الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي
جامعة وهران للعلوم و التكنولوجيا محمد بوضياف

THÈSE

En vue de l'obtention du Diplôme de Doctorat

Présentée par :

Kawther Aarizou

Intitulé

Reconstruction d'images en super-résolution basée sur des méthodes d'apprentissage artificiel

Faculté	***: Génie Électrique***
Département	***:Électronique***
Domaine	***: Sciences et Technologies***
Filière	***: Génie Électrique***
Intitulé de la Forma-	***:Systèmes Intelligents et Robotiques***

Devant le Jury Composé de :

Membres de Jury	Grade	Qualité	Domiciliation
OUAMRI Abdelaziz	Pr	Président	USTO-MB
LOUKIL Abdelhamid	Pr	Encadrant	USTO-MB
ZOUAGUI Tarik	Pr	Examinateur	USTO-MB
KECHE Mokhtar	Pr	Examinateur	USTO-MB
MERAH Mostfa	Pr	Examinateur	UMAB. Mostaganem
ZIGH Ehlem	Pr	Examinateur	INTTIC. Oran
HAMADA Aïcha		Invité(e)	USTO-MB

Année Universitaire : 2021/2022

إعادة بناء الصور فائقة الاستبانة إستنادا على أساليب التعلم الاصطناعي

ملخص

تستند أساليب إعادة بناء صور فائقة الاستبانة القائمة على التعلم الآلي إلى تطوير شبكة عصبية اصطناعية قادرة على تحويل صورة منخفضة الاستبانة إلى صورة عالية الاستبانة أكثر تفصيلا. وخلال تعلمها تتعرض الشبكة لمجموعة كبيرة من البيانات تتألف من عدة آلاف من الصور ، وعلى وجه التحديد ، ثنائيات من صور ذات استبانة عالية ومنخفضة حيث ترتبط كل صورة متدهورة بنسختها ذات النوعية العالية للإشراف على عملية التعلم.

وتشمل قواعد بيانات التدريب المكرسة لإعادة البناء فائقة الاستبانة صورا طبيعية وصورا فوتوغرافية يفترض أن تمثل أكبر عدد ممكن من الهياكل والمعلومات الحقيقية و على الرغم من العدد الكبير من الأعمال التي أُجريت مؤخراً في مجال إعادة بناء إعادة البناء الصور فائقة الاستبانة بواسطة شبكات معقدة ، فإن تعلم بارامترات الشبكة لا يزال يسترشد بمقاييس الأداء الكمية التي لا تشمل معرفة معدات الكاميرا أو المشهد الفعلي. وتستخدم هذه الشبكات مجموعات كبيرة من البيانات أثناء التعلم لمعالجة سوء تكييف إعادة بناء الصور فائقة الاستبانة. بيد أن ذلك يخلق بسرعة مفاضلة لأن مجموعات البيانات هذه لا تصف الحالات الفعلية على وجه التحديد. ويتطلب هذا النوع من التعلم معرفة مسبقة بنموذج التدهور لأنه يشارك مباشرة في توليد بيانات التعلم. ثم تتعلم الشبكة إستنادا على هذا النموذج من البيانات لإعادة بناء الصورة عالية الدقة من الصورة منخفضة الدقة ، ثم تضيف التفاصيل البصرية تدريجياً (البكسلات) لتحقيق النتيجة المرجوة.

وتسعى البحوث في هذا المجال إلى بناء تمثيلات أفضل للحالة الحقيقية وإيجاد نماذج قادرة على تعلم هذه التمثيلات من البيانات الملاحظة المتاحة .والهدف من هذه الأطروحة هو تطوير أساليب لدمج مسبق في الشبكات العصبية العميقة ، مع التركيز بشكل خاص على عامل تدهور الصورة الحقيقي الناتج عن أداة الكاميرا .وتستند مساهمتنا الأولى إلى آلية لاستغلال التشابه الذاتي للصورة في عملية تعلم آلة الشبكات العصبية .وقد تأكدت عدة ملاحظات لصالح قواعد البيانات الداخلية في هذا العمل ، وربما ملاحظات أخرى تتعلق بالتحقق الكمي والنوعي للصور الطبيعية .وفي هذا السياق نفسه ، تقترح مساهمتنا الثانية خوارزمية تقدير عشوائي لبارامترات التدهور التي تضع في الاعتبار الطابع غير المهيكل لإعادة بناء الصور فائقة الاستبانة.

Abstract

Super-resolution reconstruction methods based on deep learning are built on an artificial neural network capable of transforming a low resolution image into a more detailed high resolution image. During its training, the network is exposed to a large dataset comprising several thousand images, more precisely, combinations of high and low resolution images where each degraded image is associated with its better quality counterpart in order to supervise the learning process. The training databases dedicated to reconstruction (i.e., SR) include natural and photographic images that are supposed to provide as much structure and real-world information as possible.

Despite the large amount of recent work in the area of SR reconstruction by deep convolutional networks CNN, the learning of network parameters is still guided by quantitative performance measures that do not include the acquisition hardware or the prior real scene. These networks use large data sets during training to compensate for the poor conditioning of the SR reconstruction. However, this quickly creates a trade-off as these datasets do not particularly describe the real cases. This type of learning requires prior knowledge of the degradation model due to the fact that it is directly involved in generating the training data. The network then learns to reverse the degradation model to reconstruct the high-resolution image starting from the low-resolution image and gradually adding visual details (pixels) to achieve the desired result.

Research in this area focuses on building better representations of the real case and creating models that can learn these representations from the available observed data. The objective of this thesis is to develop methods for incorporating a priori in deep neural networks, with a particular focus on the actual blur factor generated by the acquisition device. Our first contribution is based on a mechanism allowing to exploit the self-similarity of an image in the deep neural network's automatic learning procedure. Several observations in support of internal databases have been validated in this work, eventually leading to further observations related to

the quantitative and qualitative validation of natural images. In this regard, our second contribution proposes an algorithm for stochastic estimation of degradation parameters that effectively conditions the ill-posedness of the SR reconstruction.

Résumé

Les méthodes de reconstruction en Super-Résolution basées sur l'apprentissage automatique reposent sur l'élaboration d'un réseau neuronal artificiel capable de transformer une image à faible résolution en une autre plus détaillée de haute résolution. Durant son apprentissage le réseau est exposé à un large jeu de donnée de plusieurs milliers d'images, plus précisément, de combinaisons d'images de haute et faible résolution où chaque image dégradée est associée à sa version de meilleure qualité afin de superviser le processus d'apprentissage. Les bases de données d'entraînement dédiées à la reconstruction SR (i.e. Super Résolution) comprennent des images naturelles et photographiques censés représenter un maximum de structures et d'information réelles.
Malgré le grand nombre de travaux récents dans le domaine de la reconstruction SR par réseaux convolutifs, profonds, l'apprentissage des paramètres du réseau reste guidé par des mesures de performance quantitatives n'incluant pas la connaissance du matériel d'acquisition ni de la scène réelle. Ces réseaux ont recours à de larges ensembles de données durant l'apprentissage afin de palier le mauvais conditionnement de la reconstruction SR. Néanmoins, ceci crée rapidement un compromis dans la mesure où ces ensembles de données ne décrivent pas particulièrement les cas réels. Ce type d'apprentissage requière une connaissance préalable du modèle de dégradation car il intervient directement dans la génération des données d'apprentissage. Le réseau apprend alors à inverser le modèle de dégradation pour reconstruire l'image de haute résolution en partant de l'image de basse résolution et en rajoutant progressivement des détails visuel (pixels) pour atteindre le résultat souhaité.

Les recherches dans ce domaine s'efforcent de construire de meilleures représentations du cas réel et de créer des modèles capables d'apprendre ces représentations à partir des données disponibles observées. L'objectif de cette thèse est de développer des méthodes permettant d'intégrer des a priori dans les réseaux de neurones profonds, en ciblant particulièrement le facteur de flou réel engendré par l'appareil

d'acquisition. Notre première contribution repose sur un mécanisme d'exploitation de l'autosimilarité d'une image dans la procédure de l'apprentissage automatique du réseau neuronal. Plusieurs observation en faveurs des base de données internes ont été validées dans ce travail, éventuellement d'autres qui concernent la validation quantitative et qualitative des images naturelles. Dans ce même contexte, notre deuxième contribution propose un algorithme d'estimation stochastique des paramètres de dégradation qui conditionnement efficacement le caractère mal-posé de la reconstruction SR.

Remerciements

Je tiens à remercier mon directeur de thèse, Professeur Abdelhamid LOUKIL, pour avoir accepté de diriger cette thèse, pour son encadrement, son suivi et ses conseils tout au long de ces années.
Je remercie également M. Abdelaziz OUAMRI, Professeur à l'Université des Sciences et de la Technologie d'Oran, pour avoir accepté de présider mon jury de thèse. Je le remercie également pour avoir accepté d'examiner cette thèse et de participer au jury. Je remercie vivement Mme. Ehlem ZIGH, Proefesseur à l'Institut National des Télécommunications et des TIC d'Oran, M. Tarik ZOUAGUI, Professeur à l'Université des Sciences et de la Technologie d'Oran, M. Mostefa MERAH, Professeur à l'Université de Mostaganem, et M. Mokhtar KECHE, Professeur à l'Université des Sciences et de la Technologie d'Oran pour avoir accepté d'être examinateurs dans mon jury de thèse. Je tiens également à remercie Mme Aïcha Hamada pour l'intérêt qu'elle a manifesté en participant en qualité de membre invité à ce jury.
Enfin, Pr. Z. Ahmed Foitih, responsable du Laboratoire d'Électronique des Puissances, Énergie Solaire et Automatique (LEPESA), retrouvera toute ma gratitude pour m'avoir autorisé à accéder et à travailler dans son laboratoire durant mes années de doctorat.

À mes parents et à mes sœurs . . .

Table des matières

Table des figures

Liste des tableaux

Glossaire

AdaGrad Adaptive Gradient Algorithm 77

AE Auto-Encoder 88

BR Basse-Résolution 1

BSD Berkeley Segmentation Dataset 50

BTV Bilateral Total Variation 26

CNN Convolutional Neural Network 1, 3

CS Compressed Sensing 29, 31

DCT Discrete Cosine Transform 24

DISTS Deep Image Structure and Texture Similarity 100

DIV2K DIVerse 2K 70

DL Deep Learning 3

EDSR Enhanced Deep Super Resolution 40

ELBO Evidence Lower BOund 91

EQM Erreur Quadratique Moyenne 75

ESPCN Efficient Sub-Pixel Convolutional Neural network 11, 43

FRIQUEE Feature maps based Referenceless Image QUality Evaluation Engine 69

GAN Generative Adversarial Network 46, 47

HR Haute-Résolution 1

JPEG Joint Photographic Experts Group 106

KDE Kernel Density Estimation 59

KL Kullback-Leibler 91

LLE Locally Linear Embedding 29

M-VAE Multiple-input Variational Auto-Encoder 93

MISR Multiple Image Super-Resolution 2

MOS Mean Opinion Score 50

MRF Markov Random Field 27

NOREQI NO-REference image Quality Index 69, 82

NR-IQA No Reference-Image Quality Assessment 82

PI Performance Indicator 82

PSF Point Spread Function 5

PSNR Peak Signal to Noise Ratio 44

RMSProp Root Mean Squared Propagation 77

SISR Single Image Super-Resolution 2

SR Super Résolution 1, 4

SRCNN Super-Resolution Convolutional Neural Network 38

SRGAN Super-Resolution Generative Adversarial Network 47

SRResNet Super Resolution Residual Network 11, 43

SSIM Structural SIMilarity 44

TV Total Variation 26

VAE Variational Auto-Encoder 88

VGG Visual Geometry Group 45, 49, 50

Chapitre 1

Introduction Générale

Les images à Haute-Résolution (HR) sont de rigueur dans la plupart des applications d'imagerie électronique. Une haute résolution spatiale signifie que la densité des pixels dans l'image est élevée, par conséquent une image HR offre plus de détails qu'une image à Basse-Résolution (BR). La résolution spatiale de l'image dépend de la qualité des capteurs et par le niveau atteint par la technologie de fabrication de ces derniers. Outre le coût des caméras de haute-résolution et les problèmes d'instrumentation, il arrive que l'on ne puisse pas améliorer la résolution de l'imageur dans des scénarios tels que l'imagerie satellitaire, où les dégradations introduites dans l'image durant l'acquisition sont très importantes pour pouvoir extraire correctement l'information recherchée. Par conséquent, l'utilisation de capteurs à haute-résolution n'est pas possible en raison des contraintes physiques sévères, comme la diffraction et la distorsion atmosphériques. D'autre part, certaines applications requièrent une haute résolution qui dépasse les capacités des capteurs déjà existants, comme il est le cas du scanner médical, ou dans le domaine de l'imagerie échographique ou microscopique. Une solution à ce problème consiste à accepter les limitations du capteur, et d'utiliser des techniques de traitement du signal ou de l'image afin de traiter ou d'éliminer les dégradations et augmenter la résolution sans modifier la constitution ni la technologie des capteurs. Dans le cadre particulier de ces exigences, les algorithmes de Super-Résolution (SR)

peuvent intervenir. Même si les équipements de hautes qualités visuelles sont disponibles, ces algorithmes proposent une alternative à faible coût. Ils permettent non seulement d'améliorer la finesse, mais aussi d'augmenter les performances des traitements postérieurs effectués sur les images selon l'application souhaitée, tels que la détection d'objets, l'identification, la reconnaissance, etc. L'objectif des méthodes de la SR est d'obtenir la nouvelle information vraie perdue durant l'acquisition de l'image, et d'essayer de la faire réapparaître en insérant de nouveaux pixels entre ceux déjà existants. Ces pixels sont calculés à partir d'une ou de plusieurs observations (images) qui représentent des versions dégradées de la scène. Les dégradations sont dues à divers facteurs tels que : le flou produit à cause du mouvement de la caméra ou des objets dans la scène, le flou optique de la lentille de la caméra, la perte de résolution optique, les effets du repliement spectral lors du processus de l'échantillonnage, etc. Cela provoque une perte d'informations importantes telles que les contours et la texture.

1.1 Contexte des travaux

Les méthodes de SR peuvent être scindées en deux grandes catégories : La super-résolution multi-images (i.e. Multiple Image Super-Resolution (MISR)), et la super-résolution à image unique (i.e. Single Image Super-Resolution (SISR)). L'idée de base des méthodes MISR traditionnelles est de combiner les informations non-redondantes contenues dans plusieurs images de basse résolution pour augmenter les composantes à hautes fréquences de l'image à reconstruire. La caméra capture plusieurs images BR de la scène avec des décalages sous-pixélliques. Puisque chaque image représente une version légèrement différente de la scène réelle, il est possible de fusionner les pixels non-redondants en une seule grille éparse irrégulière de haute résolution. Cependant, de nombreuses applications pratiques émergentes ne disposent que d'une seule observation de la scène, ce qui cause un manque important d'informations dans le processus de reconstruction. Dans ce contexte, le principe général des techniques de SISR est d'utiliser des informations externes pour palier le manque des observations. Dans les deux cas, la reconstruction en

SR est décrite comme un problème inverse "mal-posé", car les seules observations expérimentales (unique ou multiples) ne suffisent pas à estimer correctement l'information manquante dans l'image. Il existe plusieurs méthodes algorithmiques de résolution de problèmes inverses mal-posés appliqués à la reconstruction des images. Ces méthodes reposent sur le principe de régularisation qui requière la connaissance *a priori* sur l'information que l'on veut reconstruire. La robustesse de ces méthodes dépend essentiellement de la fiabilité de l'hypothèse de régularité, l'image reconstruite peut être loin de la réalité si cette hypothèse est erronée. Dans ce contexte, les méthodes récentes d'apprentissage artificiel profond DL (pour Deep Learning en anglais) sont les plus robustes en terme de résolution des problèmes inverses mal-posés car elles sont considérées comme des algorithmes d'approximation non-linéaires sous des hypothèses faibles compte tenu de la grande quantité de données d'apprentissage. La taille de ces données et leur diversité permettent à ces algorithmes d'approcher la solution calculée à une réalité généralisée, et par conséquent réduire le caractère mal-conditionné et mal-posé du problème de reconstruction.

Les méthodes par apprentissage artificiel ont pris de l'intérêt dans la reconstruction SR du fait de leur capacité à produire des images de qualité supérieure par rapport aux autres méthodes algorithmiques notamment dans la reconstruction en SISR. Elles sont connues par leurs robustesse à résoudre les problèmes inverses mal-posés. Leurs stratégies consistent à entraîner un réseau de neurones convolutif profond CNN (pour Convolutional Neural Network) à modéliser une relation non-linéaire entre des images BR et leur versions HR en utilisant des données d'apprentissage. De cette manière, le réseau CNN apprend dans un premier temps à extraire, à partir de ces données, l'information qui augmente la résolution spatiale de son image d'entrée, pour ensuite l'utiliser dans l'inversion du modèle de dégradation. En d'autres termes, ces réseaux réussissent à tirer l'information *a priori* nécessaire à la régularisation à partir de plusieurs données externes. Ces données sont des images photographiques de scènes naturelles de haute résolution et leurs versions BR sont synthétisées à l'aide d'un modèle de dégradation connu ou idéal, c'est-à-dire qu'il est constitué de paramètres de bruit et de flou souvent

déterministes ou faciles à estimer. Ce modèle peut être décrit comme l'ensemble des contraintes qui lient le problème mal-posé de reconstruction.

1.2 Élements de probématique

1.2.1 Objectifs de recherche

La spécification de l'ensemble des contraintes dans la génération des données d'entraînement permet de lier le problème inverse de la SR à un ensemble réduit de solutions dans un type d'application précis. Bien que, de cette manière, les méthodes d'apprentissage artificiel réduisent considérablement le caractère mal-posé du problème de reconstruction en le liant à un ensemble réduit de solutions, elles sont fortement limitées par leur dépendance aux opérateurs de dégradation utilisés dans la génération des données d'entraînement. Il est connu que ces paramètres ne sont pas adaptés aux images naturelles, car les paramètres de dégradations inhérents à l'image BR de la scène réelle sont beaucoup plus complexes que ceux simulés en phase d'entraînement. La figure 1.1 montre un tel exemple, où un réseau de reconstruction CNN échoue à traiter l'image BR lorsque son noyau de flou est complexe.

Par ailleurs, les paramètres de dégradations dans le modèle d'image, tels que le flou et le bruit, sont différents d'une image à une autre même si celles-ci représentent la même scène. Les travaux en apprentissage artificiel profond appliqué à la reconstruction SISR se sont traditionnellement concentrés sur des situations générales où les paramètres de dégradation dans les bases de données d'entraînement sont idéaux. Ce recueil de théorie suppose que, indifféremment de la scène d'étude, les paramètres de dégradation simulés dans la phase d'entraînement sont pratiquement les mêmes dans toutes les opérations d'acquisition d'image. Par conséquent, ces travaux de recherche ne sont pas adaptés pour des scénarios réels où le modèle de dégradation du processus d'acquisition est partiellement connu, complexe, ou difficile à estimer. Ce cas particulier de la SR est appelé "*Super-Résolution myope*"

(a) image BR bruitée (b) image super-résolue par le réseau RCAN [1]

FIGURE 1.1 – Exemple d'un cas de défaillance du réseau profond RCAN [1] où le bruit est amplifié causant des artéfacts indésirables, tandis que la résolution n'est pas améliorée.

et s'applique dans des contextes dits "*semi-aveugles*" où l'effet des modèles de dégradation dépend des caractéristiques intrinsèques de la caméra et de la complexité de la scène lors de la capture de l'image. Ainsi l'estimation exacte et précise de ces paramètres est souvent difficile à réaliser car elle constitue, elle-même, un problème inverse mal-posé.

1.2.2 Question et méthodologie de recherche

Les algorithmes de SR doivent être robustes face aux diverses sources de dégradation des images. Ces dégradations comprennent le bruit, les imperfections des imageurs, et les erreurs de modélisation. Nous suivons l'hypothèse que les dégradations qui altèrent l'image sont essentiellement causées par le capteur. Nous pensons que les dégradations atmosphériques sont des paramètres que d'autres techniques de reconstruction d'image peuvent régler. La reconstruction en super-résolution, quant à elle, se charge d'inverser les modèles de dégradations qu'impose l'imageur comme le sous-échantillonnage, la décimation, la PSF (i.e. Point Spread Function), le flou optique de la caméra, le bruit de la caméra, etc.

Dans ce travail, nous proposons de réduire l'utilisation des contraintes généralisées dans la résolution du problème mal-posé et mal-conditionné de la SISR, car nous considérons que la résolution de celui-ci à travers des contraintes généralisées et prédéterminées réduit considérablement le domaine d'application de ces méthodes aux critères imposées durant la régularisation. Nous visons à étudier la modélisation du paramètre de dégradation du système d'acquisition dans un contexte semi-aveugle, et nous considérons que le meilleur moyen de résoudre un tel problème inverse est de le conditionner *selon des contraintes directement liée à la donnée d'entrée (i.e. l'image BR).*

L'importance de développer des techniques DL appliquées à la SISR efficaces dans des scénarii réels nous a amené à formuler les questions suivantes :

- Quels sont les types de contraintes qui conditionnent les problèmes inverses mal-posés de la reconstruction SR, et pourquoi ne sont-elles pas adaptées aux environnements complexes/réels ?
- Comment estimer les paramètres du modèle de dégradation spécifique à l'image notamment dans le cas de la SISR ?
- Comment régulariser la solution du problème mal-posé à l'aide de la contrainte tirée à partir du modèle de dégradation estimé ?

1.2.3 Organisation de la thèse

Afin de répondre aux questions susmentionnées, nous adoptons dans cette thèse une méthodologie de recherche décrite par les chapitres suivants :

— Dans le chapitre 2 nous présentons et détaillons d'abord le formalisme mathématique de la reconstruction en super-résolution, notamment la SISR, comme étant un problème inverse mal-posé, et nous mettons ensuite en évidence les difficultés de résolution de ce type de problème.

— Le chapitre 3 est consacré à l'état de l'art des méthodes de la reconstruction MISR et SISR. Une brève étude de ces méthodes classiques dans la résolution

du problème inverse de la SR implique de manière évidente l'intérêt des techniques DL faisant ainsi l'objet du chapitre suivant.

— Le chapitre 4 est alors consacré à l'étude de ces techniques, ainsi qu'à leur robustesse dans des scénarii à contexte semi-aveugle.

— Après avoir déterminé les limites des méthodes DL dans ce domaine d'application, nous traitons dans le chapitre 5 la question de l'autosimilarité des images naturelles, et de leur rapport au conditionnement du problème de reconstruction mal-posé. Nous présentons ensuite les résultats de notre contribution à l'amélioration du comportement des techniques DL à l'aide de l'autosimilarité. Enfin, l'impact des différents éléments impliqués dans l'apprentissage des modèles DL, comme la fonction de perte utilisée, sont analysés en détails.

— Dans le chapitre 6 nous proposons un algorithme d'estimation du paramètre de flou que nous employons dans la régularisation.

— Enfin, nous clôturons notre thèse par une conclusion générale des chapitres précédents, avec une réflexion sur l'état de l'art de chacun des problèmes traités. Nous proposons ensuite des perspectives de recherche pour une régularisation plus fiable du problème inverse de la reconstruction SR.

Première partie

Contexte général de la Super-Résolution

Chapitre 2

Problème inverse appliqué à la reconstruction en SR

2.1 Introduction

Un grand nombre d'applications en reconstruction d'image et en rehaussement de la qualité se formulent comme des problèmes inverses car les quantités qui nous intéressent ne peuvent en général être observées directement. La résolution d'un problème inverse nécessite une étape initiale de modélisation mathématique du phénomène, dite modèle direct, qui décrit comment les paramètres du modèle se traduisent expérimentalement. Dans le contexte du traitement d'image, le modèle directe fournit une ou plusieurs équations mathématiques reliant les observations à la variable inconnue recherchée. Ensuite, à partir des mesures obtenues sur la scène réelle, les paramètres du modèle sont approximés à travers une résolution numérique et analytique. La résolution du problème en traitement numérique de l'image revient donc à inverser le modèle directe qui modélise le processus d'acquisition. Néanmoins, dans les applications réelles, les observations expérimentales ne suffisent pas à déterminer parfaitement tous les paramètres du modèle. En outre, le problème peut être non-linéaire où la relation entre les observations et les paramètres du modèle sont complexes, c'est-à-dire que la modélisation peut s'approcher des observations mais s'écarter des paramètres réel. Dans ce chapitre, le principe de la reconstruction SR est reformulé dans le formalisme classique des problèmes inverses mal-posés, et son comportement est étudié dans le cas de la reconstruction semi-aveugle.

2.2 Définition du problème

Problème inverse D'après J. B. Keller [11], "*deux problèmes sont dits inverses l'un de l'autre si la formulation de l'un met l'autre en cause*", c'est-à-dire qu'elle implique une solution complète ou partielle de l'autre. Une définition plus intuitive est qu'un problème inverse consiste à déterminer des causes connaissant des effets. À l'inverse, déduire les effets en connaissant les causes est appelé problème direct. Ce dernier, selon Keller, est plus étudié dans la science en raison de sa causalité : nous avons tendance à résoudre des problèmes pour lesquels les causes sont

données, puis chercher les effets. Cette logique se manifeste dans divers exemples d'application comme le calcul de la trajectoire à partir de la connaissance des forces en mécanique générale, le calcul des ondes diffractées en connaissant les points de sources et les obstacles. D'autres exemples sont tirés d'instruments physiques, tels que les dispositifs électroniques et les imageurs. Ici, le problème direct consiste à calculer la sortie de l'instrument (l'image) étant donnée l'entrée (l'objet de la scène) et les caractéristiques de l'instrument. Ainsi, le problème inverse consiste à identifier la scène capturée avec un imageur donné à partir de la connaissance de ses images.

Comme il a été expliqué précédemment, un problème direct est orienté selon une chaîne de causalité ; c'est aussi un problème orienté vers une perte d'information : sa résolution définit la transition d'une quantité physique avec un certain contenu d'information à une autre quantité avec peu d'informations. En général, cela implique que la solution est plus lisse que les données : l'image fournie par un système d'acquisition à bande limitée est plus lisse que l'objet réel correspondant sur scène. Par conséquent, résoudre le problème inverse correspondant revient à accomplir une transformation opposée qui devrait réaliser un gain d'information. Cette difficulté conceptuelle constitue une propriété mathématique typique d'un problème *mal-posé*.

Problème mal-posé Le concept de base des problèmes inverses mal-posés a été introduit par le mathématicien Jacques Hadamard dans un article publié en 1902 sur le problème des valeurs limites des équations différentielles partielles dans leurs interprétations physiques [12]. Dans sa première formulation, Hadamard définit un problème comme étant bien-posé si sa solution et unique et existe pour des données arbitraires. Dans des travaux ultérieurs, Hadamard met l'accent sur l'exigence de la dépendance continue de la solution par rapport aux données, et affirme qu'une solution qui change considérablement en présence d'une légère variation dans les données ne constitue pas une solution au sens physique du terme. Ce dernier critère représente la condition de stabilité de la solution dans un problème inverse. Ainsi,

un problème qui ne satisfait pas une de ces conditions est qualifié de mal-posé. De manière générale, un problème inverse, qui implique une inversion de la séquence cause-effet, est mal posé. Cette affirmation n'a pas de sens que si nous fournissons un cadre mathématique approprié pour la description des problèmes directs et inverses. À cette fin, nous tenons compte du fait que nous considérons principalement des problèmes d'imagerie et nous utilisons donc un lexique approprié à ces problèmes.

