cai, Y. Zhang, S.T. Xia, and L. Zhang. Second-order Attention Network for Single Image Super-Resolution. *Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Centre Canadien de Télédétection. Tutoriels : Notions Fondamentales de Télédétection. *https : //www.nrcan.gc.ca/sites/www.nrcan.gc.ca/files/earthsciences/pdf/resource/tutor/fundam/pdf/fundamentals_f.pdf*.

[21] C. Dong, C. Loy, K. He, and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on pattern analysis ans machine intelligence*, 2016.

[22] J. Duran, A. Buades, B. Coll, and C. Sbert. A non local variational model for pansharpening image fusion. *SIAM, vol. 7, no. 2, pp. 761-796*, 2015.

[23] J. Duran, A. Buades, B. Coll, C. Sbert, and G. Blanchet. A Survey of Pansharpening Methods with A New Band-Decoupled Variational Model. *ISPRS Journal of Photogrammetry and Remote Sensing, vol. 125, pp. 78-105*, 2016.

[24] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain. Fusion d'images préservant la géométrie. *GRETSI*, 2019.

[25] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain. A Residual Dense Generative Adversarial Network for Pansharpening with Geometrical Constraints. *IEEE International Conference on Image Processing*, 2020.

[26] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain. Generative Adversarial Network for Pansharpening with Spectral and Spatial Discriminators. *IEEE Transactions on Geosciences and Remote Sensing*, 2021.

[27] A. R. Gillepsie, A. B. Kahle, and R. E. Walker. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment, vol. 22, pp. 343-365*, 1987.

[28] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of Machine Learning Research, vol. 9, pp.249-256*, 2010.

[29] M. Gonzalez-Audicana, J.L. Saleta, R. Garcia Catalan, and R. Garcia. Fusion of Multispectral and Panchromatic Images Using Improved IHS and PCA Mergers based on Wavelet Decomposition. *IEEE TGRS, vol. 42, No. 6, pp.1291-1299*, 2004.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. *Neural Information Processing Systems Proceedings*, 2014.

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. *NeurIPS Conference, pp. 5767-5777*, 2017.

[32] Y. Guo, F. Ye, and H. Gong. Learning an Efficient Convolution Neural Network for Pansharpening. *Algorithms, vol. 12, pp. 16*, 2019.

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[34] X. He, L. Condat, J. Chanussot, and J. Xia. Pansharpening using Total Variation Regularization. *IEEE International Geoscience and Remote Sensing Symposium*, 2012.

[35] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot. Hyperspectral Image Super-Resolution via Deep Spatiospectral Attention Convolutional Neural Networks. *IEEE Transactions on Neural Networks and Learing Systems, pp. 1-15*, 2021.

[36] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberer. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[37] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang. A New Pan-Sharpening method with Deep Neural Networks. *IEEE Geoscience end remote sensing letters, vol. 12, no 5*, May 2015.

[38] D. P. Kingma and J. Ba. ADAM : A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[39] I. Kobyzev, S. Prince, and M. Brubaker. Normalizing Flows : An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[40] N. Kodali, J. Abernethy, J. Hays, and Z. Kira. On Convergence and Stability of GANs. *arXiv : Artificial Intelligence*, 2018.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25, NIPS*, 2012.

[42] C. Laben and B. Brower. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. *US Patent US6011875A*, 2000.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324*, 1998.

[44] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Genrative Adversarial Network. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[45] O.-Y. Lee, Y.-H. Shin, and J.-O. Kim. Multi-Perspective Discriminators-Based Generative Adversarial Network for Image Super Resolution. *IEEE*, 2019.

[46] D. Lei, H. Chen, L. Zhang, and W. Li. NLRNet : An Efficient Nonlocal Attention ResNet for Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, march/april 2021.

[47] X. Li, F. Xu, X. Lyu, Y. Tong, Z. Chen, S. Li, and D. Liu. A Remote-Sensing Image Pan-Sharpening Method based on Multi-Scale Channel Attention Residual Network. *IEEE Access,vol. 8, pp. 27163-27177*, 2020.

[48] J.G. Liu. Smoothing Filter-based Intensity Modulation : a spectral preserve image fusion technique for improving spatial details. *Remote Sensing, vol. 21, no. 18, pp. 3461-3472*, 2000.

[49] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang. PSGAN : A Generative Adversarial Network for Remote Sensing Image Pan-sharpening. *IEEE Transactions on Geoscience and Remote Sensing, vol. 14, no. 8*, 2020.

[50] X. Liu, Y. Wang, and Q. Liu. PSGAN : A Generative Adversarial Network for Remote Sensing Image Pan-sharpening. *IEEE International Conference on Image Processing*, 2018.

[51] Y. Liu, Y. Wang, N. Li, X. Cheng, Y. Zhang, Y. Huang, and G. Lu. An Attention-Based Approach for Single Image Super Resolution. *24th International Conference on Pattern Recognition (ICPR)*, 2018.

[52] L. Loncan. Fusion Hyperspectrale et Panchromatique avec des hautes résolutions spatiales. *Thèse*, 2016.

[53] L. Loncan, S. Fabre, L. Almeida, J. Bioucas-Dias, W. Liao, G. Briottet, X. Licciardi, J. Chanussot, M. Simoes, N. Dobigeon, J.-Y. Tourneret, M. Vegazones, Q. Wei, V. Gemine, and N. Yokota. Hyperspectral Pansharpening : A Review. *IEEE Goescience and Remote Sensing magazine*, Sept. 2015.

[54] J. Ma, W. Yu, C. Chen, X. Liang, P. Guo, and J. Jiang. Pan-GAN : An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion, no. 62, pp. 110-120*, 2020.

[55] G. Masi, D. Cozzolino, L. Verdolina, and G. Scarpa. Pansharpening by Convolutional Neural Network. *Remote Sensing, vol. 8, no. 7, pp. 594-616*, 2016.

[56] M. Moeller, T. Wittman, and A. Bertozzi. Variational Wavelet Pan-Sharpening. , 2008.

[57] M. Moeller, T. Wittman, and A. Bertozzi. A Variational Approach to Hyperspectral image Fusion. *Proceedings Volume 7334, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, 2009.

[58] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer, 2004.

[59] F. Palsson, J. Sveinsson, M. Ulfarsson, and J. Benediktsson. A New Pansharpening Method Using an Explicit Image Formation Model Regularized via Total Variation. *IEEE International Geoscience and Remote Sensing Symposium*, 2012.

[60] J. Park, D. Han, and H. Ko. Adaptive Weighted Multi-Discriminator CycleGAN for Underwater Image Enhancement. *Journal of MArine Science and Engineering*, 2019.

[61] S. Rahmani, M. Strait, M. Merkurjev, M. Moeller, and T. Wittman. An adaptive IHS Pan-Sharpening method. *IEEE Geoscience ans remote sensing letters, vol 7, no 4*, october 2010.

[62] R. Restaino, M. Dalla Mura, G. Vivone, and J. Chanussot. Context-adaptive Pansharpening Based on Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2*, 2017.

[63] R Restaino, G. Vivone, J. Chanussot, and M. Dalla Mura. Fusion of Multispectral and Panchroamtic Images Based on Morphological Operators. *IEEE Transactions on Image Processing, Institute of Electrical and Electronics Engineers*, 2016.

[64] R.A. Schowengerdt. *Remote Sensing : Models and Methods for Image Processing, 3rd edition.* elsevier, 2006.

[65] V. Shah, N. Younan, and R. King. An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets. *IEEE TGRS*, 2008.

[66] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is All You Need. *Advances in Neural Inforamtion Processing Systems 30 (NIPS)*, 2017.

[68] G. Vivone, J. Chanussot, and R. Restaino. A Regression-Based High-Pass Modulation Pansharpening Approach. *IEEE Transactions on Geoscience and Remote Sensing, vol.56, no. 2*, 2018.