2.3 Notation

Définissons d'abord l'ensemble des objets à observer ou à photographier ayant certaines propriétés décrites par des paramètres bien définis. Pour cet ensemble, nous déterminons une mesure de distance ε_{object} afin de décider si deux objets de cet ensemble sont proches ou non. Nous désignons cet espace par $\mathcal{X}$ que nous appellerons espace des objets ou la scène réelle. Il est question maintenant de résoudre d'abord le problème directe, c'est-à-dire, calculer pour chaque objet de l'ensemble $\mathcal{X}$ son image correspondante qu'on appelle *observation* parfaite ou *image-sans-bruit*. Vu que ce problème est direct et bien-posé, l'image sans-bruit est unique est spécifique à son objet correspondant. Comme nous l'avons noté, cette image est considérée comme une version lissée de la scène en raison du fait qu'elle contient moins d'informations que son objet réel. Cependant, cette propriété peut ne pas être vraie pour toutes les images mesurées, également appelées images bruitées, car elles correspondent à des observations parfaites corrompues par un bruit qui affecte le système de mesure. Le troisième point consiste à définir l'ensemble des images de manière à ce qu'il contienne à la fois les images sans-bruit, et les images bruitées ainsi qu'une mesure de distance ε_{image} relative à l'ensemble. Nous désignons cet ensemble par $\mathcal{Y}$ que nous appellerons espace des observations. Enfin, la solution au problème direct est de définir un modèle ou opérateur, noté A, qui transforme tout objet de l'espace $\mathcal{X}$ en une *image-sans-bruit* de l'espace $\mathcal{Y}$. L'ensemble des images sans-bruit qu'engendre A est appelé le domaine de l'opérateur A. Notons

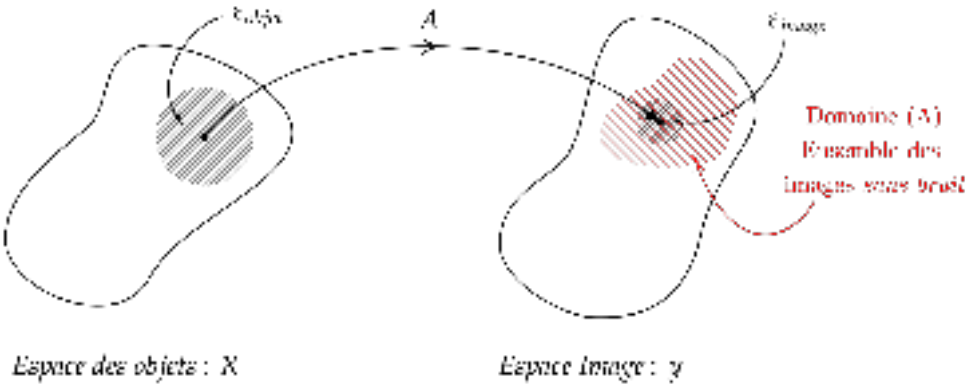


FIGURE 2.1 – Illustration schématique du problème direct bien-posé. Les zones en noir représentent la tolérence de la mesure de distance ; les éléments se trouvant dans cette zone sont considérés comme proches.

également que l'opérateur A est continu, c'est-à-dire que les images de deux objets proches sont également proches, formulant ainsi une propriété importante du problème direct bien-posé

Au moyen de ce schéma, il est possible de décrire la perte d'information typique de la solution du problème directe qui entraîne néanmoins une conséquence importante : il est possible que deux objets très éloignées aient des images très proches, en d'autres termes, il existe des ensembles très larges d'objets distincts tels que leurs ensembles d'images correspondant sont très petits.La figure 2.1 illustre schématiquement les propriétés du problème direct bien-posé.

On considère maintenant le problème inverse, c'est-à-dire la détermination de l'objet correspondant ayant une observation donnée : cette observation peut être parfaite (i.e. appartient au domaine de A) ou bruitée. On constate d'abord que ce problème est mal posé en raison de la perte d'information inhérente à la solution du problème direct. En effet, si nous avons une image correspondante à deux objets distincts, la solution n'est pas unique. Si, de plus, nous avons une image bruitée qui n'est pas dans le domaine de l'opérateur A, alors la solution au problème peut ne pas appartenir à $\mathcal{X}$. Si nous avons deux images voisines telles que leurs objets correspondants sont très éloignés, alors la solution du problème inverse ne dépend pas continuellement des données. Ces propriétés rendent par conséquent le problème inverse mal-posé au sens de Hadamard (i.e. Figure 2.2).

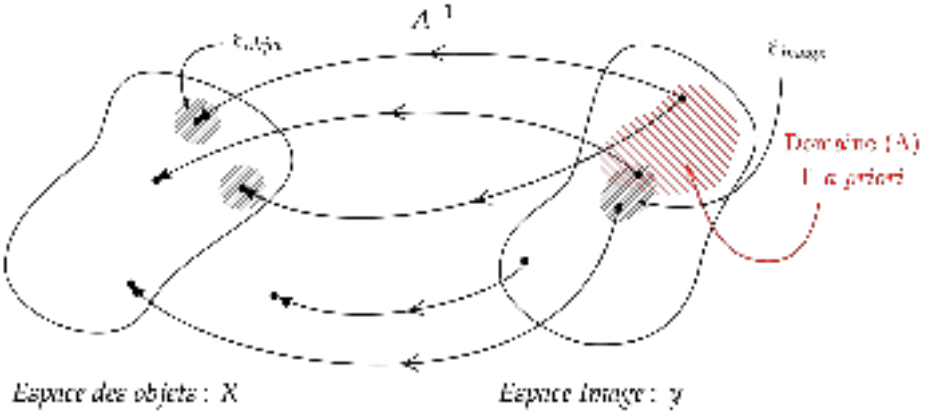


FIGURE 2.2 – Illustration schématique du problème inverse mal-posé

2.4 Modèle d'observation

Le modèle direct Dans le cas de la reconstruction en SR, l'espace $\mathcal{X}$ représente la scène continue à photographier, et ses éléments sont les images de haute résolution, que l'on appelle également images de référence. L'espace $\mathcal{Y}$ est l'ensemble des observations discrètes de la scène. Le contexte de la reconstruction d'une image discrète représentative d'une scène continue repose à la fois sur la bonne compréhension des paramètres intrinsèques de l'imageur et la nature de la scène. Pour un capteur d'image, le processus de reconstruction s'appuie sur ce que l'on appelle communément un modèle d'observation. Ce dernier décrit le lien existant entre l'image hautement résolue à reconstruire et les observations bassement résolues utilisées pour réaliser cette reconstruction. Plus précisément, il représente la notation du problème direct bien-conditionné. Le modèle d'observation comprend le paramètre de dégradation A et le paramètre additif de bruit B_k qui prend en considération les erreurs de modélisation et les variations aléatoires des mesures qui engendrent les images bruitées. Le modèle linéaire le plus répandu dans la littérature s'écrit comme suit [13] :

$$\boldsymbol{Y_k} = DH_kW_k\boldsymbol{X} + B_k, avec k = 1 \cdots K \tag{2.1}$$

où $\boldsymbol{Y_k} \in \mathcal{Y}$ est le vecteur des niveaux de gris de la k^{ieme} image bruitée BR, $\boldsymbol{X} \in \mathcal{X}$ est le vecteur des valeurs des niveaux de gris de l'image HR recherchée. K est le nombre des observations qui rentrent dans l'opération de reconstruction. W_k est la matrice qui modélise le mouvement de la caméra, la matrice H_k modélise le flou induit par l'imageur. D représente le paramètre de décimation, et B_k représente le vecteur du bruit. L'expression est souvent écrite sous la forme matricielle :

$$\boldsymbol{Y} = A\boldsymbol{X} + B \tag{2.2}$$

avec :

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}, \; A = \begin{bmatrix} DH_1W_1 \\ \vdots \\ DH_kW_k \end{bmatrix}, \; et \; B = \begin{bmatrix} B_1 \\ \vdots \\ B_k \end{bmatrix} \tag{2.3}$$

Les éléments $\boldsymbol{Y_k}$ de $\mathcal{Y}$ sont des versions dégradées et bruitées des images de $\mathcal{X}$ selon le modèle d'observation (2.2) où A est le modèle de dégradation. Notons que pour le cas d'une reconstruction SISR, nous disposons que d'une seule observation de la scène, c'est-à-dire $K = 1$. L'équation (2.2) représente une modélisation mathématique du problème direct.

Le modèle inverse À l'opposé, le problème inverse revient à déterminer une image de haute résolution $\boldsymbol{X_{SR}} \in \mathcal{X}$, à partir de k observations (images) de basses résolutions $\boldsymbol{Y_k} \in \mathcal{Y}$ par la résolution du système inverse de (2.2). Étant donné que les images Y_k et $\boldsymbol{X_{SR}}$ soient de définitions $N_1 \times N_2$ et $M_1 \times M_2$ respectivement, tel que $M_1 = lN_1$ et $M_2 = lN_2$, où l représente le facteur d'agrandissement, les matrices D, H_k, et W_k qui constituent le paramètre de dégradation A sont creuses. Le système (2.2) est alors creux de grande taille, et par conséquent impraticable car cette propriété crée une instabilité numérique de la solution recherchée. Dans ce cas discret exprimé par une écriture matricielle, on s'intéresse alors à la **singularité** et au **conditionnement** du système :

(i) Le système est dit "*singulier*" s'il n'y a pas unicité de la solution ;

(ii) Le système est "*mal-conditionné*" s'il y a une dépendance discontinue de la solution par rapport aux données.

En effet, si le système A est non singulier, il existe alors une solution unique. Si A est singulier, alors une solution qui n'est pas unique n'existe que lorsqu'une condition particulière est satisfaite. Par conséquent, il est généralement nécessaire de vérifier la singularité du système A afin de résoudre le système linéaire inverse de (2.2) en examinant son déterminant. Ce contrôle de singularité nécessite approximativement N^3 opérations si l'on considère une matrice de taille $N \times N$. Dans le cas des images, ce calcul devient rapidement coûteux en termes de temps et de ressources. Comme nous l'avons mentionné, la restauration par SR est en général un problème mal-posé. Le processus d'acquisition qui fournit les images d'observation (*Équation* 2.2) représente le problème direct. Le problème inverse correspondant est celui de déterminer l'estimée de la scène à partir des observations et la caractérisation du processus d'acquisition. Le problème direct est celui de la simulation, tandis que le problème inverse est celui de la restauration. Ce dernier peut échouer à satisfaire une ou plusieurs conditions d'Hadamard, ceci est dû au fait que :

(i) L'image $\boldsymbol{X_{SR}}$ ne soit pas unique ; la séquence $\boldsymbol{Y_k}$ d'images BR peut correspondre à deux, voire plusieurs, images SR très distinctes ;

(ii) L'image $\boldsymbol{X_{SR}}$ ne dépend pas continûment des données ; deux observations BR très proches correspondent à deux images HR très éloignées.

Le premier point se produit lorsque le nombre d'inconnus est supérieur au nombre de contraintes en raison du manque d'information sur le résultat souhaité, ou d'une mauvaise modélisation des paramètres de l'opérateur A. Le deuxième point veut dire que la méthode de reconstruction n'est pas robuste et la précision de la solution n'est pas garantie. Il ne sera pas possible d'approcher de façon satisfaisante la solution du problème inverse puisque les données disponibles sont sensibles au bruit. Ceci est expliqué par un manque de conditionnement du système de dégradation A, notamment lorsque la taille de la matrice est grande, et par le

taux du bruit additif B_k : si celui-ci est important, il devient impossible d'inverser correctement l'opérateur A car la présence de bruit peut fournir une séquence d'images observées qui n'a pas de contenu utile quelle que soit la scène. Dans ce cas, le modèle est non-inversible (système singulier ou irrégulier), et l'image résultante est erronée, fausse, ou incompréhensible.

2.5 Résolution du problème

Du point de vu analytique, la précision de la solution du problème inverse de l'équation (2.2) dépend à la fois du conditionnement de la matrice A, de sa singularité, et de la quantité du bruit B. Intuitivement, le conditionnement de la matrice (c'est-à-dire la validité de la solution par rapport aux données) devient plus élevé lorsque la taille de la matrice augmente. En raison de la nature du modèle général de dégradation des observations, la matrice A est pratiquement toujours mal-conditionnée et n'admet pas de solution numérique unique. Pour résoudre ce type de problème, la méthode des moindres carrés est souvent utilisée. Il s'agit de résoudre le problème de minimisation de la norme euclidienne de l'erreur suivante que l'on appelle "l'attache aux données" :

$$\min_X \|A\boldsymbol{X} - \boldsymbol{Y}\|_2^2 \tag{2.4}$$

L'expression exprime le degré de ressemblance entre l'image de haute résolution $\boldsymbol{X}$ et les observations $\boldsymbol{Y}$ via le modèle A. La solution du problème vérifie l'équation suivante :

$$\boldsymbol{X} = (A^T A)^{-1} A^T \boldsymbol{Y} \tag{2.5}$$

Vu que le problème est mal-posé, la matrice $A^T A$ est encore plus mal-conditionnée que la matrice A. Par suite, la solution exprimée par (2.5) ne correspond pas à la bonne solution. Afin de stabiliser le système et réduire l'espace des solutions

compatibles aux données, il faut appliquer des contraintes à la formulation du problème.

D'un point de vue méthodologique, on peut distinguer deux grandes catégories de méthodes de conditionnement de ce type de problèmes. Les techniques les plus répandues sont des approches analytiques unifiées par la théorie de la régularisation qui utilise une information additionnelle qui représente les caractéristiques désirées de la solution pour contraindre le système. Cette information est typiquement qualifiée d'information *a priori* puisqu'elle ne peut être déduite des observations ou du processus d'observation. Ainsi, il s'agit non seulement de minimiser le résidu comme dans l'équation (2.4), mais de rajouter un ou plusieurs termes de contraintes :

$$\min_{X} \|A\boldsymbol{X} - \boldsymbol{Y}\|_2^2 + \lambda \|\mathcal{R}\boldsymbol{X}\|_2 \tag{2.6}$$

$\mathcal{R}$ est l'opérateur de régularisation, $\lambda > 0$ est le paramètre de régularisation qui contrôle l'équilibre entre le terme d'attache aux données et le terme de régularisation $\|R\boldsymbol{X}\|_2$. Ce dernier permet de rejeter les solutions bruitées et de ne garder que les solutions vérifiant les connaissances *a priori* *a priori* du système [14]. La deuxième catégorie d'approches est probabiliste consistant à considérer la solution de l'équation (2.2) comme aléatoire exprimées par un modèle probabiliste bien défini.

2.6 Discussion et Conclusion

La reconstruction d'un modèle à partir des données expérimentales comporte essentiellement plusieurs étape à savoir :

- le recueil des données expérimentales à travers l'acquisition des images ;
- le choix d'un critère de qualité ;
- l'optimisation de ce critère pour obtenir une valeur numérique optimale des paramètres ;
- l'évaluation de l'incertitude sur les paramètres estimés ;

— et enfin, l'analyse critique des résultats.

Dans le but de régulariser un problème inverse mal-posé et de réduire l'espace des solutions candidates, on a besoin de connaissances *a priori* sur la structure typique de l'image originale $\boldsymbol{X}$. Cette information *a priori* correspond au caractère lisse de la solution et peut aller d'une simple hypothèse de continuité uniforme à une connaissance plus complexe de la structure géométrique et texturale de $\boldsymbol{X}$.

Dans les prochains chapitres, nous verrons comment les différentes méthodes de la SR formulent les connaissances *a priori* sur l'image HR recherchée. Nous verrons à quel degré la formulation de cette information dépend de la modélisation mathématique du modèle A et de ces paramètres. Nous verrons également l'importance de la prédétermination de ces derniers dans la fiabilité de la résolution du problème, dans son conditionnement, et dans la généralisation de la reconstruction. Exploiter leurs avantages et inconvénients compte tenu d'environnements complexes et résolution de système myope (semi-aveugle).
L'introduction de techniques de régularisation dans le modèle permet de guider la méthode de reconstruction vers une solution représentant des caractéristiques significatives. En conséquence, la résolution des problèmes mal-posé est réduite à la minimisation d'un résidu sous contraintes destinées à des applications bien spécifiques.

Chapitre 3

État de l'art

3.1 Super-résolution à image unique

Les approches précédentes de MISR se basent sur la fusion d'informations spatiales complémentaires contenues dans plusieurs images BR. Des hypothèses sur des propriétés de structure d'image sont également déployées pour régulariser correctement la solution de la reconstruction. Cette régularisation devient particulièrement cruciale lorsqu'un nombre insuffisant de mesures est fourni, comme dans le cas extrême où une seule image de basse-résolution est observée. Dans de tels cas, la qualité du résultat de reconstruction ainsi que l'efficacité de la régularisation, sont directement liée à l'information *a priori* et dépendent essentiellement de la fiabilité de l'hypothèse de régularité. Ce manque d'informations réelles rend le problème inverse de reconstruction "plus" mal-posé et accentue également le mauvais conditionnement du modèle d'observation. D'un point de vue méthodologique, la formulation et la prédéfinition de l'information *a priori* catégorise les méthodes SISR en différentes classes :

- Méthodes qui utilisent les caractéristiques locales, généralement basées sur les algorithmes d'interpolation et de déconvolution ;
- Méthodes qui utilisent des caractéristiques semi-locales, où l'information *a priori* peut correspondre par exemple aux caractéristiques de l'image d'entrée dans différentes échelles ou domaines ;
- Méthodes qui utilisent des caractéristiques externes, où l'information *a priori* est "apprise" à partir d'un ensemble d'image similaires mais non-identiques à l'image d'entrée.

Afin de remédier au problème du manque d'observations, les méthodes SISR basées sur la reconstruction utilisent un schéma particulier qui permet d'extraire l'information *a priori* à partir de :

(i) une image de référence HR ;

(ii) un modèle pyramidal de multi-résolution de l'image d'entrée ;

(iii) un domaine épars ou fréquentiel représentatif de l'image d'entrée ;

(iv) ou l'exploitation de la redondance des structures locales dans l'image d'entrée.

L'hypothèse de régularité est ensuite formulée de telle sorte à maximiser cette information et est intégrée dans la reconstruction à l'aide d'algorithmes d'optimisation et de régularisation.

Un autre type de méthodes palie le problème du manque de données observées par l'utilisation d'une source d'information alternative externe afin d'extraire l'information *a priori*. Une méthodologie répandue de régularisation du problème inverse de la SISR consiste à utiliser une source externe de données, appelée ensemble d'exemples ou dictionnaire. Il existe deux types de dictionnaires :

— Les *dictionnaires externes* qui sont constitués de plusieurs images HR naturelles $\boldsymbol{X_i}$ et de leurs versions BR $\boldsymbol{Y_i}$ synthétisées par le modèle décrit par l'équation (2.2). Le couple de données $(\boldsymbol{X_i}, \boldsymbol{Y_i})$ et appelé atome.

— Les *dictionnaires internes* qui sont générés selon un mécanisme d'exploitation des redondances des structures géométriques dans l'image (traité en chapitre 4).

À la différence des méthodes précédentes, où l'information *a priori* est de forme paramétrique qui régularise l'image entière, les méthodes basées sur l'utilisation d'exemples définissent l'information *a priori* à travers une étape d'extraction des caractéristiques des atomes. Elles apprennent ensuite à correspondre ces caractéristiques à celles de l'image BR d'entrée pour reconstruire enfin l'image HR. L'opération de la mise en correspondance des différentes caractéristiques est effectuée à l'aide d'algorithmes d'apprentissage artificiel. La figure 3.1 montre le schéma de reconstruction SISR des méthodes basées sur l'apprentissage automatique. Par ailleurs, la majorité des méthodes récentes de reconstruction SISR utilisent des mécanismes d'apprentissage artificiel, et leur robustesse est très liée à la fois au choix du modèle de dégradation et au choix du dictionnaire d'entraînement : Le modèle de dégradation intervient dans la génération des données du dictionnaire, et permet la synthétisation des images BR à partir des éléments HR du dictionnaire. Ces deux paramètres définissent également le modèle de l'information *a priori* qui contrôle la régularisation. De manière générale, les algorithmes de reconstruction en SISR basées sur l'apprentissage artificiel suivent un schéma

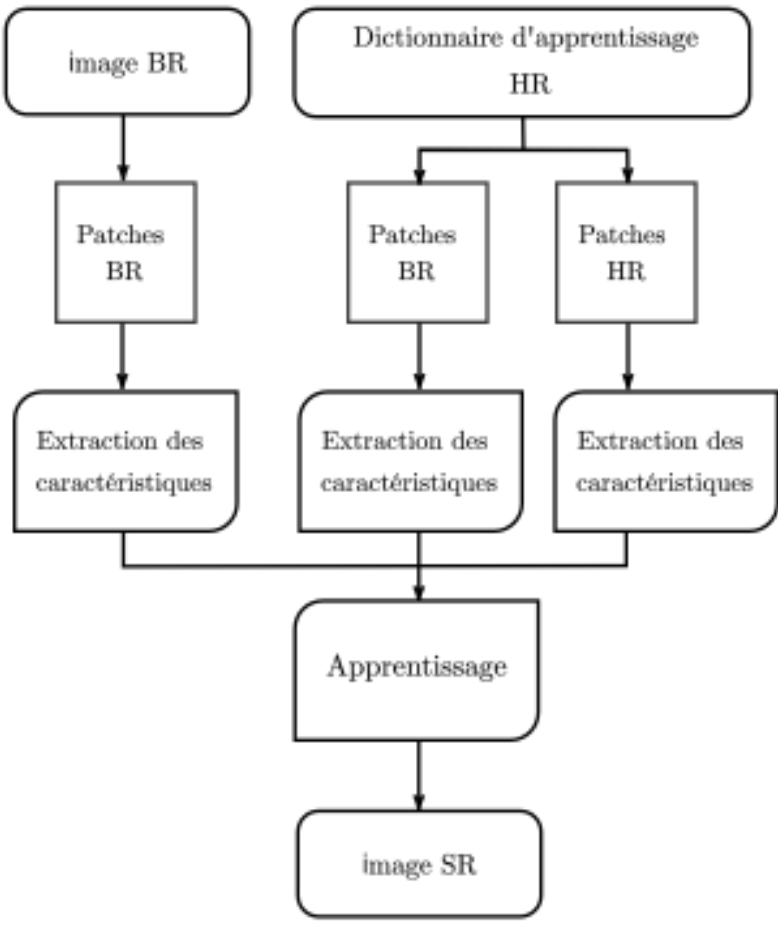


FIGURE 3.1 – SISR apprentissage

commun qui comprend les étapes suivantes : Extraction des caractéristiques, Apprentissage et Mise en correspondance.

3.1.1 Extraction des caractérisiques

Indépendamment du type de l'information *a priori* (i.e. local, semi-local, ou externe), son identification constitue l'étape initiale des différentes méthodes des catégories mentionnées précédemment. L'identification de l'information *a priori* est effectuée par des méthodes d'extraction des caractéristiques de l'image BR donnée.

Les structures primitives telles que les contours, les coins, les jonctions en T ou en croix, les crêtes, les rampes concaves et convexes représentent les caractéristiques

les plus pertinentes dans une image dans la mesure où l'œil humain est plus sensible à leur perception. Dans le même contexte, certains travaux ont considéré la luminance comme caractéristique représentative du contenu de l'image car l'œil humain est également susceptible aux changements de luminance qu'aux changements de couleurs [15–26]. Plusieurs travaux ont proposé de raffiner la qualité visuelle de l'image en améliorant l'intensité et la continuité de ces structures primitives. Ils ont cenpendant exprimé des termes de régularité liés à ces structures, notamment en ce qui concerne la netteté des contours [27–29], leurs gradients [30–32], et leur orientation [33, 34] dans la régularisation des problèmes inverses de la reconstruction SR [35–37], également dans les techniques d'interpolation [38–42]. Sun et al ont utilisé un algorithme de dérivatives gaussienne afin d'extraire les structures primitives de l'image [35]. Un autre modèle utilisé destiné à décrire les contours s'appuie sur un filtrage linéaire passe-haut des fréquences contenues dans l'image [15, 43–47].
Dans d'autres travaux tels que [48–53, 53–56], les auteurs ont procédé à une analyse multi-résolution de l'image d'entrée en utilisant un modèle pyramidal Gaussien et Laplacien. Les caractéristiques primitives de l'image sont définies sous forme de vecteur de dérivées Gaussiennes et Laplaciennes multi-échelle, ou bien de blocs d'images extraits à partir de la soustraction des différents niveaux de la pyramide. Un autre type de caractéristiques également utilisé dans la formulation de la connaissance *a priori* est les composante fréquentielles [57, 58], ou plus particulièrement dans le domaine de la DCT [59].

L'application de ces méthodes destinées à extraire les caractéristiques détermine la stratégie ultérieure de la reconstruction en SR :

- D'une part, pour le cas des méthodes **SISR basées sur la reconstruction**, ces caractéristiques sont exprimées sous forme de termes de régularisation intégrés à la fonction de coût du problème de reconstruction ;
- D'autre part, pour le cas des méthodes **SISR basées sur des exemples**, l'opération d'extraction des caractéristiques est appliquée séparément à l'image

d'entrée BR ainsi qu'aux atomes du dictionnaire. Ces caractéristiques sont ensuite mises en correspondances à l'aide de modèles d'apprentissages.

3.1.2 SISR basée sur la reconstruction

Selon le théorème de Bayes, le terme de régularisation qui stabilise la solution du problème inverse mal-posé de la SR représente le modèle des connaissances *a priori* sur l'image HR recherchée [60]. Dans le cas particulier de la SISR, l'énergie du terme de l'attache aux données est faible en raison du manque d'observations. Par conséquent, la reconstruction SR dépend essentiellement de l'information *a priori*, de sa formulation, et de son déploiement dans la reconstruction. Dans la section 2.5, nous avons arboré les différents types d'algorithmes de régularisation pour la résolution des problèmes inverses mal-posés. Appliquées à la SISR, la plupart des méthodes suggérées dans la résolution de ce problème sont basées sur des implémentations itératives d'algorithmes d'optimisations. Malgré le fait que ces algorithmes de résolution évitent l'utilisation de matrices inverses, elles restent néanmoins gourmandes en termes de temps de calculs.

3.1.2.1 Régularisation par analyse de l'*a priori* local

Régularisation par lissage : Dans les premiers travaux tels que [14, 61], le lissage des images naturelles était souvent considéré dans la SR puisque l'opération de lissage permet de réduire considérablement le bruit du capteur, ce qui mène à la propriété quadratique des formules de régularisations de Tikhonov [14] qui utilise la norme $\mathfrak{l}_2$, $(\|\cdot\|_{p=2})$, et dont le terme de régularisation est : $\|\Gamma x\|_2^2$, où Γ est généralement un filtre passe-haut , ou un opérateur Laplacien 2-D [61]. Une autre catégorie de régularisation est basée sur la théorie de Markov [62]. Ces méthodes régularisées lissent les images restaurées en pénalisant les composantes des hautes fréquences, et atténuent inévitablement les contours. Certains travaux ajoutent d'autres termes de contraintes sur la préservation des contours dans la formule de

régularisation, ou utilisent des modèles des champs aléatoires gaussiens de Markov [63].

Régularisation par accentuation des contours : Le caractère quadratique des formules de régularisation par lissage d'image produit des images floutées car une moyenne (pondérée ou non) est appliquée sur les pixels. La norme l_2 a été remplacée par la norme l_1 pour sa propriété à préserver l'intensité des contours. La méthode par variation totale (TV pour Total Variation), combinée à la l_1, a été largement utilisée dans la reconstruction SR [64–75]. D'autres approches utilisent la norme l_p comme fonction de contrainte sur le gradient dans la formule de la TV [76–78]. D'autres amélioration à la TV ont été proposées et appliquées à la SISR comme la BTV [69, 79–85], la BTV adaptative [77], et la high-order TV [86].

3.1.2.2 Régularisation par analyse de l'*a priori* semi-local

Les méthodes de régularisation basée sur l'*a priori* semi-locale effectuent l'opération de l'approximation des nouveaux pixels en considérant des régions différentes du voisinage de ces derniers. Cette stratégie est basée sur l'assomption que chaque caractéristique dans une image donnée peut réapparaître de manière redondante dans différentes parties de l'image, ou dans des représentations multi-résolution de celle-ci. On appelle cette propriété la redondance de patchs.

3.1.3 SISR basée sur l'analyse d'exemples

3.1.3.1 Préparation du Dictionnaire

Étant donné un ensemble d'images naturelles HR à partir des quelles leur versions BR sont générées en utilisant le modèle de dégradation défini par l'équation 2.2. Dans la majorité des méthodes de la SISR, ce modèle de dégradation est défini par un opérateur de décimation selon un facteur $'r'$, un noyau idéal 'bicubique' pour modéliser le flou, et d'un bruit gaussien blanc additif. Les paires d'images

sont ensuite découpées en P patchs partiellement superposés notés $\{p_{\boldsymbol{X},i}\}_{1\leq i\leq P}$ et $\{p_{\boldsymbol{Y},i}\}_{1\leq i\leq P}$ pour les images HR et BR respectivement. L'ensemble des paires $(p_{\boldsymbol{X},i}, p_{\boldsymbol{Y},i})$ constituent le Dictionnaire d'apprentissage D, où chacune des paires de patchs vérifie l'équation (2.2) liée au modèle d'observation. Notons que les images HR et BR du dictionnaire sont d'abord soumises à l'opération d'extraction des caractéristiques avant l'extraction des patchs. Ainsi, ces derniers sont généralement représentés par leurs caractéristiques locales. De façon similaire, l'image BR d'entrée est également découpée en patchs notés $\{p^t_{\boldsymbol{Y},i}\}_{1\leq i\leq P}$. L'opération de correspondance consiste à trouver les similarités entre les patchs $\{p^t_{\boldsymbol{Y},i}\}_{1\leq i\leq P}$ et $\{p_{\boldsymbol{Y},i}\}_{1\leq i\leq P}$ pour ensuite les appliquer à la restauration par bloc de la nouvelle image HR, par prédiction des patchs HR correspondants, notés $\{p^t_{\boldsymbol{Y},i}\}_{1\leq i\leq P}$, à l'aide d'un mécanisme de *Matching*.

3.1.3.2 Méthoes basées sur les plus proches voisins

L'approche la plus basique consiste à trouver, pour l'ensemble des patchs $\{p^t_{\boldsymbol{Y},i}\}$ d'une image d'entrée BR, les patchs HR correspondants qui existent dans ce dictionnaire au sens du plus proche voisin. Freeman et al. [15] ont proposé une approche basée sur les k plus proches voisins qui maximisent la compatibilité entre patchs HR adjacents en utilisant un champs aléatoire de Markov (MRF pour Markov Random Field), où le meilleur patch $\{p_{\boldsymbol{Y},i}\}$ est sélectionné en utilisant l'algorithme de propagation de croyance [87].

D'autre travaux ont étendu cette méthode en utilisant les structures primitives de l'image comme connaissance *a priori* afin de préserver la netteté des contours [88]. Irani et Peleg ont par la suite proposé d'intégrer le processus IBP dans la reconstruction [89] (i.e. Section 3.1.3.3). Dans une version ultérieure, Wang et al. ont étendu la méthode en utilisant un modèle statistique pouvant traiter les PSF inconnues, rendant l'algorithme plus efficace [90].

L'efficacité de ces méthodes est étroitement liée à la fiabilité de l'algorithme de la mise en correspondance des patches. En effet, comme le montre la figure 3.2, il se peut que des patches HR sélectionnés ne correspondent pas forcément au patch HR recherché.

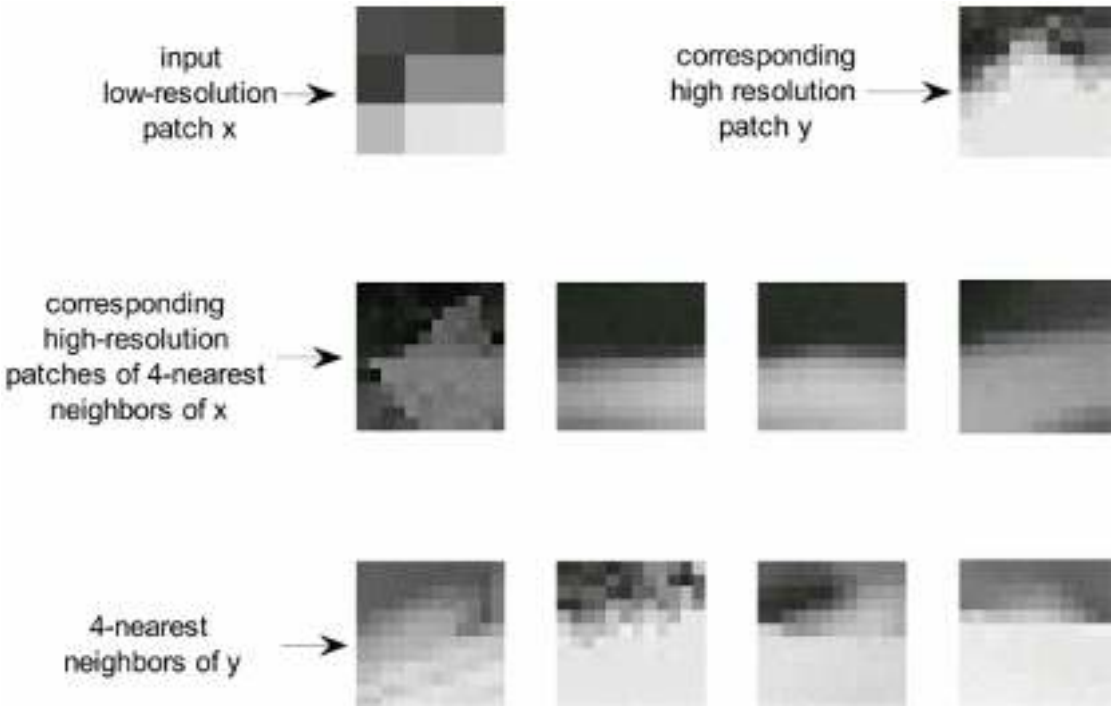


FIGURE 3.2 – Les 4 patchs HR correspondants aux 4 patchs BR voisins du patch initial ne sont pas proches du patch HR recherché.

3.1.3.3 Méthoeds basées sur les variétés (Manifolds)

Les méthodes susmentionnées se basent principalement et directement sur les patchs de l'image, ce qui implique l'utilisation d'un large ensemble de donnée d'apprentissage afin d'inclure tout les modèles susceptibles d'être rencontrés lors des tests. De nombreuses approches ont étudié la possibilité d'utiliser un volume de donnée d'apprentissage réduit en définissant l'ensemble des patchs BR comme étant un espace de dimension moindre. L'algorithme de réduction de dimension pionnier dans cette voie est l'algorithme d'intégration des voisins (en anglais *Neighbors Embedding*) [91]. Ce dernier stipule qu'un atome donné (i.e. patch) peut être reconstruit à l'aide d'une combinaison linéaire pondérée de ces proches voisins, où les poids représentent la similarité relative entre l'atome original et ses atomes voisins. De façon similaire, ces même poids, une fois calculés, peuvent être utilisés dans la reconstruction d'atome de dimension supérieure à partir de ces voisins proches à l'aide d'une combinaison linéaire (Figure 3.3).

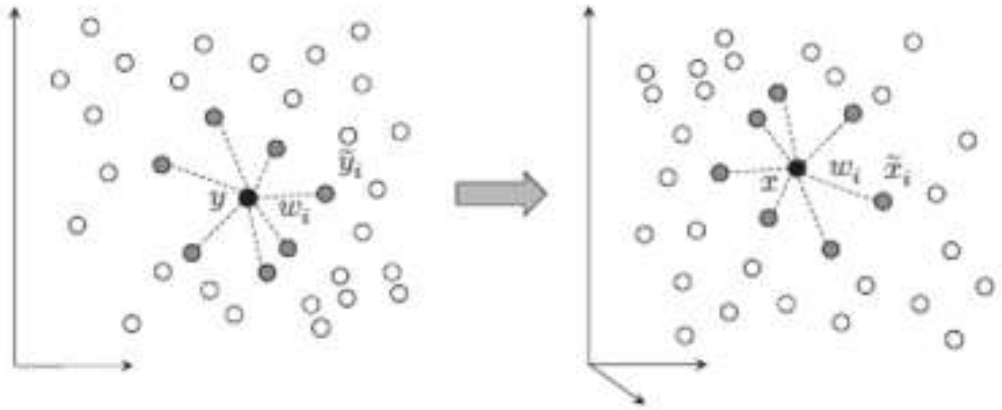


FIGURE 3.3 – À gauche, un *atome* y de basse résolution (de basse dimensionalité) est approximé en termes de combinaisons pondérées avec ces voisins $\tilde{y}_i$. À droite, les mêmes poids peuvent être utilisés sur les homologues à haute résolution $\tilde{x}_i$ des atomes $\tilde{y}_i$ dans un espace de haute dimension pour approximer la donnée HR (de haute dimensionnalité) correspondante.

Ce principe fait également référence aux algorithmes d'intégration linéaire locale LLE (pour Locally Linear Embedding) qui calcule une projection de dimension inférieure des données en préservant les distances (ou structures géométriques locales) dans les voisinages locaux [16, 92–94] (Figure 3.4. Le principe de cette méthode de reconstruction est dérivé de la théorie de l'échantillonnage compressé (CS pour *Compressed Sensing*) qui garantit que les relations linéaires entre les données de hautes résolutions peuvent être récupérées avec précision à partir de leur projection à faible dimension [95]. Les patchs des images à haute résolution sont représentées par des points dans un espace de données à haute dimension, et les patchs d'images à basse résolution correspondants sont des points dans un espace de données à basse dimension.

L'applicabilité de ce principe à la super-résolution exige que les patchs HR correspondants aux voisins les plus proches du patch que l'on souhaite reconstruire soient connus. Pour chacun des patchs $\{p^t_{\boldsymbol{Y},i}\}$ de l'image à reconstruire, l'algorithme développé par Chang et al [43] trouve ses k plus proches voisins notés $\mathtt{N}_t$ à partir des $\{p^t_{\boldsymbol{Y},i}\}_{1\leq i\leq P}$ d'une seule image BR, et calcule les poids de reconstruction $\hat{W}_r$ par :

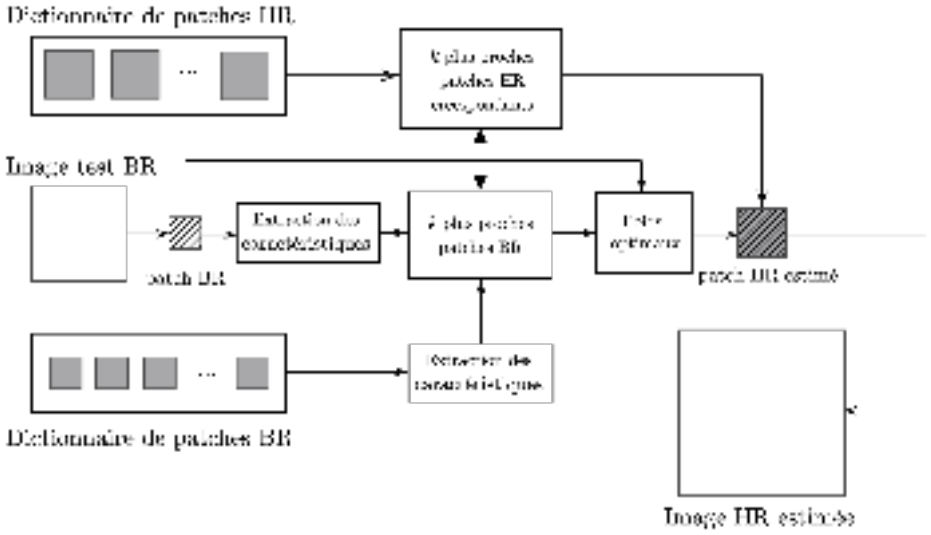


FIGURE 3.4 – LLE Framework

$$\hat{W}_r = \arg\min_{W_r} \|p^t_{\boldsymbol{Y},i} \; - \sum_{P_{\boldsymbol{Y},i} \in \mathtt{N}_t} W_r P_{\boldsymbol{Y},i}\|_2^2, \; tel\ que \sum_{P_{\boldsymbol{Y},i} \in \mathtt{N}_t} W_r = 1. \tag{3.1}$$

Les poids de reconstruction sont ensuite utilisés pour générer le patch de haute résolution $p^t_{\boldsymbol{Y},i}$ correspondant :

$$p^t_{\boldsymbol{Y},i} = \sum_{P_{\boldsymbol{X},i} \in \mathtt{N}_t} \hat{W}_r P_{\boldsymbol{X},i} \tag{3.2}$$

L'implémentation de cet algorithme s'est avérée robuste même avec l'utilisation de patchs de taille relativement petite que ceux utilisé dans [96] et [15]. Le rôle des poids de reconstruction W_r est de caractériser les contraintes locales d'un ensemble de dimension donnée, leur détermination varie d'un algorithme à un autre. L'algorithme d'intégration des voisins préserve la cohérence géométrique des structures locales des ensembles à travers différentes dimensions : les versions correspondantes aux voisins les plus proches d'un point donné dans un espace de haute dimension sont les voisins les plus proches dans un espace de basse dimension. Néanmoins, l'efficacité de cette propriété dépend à la fois de la tailles des patchs et du vecteur de caractéristiques de ces derniers.

3.1.3.4 Méthodes basées sur la représentations éparse des données

Un autre type de méthode appelée *Codage Parcimonieux* (ou *Sparse Coding* en anglais) suggère que les patchs des images naturelles sont représentés par une combinaison linéaire de seulement quelques atomes provenant d'un dictionnaire extensible et choisi de manière appropriée. De la même manière que les méthodes basées sur les variétés, cette méthode aborde le problème de la SR sous le principe de la CS, et utilise en plus la *'sparsité'* comme information *a priori* pour régulariser le problème inverse de la SR. L'idée de base de l'algorithme est très semblable à l'hypothèse de l'algorithme LLE qui stipule que les patches $\{P_{\boldsymbol{Y},i}\}_{1<i<p}$ ont des propriétés géométriques locales similaires à leurs patchs $\{P_{\boldsymbol{X},i}\}_{1\leqslant i\leqslant p}$ correspondants. La différence tient seulement au fait que les voisins les plus pertinents sont sélectionnés de manière adaptative selon la connaissance *a priori* imposée :

$$p^t_{\boldsymbol{X},i} \approx D_h \underline{\alpha} \qquad avec\ \alpha \in \mathbb{R}^k \qquad tel\ que \quad \|\alpha\|_0 \ll k \tag{3.3}$$

où $D_h \in \mathbb{R}^{n\times K}$ est un dictionnaire dit quasi-complet contenant la représentation vectorielle des patchs HR $\{p_{\underline{\boldsymbol{X}},i}\}_{1\leqslant i\leqslant p}$, k est le nombre d'atomes de H_h, et α est la représentaion éparse de $p^t_{\boldsymbol{X},i}$ qui décrit la connaissance *a priori*.
Le but est de trouver le vecteur α le plus épars possible qui représente le patch $p^t_{\boldsymbol{X},i}$ de l'image HR recherchée de façon à ce que, la version dégradée de celui-ci se rapproche le plus possible du patch $p^t_{\boldsymbol{Y},i}$ de l'image BR mesurée :

$$min\|\alpha\|_1 \qquad tel\ que: \quad \|FD_l\underline{\alpha} - Fp^t_{\boldsymbol{Y},i}\|_2^2, \tag{3.4}$$

où F est l'opérateur d'extraction des caractéristiques, généralement un filtre passe-haut [44, 45, 97, 98]. De manière générale, les algorithmes de Codage parcimonieux appliqué à la SR modélisent les données d'apprentissage par deux dictionnaires $D_h = [\{p_{\boldsymbol{X},i}\}_{1\leqslant i\leqslant p}]$ et $D_l = [\{p_{\boldsymbol{Y},i}\}_{1\leqslant i\leqslant p}]$, où les patchs se chevauchent d'un pixel dans les deux directions, et sont représentés par leur changement d'intensité ou leurs structures primitives. Ensuite, ils reconstruisent le patch HR en estimant sa

représentation éparse α à partir de l'image d'entrée $\boldsymbol{Y}$ (l'image BR) par rapport au dictionnaire D_l. Afin de forcer la compatibilité entre les patchs adjacents, les auteurs ont procédé à une optimisation conjointe sur les deux dictionnaires D_h et D_h. L'équation (3.4) devient :

$$\hat{\underline{\alpha}} = \min_{\underline{\alpha}} \lambda \|\underline{\alpha}\|_1 + \frac{1}{2}\|\tilde{D}\underline{\alpha} - \tilde{p}_{\boldsymbol{Y},i}\|_2^2 \tag{3.5}$$

avec $\tilde{D} = \begin{bmatrix} FD_l \\ \beta P_c D_l \end{bmatrix}$ et $\tilde{p}_{\boldsymbol{Y},i} = \begin{bmatrix} Fp_{\boldsymbol{Y},i} \\ \beta w \end{bmatrix}$. La matrice P_c extrait la région de chevauchement entre les patchs cibles actuels et l'image HR précédemment reconstruite, et w contient les valeurs de l'image HR précédemment reconstruite sur cette région [44]. La résolution du problème d'optimisation de l'équation (3.5) permet de reconstruire les patchs HR en appliquant un codage parcimonieux comme suit :

$$p^t_{\boldsymbol{X},i} = D_h \hat{\underline{\alpha}} \tag{3.6}$$

Les méthodes de codage parcimonieux appliquées à la SR reposent sur l'hypothèse qui établie que chaque paire de patchs $(p_{\boldsymbol{X},i}, p_{\boldsymbol{Y},i})$ a les mêmes représentations éparses par rapport aux deux dictionnaires D_h et D_l. La fiabilité de ces méthodes est alors étroitement liée aux choix de ces derniers. Une façon directe d'obtenir deux dictionnaires de ce type consiste à extraire directement les patchs à partir de paires d'image $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, ce qui permet de préserver la correspondance entres les deux espaces HR et BR. Cependant, une telle stratégie entraîne des dictionnaires de grandes tailles et par conséquent des calculs coûteux. Des méthodes, pour pallier ce problème, consistent à l'apprentissage de ces dictionnaires de façon à ce qu'ils soient plus compacts [46, 99].

3.2 Conclusion

Dans ce chapitre, nous avons énoncé les différentes et les plus intéressantes méthodes de la reconstruction SR dans la littérature. Dans l'optique de travailler sur une amélioration en concordance avec la scène réelle, nous avons fait état des recherches réalisées sur ce sujet. Il s'avère que la plupart des travaux classiques visent à une amélioration quantitative standard plutôt que de synthétiser les détails réels de l'image. Pour cause, les paramètres de dégradation du modèle d'acquisition sont définis comme idéaux. Or, étant donné que notre principal défi concerne un conditionnement réel de la reconstruction, nous ne sommes pas parvenus à distinguer des pistes de recherches suffisamment prometteuses pour les méthodes de reconstructions classiques.

Chapitre 4

SISR par apprentissage artificiel

4.1 Réseaux convolutifs profonds

Un réseau de neurones convolutif CNN (pour Convolutional Neural Networks, ou ConvNets) est un type de réseaux de neurones artificiels acyclique qui analyse plusieurs caractéristiques de l'image d'entrée en utilisant des opérations mathématiques linéaires discrètes appelées convolution ou produit de convolution noté $*$. Ces opérations sont effectuées par plusieurs strates empilées qui forment ce que l'on appelle une couche de convolution. Le nombre de ces couches dans un CNN définissent sa profondeur, qui va dépendre également de la taille du dictionnaire d'entraînement. La formule de convolution utilisée par un CNN s'écrit :

$$S(i,j) = (X * K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \tag{4.1}$$

où l'image X est une fonction définie sur R assimilée à l'entrée de la couche, et K représente le noyau de convolution. Chaque noyau, une fois convolué à la donnée d'entrée de la strate, génère une carte de caractéristiques $S(i,j)$ appelée descripteur (en anglais *feature map*) qui sera ensuite normalisé par des fonctions linéaires ou non-linéaires. Contrairement aux réseaux neuronaux classiques, les couches du CNN ont une interaction clairsemée entre leurs unités d'entrée et de sortie. Ceci est accompli lorsque la taille du noyau de convolution est plus petite que celle de l'entrée. Par exemple, l'image d'entrée du réseau peut comporter des millions de pixels, mais grâce aux couches de convolution, nous pouvons détecter les petites caractéristiques utiles, telles que les contours avec des noyaux n'occupant que des dizaines de pixels. Cela permet de stocker moins de paramètres ce qui améliore l'efficacité de l'algorithme. Par conséquent, et à la différence des autres modèles d'apprentissage automatique, les CNN effectuent implicitement l'opération d'extraction des caractéristiques durant leur entraînement.

Les couches de convolution sont introduites au début du réseau CNN afin d'extraire les caractéristiques de l'image de manière pertinente à travers les noyaux de convolution.Á noter que la pertinence des caractéristiques est proportionnelle à l'intervention tardive des noyaux de convolution : plus la convolution intervient

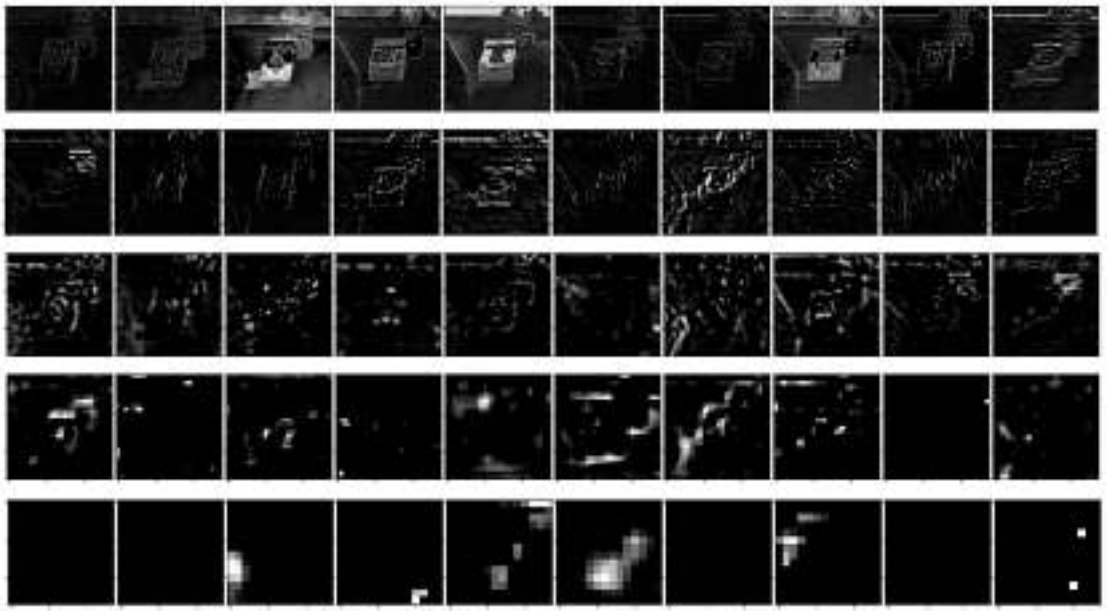


FIGURE 4.1 – Feature-maps des 5 blocs du réseau VGG16 [2] pré-entraîné. Chaque ligne représente les 10 premiers descripteurs de chaque bloc. La profondeur du réseau détermine avec une complexité progressive d'abord, les contours, puis les textures, enfin les formes géométriques et les objets au sein d'une image. Plus la couche considérée est profonde, plus les caractéristiques extraites sont génériques et de haut-niveau.

tard dans le réseau, plus les noyaux détectent des formes plus détaillées (Figure 4.1). Ceci est dû à l'initialisation aléatoire des noyaux de convolution au début de l'apprentissage du réseau afin d'améliorer les résultats d'extraction des caractéristiques.

Le processus d'apprentissage consiste ensuite à mettre à jour ces noyaux par la minimisation de la fonction de coût du réseau (Équation 4.2) également appelée fonction de perte. Cette fonction est essentielle à tout processus d'apprentissage dans la mesure où elle guide le réseau vers la reconstruction d'une image aussi fidèle que possible à l'image de vérité-terrain. Elle règle le degré d'ajustement des données en quantifiant l'écart entre la sortie $\tilde{X}_i$ prédite par le réseau (l'image SR reconstruite) et l'image cible $\boldsymbol{X_i}$ attendue (image de référence HR) extraite des données de la base d'apprentissage. La fonction de coût est souvent exprimée par l'erreur quadratique $\|\tilde{X}_i - X_i\|^2$ associée à un terme de régularisation $\mathcal{R}(\theta) = \lambda\|W\|^2$

qui permet d'éviter un sur-apprentissage du réseau. Soit la base de données d'entraînement de taille N composées de paires $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i\in\{1,\cdots,N\}}$, la fonction générale de coût du réseau est décomposées en la somme des fonctions coût $J_{\theta,i}, i \in \{1, \cdots, N\}$ appliquées aux paires des données d'entraînement une par une :

Le processus d'apprentissage consiste à optimiser les paramètres $\theta = (W_i, b_i)_{i\in\{1,\ldots,l\}}$ où l est la profondeur du réseau ;

$$J_\theta = argmin_\theta \sum_{i=1}^{N} J_{\theta,i} = argmin_\theta \sum_{i=1}^{N} \|\tilde{X}_i - X_i\|^2 + \mathcal{R}(\theta) \tag{4.2}$$

où $\tilde{X}_i$ est la reconstruction SR de l'image y_i, et $\theta = (W_i, b_i)_{i\in\{1,\ldots,l\}}$, sont les paramètres à optimiser où l représente la profondeur du réseau et W_i,et b_i sont le poids et le biais des neurones de la couche i respectivement.
De nombreux CNN actuels sont conçus pour des tâches de classification d'images. Ces réseaux transforment l'image d'entrée en un vecteur de caractéristiques à 1-dimension, ce qui rend le maintien d'information spatiale de l'image d'entrée dans le vecteur des caractéristiques très difficile. Pour certaines applications, telles que la restauration d'image, l'information spatiale est essentielle car la nouvelle valeur du pixel est déterminée principalement par l'information locale à cette position spécifique. Les réseaux profonds entièrement convolutifs sont mieux adaptés à ce type d'applications dans la mesure où la dimensionnalité des données d'entrée et de sortie est préservée.

4.2 CNN appliqués à la SR

Nous avons constaté qu'il est envisageable de catégoriser les réseaux CNN appliqués à la SR en (04) générations :

4.2.1 Génération 1 : Le premier CNN appliqué à la SR

Dong et al ont proposé le premier CNN profond appliqué à la reconstruction SR, appelé SRCNN (Pour Super-Resolution Convolutional Neural Network) [3]. Leur objectif est de récupérer, à partir d'une image d'entrée BR $\boldsymbol{Y}$, les hautes fréquences perdues durant le processus de dégradation décrit par l'équation (2.2) pour reconstruire $\tilde{\boldsymbol{X}}$ proche de l'image $\boldsymbol{X}$. Pendant l'entraînement, des images $\boldsymbol{X}$ d'une base de données sont artificiellement dégradées pour créer des paires $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$. Pour chaque paire, l'algorithme du SRCNN consiste d'abord à augmenter la taille de $\boldsymbol{Y}_i$ d'un facteur de grossissement $r = 2$ à l'aide d'une opération d'interpolation bicubique pour atteindre la taille d'image souhaitée. L'image résultante interpolée, notée $\boldsymbol{Y}_i^{\uparrow}$ est une image de basse résolution ayant les même dimensions que $\boldsymbol{X}_i$. Ensuite, un CNN de trois couches de convolution est employé pour apprendre de bout en bout, de manière supervisée, une correspondance non-linéaire entre $\boldsymbol{Y}_i^{\uparrow}$ et $\boldsymbol{X}_i$.

La première couche de convolution du SRCNN est consacrée à l'extraction des caractéristiques, et consiste à l'application de la convolution à la donnée d'entrée $\boldsymbol{Y}_i^{\uparrow}$ en vue d'extraire 64 descripteurs distincts de celle-ci. Une fonction d'activation de type ReLU est ensuite appliqué à ces descripteurs afin d'introduire la non-linéarité dans le modèle. Ces descripteurs sont injectés à la deuxième couche de convolution afin de projeter les caractéristiques BR dans un espace HR. Enfin, la dernière couche de convolution reconstruit l'image finale $\tilde{\boldsymbol{X}}$ à partir des descripteurs résultants de la deuxième couche. L'architecture générale du réseau SRCNN est illustrée par la figure 4.2.

Malgré sa structure simple, le SRCNN a montré une efficacité supérieure à celle des méthodes traditionnelles susmentionnées basées sur des exemples, dans la mesure où les étapes de la reconstruction SISR, à savoir l'extraction des caractéristiques, le *Matching*, et la reconstruction, sont implémentées conjointement en un seul bloc. Dans le domaine de la reconstruction SISR, les études de recherches récentes ont principalement amélioré la structure de base du SRCNN en rajoutant diverses stratégies de conception et d'apprentissage. Ces améliorations se basent

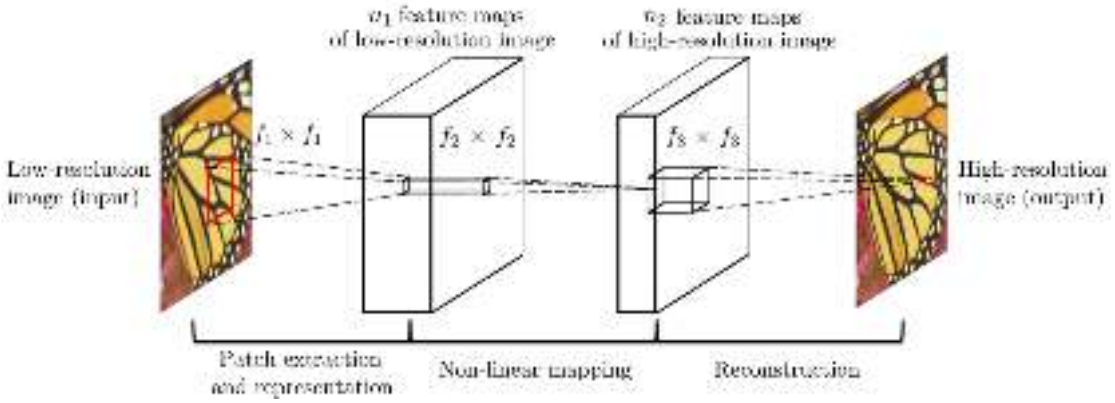


FIGURE 4.2 – Architecture du SRCNN [3]

fréquemment sur la proposition d'une meilleure architecture du CNN, sur une fonction de coût plus effective, sur des bases de données plus larges, plus adaptées, ou encore sur une manière plus pertinente d'évaluer la qualité de l'image reconstruite.

4.2.2 Génération 2 : Amélioration de l'architecture

Le véritable pouvoir des CNN profonds réside dans la capacité de capturer les caractéristiques abstraites de l'image à mesure que ce signal se déplace plus profondément dans les couches. Cependant, les réseaux de neurones trop profonds sont souvent victimes des fameux disparition/explosion du gradient. À mesure que la profondeur du réseau augmente, la précision sature puis se dégrade rapidement. Pour pallier ce problème, He et al. ont conçu une architecture profonde résiduelle pour un réseau allant jusqu'à 152 couches, définie de telle manière que la sortie d'une couche soit reprise par une autre couche plus profonde dans le réseau [100]. Ce «raccourci» créé par les connexions résiduelles permet l'ajout de couches de convolution supplémentaires dans le réseau sans affecter les performances. Cette stratégie a rapidement été généralisée et intégrée dans plusieurs modèles CNN de la SR.