[69] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, M. Scarpa, G. Ulfarsson, L. Alparone, and J. Chanussot. A New Benchmark Based on Recent Advances in Multispectral Pansharpening. *IEEE Geoscience and Remote Sensing Magazine*, 2021.

[70] G. Vivone, R. Restaino, and J. Chanussot. Full Scale Regression-based Injection Coefficients for Panchromatic Sharpening. *IEEE Transactions on Image Processing, vol. 27, no. 7*, 2018.

[71] L. Wald. Data Fusion : Definitions and Architectures - Fusion of images of different spatial resolutions. *Les Presses de l'Ecole des Mines*, 2002.

[72] L. Wald, T. Ranchin, and M. Mangolini. Fusion of satellite images of different spatial resolution : Assessing the quality of resulting images. *Photogramm. Eng. Remote Sens., vol. 63, no. 6*, 1997.

[73] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local Neural Networks. *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, pp.7794-7803*, 2018.

[74] X. Wang, K. Yu, C. Dong, and C.-C. Loy. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. *IEEE Conference on Computer Vision Pattern Recognition, pp. 606-615*, 2018.

[75] Q. Wei, J.M. Boucas Dias, N. Dobigeon, and J.Y. Tourneret. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 7, pp. 3658-3668*, 2015.

[76] Y. Wei, H. Yuan, H. Shen, and L. Zhang. Boosting the Accuracy of Multispectral Image Pansharpening by Learning a Deep Residual Network. *IEEE GRSL, pp. 99*, 2017.

[77] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon. CBAM : Convolutional Block Attention Module. *ECCV*, 2018.

[78] H. Wu, J. Zou, Z. Gui, W.J. Zeng, J. Ye, J. Zhang, H. Liu, and Z. Wei. Multi-grained Attention Networks for Single Image Super-Resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[79] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley. PanNet : A Deep Network Architecture for Pan-Sharpening. *IEEE International Conference on Computer Vision*, October 2017.

[80] X. Yang, X. Li, Z. Li, and D. Zhou. Image super-resolution based on deep neural network of multiple attention mechanism. *Journal of Visual Communication and Image Recognition*, 2021.

[81] D.A. Yocky. Multiresolution Wavelet Decomposation Images merger of Landsat Thematic Mapper ans SPOT Panchromatic Data. *Photogrammetric Engineering and Remote Sensing, vol. 62, no. 9*, September 1996.

[82] D. Zhang, J. Shao, Z. Liang, L. Gao, and H.T. Shen. Large Factor Image Super-Resolution with Cascaded Convolutional Neural Networks. *IEEE Transactions on Multimedia*, 2020.

[83] W. Zhang, J. Li, and Z. Hua. Attention-based Tri-UNet for Remote Sensing Image Pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 3719-3732*, March 2021.

[84] Y. Zhang. Understanding image fusion. *Photogrammetric Engineering & Remote Sensing, vol. 70, no. 6, pp. 657-661*, 2004.

[85] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image Super-Resolution Using Very Deep Residual Attention Networks. *ECCV*, 2018.

[86] Y. Zhang, X. Li, and J. Zhou. SFTGAN : a generative adversarial network for pan-sharpening equipped with spatial feature transform layers. *Journal of Applied Remote Sensing, vol.13, no. 2*, 2019.

[87] Y. Zheng, J. Li, Y. Li, J. Guo, X Wu, and J. Chanussot. Hyperspectral Pansharpening Using Deep Prior and Dual Attention Residual Network. *IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 8059-8076*, 2020.

[88] H. Zhou, Q. Liu, and Y. Wang. PGMAN : An Unsupervised Generative Multi-adversarial Network for Pan-sharpening. *in review process*, 2020 sur arxiv.

[89] L. Zhou, X. Luo, J. Yin, and X. Shi. Spectral Diversity Enhancement for Pansharpening. *25th IEEE International Conference on Image Processing*, 2018.

[90] X. Zhu, Y. Cheng, J. Peng, M. Wang, R. Le, and X. Liu. Super-Resolved Image Peceptual Quality Improvement via Multi-Feature Discrimiantors. *CoRR, abs/1904.10654*, 2019.