La façon d'effectuer le sur-échantillonnage dans le réseau est une opération non-triviale. En fonction de cette dernière et de sa position dans la structure du modèle,

les méthodes de reconstruction SR peuvent être attribuées à quatre catégories de modèles :

- *Pré-échantillonnage :* L'opération d'augmentation de la résolution de l'image est réalisée à l'aide d'une opération d'interpolation algorithmique classique qui génère une image de haute définition avec une résolution grossière. Ensuite, la SR se trouve ici être principalement une opération de filtrage visant à corriger les artéfacts obtenus après augmentation de la définition de l'image. Ceci va permettre au réseau d'effectuer un apprentissage de bout en bout entre son image d'entrée et son image de sortie (Figure 4.3) ;
- *Post-échantillonnage :* Les images BR sont directement injectées dans le CNN sans passer par une étape d'augmentation de leurs tailles. Le sur-échantillonnage est intégré dans la dernière couche du réseau. Les opérations de convolution pour extraire les caractéristiques sont alors effectuées dans l'espace à faible dimension, ce qui réduit considérablement la complexité de mappage sur une image de grande définition (Figure 4.4). Néanmoins, l'inconvénient rencontré par ce type de principe est que l'agrandissement est toujours réalisé en une seule étape, ce qui implique des difficultés de performances pour les grands facteurs d'agrandissement comme $\times 4$ et $\times 8$. Cela implique également la nécessité de disposer de réseaux différents pour divers facteurs d'agrandissement, ou du moins de blocs de sortie différents comme il est le cas pour le réseau EDSR [8] illustré dans la figure (4.4(c)).
- *Sur-échantillonnage progressif :* Il s'agit d'un modèle multi-résolution ou multi-échelle similaire à la structure pyramidale de Laplace. Son principe est basé sur une cascade de CNN où l'image est reconstruite progressivement. À chaque étape, les images sont ré-échantillonnées à une résolution supérieure, où une image exploitable est créée à l'issu de chaque niveau d'agrandissement. Cette stratégie est proposée afin de palier la lacune rencontrée dans les réseaux à post-échantillonnage lorsque le facteur d'échelle est grand ;
- *Sur-échantillonnage itératif* : afin de mieux restituer les dépendances mutuelles des paires BR-HR, une procédure itérative effective de rétroprojection

est intégrée au modèle profond [101]. Cette stratégie, à savoir l'échantillonnage itératif ascendant et descendant, tente de raffiner la rétroprojection en calculant l'erreur de reconstruction puis de la fusionner afin d'ajuster le résultat de la reconstruction. La stratégie du sur-échantillonnage d'un modèle définit implicitement le type d'analyse des caractéristiques de son image d'entrée, et influe également sur la complexité du calcul de l'entraînement et de test. Ce procédé bien que très prometteur pose encore de nombreux soucis de conception et reste relativement peu renseigné à ce jour.

Les figure 4.3 et 4.4 illustrent quelques différents réseaux profonds très utilisés appliqués dans ces stratégies. À noter que la stratégie du sur-échantillonnage d'un modèle définit implicitement le type d'analyse des caractéristiques de son image d'entrée, et influe également sur la complexité du calcul.

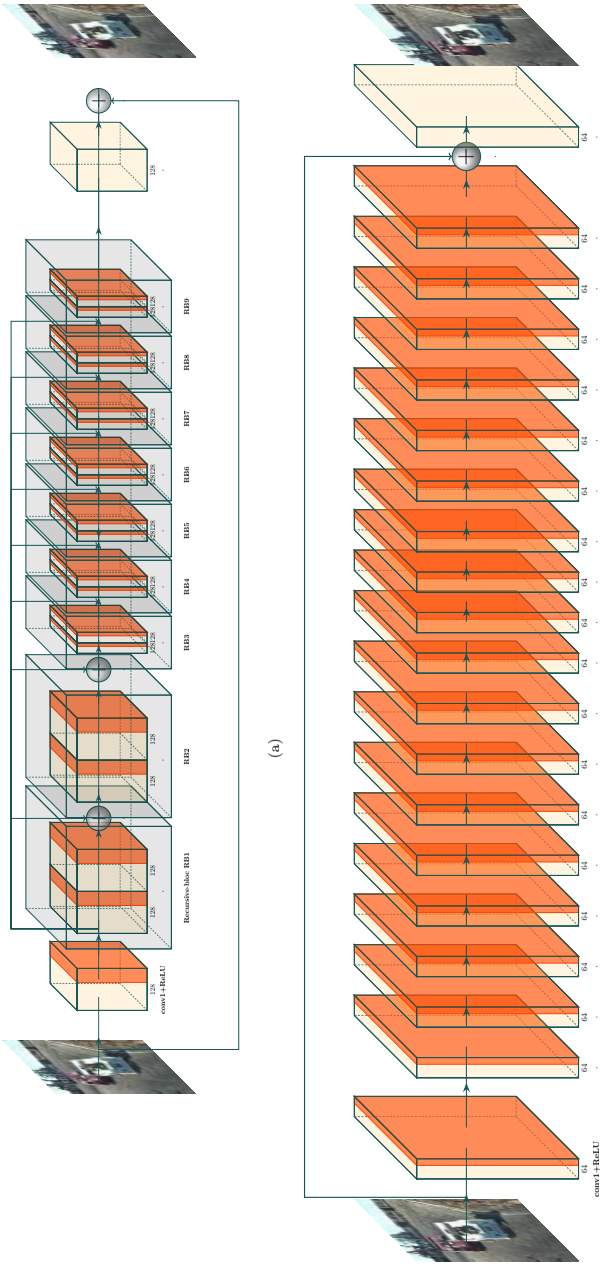


FIGURE 4.3 – Pré-échantillonnage : (a) VDSR [4]. (b) drrn [5]

(b)

(c)

FIGURE 4.4 – Post-échantillonnage : (a) ESPCN [6], (b) SRResNet [7], (c) edsr [8]

4.2.3 Génération 3 : Amélioration de la fonction de coût

Fonction de perte orientée-pixel : La majorité des modèles profonds de la SISR emploient des fonctions de perte dites «orientées pixel» (ou *pixel-wise* en anglais) qui comparent directement l'image reconstruite à la vérité terrain au niveau de la valeur des pixels. Ce type de perte intègre l'erreur quadratique moyenne (Norme $L2$ ou distance Euclidienne), ou l'erreur absolue (Norme $L1$ ou distance de Manhattan) dans l'optimisation de l'équation (4.2) en vue de sa relation avec la définition de l'indice PSNR (i.e. Peak Signal to Noise Ratio). Cependant, les majeures limitations de ces fonctions viennent de leur caractère global, dans la mesure où les différences entre pixels individuels sont prises en compte et comptabilisées par ces mesures. Cela peut conduire dans certains cas à des résultats aberrants. Par exemple, si l'on décale la valeur de chaque pixel de 3 unités, le résultat va paraître comme identique à la vérité terrain pour l'œil humain, mais sera fortement pénalisé par ce type de fonction.
De plus, cette fonction d'erreur ne tient pas compte de la qualité perceptive et texturale de l'image, et produit par conséquent des résultats flous ou perceptivement insatisfaisants avec des textures trop lisses. Les auteurs dans [102] mettent en avant la faculté de l'œil humain dans l'interprétation des informations structurelles d'une scène. Ainsi, il serait bien plus sensible aux défauts de structure dans une image altérée plutôt qu'à des variations de valeurs des pixels individuels. Dans ce contexte, les auteurs proposent un système d'évaluation alternatif au PSNR, nommé SSIM (pour Structural SIMilarity), permettant une bien meilleure fidélité quant à l'évaluation de la qualité d'une image par le système visuel humain. Cette fonction est alors définie comme décrit par l'équation (4.3) suivante :

$$Loss_{SSIM} = 1 - SSIM(\tilde{X}_i, X_i) \tag{4.3}$$

Les fonctions de perte orientées-pixel décrites par les équations (4.2) et (4.3) permettent d'évaluer rapidement les résultats de reconstruction, et fournissent un moyen simple et direct afin d'effectuer des comparaisons entre différentes méthodes

pour des objectifs d'étalonnage de l'ensemble des nombreux réseaux CNN proposés dans la littérature. Néanmoins, malgré la première avancée du SSIM, ces méthodes d'évaluation restent encore très objectives et sont encore très loin dans la retranscription de la qualité d'une image du point de vue de la perception humaine.

Fonction de perte perceptuelle : Étant donné que la perception visuelle humaine n'analyse pas les images pixel par pixel, une autre fonction de coût plus appropriée a été proposée par [103]. Le principe sous-jacent de cette fonction est de réaliser l'évaluation de la qualité des images selon leur qualité perceptive dans le domaine des paramètres du réseau au lieu du domaine spatial. Plus précisément, cette erreur mesure les différences sémantiques entre les images en utilisant un CNN profond de classification d'image pré-entraîné [2] [100]. L'exemple le plus utilisé, en raison de son efficacité et de sa disponibilité, est le réseau pré-entraîné VGG (i.e. *Visual Geometry Group*) illustré dans la figure 4.5. Ce réseau s'appuie sur une architecture de 5 blocs contenant des couches de convolution, suivies par des couches denses qui sont chargées de l'opération finale de classification. De leur côté, les couches de convolution réalisent des tâches d'extraction de caractéristiques (Figure 4.1), et agissent par conséquent comme un puissant extracteur de formes. Ainsi, il est possible de définir une fonction de coût perceptuelle en calculant la distance Euclidienne entre les caractéristiques obtenues au niveau des différentes couches d'un réseau VGG et celle obtenues par le réseau à entraîner, et l'erreur quadratique moyenne est calculée entre les *feature-maps* plutôt que les images naturelles. De cette manière, l'image reconstruite est contrainte à être visiblement similaire à l'image de référence. Une autre stratégie d'entraînement propose de minimiser l'écart entre les styles comme les textures et contrastes [104, 105]. Le résultat d'une telle optimisation présente des textures réalistes plus satisfaisantes visuellement. Néanmoins, la détermination de la taille du patch pour faire correspondre les textures reste empirique : des patchs trop petits entraînent des artefacts dans les régions texturées, tandis que des patchs trop grands entraînent des artefacts dans l'ensemble de l'image.

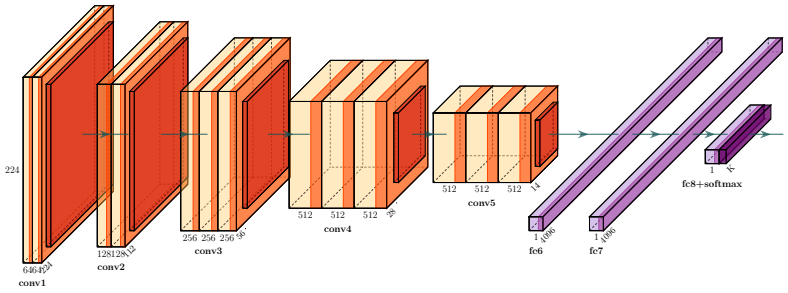


FIGURE 4.5 – Architecture du réseau VGG16 [2]

4.2.4 Génération 4 : Les réseaux génératifs et apprentissage non-supervisé

Les réseaux antagonistes génératifs (GAN) : Les réseaux antagonistes génératifs (en anglais GAN pour *Generative Adversarial Network*) sont un type d'architecture de réseaux de neurones offrant des performances dépassant complètement celles obtenues via les réseaux de neurones traditionnels. Ils ont été introduits par I. Goodfellow [106] et leur principe repose sur la minimisation de multiples fonctions avec des objectifs antagonistes. Deux réseaux CNN fonctionnent en parallèle et sont mis en concurrence. Un premier appelé le générateur, a pour objectif de générer, à partir d'un vecteur de variables aléatoires fourni comme entrée, une «fausse» image, qui est sensée approcher au maximum les images de l'ensemble d'apprentissage considéré. Le générateur vise ainsi à maximiser la probabilité que l'image synthétisée appartienne à la base des données d'apprentissage. Le second réseau appelé discriminateur est un classifieur binaire qui prend en entrée les données de la base d'apprentissage et des données de la sortie du générateur. Son objectif est de minimiser la probabilité d'appartenance à l'ensemble d'apprentissage pour les images synthétisées par le générateur. Intuitivement parlant, le réseau discriminateur fait office de l'expert qui s'entraîne à distinguer le vrai du faux, et le réseau générateur est un faussaire qui s'entraîne à «duper» l'expert. Les deux réseaux sont entraînés simultanément sur une même base d'apprentissage. Dans sa version originale, le réseau GAN produit des images à partir d'un bruit aléatoire. Mais il est tout à fait possible d'orienter son fonctionnement selon la tâche considérée en fournissant comme entrée des informations liées à l'image

que l'on souhaite générer. Ce mécanisme peut être perçu comme une opération de conditionnement du problème de reconstruction SR.

GAN appliqué à la SR : Si l'on considère le cas de la SR, l'entrée du générateur va représenter une version basse résolution de l'image à traiter. La première approche proposée dans ce contexte est dite SRGAN décrite dans [7]. Ici, les auteurs introduisent un réseau réalisant une super-résolution par apprentissage antagoniste d'un GAN (Figure 4.6).

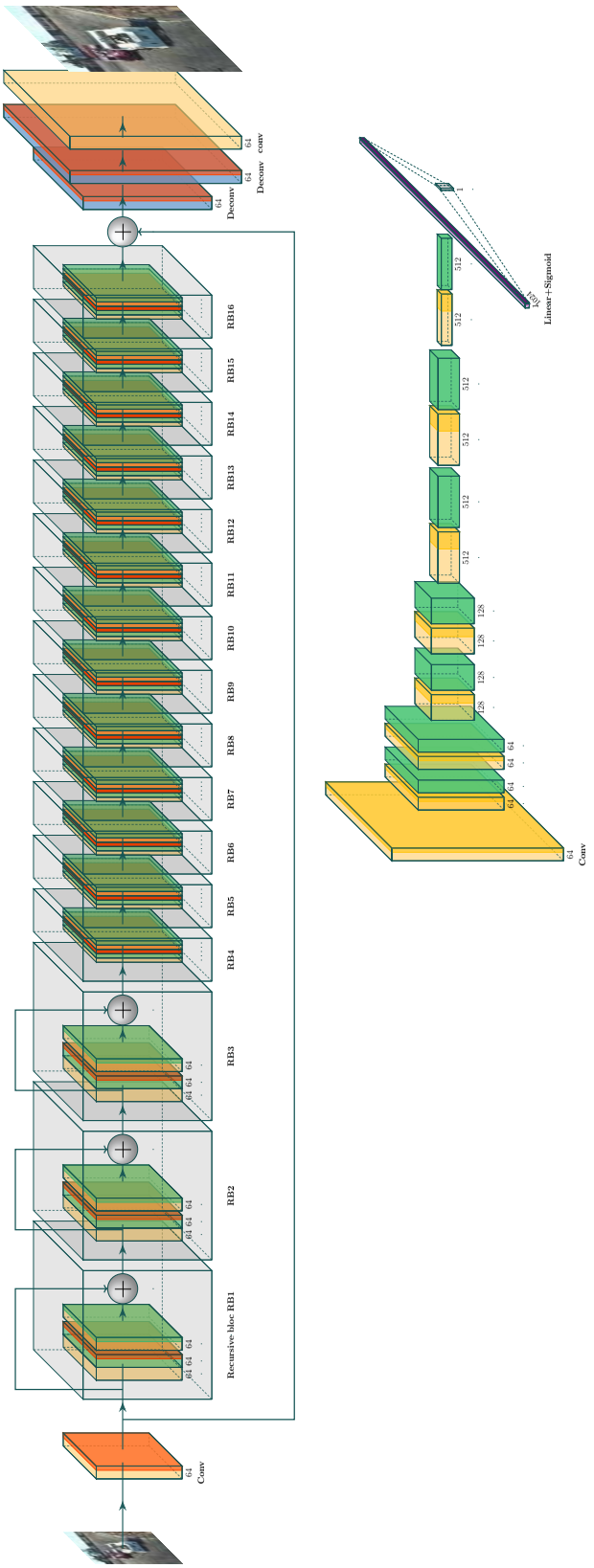


FIGURE 4.6 – Architecture de SRGAN [7]

Le générateur G possède une architecture très profonde avec un nombre donné de blocs résiduels composés d'une couche de convolution ayant 64 noyaux de taille 3×3 qui est suivi d'une couche de normalisation de batch puis enfin d'une fonction d'activation. Nous pouvons noter qu'une structure de post-échantillonnage est intégrée dans ce réseau, et que l'agrandissement est réalisé par deux couches de convolution sub-pixeliques successives.

Le discriminateur D possède une architecture de 8 couches de convolution avec un nombre de noyaux de taille 3×3 allant de 64 à 512 en étant multiplié par 2 à chaque fois en analogie au réseau VGG. Ici, une fonction de type LeakyRelu est utilisée pour l'activation. Après la succession de convolutions, deux couches denses sont appliquées suivies d'une couche d'activation sigmoïde qui permet d'obtenir la décision finale du réseau. La fonction de perte du réseau SRGAN intègre deux composantes distinctes propres aux deux réseaux G et D : la première étant la perte de contenu, en rapport avec le contenu traité par le générateur, et la seconde concernant la perte antagoniste, en rapport avec la décision du discriminateur :

$$Loss^{SR} = \underbrace{Loss^{SR}_{content}}_{\text{perte de contenu}} + \underbrace{10^{-2} Loss^{SR}_{Gen}}_{\text{perte antagoniste}} \tag{4.4}$$

Pour la perte du contenu, les auteurs proposent plusieurs versions. La MSE étant la mesure classique peut être utilisée comme suit :

$$Loss^{SR}_{content} = Loss^{SR}_{MSE} = \frac{1}{r^2 W H} \sum_{i=1}^{rW} \sum_{j=1}^{rH} \left(I^{HR}_{i,j} - G_{\theta_G} \left(I^{BR} \right)_{i,j} \right)^2 \tag{4.5}$$

Où r est le facteur d'agrandissement. La perte antagoniste est quant à elle définie par :

$$Loss^{SR}_{Gen} = \sum_{n=1}^{N} - \lg D_{\theta_D} \left(G_{\theta_G} \left(I^{BR} \right) \right) \tag{4.6}$$

Le réseau a été entraîné durant 10^5 itérations sur un ensemble de 350000 images extraites de la base de données ImageNet [107], et le facteur d'agrandissement

étudié est de $r = 4$.

Pour le modèle proposé, un important ensemble d'images d'évaluation a été considéré, incluant les bases d'évaluation Set5, Set14, BSD100 et BSD300 [108]. Enfin, les méthodes retenues pour comparaisons sont : interpolation par proches voisins, interpolation bicubique, SRCNN, DRCN [109], ESPCN, SRResNet-MSE, SRResNet-VGG22, SRGAN-MSE, SRGAN-VGG22, et SRGAN-VGG54 (le terme MSE dans la nomination signifie que le réseau a été entrainé avec MSE comme composante perte de contenu, tandis que VGG22 et VGG54 signifient que la perte a été calculée sur des couches du réseau VGG bas-niveau ($22^{iéme}$ couche) et plus profondes ($54^{iéme}$ couche) respectivement). L'intérêt d'entraîner les réseaux SRGAN et SRResNet avec des fonctions de pertes perceptuelles qui exploitent différents niveaux des réseaux VGG est de souligner la pertinence de la profondeur dans laquelle la convolution a été appliquée. L'évaluation des résultats a été réalisée selon trois critères : le PSNR, le SSIM, et une évaluation subjective notée MOS (pour Mean Opinion Score) conduite sur un ensemble de 26 personnes ayant été amenés à évaluer 1128 images chacune : pour cette étude, un score entre 1 (mauvaise qualité) et 5 (excellente qualité) a été attribué à chaque image. D'après les auteurs, les résultats les plus visuellement convaincants sont effectivement obtenus en utilisant un VGG54 plutôt que le VGG22, comme illustré dans la figure 4.7. Ce résultat est d'autant plus vérifié significativement par la MOS (Tableau4.1) où l'effet des fonctions de perte perceptuelle est notable. Néanmoins, les résultats montrent également des scores de PSNR et SSIM relativement élevés pour les SRResNet et SRGAN entraînés avec MSE comme fonction de perte. Nous reviendrons en détails sur l'explication de ce phénomène dans le Chapitre 5.

Les réseaux de type GAN surpassent quantitativement et qualitativement les réseaux CNN classiques du benchmark de la reconstruction SR (Tableau 4.2). Les méthodes de synthèse d'images naturelles, notamment les images générées par les réseaux GAN représentent cependant quelques faiblesses lorsqu'il s'agit d'entraîner le modèle à synthétiser des échantillons de haute qualité à partir de données complexes. D'une part, les SRGAN s'avèrent en pratique plus difficile à entraîner dans

	SRResNet-		SRGAN-		
Set5	MSE	VGG22	MSE	VGG22	VGG54
PSNR	**32.05**	30.51	30.64	29.48	29.40
SSIM	**0.9019**	0.8803	0.8701	0.8468	0.8472
MOS	3.37	3.46	3.77	**3.78**	3.58
Set14					
PSNR	**28.49**	27.19	26.92	26.44	26.02
SSIM	**0.8184**	0.7807	0.7611	0.7518	0.7397
MOS	2.98	3.15	3.43	3.57	**3.72**

TABLEAU 4.1 – Performance of different loss functions for SRResNet and the adversarial networks on Set5 and Set14 benchmark data. MOS score significantly higher (p ¡ 0.05) than with other losses in that category

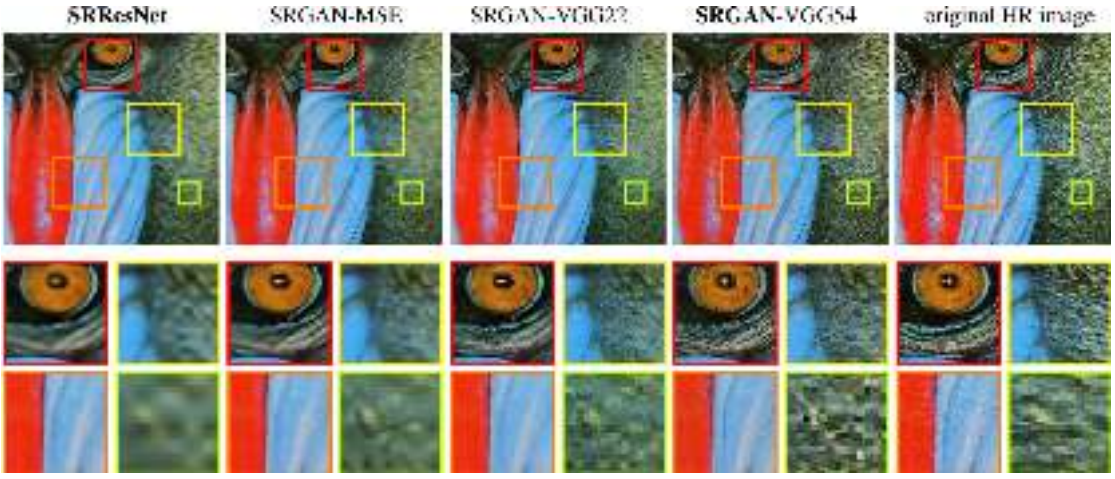


FIGURE 4.7 – Comparaison visuelle des différences de détails de reconstruction SR obtenues avec le SRGAN par calcul de coût respectif à l'utilisation de la MSE, et l'appel des réseau VGG22 et VGG54. Source [7]

la mesure où la convergence de l'algorithme d'optimisation durant l'apprentissage n'est pas stable à cause du caractère antagoniste de celui-ci. D'autre part, ces réseaux rencontrent des problèmes majeurs qui concernent les aspects de fidélité du rendu. En effet, même si d'un point de vu visuel l'image est beaucoup plus plaisante, ce type de traitement peut facilement conduire à la destruction ou au remplacement d'éléments critiques présents dans l'image.

Set5	ppv	bicubique	SRCNN	DRCN	ESPCN	SRResNet	SRGAN	HR
PSNR	26.26	28.43	30.07	31.52	30.76	**32.05**	29.40	
SSIM	0.7552	0.8211	0.8627	0.8938	0.878	**0.9019**	0.8472	1
MOS	1.28	1.97	2.57	3.26	2.89	3.37	**3.58**	4.32
Set14								
PSNR	24.64	25.99	27.18	28.02	27.66	**28.49**	26.02	
SSIM	0.7100	0.7486	0.7861	0.8074	0.8004	**0.8184**	0.7397	1
MOS	1.20	1.80	2.26	2.84	2.52	2.98	**3.72**	4.32
BSD100								
PSNR	25.02	25.94	26.68	27.21	27.02	**27.58**	25.16	
SSIM	0.6606	0.6935	0.7291	0.7493	0.7442	**0.7620**	0.6688	1
MOS	1.11	1.47	1.87	2.12	2.01	2.29	**3.56**	4.46

TABLEAU 4.2 – Résultats des évaluations quantitatives des modèles profonds de reconstruction SISR sur les bases Set5, Set14, et BSD100. Les meilleures performances sont indiquées en gras.

4.3 Synthèse et conclusion

Si l'on revient à notre axe de recherche, compte tenu des observations faite sur l'analyse de l'état de l'art des réseaux CNN appliqués à la SR, nous remarquons que les éléments du conditionnement du problème mal-posé de la SR qui influent sur le résultat sont la base de données d'apprentissage et la stratégie d'entraînement du réseau considérées. Bien que l'opération d'extraction des caractéristiques soit apprise durant la phase d'entraînement, les CNNs que nous avons vu ont tous une architecture générique, ce qui signifie que la connaissance à priori sur l'image à reconstruire est déduite seulement à partir des données d'apprentissage. En plus, la dégradation d'une image dépend du capteur utilisé lors de l'acquisition, par conséquent, différents modèles peuvent s'avérer plus adaptés que d'autres selon le type de dégradation du capteur. Si l'on se remet à la formulation générale de la reconstruction SR, exprimée par l'équation (1.2), nous remarquons que les propriétés de la base de données d'entrainement interviennent directement dans le choix des termes de cette formule. D'une part, le terme de l'attache des données est lié au modèle de dégradation A imposé durant la génération de la base de données d'entraînement, qui, dans la plupart des modèles, ne respecte pas les caractéristiques du capteur ni la nature de la scène au moment

(a) (b)

FIGURE 4.8 – Base de donnée d'images externes (b) qui ne correspondent pas au contenu de l'image de test (a)

de l'acquisition, et d'autre part, le terme de régularisation, défini par l'information *a priori*, contraint la solution de la reconstruction uniquement aux images de la base de donnée d'entraînement. Ces deux problématiques relèvent le caractère générique des approches basées sur l'apprentissage profond, ainsi que leur dépendance aux données d'entraînement et à l'information à priori extraite à partir de celles-ci. Plus précisément, quelque soit la disparité des données d'apprentissage, ces réseaux CNN apprennent une «manière» d'inverser le modèle général de dégradation prédéfini durant la génération de celles-ci. L'intégralité des travaux vus précédemment s'appuient sur des bases de données externes dans l'étape d'apprentissage de leurs réseaux CNN. Ces dernières sont constituées d'un grand nombre d'images naturelles telles que des photographies, des images de paysages naturels, d'animaux, etc. Malgré leurs tailles et leurs disparités, l'inconvénient majeur de ces bases de données d'entraînement des réseaux profond est leur non-exhaustivité ; La non-congruence entre l'information HR des images de la base de données, et celle de l'image de test mène à une extraction d'information à priori qui peut ne pas correspondre à la structure de l'image de test. La figure 4.8 montre un exemple où le contenu d'une image de test ne correspond pas au contenu de la base d'entraînement.

Pour pallier partiellement ce problème, des méthodes récentes ont eu recours à la

collecte de très larges base de données d'entrainement, couvrants une variété de scènes différentes [Timofte et al, Dong et al]. Bien que cela puisse améliorer la qualité de la reconstruction, l'utilisation de grandes base de données d'entraînement externes implique souvent l'emploi d'algorithmes d'apprentissage élaborés et coûteux en terme de calcul. L'exemple illustré dans la figure 4.8 montre l'intérêt de l'utilisation d'une base de données interne dans le processus d'apprentissage. Une autre technique consiste à générer une base à part entière à partir de l'image de test, éventuellement à partir de plusieurs échelles de celle-ci. Son principe est similaire à celui de l'extraction de l'a priori semi-local discuté dans le paragraphe 3.1.2.2 où redondance de l'information à travers différentes échelles suggère que la version HR d'un patch d'une image existe également dans la même image. Cela donne lieu à un a priori puissant qui peut être exploité par la SR sans avoir besoin d'une base de données d'entraînement externe [9].

Deuxième partie

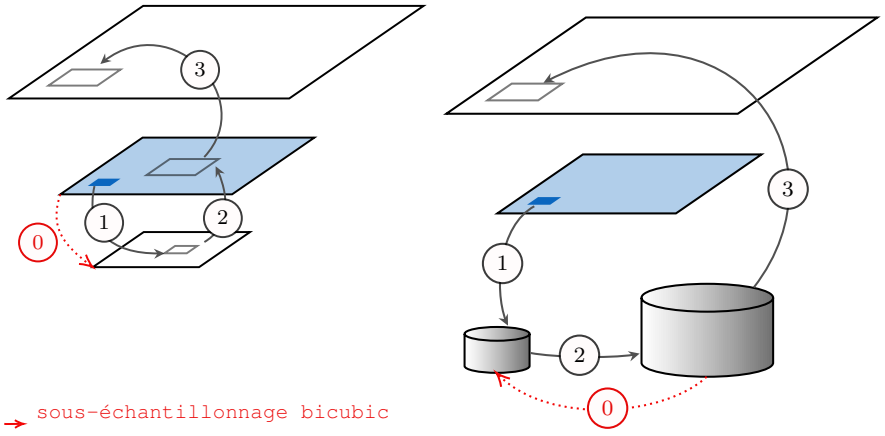
Contribution et Réalisation

Chapitre 5

Super-Résolution par autosimilarité

5.1 Introduction

Les techniques de super-résolution basées sur l'autosimilarité sont régies par le fait que les images ont tendance à contenir des patchs redondants à travers différentes échelles. Cette récurrence de patches suggère que la version SR d'un patch donné de l'image d'entrée peut être obtenue directement à partir de l'image elle-même (Figures 5.1(a)). Cela donne lieu à une propriété statistique puissante qui peut être exploitée pour la reconstruction SR sans avoir recours à une base de données d'entraînement externe. L'exploitation de ce type d'information dans les techniques du DL est possible à condition que l'apprentissage automatique se fasse sur la redondance interne de l'image de test sans utiliser des images venues de l'extérieur (dictionnaires externes). Dans ce chapitre, nous nous proposons d'étudier l'effet des dictionnaires d'entrainement internes et externes sur les performances des réseaux SRCNN. Pour ce faire, une première piste d'étude concerne d'abord la description de la densité interne des patchs dans une image naturelle afin de quantifier leur récurrence à travers différentes échelles. Ensuite, les propriétés de la récurrence interne sont exploitées dans l'apprentissage d'un réseau profond de la reconstruction SR [110]. Les résultats de validation sont présentés dans la fin de ce chapitre à travers une étude comparative à visée explicative qui fait également intervenir d'autres paramètres de l'apprentissage profond d'un CNN, à savoir sa fonction objective et sa base de données d'apprentissage ainsi que le noyau de dégradation intervenant dans la génération de celle-ci.



(a) (b)

FIGURE 5.1 – Reconstruction SR par autosimilarité multi-échelle (a) ; reconstruction SR par analyse d'exemples de Dictionnaire externe (b).

5.2 Quantification des statistiques internes

Selon [9], les images naturelles ont une forte redondance interne des données. Par exemple, des patchs de taille 5×5 admettent plusieurs patches similaires à la même échelle ainsi qu'à des échelles plus grandes ou plus petites de l'image. L'image naturelle est majoritairement constituée de patches uniformes dont la structure est plutôt lisse (Figure 5.2(a)). Ces patches contiennent les zones homogènes de l'image, comme les éléments de surface, et ont une continuité dans l'intensité de leurs pixels. Parallèlement, seulement un faible nombre de patches représente la discontinuité de l'irradiance de l'image et des intensités de leurs pixels. Ces patches sont appelés patches contrastés (Figure 5.2(b)) et contiennent des détails texturaux de l'image comme les contours, et leur discontinuité de l'irradiance s'exprime par la magnitude du gradient (i.e. première dérivée). De manière générale, les patches uniformes réapparaissent beaucoup plus fréquemment dans l'image que les patches contrastés.

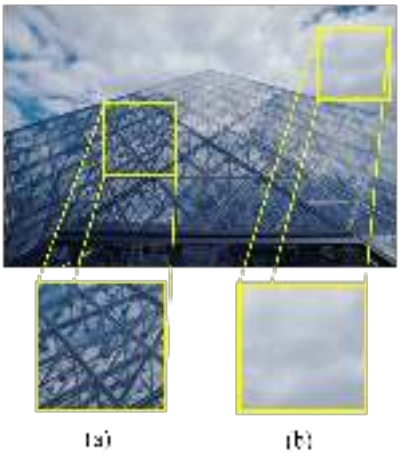


FIGURE 5.2 – Patch uniforme (b) et patch contrasté (a)

Ces observations sont empiriquement vérifiées à travers une quantification de la récurrence des patches lisses et contrastés. Dans un premier temps, la densité d'un patch donné $\boldsymbol{p}$ de taille 5×5 est estimée par rapport à tout patch de son voisinage par la méthode non-paramétrique d'estimation par noyau (ou encore méthode de Parzen-Rosenbaltt, en anglais Kernel Density Estimation ou KDE), donnée par :

$$d(p, dist) = \frac{1}{N} \sum_{p_k \in \mathcal{N}_{dist}} \mathcal{K}_h \left(\|\mathrm{p} - \mathrm{p_k}\|_2^2 \right) = \frac{1}{hN} \sum_{p_k \in \mathcal{N}_{dist}} \mathcal{K} \left(\frac{\|\mathrm{p} - \mathrm{p_k}\|_2^2}{h} \right) \tag{5.1}$$

Où $\boldsymbol{p_k}$ représente les patches du voisinage $\boldsymbol{N_{dist}}$ de $\boldsymbol{p}$ de taille $\boldsymbol{N}$ et de diamètre $\boldsymbol{dist}$. $\boldsymbol{\mathcal{K}}$ est le noyau de l'estimation communément choisi comme la densité d'une fonction gaussienne centrée réduite donnée par : $\boldsymbol{\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}}$. La constante $\boldsymbol{h}$, appelée la fenêtre du noyau, contrôle le degré de lissage de l'estimation. En faisant varier le nombre d'échantillons inclus dans l'estimation (c'est-à-dire le diamètre $\boldsymbol{dist}$) de part et d'autre du patche, nous remarquons que la densité estimée de celui-diminue à mesure que la distance augmente (i.e. Figure 5.3).

Ce résultat est intuitif dans la mesure où, dans une image naturelle, une grande distance spatiale entre les patches implique une distance quantitative réduite. Si en plus la structure du patch (uniforme, contrasté) est considérée dans la

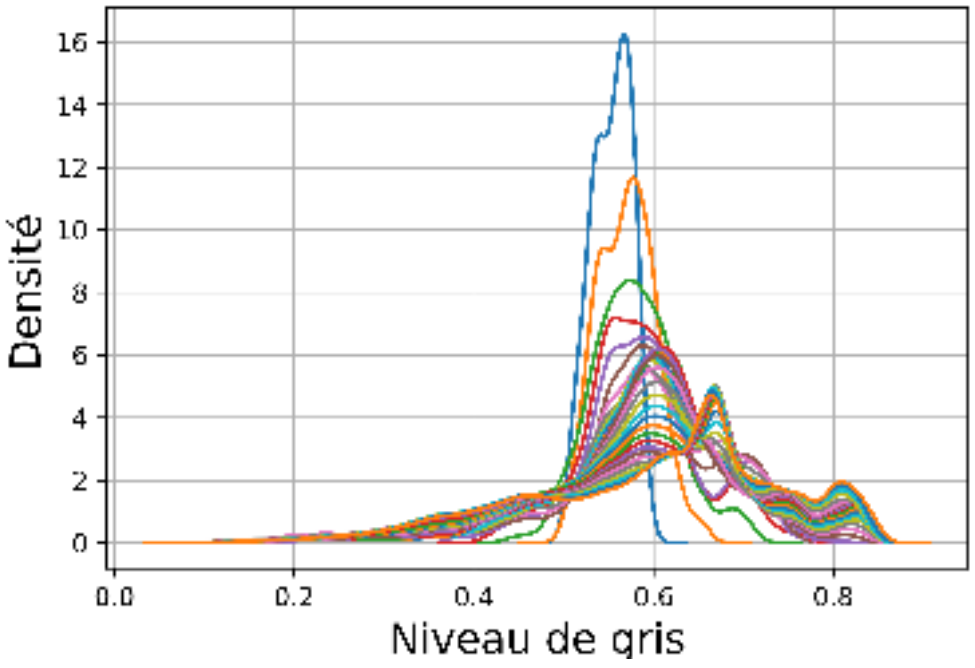


FIGURE 5.3 – Tracé des densité $d(p, dist)$ d'un patch avec variation de la distance spatiale $dist$.

représentation de la densité par la magnitude de son gradient, la formule de la densité devient :

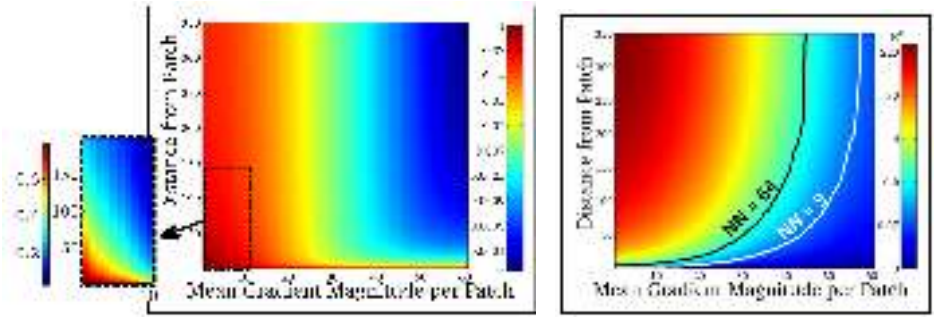
$$D(dist, |grad|) = Moyenne_{p_j \in N_{|grad|}} d(p_j, dist) \tag{5.2}$$

Où $\boldsymbol{N_{|grad|}}$ est l'ensemble des patches ayant la même magnitude du gradient. Ainsi, le nombre moyen des voisins les plus proches NN peut être défini par :

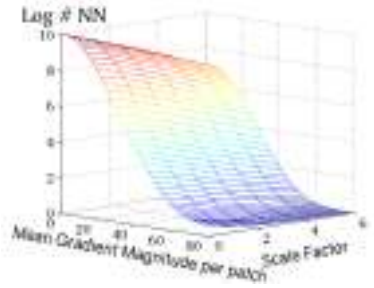
$$NN(dist, |grad|) = D(dist, |grad|) \times N_{p_j} \tag{5.3}$$

Où N_{p_j} représente le nombre de pixels dans un voisinage j. Les figures 5.4(a) et 5.4(b) montrent le tracé des valeurs $\boldsymbol{D(dist, |grad|)}$ ainsi que le nombre des plus proches voisins $\boldsymbol{NN}$ en fonction de la moyenne des magnitudes du gradient des patches ainsi que de la distance spatiale qui définit le rayon de l'espace de recherche.

Les résultats montrent que la densité moyenne des patches augmente à mesure que la distance spatiale est la moyenne des magnitudes du gradient diminuent. Ceci indique qu'un patch uniforme possède une plus forte récurrence, notamment dans son voisinage proche, qu'un patch contrasté, et que sa densité diminue rapidement à mesure que l'on s'éloigne de son voisinage. Les patches contrastés (i.e. de $|grad|$ important) ne bénéficient pas du pouvoir de récurrence présent dans l'image, ou dans un ensemble d'images externes. Dans le cas où plusieurs échelles (Scale factor r) des images sont considérées dans l'estimation 5.4(c), on remarque que le nombre de patches identiques et contrastés tend vers zéro lorsqu'on considère une seule échelle. Tandis que ce nombre augmente à mesure que le facteur d'échelle change. Nous pouvons conclure que les bases de données internes qui considèrent plusieurs facteurs d'échelle possèdent plus d'information utile à la reconstruction que d'autres bases externes notamment pour les patches contrastés.



(a) $D(dist, |grad|)$ (b) $NN(dist, |grad|)$



(c) $NN(r, |grad|)$

FIGURE 5.4 – Tracé des valeurs de densité (a) et des NN plus proches voisins (b) de patchs en fonction de la distance spatiale *dist* et de la moyenne de la magnitude du gradient. (c) représente le tracé du nombre des plus proches voisin en fonction de la moyenne de la magnitude du gradient avec variation d'échelle. Les patchs sont extraits de la base BSD300. Source [9]

5.3 Dictionnaire Interne Vs. Externe

Basée sur les observations faites dans la section 5.2 qui stipulent que la densité du patch est significative dans son voisinage plutôt que dans des images externes, et que l'information texturale est mieux préservée dans la structure multi-échelle de l'image, l'expérimentation suivante tente d'introduire uniquement l'information locale multi-échelle dans les dictionnaires d'entraînement. Ces observations sont davantage intéressantes pour l'entraînement d'un réseau profond. Il convient de constater que l'information complexe comme la texture et les contours est mieux préservée dans les bases de données internes. Cette propriété peut être considérée comme un a priori puissant sur la texture de l'image, et peut également régulariser l'apprentissage d'un réseau CNN de reconstruction. Nous allons, dans ce qui suit, décrire la chaîne algorithmique qui permet d'exploiter une telle propriété.

Nous proposons un scénario de génération d'un tel ensemble qui fera intervenir

une technique appelée « augmentation des données » permettant de construire une large base d'image à partir d'une seule entrée en créant des versions légèrement modifiées de celle-ci. Ainsi, plusieurs variations sont rajoutées dans le jeu de données. Les différences entre les images générées sont exprimées par des redimensionnements aléatoires, des symétries verticales et horizontales, des translations, des rotations, et des changements de luminosité. La figure 5.5 montre des exemples d'images obtenues par augmentation des données à partir d'une seule observation.

FIGURE 5.5 – Exemple d'augmentation des données à partir d'une seule image de test

5.4 Architecture proposée

Comme il a été mentionné précédemment, la majorité des méthodes du benchmark des réseaux CNN appliqués à la SR tentent de reconstruire l'image de test en ayant recours, durant son apprentissage, à des bases de données externes sans accéder aux caractéristiques internes réelles de celle-ci lors de la phase de test. Toutefois, nous avons pu démontrer dans les sections 5.2 et 5.3 que l'a priori semi-local est plus puissant que celui apporté par les bases de données externes, et ce malgré leurs tailles. Une autre constatation intéressante est faite autour de l'emploi d'un paramètre de dégradation estimé lors du processus de génération des données : Un noyau de dégradation adapté représente mieux les structures internes de l'image, notamment les textures et les contours, plutôt qu'un noyau de

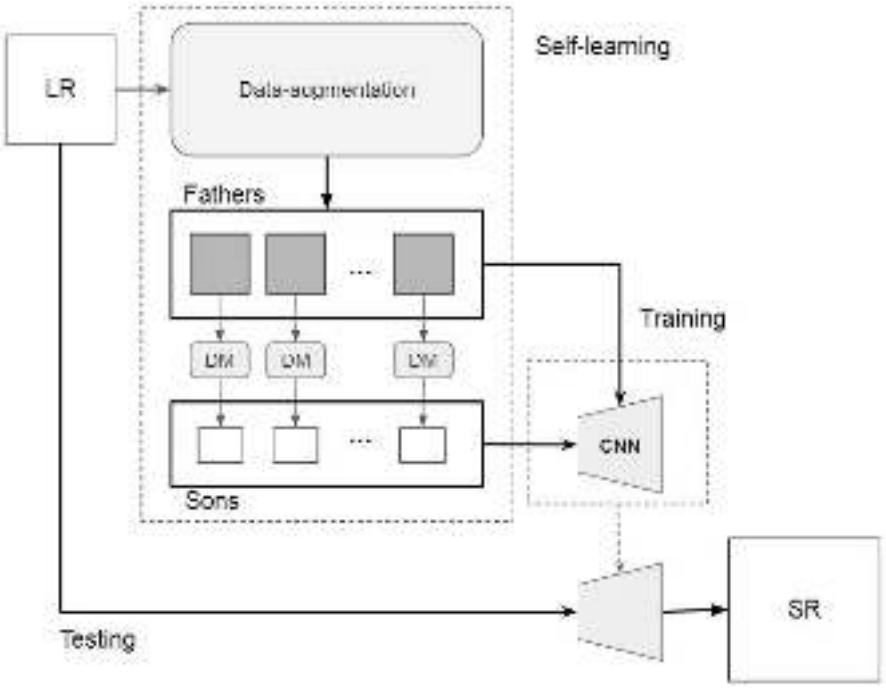


FIGURE 5.6 – Schémas synoptique de notre architecture proposée.

dégradation idéal (i.e. un noyau bicubique). Son estimation permet donc de tirer davantage profit de la récurrence interne des patches, pour ensuite l'introduire dans le processus d'apprentissage du réseau. Dans cette section, nous présenterons dans un premier temps le modèle SR proposé comme le montre la figure 5.6. Notre méthode est principalement divisée en trois étapes : La première consiste à estimer la dégradation de l'image de test et de générer sa version sous-échantillonnée. La deuxième étape consiste à considérer la dégradation spécifique à l'image dans le processus d'apprentissage du réseau en incorporant le noyau estimé dans le pipeline de l'augmentation des données. Enfin, la troisième étape consiste à entrainer un réseau CNN sur ces données reconstruites.

Pour obtenir des performances sur la tâche de reconstruction SR par autosimilarité, il est nécessaire d'effectuer d'abord une mise-en-place d'un protocole expérimental par la définition de la topologie du réseau et des paramètres qui conditionnent son apprentissage.

5.4.1 Préparation des données

5.4.1.1 Estimation du noyau de dégradation

Selon Shocher et al [111], un noyau de dégradation correcte est celui qui maximise la distribution des patches dans une image à travers ces différentes échelles. Dans ce contexte, nous employons le réseau KernelGAN [112] pour estimer directement le noyau spécifique à l'image d'entrée, que l'on note $\boldsymbol{K_{SR}}$. Ce réseau utilise un GAN pour estimer $\boldsymbol{K_{SR}}$ en maximisant la distribution des patches LR internes dans une structure multi-échelles.

L'approche de l'estimation du noyau $\boldsymbol{K_{SR}}$ par un CNN peu profond entrainé uniquement sur l'image de test. Le réseau KernelGAN est composé d'un réseau générateur $\boldsymbol{G}$ réducteur d'échelle et d'un réseau discriminateur $\boldsymbol{D}$. $\boldsymbol{G}$ et $\boldsymbol{D}$ sont tous les deux des CNN entièrement convolutifs, ce qui suggère que c'est derniers traitent des patches plutôt que des images entières. Soit une image d'entrée de basse résolution notée $\boldsymbol{I_{BR}}$, le réseau $\boldsymbol{G}$ apprend à réduire son échelle pour produire une image dégradée $\boldsymbol{I_{BR\downarrow}}$ de sorte à ce que le discriminateur $\boldsymbol{D}$ ne puisse pas à faire la distinction entre les vrais patches, notés $\boldsymbol{p_y^{test}}$, et ceux de l'image $\boldsymbol{I_{BR\downarrow}}$ notés $\boldsymbol{p_y^G}$ (Figure 5.7). Contrairement à l'implémentation classique de réseau SRGAN, le réseau $\boldsymbol{D}$ ne génère pas une image super-résolue $\boldsymbol{I_{SR}}$, mais s'entraîne à générer une matrice appelée carte de référence (*Reference-map* en anglais) que l'on note $\boldsymbol{D_{map}}$ et qui indique pour chaque pixel la probabilité que le patch voisin soit de même distribution. Le réseau KernelGAN tente de maximiser la similarité entre les patches de deux différentes échelles en minimisant la distance Euclidienne entre la $\boldsymbol{D_{map}}$ et une matrice de labélisation binaire : $'1'$ pour les patches extraits à partir de l'image d'entrée $\boldsymbol{I_{BR}}$, et $'0'$ pour les patches extraits de $\boldsymbol{I_{BR}}$ dans une échelle réduite.

La fonction objective du réseau KernelGAN est donnée par :

$$\mathrm{G}^*(\mathrm{I}_{BR}) = arg \min_G \max_D \left\{ \mathop{\mathrm{E}}_{X \sim patches(\mathrm{I}_{BR})} [|D(X) - 1| + |D(G(X))|] \; + \; \mathcal{R} \right\} \quad (5.4)$$

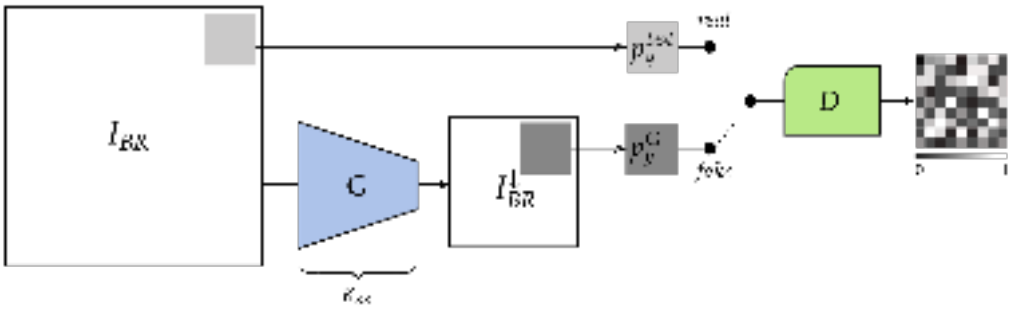


FIGURE 5.7 – Architecture du réseau KernelGAN.

Où $\mathcal{R}$ est un terme de régularisation qui intervient sur le générateur $\boldsymbol{G}$. Le noyau de dégradation $\boldsymbol{K_{SR}}$ qui préserve le mieux la distribution des patchs sur une image $\boldsymbol{I_{BR}}$ est implicitement déterminé à partir des couches de convolution du réseau G après sa convergence.

5.4.1.2 Génération du jeu de données à partir de l'image

Les bases de données d'apprentissage existantes destinées à la reconstruction SR (DIV2k [113], BSD300 [108], BSD500 [114], Net14 [46]) représentent de larges ensembles de paires d'images : des images BR de basse résolution, et leurs versions de haute résolution HR qui supervisent l'apprentissage. Ces dernières sont des images naturelles et photographiques généralement suffisantes pour supporter la diversité des images de tests. Elles sont artificiellement dégradées pour générer les images BR par un noyau de dégradation bicubique, communément utilisé dans cette opération (Figure 5.8(a)). Les paires (HR, BR) sont ensuite utilisées dans l'apprentissage automatique du réseau profond. Notre base de données d'apprentissage est créée à partir de l'image de test $\boldsymbol{I_{BR}}$ uniquement. Cette dernière est sujette à une procédure d'augmentation des données selon un pipeline défini afin d'ajouter certaines variétés aux données sans affecter la sémantique de l'image. Des opérateurs de rotations, de symétrie verticale et horizontale, des zooms et des recadrages aléatoires sont introduits dans ce processus d'augmentation des données.

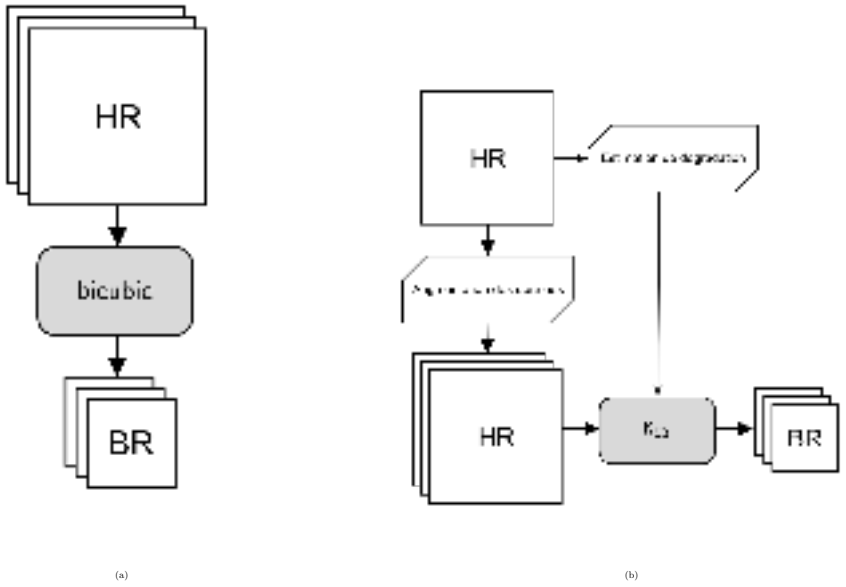


FIGURE 5.8 – Mécanismes de préparation des données d'apprentissage des bases de données externes. Les images HR sont dégradées par : un noyau bicubique idéal pour générer des paires d'image (a) ; par un noyau de dégradation adapté estimé directement à partir de l'image de test (b).

Ces opérations n'effectuent aucune interpolation sur l'image, et l'intensité effective reste préservée. Les images résultantes de cette chaine algorithmique sont appelées Pères HR et sont définies par l'ensemble :

$$\mathrm{I}^{BR}_{pères} = (\mathrm{I}^{BR}_0, \mathrm{I}^{BR}_1, \mathrm{I}^{BR}_2, \cdots, \mathrm{I}^{BR}_n)$$

Les images pères $\boldsymbol{I^{BR}_{pères}}$ sont considérées comme les images de référence qui assurent la supervision de l'apprentissage du réseau. À noter que chaque image de cet ensemble a la même taille que l'image d'entrée $\boldsymbol{I_{BR}}$.

chaque image de $\boldsymbol{I^{BR}_{pères}}$ est dégradée à l'aide d'un noyau estimé spécifique à l'image père ; appelé $\boldsymbol{K_{pères}}$ pour obtenir les images filles qui constituent les images d'entrée du réseau, à savoir :

$$\mathrm{I}^{BR}_{filles} = (\mathrm{I}^{BR\downarrow}_0, \mathrm{I}^{BR\downarrow}_1, \mathrm{I}^{BR\downarrow}_2, \cdots, \mathrm{I}^{BR\downarrow}_n)$$

Le noyau $\boldsymbol{K_{pères}}$ est directement estimé à partir de chaque image père $\boldsymbol{I_i^{BR}}$ à l'aide d'un KernelGAN défini dans la section 5.4.1.1 Le processus de dégradation est effectué en convoluant chaque image père avec son noyau estimé en veillant à ce que la sortie de l'opération de convolution s'adapte bien au facteur d'échelle défini sans interpolation afin de préserver l'intensité réelle du pixel. La base de données d'apprentissage résultante consiste en un ensemble de paires $\boldsymbol{(I_i^{BR\downarrow}, I_i^{BR})}$ adaptées à l'image de test (Figure 5.8(b)). Le réseau CNN sera ensuite entraîné de manière stochastique sur ces paires. L'algorithme 1 montre la génération des données $\boldsymbol{(I_i^{BR\downarrow}, I_i^{BR})}$ à partir de l'image $\boldsymbol{I^{BR}}$.

Algorithm 1 Génération du jeu de données à partir d'une seule image d'entrée I^{BR}

$rand(a, b)$ renvoie une valeur tirée aléatoirement selon une loi uniforme entre a et b
applyTransform(T, I) renvoie une image résultante de l'application du vecteur de transformation T sur l'image I
$KernelGAN(I)$ renvoie un noyau (matrice de taille 17×17) estimé à partir de l'image I
Paramètre p indique la probabilité d'application de la transformation sur l'image

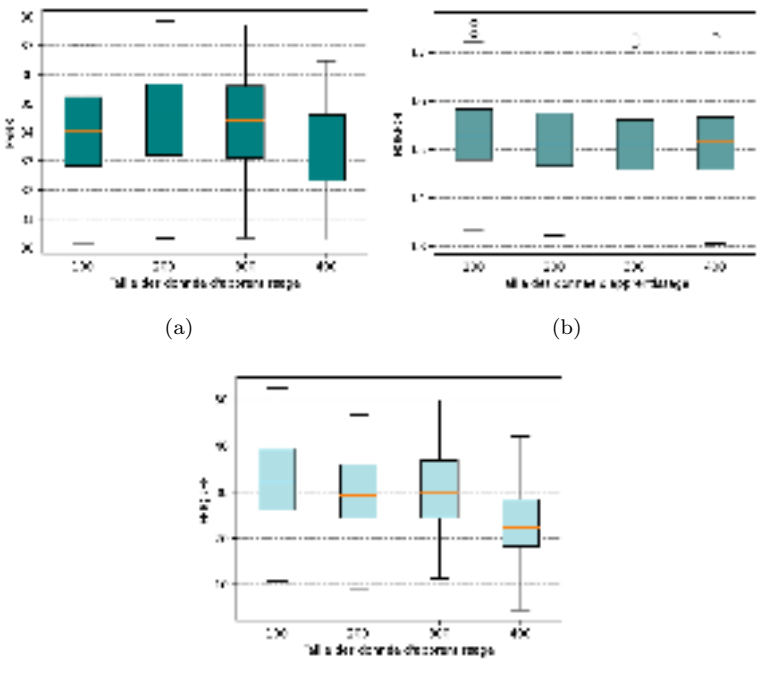
$degree1 \leftarrow rand(10, 100)$
$degree2 \leftarrow rand(0.3, 0.7)$
$trfm \leftarrow [rotate(degree1, p = 0.75), flipLr(), brightness(p = 0.3), shear(degree2, p = 0.6)]$
Entrée: $Nb_{pères}$, I^{BR} ▷ Nombre maximal des données
Sortie: $I_i^{BR\downarrow}$ ▷ image sous-échantillonnée
for $i \leftarrow 1$ to $Nb_{pères}$ **do**
$I_i^{BR} \leftarrow applyTransform(trfm, I^{BR})$
$K_i^{père} \leftarrow kernelGAN(I_i^{BR})$
$I_i^{BR\downarrow} \leftarrow K_i^{père}$
end for
return $I_i^{BR\downarrow}$

5.4.1.3 Variation de la taille des données

La généralisation des réseaux CNN profonds est principalement liée à la taille du jeu de donnée d'entraînement, dans la mesure où la disponibilité d'un nombre important et suffisant des paires $(\boldsymbol{I_i^{BR\downarrow}, I_i^{BR}})$ durant l'apprentissage ramène le

réseau à mieux reconstruire les images en SR : Une quantité de données d'entraînement trop importante rend le modèle profond particulièrement sujet au sur-apprentissage, tandis que la situation inverse provoque une mauvaise adaptation du modèle aux données. Dans cette mesure, nous élaborons une étude empirique qui explore l'impact de la variation de la taille du jeu de données d'apprentissage sur le comportement du modèle.

Dans un premier temps, nous avons généré selon l'Algorithme 1 quatre jeux de données de tailles différentes avec $\boldsymbol{Nb_{pères}} = 100, 200, 300, et 400$. Notre première expérimentation consiste en l'apprentissage du réseau CNN sur chacun de ces jeux séparément. Notons que la stratégie d'apprentissage de ce dernier ainsi que sa profondeur et son paramétrage correspondent, dans cette étude, à la fixation faite et décrite dans les parties 5.4.2 et 5.4.4 Nous avons également généré 200 paires d'images de test selon le même algorithme de génération. Les scores PSNR, NOREQI (NO-REference image Quality Index) [115] et FRIQUEE (Feature maps based Referenceless Image QUality Evaluation Engine) [116] ont été calculés pour évaluer la prédiction des quatre modèles sur ces 200 images de test (Figure 5.9).



(a) (b)

(c)

FIGURE 5.9 – Résultats des scores PSNR, NOREQI et FRIQUEE sur 200 images de test des modèles entraîné sur des jeux de données de différentes tailles

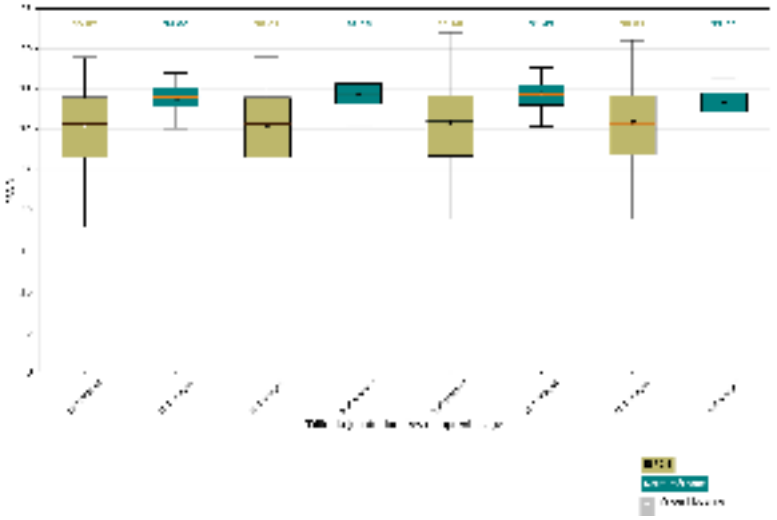
Le modèle entraîné sur un jeu de taille réduite (100 images) parvient parfaitement à s'adapter aux données d'apprentissage aussi bien que le modèle entraîné sur un jeu de 400 images. Selon la figure 5.9, et indépendamment de la taille des données d'entrainement, les scores présentés sont bons, et le réseau admet une convergence précoce ainsi qu'une adaptation rapide aux données. Ce comportement équivoque peut révéler l'un des cas suivants :

(i) Soit la régularisation faite par la fonction de coût et l'intégration de l'a priori du noyau de dégradation (selon notre stratégie d'apprentissage proposée) a permis de contraindre le modèle et ainsi alléger le besoin aux données ;

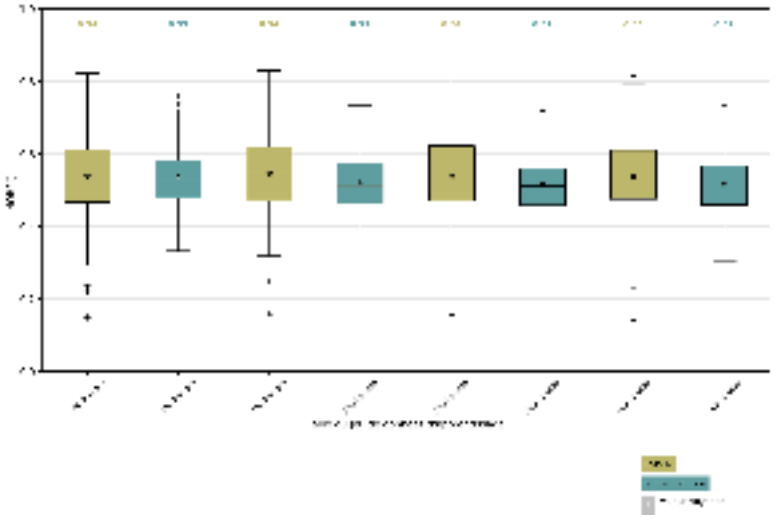
(ii) Ou alors ce résultat suggère une situation de sur-apprentissage du réseau ou d'une erreur de généralisation trop importante causée par le nombre réduit des données d'apprentissage.

Vérifions d'abord que le réseau n'ait pas simplement appris les données "par cœur". Ainsi, la deuxième expérience à réaliser a pour but de déterminer les causes qui ont abouties à de tels résultats, en exploitant particulièrement le cas (ii), par un apprentissage classique du réseau ESPCN du benchmark de la SISR sur des jeux de données externes (DIV2K) de tailles 100, 200, 300, et 400 images respectivement. Nous avons choisi le modèle ESPCN car son architecture correspond à celle de notre modèle tel qu'il est expliqué dans la section 5.4.2. La MSE représente la fonction coût qui mène l'apprentissage. De cette manière, on pourra mesurer et confronter les résultats de performance d'un réseau en fonction de sa stratégie de régularisation. Nous nous fions au paramétrage de base du modèle ESPCN fixé de telle sorte à ce qu'il n'y ait pas de sur-apprentissage du modèle. Les résultats d'évaluation sur 200 images de tests (issues de la base DIV2K) sont présentés dans la figure 5.10. Ces derniers montrent que ce n'est pas la taille du jeu de donnée qui influe sur la stabilité du modèle durant son apprentissage mais la stratégie avec laquelle ce dernier est mené. En effet, nous remarquons que les performances des différents modèles des deux expérimentations sont proches en sachant que le modèle ESPCN soit parfaitement généralisé. On peut donc rejeter l'hypothèse du sur-apprentissage et affirmer que l'*a priori* intégré dans la régularisation est

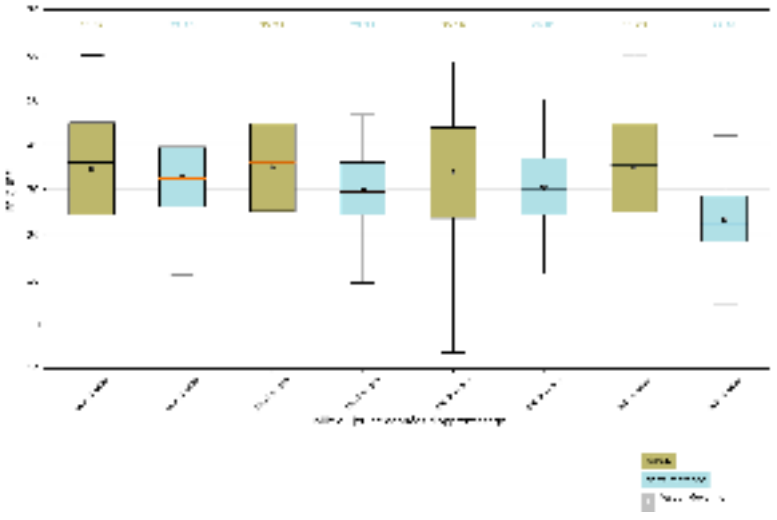
responsable de la stabilité de l'entraînement à travers un nombre réduit de données. La taille du jeu de données retenue dans nos expérimentations est fixée à 400 images. Les images générées du corpus de données d'entrainement sont de largeur et hauteur fixes de 100×100 pour les images BR, et de 200×200 pour les images HR, ce qui correspond à un facteur d'agrandissement de 2.



(a)



(b)



(c)

FIGURE 5.10 – Statistiques des scores PSNR, NOREQI et FREQUEE de notre modèle proposé entraîné sur 400 images, contre le réseau ESPCN [6] entraîné sur 50000 images

5.4.2 Architecture du réseau CNN

Selon la taille retenue de l'ensemble généré des données d'entraînement, nous avons conclu qu'un CNN classique de profondeur réduite présente un résultat optimal et satisfaisant dans la restitution des caractéristiques pertinentes à la reconstruction SR. Nous avons en amont implémenté un CNN à 5 couches de convolution entièrement connectées. Sa structure globale se compose de deux parties qui sont analogues à celle du réseau ESPCN [6]. L'opération d'extraction des caractéristiques est effectuée par la première couche de convolution, tandis que la projection non-linéaire est gérée par la deuxième et la troisième couche convolutives. L'augmentation de l'échelle est traitée par la dernière couche du réseau dite couche de déconvolution sous-pixellique (Figure 5.11). La structure de notre réseau CNN est présentée en détail dans le Tableau 5.1.

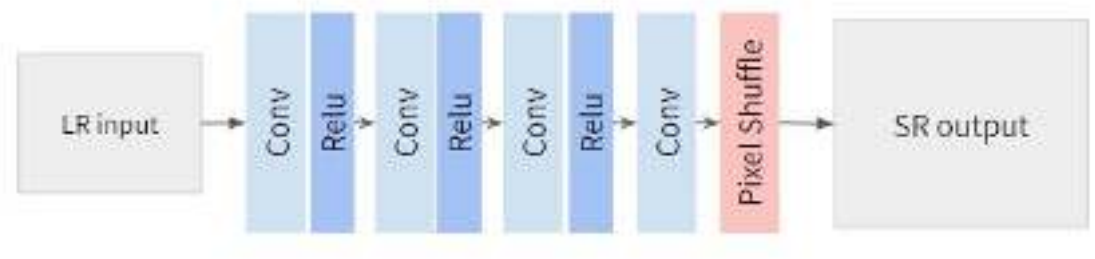


FIGURE 5.11 – Architecture proposée de notre CNN

Couches	Noyau			Taille de l'entrée	Taille de la sortie
	$w \times h$	str	pad	$bs \times w \times h \times c$	$bs \times w \times h \times c$
Conv_1	5×5	1	2	$20 \times 100 \times 100 \times 1$	$20 \times 100 \times 100 \times 64$
ReLu				$20 \times 100 \times 100 \times 64$	$20 \times 100 \times 100 \times 64$
Conv_2	3×3	1	1	$20 \times 100 \times 100 \times 64$	$20 \times 100 \times 100 \times 64$
ReLu				$20 \times 100 \times 100 \times 64$	$20 \times 100 \times 100 \times 64$
Conv_3	3×3	1	1	$20 \times 100 \times 100 \times 64$	$20 \times 100 \times 100 \times 32$
ReLu				$20 \times 100 \times 100 \times 32$	$20 \times 100 \times 100 \times 32$
Conv_4	3×3	1	1	$20 \times 100 \times 100 \times 32$	$20 \times 100 \times 100 \times r^2$
Pixel_shuffle				$20 \times 100 \times 100 \times r^2$	$1 \times 200 \times 200 \times 1$

TABLEAU 5.1 – Description de la structure de notre réseau CNN, w = largeur de l'image, h = hauteur de l'image, str = stride, pad = padding, bs taille du batch des images, c = nombre de canaux, et r = le facteur d'agrandissement (= 2).

5.4.3 Régularisation

Comme nous l'avons vu, les méthodes de régularisation représentent la solution la plus communément employée pour réduire l'erreur de généralisation à l'aide d'un terme de pénalité appliqué aux paramètres du réseau. Elle permet également l'adaptation rapide du modèle aux données quelles que soit leurs tailles.À travers l'expérimentation faite dans la section 5.4.1.3, on a pu considérer deux types de régularisation :

— La régularisation formelle, où l'on cherche à améliorer la capacité du modèle par l'ajout d'une pénalité à la fonction de coût pour obtenir une fonction coût régularisée ;

— La régularisation implicite, où l'on cherche à améliorer la capacité du modèle par la réforme de certains éléments du réseau sans toutefois introduire un terme de régularisation ajouté à la fonction de coût.

Les méthodes de régularisation du deuxième type peuvent être considérées comme des méthodes «manuelles ou assistées» car elles n'interviennent pas explicitement dans l'amélioration de la complexité du modèle. Nous pouvons citer quelques méthodes de ce type de régularisation comme l'augmentation de données, le transfert d'apprentissage, l'arrêt anticipé de l'apprentissage, le dropout, et la *batch-normalization*. Notre méthode de régularisation proposée s'inscrit en partie dans la régularisation implicite par augmentation de données. Toutefois, l'estimation du paramètre de dégradation spécifique à l'image constitue en soi un a priori puissant pour la reconstruction SR. Nous proposons également de rajouter un terme de régularisation formelle pour conditionner au mieux le comportement du modèle.

5.4.4 Stratégie d'apprentissage

5.4.4.1 Fonction de coût

Nous avons vu dans le chapitre précédent que la majorité des modèles profonds de la SISR intègrent l'erreur quadratique moyenne EQM (Equation 5.5) (ou MSE en anglais) dans la mesure de l'erreur entre le modèle et la base de données puisqu'elle est fortement corrélée à l'indice PSNR standard qui mesure les performances finales du réseau. Cependant, la SR est principalement liée à la perception humaine plutôt qu'aux changements de rang pixellique, ce qui rend l'indicateur PSNR décorrélé à la perception visuelle.

$$\mathcal{L}_{pixel}(\mathrm{I}^{SR}, \mathrm{I}^{BR}) = \frac{1}{hwc} \sum_{i,j,k} (\mathrm{I}^{SR}_{i,j,k} - \mathrm{I}^{BR}_{i,j,k})^2 \tag{5.5}$$

Où $\mathbf{I}^{\boldsymbol{SR}}$ est l'image reconstruite et $\mathbf{I}^{\boldsymbol{BR}}$ représente l'image de référence assimilée à l'image $\boldsymbol{BR}$ de la base de donnée. h, w et c représentent la hauteur, la largeur et le nombre de canaux des images évaluées, respectivement.
Pour pallier le problème du lissage de l'image résultante engendré par le caractère quadratique de la fonction MSE, nous employons également une fonction de coût perceptuelle, notée $\boldsymbol{\mathcal{L}_{content}}$, liée aux caractéristiques de l'image (Equation 5.6). Plus précisément, cette fonction mesure les différences sémantiques entre les couches convolutives d'un réseau extracteur de caractéristiques déjà entraîné : cette procédure force l'image générée à être perceptuellement similaire à l'image cible. Si l'on considère le réseau pré-entraîné VGG19, noté $\boldsymbol{\phi}$, comme extracteur de caractéristiques, et les représentations de haut niveau extraites depuis la $\boldsymbol{l^{ième}}$ couche convolutive par $\boldsymbol{\phi^{(l)}(\mathbf{I})}$, alors :

$$\mathcal{L}_{content}(\mathrm{I}^{SR}, \mathrm{I}^{BR}) = \frac{1}{h_l w_l c_l} \sqrt{\sum_{i,j,k} (\phi^{(l))}_{i,j,k}(\mathrm{I}^{SR}) - \phi^{(l))}_{i,j,k}(\mathrm{I}^{BR}))^2} \tag{5.6}$$

Pour notre expérimentation, nous avons sélectionné la 1ère couche convolutive du réseau VGG19 (Figure 5.12). À noter que les descripteurs du premier bloc

de couches convolutives détectent les caractéristiques à forte intensité comme les contours complets, tandis que les descripteurs situés près de la sortie du modèle VGG19 capturent des détails de haut-niveau. De ce fait, la sélection de ce bloc ainsi que la minimisation de l'équation (5.6) ont tendance à préserver le contenu de l'image et la structure spatiale, alors que les couleurs, les textures, et les contrastes ne sont pas préservés. Pour pallier ce problème, nous proposons de rajouter un terme régulateur supplémentaire, noté $\boldsymbol{\mathcal{L}_{texture}}$ (Equation (3.6)), communément connu sous le nom de la fonction coût de style qui, à l'instar de la fonction de coût perceptuelle, effectue son évaluation dans le domaine des paramètres du réseau VGG19.

Dans le calcul de la fonction $\boldsymbol{\mathcal{L}_{texture}}$, plusieurs couches convolutives $\boldsymbol{l_i}$ entrent en jeu (Figure 5.12). La perte est obtenue en calculant la distance Euclidienne moyenne des matrices Gram associées aux couches $\boldsymbol{l_i}$:

$$\mathcal{L}_{texture}(\mathrm{I}^{SR}, \mathrm{I}^{BR}) = \sum_{l \in \mathcal{L}} \left(\frac{1}{c_2^l} \sqrt{\sum_{i,j} \left(\mathcal{G}_{i,j}^{(l)}(I_{SR}) - \mathcal{G}_{i,j}^{(l)}(I_{BR}) \right)^2} \right) \tag{5.7}$$

Où $\mathcal{G}_{i,j}^{(l)}(.)$ est la matrice *Gram* décrite par le produit scalaire des descripteurs i et j sous leurs formes vectorielles :

$$\mathcal{G}_{i,j}^{(l)}(I) = vec\left(\phi_i^{(l)} - \phi_j^{(l)} \right)$$

Enfin, la fonction de coût générale, notée $\boldsymbol{Loss_{train}}$, qui optimise les paramètres de notre réseau CNN durant son apprentissage s'exprime par les termes des équations (5.5), (5.6), et (5.7) telle que :

$$Loss_{train} = \mathcal{L}_{pixel} + \mathcal{L}_{content} + \mathcal{L}_{texture} \tag{5.8}$$

L'apprentissage de notre réseau CNN de reconstruction a pour but de minimiser la fonction de coût totale, et donc de converger vers une reconstruction balancée entre les caractéristiques spatiales, perceptuelles, et texturales de l'image.

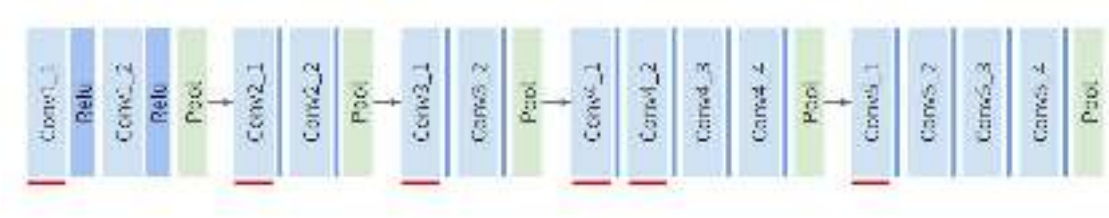


FIGURE 5.12 – Architecture du réseau extracteur de caractéristiques VGG19. Les couches soulignées en rouge sont celles utilisées dans la fonction de perte

5.4.4.2 Taux d'apprentissage et optimiseur

Après quelques lancements d'expériences et comparaisons de performances, nous choisissons d'utiliser l'algorithme d'optimisation Adam [117], une extension de la descente de gradient stochastique, qui combine deux autres méthodes issues de la descente du gradient stochastique à savoir l'algorithme de gradient adaptatif (AdaGrad [118]) et la propagation quadratique moyenne (RMSProp [119]), avec un taux d'apprentissage de 0.001, et des taux de décroissance exponentielle du premier et second moment fixés à 0.9 et 0.999 respectivement.

5.4.4.3 Pertinence et arrêt de l'apprentissage

Il est important de choisir les paramètres du réseau de manière à ce qu'ils soient en accord avec les données d'entrée, la profondeur du réseau et l'objectif de l'apprentissage. Au cours de nos expérimentations, nous avons constaté que la pertinence de la reconstruction du réseau est atteinte au bout de 500 itérations.

5.5 Validation des résultats

5.5.1 Étude comparative

On cherche à étudier le comportement des réseaux CNN de reconstruction SR en fonction de trois facteurs : la nature du jeu de donnée (interne ou externe), le noyau de dégradation (idéal ou estimé), et la fonction de coût de l'apprentissage. Pour

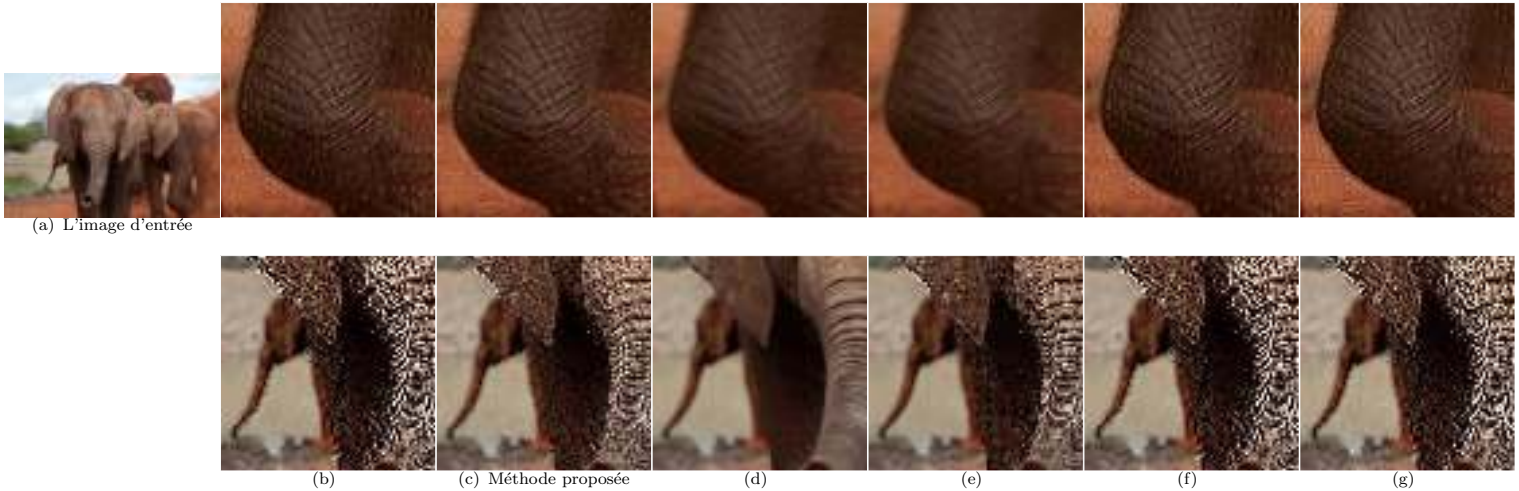
ce faire, nous réalisons un ensemble d'expériences, structurées dans le Tableau 5.2, afin de révéler l'effet de l'estimation du noyau de dégradation (Section 5.4.1.1), de l'utilisation d'une fonction de coût sophistiquée (Section 5.4.4.1), ainsi que l'usage d'une base interne de données (Section 5.4.1.2) sur le résultat du réseau CNN. Le même réseau est sujet à ces expériences afin de confronter les résultats. Enfin, l'évaluation est quant à elle subjective et s'appuie sur le constat visuel humain (Figure 5.13). L'image de test provient de la base de données publique BSD300.

Fixation du noyau de dégradation : Dans le choix du noyau de dégradation, nous appliquons à la fois le $\boldsymbol{k_{SR}}$ adapté à l'image de test estimé par le réseau KernelGAN, et le noyau de dégradation bicubique représentatif de l'opération de réduction d'échelle idéale utilisée par la majorité des méthodes SR par CNN du benchmark. Ces deux noyaux sont employés pour dégrader la résolution des images HR des bases de données d'entraînement (i.e. externe et interne).

Fixation de la base de données d'entraînement : Ce paramètre est défini pour observer l'effet de la reconstruction du jeu de données sur la reconstruction SR. Dans cette étude comparative, BSD300 représente la base de données externe d'images HR. Les images BR sont synthétisées à partir de celle-ci en employant soit une dégradation bicubique, ou une dégradation adaptée ($\boldsymbol{k_{SR}}$ estimé).

Fixation de la fonction de coût : Enfin, nous avons établi ce paramètre afin de déterminer la pertinence d'une fonction de coût sophistiquée basée sur la perception et la texture (Équation (5.8)) par rapport à une fonction de coût classique MSE (Équation (5.5)) utilisée dans le benchmark de la SR.

Dans un premier temps, nous remarquons que l'optimisation de la fonction objective MSE à travers une base de données externe tend à générer une image relativement floue, notamment lorsque le paramètre de dégradation utilisé est le noyau bicubique (Figure 5.13(d)). Ce cas particulier représente la stratégie commune adoptée par la majorité des travaux de recherche de la SR. Comme nous



(a) L'image d'entrée (b) (c) Méthode proposée (d) (e) (f) (g)

FIGURE 5.13 – 3.16 Comparaison visuelle des résultats de nos expérimentations. Les cas (b) –(g) correspondent aux descriptions exprimées par le Tableau 5.2

Base d'entrainement	**Noyau de dégradation**	**Fonction de coût**	**Cas #**
interne	k_{SR}	Éq 5.5	(b)
interne	k_{SR}	Éq 5.8	(c)
externe	bicubique	Éq 5.5	(d)
externe	bicubique	Éq 5.8	(e)
externe	k_{SR}	Éq 5.5	(f)
externe	k_{SR}	Éq 5.8	(g)

TABLEAU 5.2 – Labélisation des expériences de l'étude comparative

l'avons vu au chapitre 4 dans une étude menée dans [7], une fonction objective de type MSE n'est pas liée à la perception visuelle humaine. Pour des applications photo-réalistes, utiliser une fonction de pertes de type MSE est loin d'être la meilleure solution. Cependant, si l'opérateur de dégradation est estimé, le flou de l'image est moins prononcé, même si les résultats montrent une déformation dans la texture de l'image (Figure 5.13(f)). Ceci prouve que le caractère quadratique de la fonction MSE n'est pas le seul facteur à l'origine du flou dans la reconstruction des textures. En outre, la combinaison des données externes avec la fonction de coût conceptuelle et texturale donne de meilleurs résultats visuels que dans les cas (5.13(d)) et (5.13(d)).

Dans un deuxième temps, les résultats de l'expérience (5.13(e)) montrent des effets de blocs dans le contenu de l'image reconstruite (Figure 5.13(d)). En revanche, si le noyau bicubique est substitué par un noyau estimé, les textures semblent visuellement corrompues (Figure 5.13(g)), et ce malgré que la fonction objective utilisée dans ce cas soit adaptée à la reconstitution perceptuelle et texturale. En résumé, nous obtenons des textures trop lisses ou corrompues lorsque le réseau est entraîné sur un ensemble de données externes qui ne contient pas les informations souhaitées sur l'image de test. En revanche, il est clair que les résultats de reconstruction des cas (f) et (d) sont visuellement proches, ce qui signifie que l'emploi d'un noyau estimé et d'une fonction de coût sophistiquée n'est pas encore totalement suffisant pour restaurer correctement et entièrement les détails de l'image. Cela souligne également l'avantage d'utiliser une stratégie d'augmentation des données d'apprentissage.

Enfin, nous comparons les résultats de notre méthode proposée (Figure 5.13(c)) avec ceux de l'expérience (b). L'utilisation d'une fonction de perte perceptuelle réduit considérablement la formation d'artéfacts indésirables de la texture retrouvés dans l'image (Figure 5.13(b)). Bien que le noyau de dégradation estimé favorise la restitution des détails texturaux, la fonction de perte vient la régulariser pour améliorer la qualité finale de la reconstruction. Notre stratégie d'apprentissage génère des images satisfaisantes sur le plan perceptif, avec des textures nettement meilleures, tandis que les résultats générés à l'aide d'une fonction de perte MSE sont moins fidèles à la réalité terrain.

De manière générale, les fonctions de pertes sophistiquées basées sur la perception ne sont pas suffisantes pour reconstruire une meilleure texture. Les informations statistiques internes de l'image sont indispensables à la reconstruction perceptive. Nous rappelons également le comportement de la minimisation de la fonction MSE qui conduit à la suppression des hautes fréquences, ce qui entraîne la génération d'images floues et lisses. Néanmoins, nous supposons qu'il ne s'agit pas du seul facteur responsable du manque de netteté des images résultantes. En fait, l'utilisation d'une fonction de perte qui favorise la texture comme fonction objective de l'apprentissage sans s'appuyer sur les statistiques internes de l'image peut conduire à la même image floue avec –en plus- des artefacts de textures. Par conséquent, la constitution d'une base de données d'entraînement à travers une augmentation des données en prenant en compte l'autosimilarité de l'image de test, permet d'améliorer la reconstruction perceptive de la SR. **Il faut donc une utilisation conjointe de régularisation formelle et régularisation implicite** : une fonction de coût sophistiquée et une base interne d'apprentissage extraite à partir de l'image de test.

5.5.2 Évaluation et résultats

Pour nos expériences, nous avons retenus et comparé plusieurs méthodes de reconstruction SISR décrits dans le tableau 5.3.

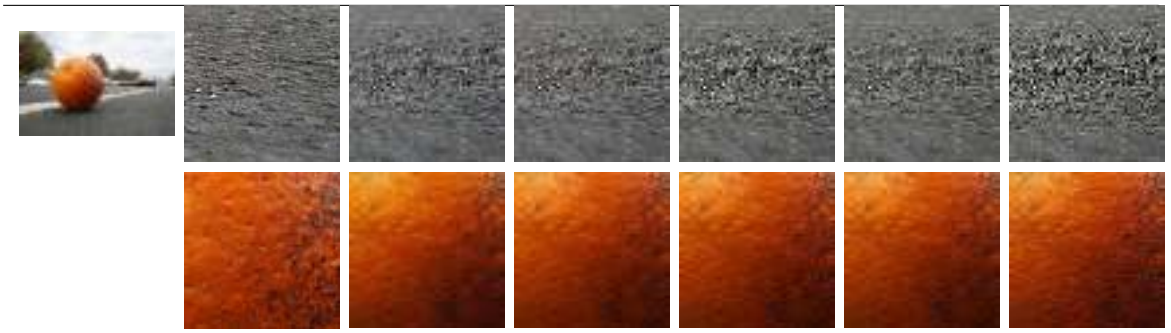
Méthode	Jeu de données d'apprentissage (nombre total d'images)	Fonction coût	Réalité terrain
Interpolation bicubique	–	–	–
SRGAN [7]	ImageNet (350000)	Pixellique (MSE) + perceptuelle	Oui (Dég idéale)
ESPCN [6]	ImageNet (50000)	Pixellique (MSE)	Oui (Dég idéale)
ZSSR [111]	Générée	Pixellique (L1)	Non
VDSR [4]	BSDS100 + Set14 + Set5 (291)	Pixellique (MSE)	Oui (Dég idéale)
Méthode proposée	Générée (400)	**Éq 5.8**	Non

TABLEAU 5.3 – Modèles des réseaux CNN appliqués à la SISR retenus pour comparaison et leurs paramétrages respectifs

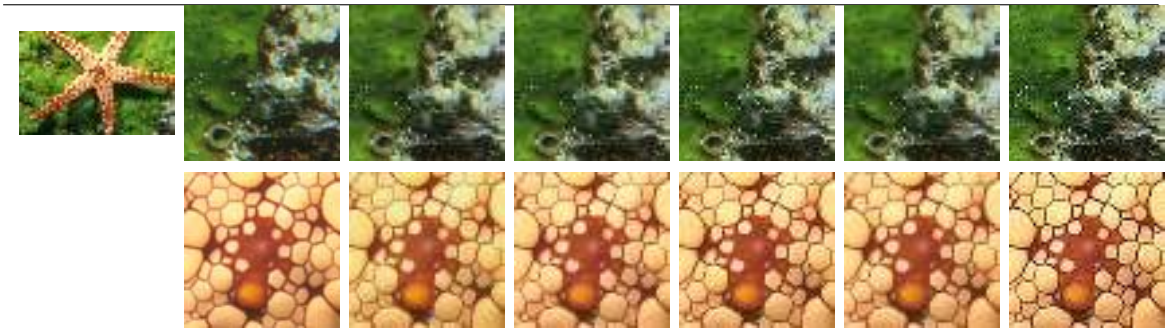
Notre approche est validée sur plusieurs images de test provenant de la base de données BSD300. Ces images ne possèdent pas de réalité terrain conduisant ainsi la validation à s'appuyer sur des métriques d'évaluation sans-référence dites NR-IQA (pour No Reference-Image Quality Assessment). Dans cette expérience, nous mesurons les performances en termes qualitatif et quantitatif à savoir les index PI (i.e. Performance Indicator) et NOREQI (i.e. NO-REference image Quality Index). Les résultats sont montrés dans la figure 5.14

Bien que le gain sur les scores PI et NOREQI soient très peu visible à l'égard de certains modèles (à l'ordre du 10ème de l'unité), les résultats obtenus par notre méthode proposée se trouvent être bien meilleurs visuellement par rapport aux méthodes retenues pour comparaison. Ainsi les résultats montrent la pertinence quant à l'utilisation d'un KernelGAN pour l'estimation de dégradation, et à la fois l'intérêt d'utiliser une fonction de perte composée et de retirer complétement l'‘usage d'une base externe d'entraînement.

	BR	SRGAN	ZSSR	ESPCN	VDSR	**méthode proposée**
PI↓		6.5333	6.8652	6.8756	6.8576	6.7693
NOREQI↑		0.7228	0.7875	0.7989	0.8324	0.8310



	BR	SRGAN	ZSSR	ESPCN	VDSR	**méthode proposée**
PI↓		7.4514	7.6725	6.9813	6.9516	6.5977
NOREQI↑		0.5611	0.6279	0.6648	0.6181	0.6504



	BR	SRGAN	ZSSR	ESPCN	VDSR	**méthode proposée**
PI↓		6.5434	6.7997	6.9181	6.6787	6.9196
NOREQI↑		0.5256	0.5334	0.5322	0.5338	0.5595

FIGURE 5.14 – Évaluations qualitatives et quantitatives aveugles des modèles SISR sur des images de la base de test BSD300.

5.6 Conclusion

Dans la première partie de ce chapitre, nous avons montré que, bien que les images possèdent beaucoup d'informations, elles ont la particularité d'être fortement corrélées localement. Les pixels proches dans une image ont une forte corrélation spatiale et cela signifie également que cette corrélation est préservée à travers différentes échelles de l'image.

À travers les contributions présentées précédemment, nous avons établi un certain nombre d'observations autours du gain qu'apporte la similarité interne de l'image (ou autosimilarité) dans la régularisation d'un modèle CNN profond. La piste de régularisation qui concerne l'intégration de connaissances à priori dans les modèles permet de limiter le nombre d'exemples nécessaires à l'apprentissage des architecture CNN. En effet, certaines règles simples valent mieux qu'un grand nombre d'exemples. Ceci est d'autant plus vrai lorsque ces exemples sont extraits à partir de l'image elle-même, et que la condition à priori soit formulée à partir des structures internes de celle-ci. Une autre constatation importante est celle de l'amélioration de la reconstruction des textures de l'image. L'élimination de l'usage d'une perte MSE a également contribué à la clarté des textures reconstruite.

Bien que les techniques de DL du benchmark actuel permettent d'avoir des résultats concluants dans des scénarios idéaux, la reconstitution des détails (conceptuels et texturaux) réels, quant à elle, reste insatisfaisante sur la base de la source de données utilisée. La dégradation par noyau bicubique s'est implantée comme unique standard dans la génération des données d'apprentissage. Pour cause, les techniques du benchmark réussissent à inverser ce modèle de dégradation pour prétendre augmenter la résolution de l'image d'entrée. L'étude effectuée dans ce chapitre a mis en avant l'intérêt du noyau estimé. Ainsi, un pipeline alternatif et homologue au KernelGAN sera proposé dans le prochain chapitre et fera office de notre seconde contribution, alliant le pouvoir de l'autosimilarité d'image et la reconstruction SR aveugle. La validation de notre approche dans ce chapitre nous a fourni un cadre expérimental aux améliorations que l'on peut espérer atteindre dans l'estimation du noyau de dégradation.

Chapitre 6

Super-résolution semi-aveugle

6.1 Introduction

Comme nous l'avons montré dans le chapitre précédent, les modèles CNN appliqués à la SR nécessitent de grands ensembles de paires d'images HR et BR pour leurs apprentissages. Cependant, l'acquisition de telles paires d'images de scènes réelles n'est pas triviale. Ainsi, l'apprentissage des réseaux SR actuels du benchmark repose sur des corpus d'images générés synthétiquement. Une première piste de recherche nous a permis de conclure qu'un a priori adapté régularisant l'entrainement du modèle peut alléger ce besoin de données. Cet à priori, qui concerne les paramètres constituant le modèle d'acquisition, permet de compenser l'information "réelle" recherchée, conduisant ainsi au renoncement à l'usage des jeux de données classiques établis par la communauté scientifique de la SR.
Notre contribution faite dans le chapitre précédent montre également que même si les réseaux SR sont performants sur les images BR synthétisées par sous-échantillonnage bicubique, leurs performances restent limitées quant à la reconstitution des textures réelles car leur fonctionnement est régi par des hypothèses de noyaux erronés [110]. Ceci est expliqué par le fait que les modèles SR sont sensibles aux noyaux de dégradation. Les artéfacts les plus courants qui apparaissent dans les images reconstruites sont ceux produits par des modèles dont l'information sur la dégradation n'est pas la même dans l'image de test et les images du jeu de donnée d'apprentissage. La figure 6.1 montre un tel phénomène, où $\boldsymbol{K_{SR}}$ désigne le noyau réel utilisé sur les images de test, et $\boldsymbol{K_{train}}$ désigne un noyau simulé appliqué sur les images d'entrainement. Lorsque le noyau $\boldsymbol{K_{train}}$ est plus net que le noyau réel $\boldsymbol{K_{SR}}(\sigma < \sigma_{SR})$, les images résultantes sont trop lisses et les textures à hautes fréquences sont considérablement floues. Tandis que les résultats présentent des artéfacts de *ringing* synthétisés lorsque le noyau simulé est plus lisse que le noyau réel ($\sigma > \sigma_{SR}$). En revanche, les images de la diagonale, où les noyaux de dégradation $\boldsymbol{K_{train}}$ et $\boldsymbol{K_{SR}}$ sont identiques ($\sigma = \sigma_{SR}$), semblent naturelles sans artefacts ni flou très prononcés. Ce phénomène révèle que l'erreur d'estimation du noyau k sera considérablement amplifiée par le modèle SR conduisant ainsi à une reconstruction d'images insatisfaisante. Afin de résoudre le problème de disparité

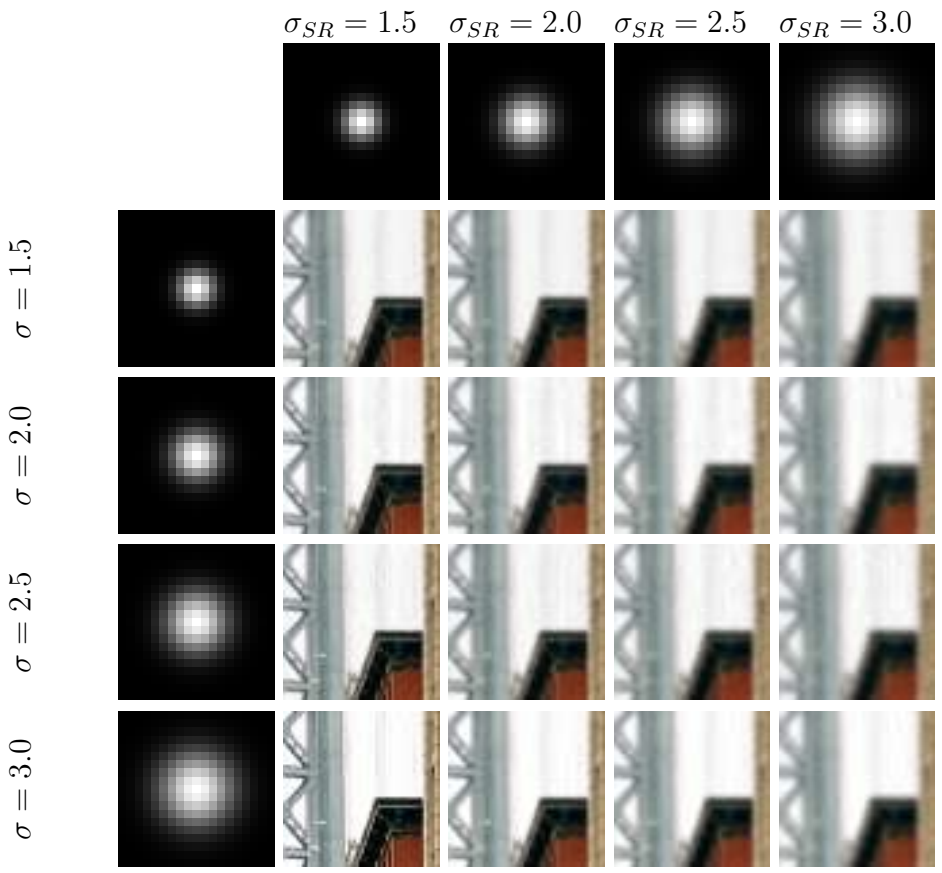


FIGURE 6.1 – Résultats de simulation du réseau ESPCN. La première ligne représent les noyaux réels de dégradation K_{SR} utilisés durant la phase de test, tandis que les noyaux de la première colonne sont les noyaux sur lesquels le réseau s'est entraîné. Les images de la diagonale sont les résultats de reconstruction où les noyaux vus en test sont les même que ceux vus en entraînement.

des noyaux de dégradation, nous proposons dans ce chapitre un modèle destiné à estimer le noyau k.

6.2 Estimation de la dégradation

6.2.1 Hypothèse

Le verrou major lié à l'estimation de la dégradation dans une reconstruction SISR concerne la disponibilité d'une seule observation de la scène. Ainsi, à l'instar du KernelGAN (traité en section 5.4.1.1), notre méthode d'estimation du noyau est

basée sur une hypothèse importante faite par de Bell-Kligler [112] qui stipule qu'un noyau de dégradation correct est celui qui *maximise la similarité des différents patches de l'image d'entrée à travers plusieurs échelles*. Le but est donc de créer un modèle stochastique capable de "représenter" les patches de l'image en termes de leurs distributions en maintenant une structure multi-échelle de l'image, afin de pouvoir minimiser la distance entre ces représentations [10]. Pour y arriver, nous utilisons un type spécifique de réseau de neurones : un auto-encodeur variationnel (VAE pour Variational Auto-Encoder), un modèle qui apprend à modéliser une représentation plus simple des données (des images) sur lesquelles nous appliquerons une mesure de distance permettant au réseau général à apprendre de maximiser la similarité entre ces représentations.

6.2.2 Modélisation avec les Auto-Encodeur Variationnels

Les auto-encodeurs (pour Auto-Encoder en anglais, AE) sont des réseaux CNN de modélisation non-supervisée destinés à encoder des structures de données complexes dans des représentations latentes.

6.2.2.1 Les auto-encodeur (AE)

Les premiers travaux se rapprochant de l'auto-encodeur connu aujourd'hui remontent aux années 1980 [120] où le principe était d'apprendre une représentation cachée en utilisant l'entrée elle-même comme variable à prédire. Historiquement, les auto-encodeurs étaient vus comme une méthode de réduction de dimensionnalité, mais désormais, ceux-ci ont davantage d'applications liées au fait qu'ils peuvent apprendre des variables latentes riches en information [121, 122], et sont fréquemment utilisés pour le débruitage [123]. Avec la montée en popularité des réseaux de neurones au début des années 2000, une nouvelle génération d'auto-encodeur a vu le jour. Ces derniers possédaient désormais plusieurs couches cachées superposées permettant ainsi de faire l'apprentissage de données plus complexes

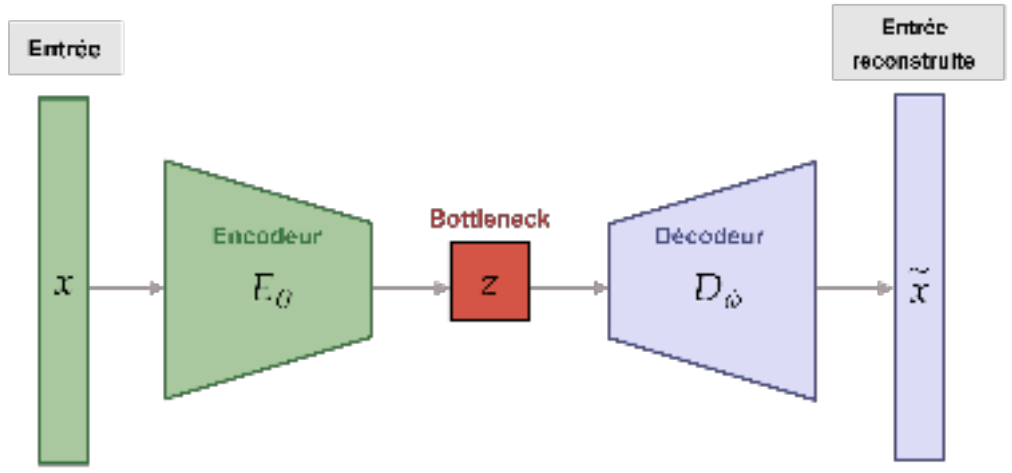


FIGURE 6.2 – Schémas de l'architecture de base d'un réseau AE

comme les images. Un auto-encodeur est un réseau de neurones qui a comme objectif d'apprendre une représentation intermédiaire d'une entrée de manière non-supervisée [124]. Pour réaliser cet objectif, l'auto-encodeur se compose de deux unités : un encodeur et un décodeur. L'encodeur reçoit en entrée la donnée $\boldsymbol{x_i}$ et la réduit en une représentation latente $\boldsymbol{z}$. Le décodeur prend en entrée cette représentation latente z et la décode vers une sortie qui doit se rapprocher le plus possible de l'entrée $\boldsymbol{x}$. Techniquement, le décodeur et l'encodeur sont des réseaux CNN entièrement connectés dont les couches cachées sont de petites tailles que les couches d'entrée et de sortie réduisant (ou augmentant) ainsi progressivement la taille de la donnée. Pour empêcher le système d'apprendre une identité triviale de l'entrée, la couche cachée au milieu du modèle (interprétée comme espace latent) est généralement contrainte d'être de dimension inférieure (Goulot ou *Bottleneck* en anglais) de sortie non-linéaire afin de définir l'information compressée. Cette structure de base est illustrée dans la figure 6.2

L'intuition derrière les auto-encodeurs est de reconstruire une entrée $\boldsymbol{x}$ en passant par deux fonctions (l'encodeur $\boldsymbol{E_\theta(\cdot)}$ et le décodeur $\boldsymbol{D_\phi(\cdot)}$) apprises par le modèle. Ces deux fonctions vont permettre d'obtenir une sortie $\tilde{\boldsymbol{x}}$ donnée par l'équation :

$$\tilde{x} = p_\phi q_\theta(x) \tag{6.1}$$

Ainsi, ce genre de méthodes ne dispose pas d'un label (y) car son objectif est de produire la valeur (x) elle-même ; d'où son apprentissage non-supervisé. Ce dernier est d'ailleurs effectué en minimisant l'erreur de reconstruction en s'assurant que les unités cachées capturent seulement les aspects les plus pertinents des données. La perte est définie par une fonction de la forme :

$$\mathcal{L}(x, \tilde{x}) = \mathcal{L}\left(x, p_\phi q_\theta(x)\right) \tag{6.2}$$

Où la fonction de coût $\boldsymbol{\mathcal{L}}$ peut être la fonction MSE par exemple. L'optimisation est faite par descente du gradient en prenant en compte la dérivée de la perte par rapport aux paramètres $\boldsymbol{\theta}$ de l'encodeur et $\boldsymbol{\phi}$ du décodeur simultanément, telle que :

$$\Theta' \longleftarrow \Theta - \varepsilon \times \frac{\delta \mathcal{L}}{\delta \Theta} \tag{6.3}$$

Où $\boldsymbol{\varepsilon}$ est le taux d'apprentissage qui module la vitesse de l'apprentissage. $\boldsymbol{\Theta = \theta, \phi}$ comprend les paramètres de l'encodeur et du décodeur, et $\boldsymbol{\Theta'}$ correspond aux valeurs des paramètres du modèle complet à la suite d'une itération d'optimisation. Les auto-encodeurs modernes ont généralisé l'idée d'un encodeur et d'un décodeur au-delà des fonctions déterministes en allant vers des représentations stochastiques telle que $\boldsymbol{q_\theta(z|x)}$ pour l'encodeur, et $\boldsymbol{p_\phi(x|z)}$ pour le décodeur dont le modèle graphique est présenté par la figure 6.3.

6.2.2.2 Les auto-encodeurs variationnels (VAE)

Une variante intéressante des AE est l'auto-encodeur variationnel (VAE) proposée par [125] dont le principe est de contraindre la représentation latente et ses attributs à suivre une distribution fixe souvent définie par une loi normale centrée réduite. Plus précisément, les données sont encodées dans deux vecteurs : un vecteur de moyennes μ et un vecteur d'écarts-types $\boldsymbol{\sigma}$. Ces derniers sont ensuite utilisés comme paramètres d'une distribution paramétrique servant à générer

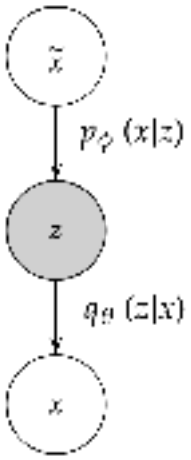


FIGURE 6.3 – Modèle graphique d'un réseau AE

aléatoirement une représentation latente $\boldsymbol{z} \sim \boldsymbol{p(z)} = \boldsymbol{\mathcal{N}(0,1)}$. L'intérêt est qu'avec une telle contrainte, chaque donnée $\boldsymbol{x}$ permet d'obtenir un espace latent continu, et tout échantillon tiré selon cette distribution pourra être décodé en une image plausible. De cette manière, chaque dimension de l'espace latent représente une caractéristique qui varie continuellement selon les échantillons de la base de données. La méthode de contraindre l'espace latent dans les VAE est la minimisation de la divergence de Kullback-Leibler (KL), une mesure de dissimilarité entre deux distributions de probabilité, telle que :

$$KL(q(z|x)\|p(z|x)) = -\sum q(z|x)\log\left(\frac{p(z|x)}{q(z|x)}\right) \tag{6.4}$$

Par simplification, le problème de minimisation de l'équation ci-dessus est équivalent au problème de maximisation de la borne inférieur ELBO pour Evidence Lower BOund :

$$max\left(ELBO = E_{q(z|x)}[logp(x|z)] - KL(q(z|x)\|p(z|x))\right) \tag{6.5}$$

Où le premier terme représente la vraisemblance de reconstruction, et le deuxième garantit que la distribution $\boldsymbol{q(z|x)}$ apprise par l'encodeur soit similaire à la distribution à la distribution à priori $\boldsymbol{p(z|x)}$.

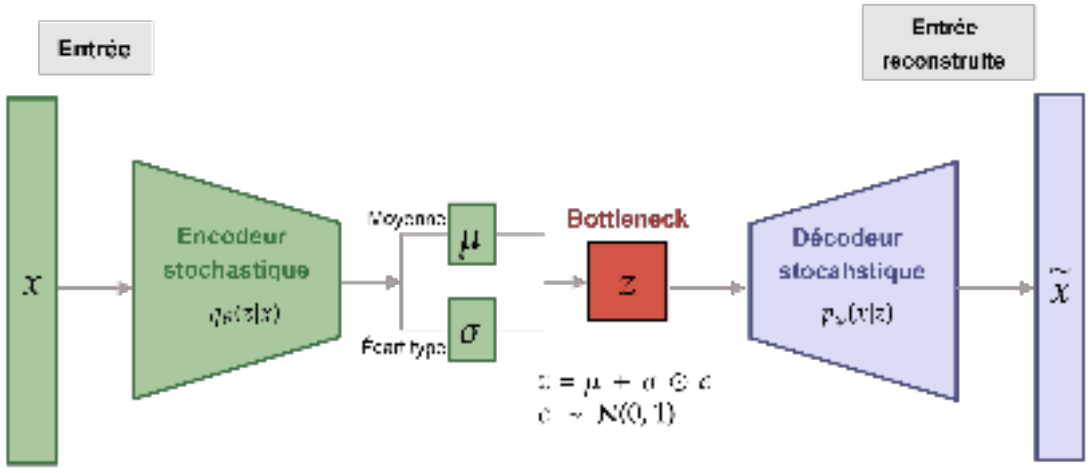


FIGURE 6.4 – Schémas de l'architecture d'un réseau VAE

Ainsi, si l'on contraint le réseau VAE durant son apprentissage à minimiser la divergence dans l'équation (6.5), alors le premier terme de **ELBO** est considéré comme l'erreur de reconstruction de la fonction objective du modèle. En effet, si par exemple la fonction de distribution q est gaussienne, alors la minimisation de l'espérance $\boldsymbol{E_{q(z|x)}}$ revient à minimiser l'erreur de reconstruction MSE. Une fois le modèle entraîné, la sortie de celui-ci est stochastique dans le sens où deux encodages d'une même entrée $\boldsymbol{x}$ peuvent donner deux sorties différentes. Cependant, une donnée $\boldsymbol{x}$ va toujours mener aux mêmes valeurs $\boldsymbol{\sigma}$ et $\boldsymbol{\mu}$ ce qui ne la rend pas une composante stochastique. La figure 6.4 illustre la structure de base des VAE. Là aussi, comme dans le cas d'un AE traditionnel, les données en entrée passent par des couches convolutives cachées. La sortie du réseau encodeur se divise en deux composantes à partir desquelles la représentation latente est générée par la simulation d'une réalisation d'une loi normale. Une fois la représentation $\boldsymbol{z}$ générée, celle-ci est décodée jusqu'au format original par des couches de convolution transposées entièrement connectées.

6.2.2.3 Architecture proposée

Les modèles VAE fournissent un cadre important pour simuler la façon dont les données sont représentées de manière stochastique, et constituent une piste à explorer pour estimer le noyau de dégradation correcte de l'image. Nous proposons de combiner deux tâches mutuellement supervisées pour l'apprentissage des opérations de dégradation. La première tâche est instanciée par un réseau CNN ($\boldsymbol{D_w}$) réducteur d'échelle qui apprend à sous-échantillonner son image d'entrée, et à créer une image d'une échelle moindre de sorte à maintenir la distribution de ses patches. La deuxième tâche consiste en la mise en correspondance des espaces latents obtenus à partir des patches de l'image, et d'assurer la similarité de leurs distributions. Pour cette dernière tâche, nous avons implémenté un VAE à entrées multiples destiné à apprendre la représentation latente des patches dans deux échelles. Plus précisément, deux encodeurs ($\boldsymbol{E_1}$ et $\boldsymbol{E_2}$) sont employés simultanément quant à la structure multi-échelle de l'image où chaque encodeur traite une échelle donnée. De cette manière, deux espaces latents indépendants sont générés à partir des deux encodeurs, et constituent respectivement les représentations de patches d'une même image dans deux échelles différentes. Ces deux représentations sont ensuite concaténées puis décodées par deux réseaux décodeurs indépendants ($\boldsymbol{D_1}$ et $\boldsymbol{D_2}$) vers les dimensions initiales des images d'entrée (image de test, et image de test sous-échantillonnée). Cette structure à entrées multiples du VAE est dite M-VAE (pour Multiple-input Variational Auto-Encoder). La figure illustre le schéma d'une telle architecture, où les réseaux $\boldsymbol{D_w}$, $\boldsymbol{E_1}$, $\boldsymbol{E_2}$, $\boldsymbol{D_1}$ et $\boldsymbol{D_2}$ sont des CNN.

Apprentissage du M-VAE Si l'on revient au modèle graphique du M-VAE, on considère les observations $\boldsymbol{X}$ et $\boldsymbol{Y}$, telles que $\boldsymbol{(X, Y) = (x_1, y_1), ..., (x_n, y_n)}$, comme indépendantes (i.e. $\boldsymbol{x, y \sim p(x, y|z) = p_{\theta_x}(x|z)p_{\theta_y}(y|z)}$) où $\boldsymbol{\theta_x}$ et $\boldsymbol{\theta_y}$ sont les paramètres des deux décodeurs respectivement. On considère également que les représentations latentes respectives de ces observations, notées $\boldsymbol{z_1}$ et $\boldsymbol{z_2}$, sont échantillonnées à partir d'une distribution à priori $\boldsymbol{p(z)}$ définie par $\boldsymbol{p(z) = \mathcal{N}(0, 1)}$.

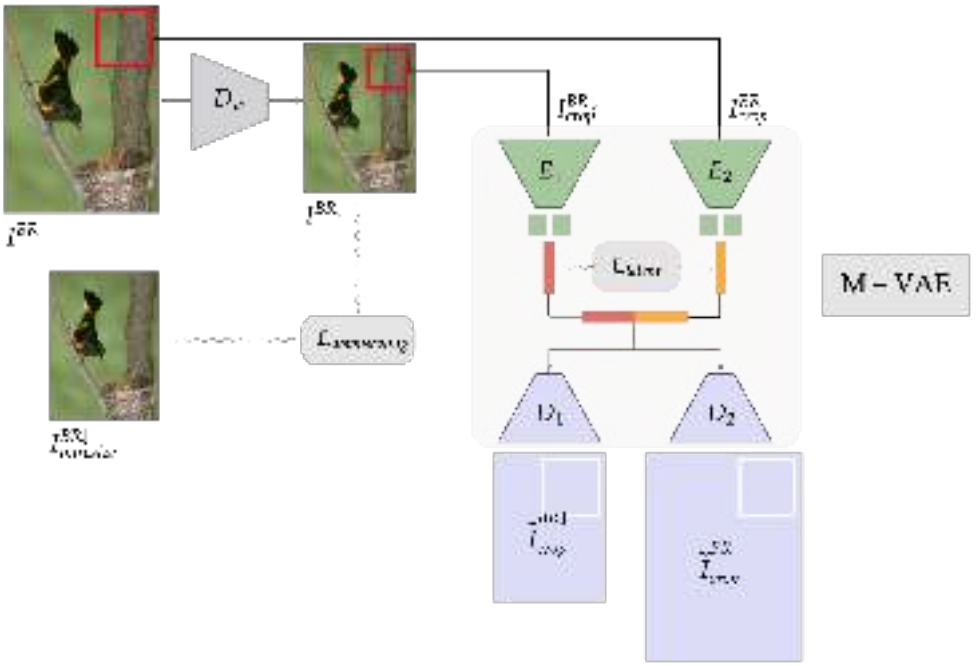


FIGURE 6.5 – Architecture proposée du modèle M-VAE [10] et du réseau D_w réducteur d'échelle

Étant donné que les observations sont statistiquement indépendantes et conditionnées sur le même espace latent, l'objectif du M-VAE est donc de maximiser la valeur **ELBO** de la distribution marginale des variables $\boldsymbol{x}$ et $\boldsymbol{y}$ telle que :

$$ELBO = \underbrace{E_{q_\phi(z|x,y)}\left[logp_{\phi_x}(x|z)\right]}_{\text{(I)}} + \underbrace{E_{q_\phi(z|x,y)}\left[logp_{\phi_y}(y|z)\right]}_{\text{(II)}} - \underbrace{KL\left(q_\phi(z|x,y)\|p(z|x,y)\right)}_{\text{(III)}} \tag{6.6}$$

Où le terme (I) et (II) représentent les erreurs de reconstruction des décodeurs $\boldsymbol{D_1}$ et $\boldsymbol{D_2}$ (i.e. les logarithmes de vraisemblance des variables $\boldsymbol{x}$ et $\boldsymbol{y}$) définies respectivement par la MSE entre l'image d'entrée de l'encodeur $\boldsymbol{E_1}$ et l'image de sortie du décodeur $\boldsymbol{D_1}$, et la MSE entre l'image d'entrée de l'encodeur $\boldsymbol{E_2}$ et l'image de sortie du décodeur $\boldsymbol{D_2}$. Le terme (III) est la divergence de Kullback-Leibler entre $\boldsymbol{q_\phi(z|x,y)}$ et $\boldsymbol{p(z|x,y)}$ et qui représente le terme de régularisation du modèle M-VAE. L'équation 6.6 définie ainsi la fonction objective du modèle

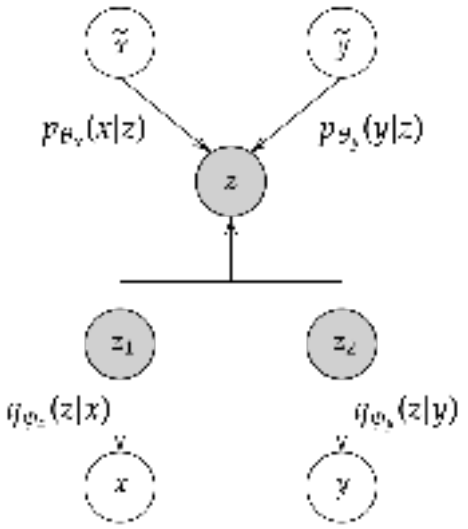


FIGURE 6.6 – Modèle graphique du réseau M-VAE. Les nœuds blancs représentent les observations, tandis que les nœuds gris sont les représentation latentes.

composée de deux erreurs de reconstruction et un terme de régularisation. La figure 5.6 montre le modèle graphique du réseau M-VAE.

Apprentissage de D_w Le réseau $\boldsymbol{D_w}$ est destiné à apprendre à sous-échantillonner l'image de manière à préserver la distribution des patches de son entrée. Comme mentionné précédemment, cette tâche est réalisable si l'apprentissage de $\boldsymbol{D_w}$ est effectué conjointement avec celui du modèle M-VAE. Il est clair que certaines observations du réseau M-VAE sont générées par $\boldsymbol{D_w}$, ainsi, ces deux réseaux convergents conjointement vers un minimum global de la fonction de coût notée $\boldsymbol{\mathcal{L}_{deg}}$ par :

$$\mathcal{L}_{deg} = \mathcal{L}_{MVAE} + \mathcal{L}_{latent} + \mathcal{L}_{downscaling} \tag{6.7}$$

Où $\boldsymbol{\mathcal{L}_{MVAE}}$ représente la fonction objective du réseau M-VAE définie par l'équation (6.6) qui optimise les réseaux encodeurs et décodeurs en fonction de leurs paramètres. $\boldsymbol{\mathcal{L}_{latent}}$ est la norme $L1$ entre les distribution marginales $\boldsymbol{q_{\phi_x}(z|x)}$

$\boldsymbol{q_{\phi_y}(z|y)}$. Sa minimisation contraint le modèle à faire correspondre les distributions des patches d'un "crop" et celle d'un "crop" d'image redimensionnée. $\boldsymbol{\mathcal{L}_{latent}}$ et $\boldsymbol{\mathcal{L}_{MVAE}}$ permettent de déterminer comment la sortie du réseau $\boldsymbol{D_w}$ doit être ajustée de manière à optimiser le score. Enfin, $\boldsymbol{\mathcal{L}_{downscaling}}$ est la distance MSE entre la sortie du réseau réducteur d'échelle $\boldsymbol{D_w}$ et une image redimensionnée par un changement d'échelle avec une interpolation bicubique. La rétropropagation de cette erreur concerne seulement les paramètres du réseau $\boldsymbol{D_w}$ afin de le conduire à reconstruire une image sous-échantillonnée idéale avant de l'ajuster selon la perte du réseau M-VAE et la perte $\boldsymbol{\mathcal{L}_{latent}}$.

6.3 Expérimentations

6.3.1 Détails d'implémentation

Durant sa phase d'apprentissage, le modèle ($Dw + M - VAE.$) est entraîné uniquement sur son image de test pendant 7000 itérations. À chaque itération, deux crops sont extraits aléatoirement à partir de l'image d'entrée $\boldsymbol{I^{BR}}$ et de sa version sous-échantillonnée $\boldsymbol{I^{BR\downarrow}}$. Ces derniers sont de tailles 32×32 pixels et 64×64 pixels respectivement, induisant ainsi le modèle à apprendre une reconstruction avec un facteur d'agrandissement fixé à 2. De cette manière, le modèle s'entraîne sur un ensemble de 5000 crops suivant le nombre maximal d'itérations. Les cinq réseaux du modèle (Voir tableau 6.1 pour l'architecture détaillée des réseaux) sont mis en œuvre en utilisant la fonction (6.7) comme critère d'optimisation. Dans un premier temps, cette fonction va favoriser l'optimisation de l'erreur de reconstruction de $\boldsymbol{D_w}$ pondérée par un facteur fixé à 0.96 pendant les 2000 premières itérations, tandis que les autres termes de la fonction évolueront de manière identique. La fonction de coût en termes pondérés et donnée par :

$$\mathcal{L}_{deg} = 0.03 \times \mathcal{L}_{MVAE} + 0.01 \times \mathcal{L}_{latent} + 0.96 \times \mathcal{L}_{downscaling} \tag{6.8}$$

Ceci permettra au modèle d'apprendre d'abord à dégrader l'image, puis à se régulariser progressivement par rapport à la préservation de la densité des patchs. De plus, à cause du phénomène observé de l'effondrement de l'a posteriori du M-VAE (*posterior collapse*), l'importance de sa fonction objective sera tempérée durant les premières itérations de l'entrainement. Ce phénomène sera traité en partie plus tard par un ajustement du pas d'apprentissage. Enfin, et à partir de la $2000^{iéme}$ itération, les termes de la fonction objective sont mis-à-jour sans compromis telle définie par l'équation (6.7).

Les hyper-paramètres des cinq réseaux sont fixés empiriquement en exécutant le processus d'apprentissage avec des valeurs différentes, et en choisissant celles qui allient meilleure vitesse de convergence et performances. En amont, nous avons retenu le paramétrage suivant :

L'algorithme Adam [117] est utilisé pour la rétropropagation du gradient en utilisant un taux d'apprentissage initial de 0.001. L'ajustement du pas d'apprentissage est effectué selon l'algorithme du *warm-up* suivant la formule du *Cosine Annealing* [126]. Cette stratégie permet d'augmenter de façon cyclique le pas d'apprentissage et va aider le modèle à quitter le point-selle rapidement (i.e. le point où le gradient est proche de zéro ce qui va ralentir la progression de l'apprentissage). À la fin de son entrainement, le réseau $\boldsymbol{D_w}$ est employé dans le processus de génération du jeu de donnée (tel décrit par l'Algorithme 1) comme opérateur de dégradation des images afin de synthétiser des images $\boldsymbol{BR}$. Les différents réseaux sont implémentés en utilisant le langage de programmation Python et la bibliothèque d'apprentissage automatique *Pytorch*. L'évaluation de la reconstruction pose également la question quant au sur-ajustement du modèle. Là encore, le corpus d'image d'entraînement (à savoir les paires d'images $\boldsymbol{HR}$ et $\boldsymbol{BR}$) sont synthétisées à partir de l'image de test. Il est vrai que le réseau, selon cette stratégie, s'ajuste exclusivement aux données générées à partir d'une seule image, néanmoins, la tâche de son entraînement se fait pour chaque nouvelle donnée de test. Au lieu d'entraîner un réseau une seule fois de manière générale de sorte à ce qu'il s'adapte aux nouvelles images de tests (ce qui est difficile dans la pratique), l'idée est de le ré-entraîner sur chaque image d'entrée. Ainsi, la possibilité du sur-apprentissage du modèle est

écartée.

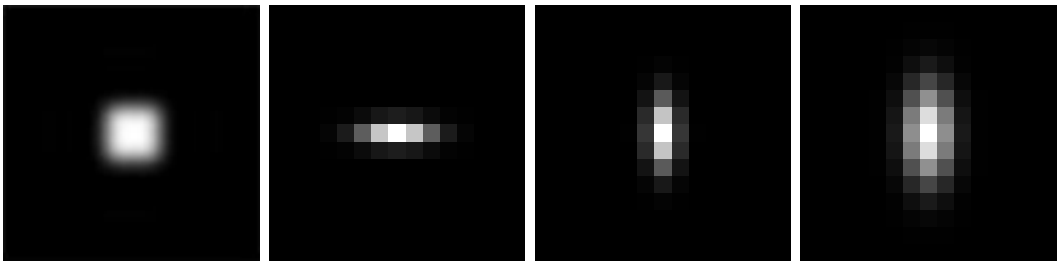
D_w	Noyau			Taille de l'entrée	Taille de la sortie
	$w \times h$	*str*	*pad*	$bs \times w \times h \times c$	$bs \times w \times h \times c$
Conv_1	3×3	2	1	$1 \times 64 \times 64 \times 1$	$1 \times 32 \times 32 \times 1$
Conv_2	3×3	1	1	$1 \times 32 \times 32 \times 1$	$1 \times 32 \times 32 \times 1$
Conv_2	3×3	1	1	$1 \times 32 \times 32 \times 1$	$1 \times 32 \times 32 \times 1$
E_1	Noyau			Taille de l'entrée	Taille de la sortie
	$w \times h$	*str*	*pad*	$bs \times w \times h \times c$	$bs \times w \times h \times c$
Conv_1	4×4	2	0	$1 \times 64 \times 64 \times 1$	$1 \times 31 \times 31 \times 32$
Relu				$1 \times 31 \times 31 \times 32$	$1 \times 31 \times 31 \times 32$
Conv_2	4×4	2	0	$1 \times 31 \times 31 \times 32$	$1 \times 14 \times 14 \times 64$
Relu				$1 \times 14 \times 14 \times 64$	$1 \times 14 \times 14 \times 64$
Conv_3	4×4	2	0	$1 \times 14 \times 14 \times 64$	$1 \times 6 \times 6 \times 128$
Relu				$1 \times 6 \times 6 \times 128$	$1 \times 6 \times 6 \times 128$
Conv_4	4×4	2	0	$1 \times 6 \times 6 \times 128$	$1 \times 2 \times 2 \times 256$
Relu				$1 \times 2 \times 2 \times 256$	$1 \times 2 \times 2 \times 256$
Flatten				$1 \times 2 \times 2 \times 256$	1×1024
Linear_1				1×1024	1×32
Linear_2				1×1024	1×32
E_2	Noyau			Taille de l'entée	Taille de la sortie
	$w \times h$	*str*	*pad*	$bs \times w \times h \times c$	$bs \times w \times h \times c$
Conv_1	4×4	2	0	$1 \times 32 \times 32 \times 1$	$1 \times 15 \times 15 \times 64$
Relu				$1 \times 15 \times 15 \times 64$	$1 \times 15 \times 15 \times 64$
Conv_2	4×4	2	0	$1 \times 15 \times 15 \times 64$	$1 \times 6 \times 6 \times 28$
Relu				$1 \times 6 \times 6 \times 28$	$1 \times 6 \times 6 \times 28$
Conv_3	4×4	2	0	$1 \times 6 \times 6 \times 28$	$1 \times 2 \times 2 \times 256$
Relu				$1 \times 2 \times 2 \times 256$	$1 \times 2 \times 2 \times 256$
Flatten				$1 \times 2 \times 2 \times 256$	1×1024
Linear_1				1×1024	1×32
Linear_2				1×1024	1×32

suite dans page suivante

Tableau 6.1

	Noyau			Taille de l'entrée	Taille de la sortie
	$w \times h$	*str*	*pad*	$bs \times w \times h \times c$	$bs \times w \times h \times c$
D_1	Noyau			Taille de l'entrée	Taille de la sortie
	$w \times h$	*str*	*pad*	$bs \times w \times h \times c$	$bs \times w \times h \times c$
Linear				1×64	1×1024
Unflatten				1×1024	$1 \times 1 \times 1 \times 1024$
TransConv_1	5×5	2	1	$1 \times 1 \times 1 \times 1024$	$1 \times 5 \times 5 \times 128$
Relu				$1 \times 5 \times 5 \times 128$	$1 \times 5 \times 5 \times 128$
TransConv_2	5×5	2	1	$1 \times 5 \times 5 \times 128$	$1 \times 13 \times 13 \times 64$
Relu				$1 \times 13 \times 13 \times 64$	$1 \times 13 \times 13 \times 64$
TransConv_3	6×6	2	1	$1 \times 13 \times 13 \times 64$	$1 \times 30 \times 30 \times 32$
Relu				$1 \times 30 \times 30 \times 32$	$1 \times 30 \times 30 \times 32$
TransConv_4	6×6			$1 \times 30 \times 30 \times 32$	$1 \times 64 \times 64 \times 1$
Relu				$1 \times 64 \times 64 \times 1$	$1 \times 64 \times 64 \times 1$
D_2	Noyau			Taille de l'entrée	Taille de la sortie
	$w \times h$	*str*	*pad*	$bs \times w \times h \times c$	$bs \times w \times h \times c$
Linear				1×64	1×1024
Unflatten				1×1024	$1 \times 1 \times 1 \times 1024$
TransConv_1	5×5	2	1	$1 \times 1 \times 1 \times 1024$	$1 \times 5 \times 5 \times 128$
Relu				$1 \times 5 \times 5 \times 128$	$1 \times 5 \times 5 \times 128$
TransConv_2	5×5	2	1	$1 \times 5 \times 5 \times 128$	$1 \times 13 \times 13 \times 64$
Relu				$1 \times 13 \times 13 \times 64$	$1 \times 13 \times 13 \times 64$
TransConv_3	6×6	2	1	$1 \times 13 \times 13 \times 64$	$1 \times 30 \times 30 \times 1$
Relu				$1 \times 30 \times 30 \times 1$	$1 \times 30 \times 30 \times 1$

TABLEAU 6.1 – Architecture des réseaux D_w, E_1, E_2, D_1, et D_2. *bs* Représente le batch-size, w et h sont la largeur et la hauteur respectivement, et c représente le nombre de canaux. *str* représente le paramètre *stride* de la couche de convolution, et *pad* est le *padding* de celle-ci.



(a) bicubic (b) k_1 (c) k_2 (d) k_3

FIGURE 6.7 – les noyaux de dégradation utilisée dans la simulation. (a) est un noyau bicubique idéal. (b), (c) et (d) sont des noyaux gaussiens non-isotropiques de moyenne nulle et d'écart-types selon leurs deux axes $(\sigma_1, \sigma_2) = (0.4797, 1.411)$, $(\sigma_1, \sigma_2) = (1.3825, -0.5692)$, et $(\sigma_1, \sigma_2) = (1.8705, -0.5692)$ respectivement. Les différentes valeurs de σ des axes de chaque noyau sont générées aléatoirement à partir d'une loi normale centrée.

6.3.2 Expérience sur des données synthétisées : Noyau bicubique Vs. Noyau gaussien

L'objet de cette expérience est d'étudier le comportement de notre méthode SR dans la reconstruction d'image dégradée avec des noyaux connus et prédéfinis. Nous avons appliqué 4 différents noyaux de dégradation sur un ensemble d'images de la base de données Div2k pour générer 4 jeux de données de test. La figure 6.7 montre les 4 noyaux utilisés dans cette expérimentation.

Les méthodes de reconstruction SISR retenues pour comparaison sont SRGAN, ESPCN, ZSSR, KernelGAN, ainsi qu'une interpolation bicubique classique. L'intérêt principal de cette simulation réside dans la prédisposition d'images de références. Ainsi, les critères d'évaluation choisis sont le PSNR et le SSIM, en plus des métriques NR-IQA Les résultats NOREQI, DISTS[127], et FREQUEE sont présentés dans le tableau 6.2.

Nous remarquons que le modèle basé M-VAE produit de moins bons résultats sur les images BR générées avec le noyau bicubique. Il surpasse néanmoins les réseaux sur les autres paramètres expérimentaux. Les performances des réseaux entraînés sur des images BR bicubiques sont limitées lorsque leur noyau s'écarte du véritable noyau de flou. En termes des scores FR-IQA PSNR et SSIM, le modèle ZSSR gagne pas plus de 0.02dB et 0.01 par rapport à une interpolation bicubique classique.

Même avec des couches plus profondes, les réseau profond ZSSR n'est pas significativement performant que les réseaux peu profonds ESPCN et SRGAN. Les scores FR-IQA les plus élevés sont marqués par l'interpolation bicubique à cause de sa corrélation avec le PSNR. En modélisant les noyaux réels, les modèles basés KernelGAN [110] et M-VAE [10] surpassent les modèles en terme de scores NR-IQA, notamment lorsque le noyau réel est différent de celui utilisé dans l'apprentissage.

Dans cette expérience, Les méthodes basées sur l'estimation de la dégradation obtiennent un gain en termes de qualité visuelle et des scores FRIQUEE, DISTS et NOREQI. Néanmoins, elles aboutissent à une dégradation du PSNR malgré une reconstruction bien meilleure visuellement (Figure 6.8). Les métriques orientées pixel (i.e. PSNR/SSIM) ne sont alors pas un indicateur de performances fiable.

	Noyau	Bicubique	SRGAN	ESPCN	ZSSR	kernelGAN [110]	M-VAE [10]
PSNR ↑	Bicubique	38.4313	31.1594	31.0365	38.9892	**27.3450**	**23.2428**
SSIM ↑		0.9764	0.7651	0.9092	0.9787	**0.5021**	**0.6690**
NOREQI ↑		0.7828	0.8299	0.8074	0.8067	**0.9605**	**0.8039**
DISTS ↓		0.0005	0.0338	0.0063	0.0006	**0.0004**	**0.0001**
FRIQUEE ↑		61.5736	63.4804	65.3035	62.0392	**64.8501**	**65.3244**
PSNR ↑	k_1	27.0622	26.3064	25.3168	26.8106	**18.9931**	**16.9371**
SSIM ↑		0.84977	0.8476	0.8348	0.8562	**0.4302**	**0.3467**
NOREQI ↑		0.5003	0.5115	0.5064	0.5079	**0.5255**	**0.5300**
DISTS ↓		0.0150	0.0331	0.0142	0.0150	**0.0209**	**0.0142**
FRIQUEE ↑		61.7317	65.0912	67.3884	63.9991	**72.1176**	**70.5314**
PSNR ↑	k_2	32.6604	30.1380	31.5272	32.6838	**27.8695**	**28.2000**
SSIM ↑		0.9477	0.8962	0.9376	0.9507	**0.8689**	**0.8690**
NOREQI ↑		0.7747	0.8015	0.7935	0.7540	**0.7832**	**0.7963**
DISTS ↓		0.0083	0.0403	0.0128	0.0198	**0.0201**	**0.0111**
FRIQUEE ↑		54.5632	55.4473	57.4766	56.0574	**57.4922**	**60.0617**
PSNR ↑	k_3	38.6568	34.1384	36.6918	38.3651	**35.1742**	**33.8648**
SSIM ↑		0.9521	0.9165	0.9372	0.9525	0.9209	**0.9209**
NOREQI ↑		0.6099	0.6128	0.6186	0.5944	**0.6305**	**0.6100**
DISTS ↓		0.0129	0.0778	0.0396	0.0332	0.0164	0.0148
FRIQUEE ↑		48.9861	46.1674	50.7485	49.0308	**51.0443**	**54.8023**

TABLEAU 6.2 – Tableau des scores de reconstruction des méthodes retenus de l'état de l'art dans l'évaluation des noyaux synthétisé en termes de métrique NR-IQA (NOREQI-FREQUEE-DISTS) et FR-IQA (PSNR-SSIM). Les meilleurs scores sont mis en évidence en rouge, et les seconds en bleu. L'évaluation qualitative est présentées dans la figure

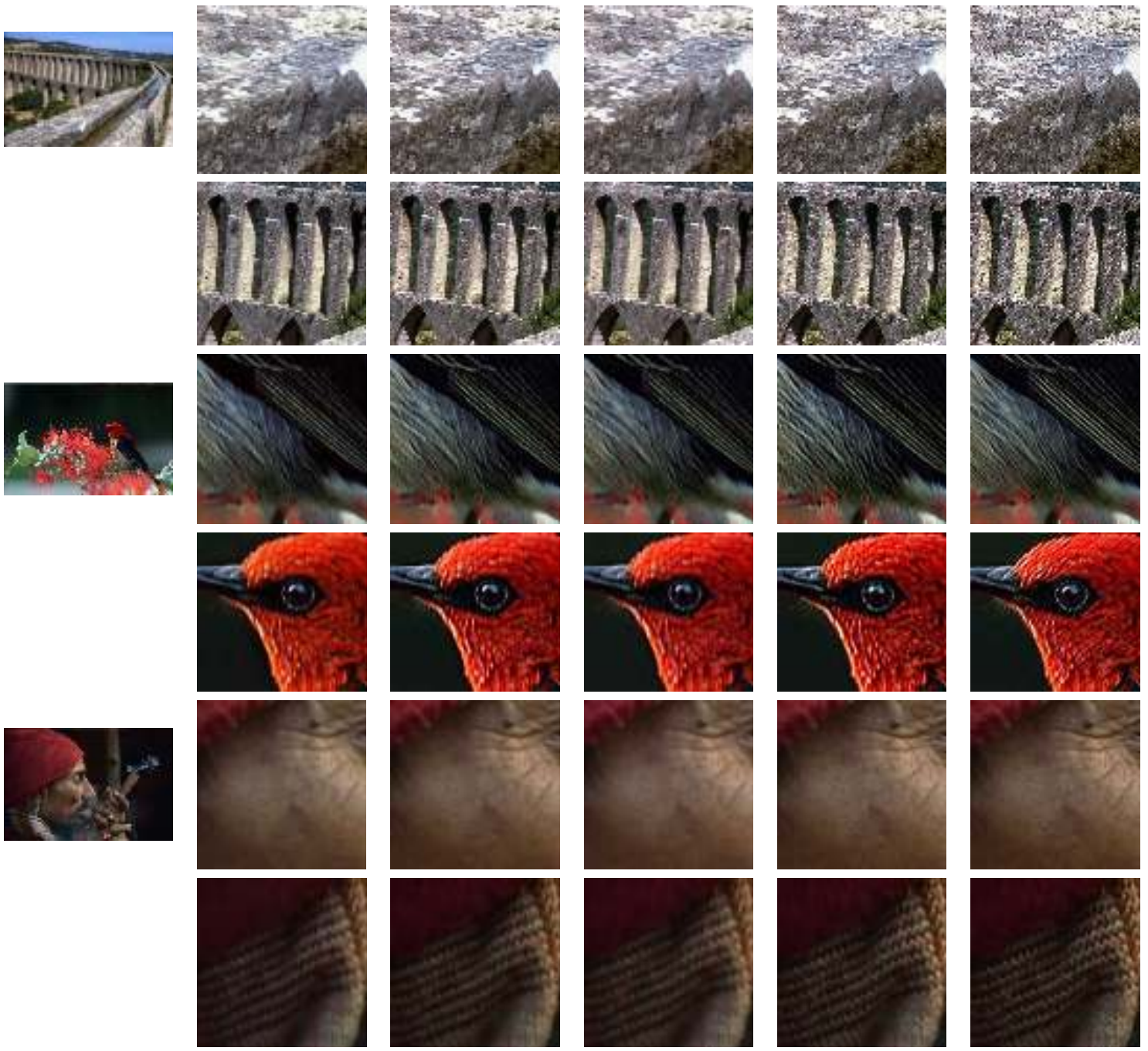
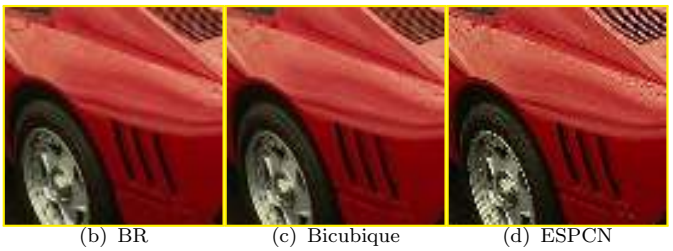


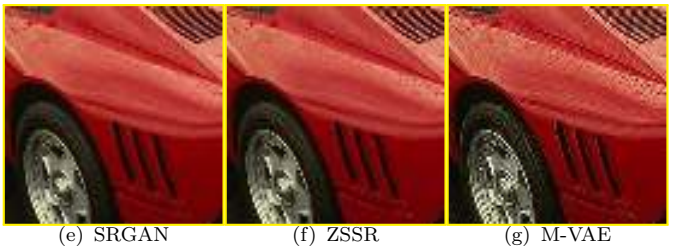
FIGURE 6.8 – Comparaison visuelle des performances des réseaux du benchmark de la reconstruction SISR sur des images synthétisées avec des noyaux bicubique et gaussiens

6.3.3 Expériences sur des données réelles : Noyaux réels

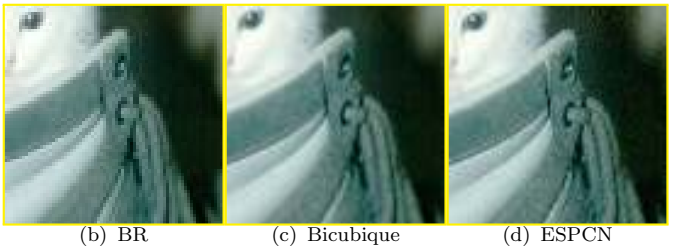
Cette expérience a pour but de s'assurer que la méthode fonctionne correctement dans un contexte semi-aveugle. Les images de tests sont issues de la base BSDS300. Dans ce cas particulier, l'image de réalité terrain n'existe pas et le modèle de dégradation réel n'est pas connu étant donné que l'entrée BR constitue l'unique information disponible sur la scène. Ainsi, en plus de l'usage de métriques NR-IQA, l'évaluation de cette expérience est qualitative et est présentée dans la Figure 6.8.



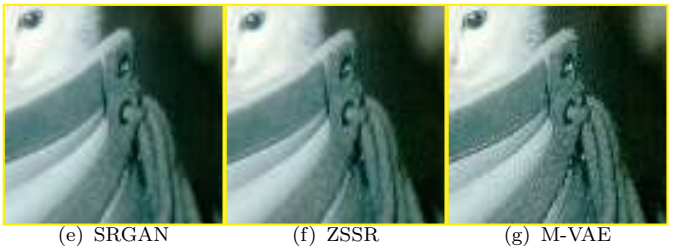
(b) BR (c) Bicubique (d) ESPCN



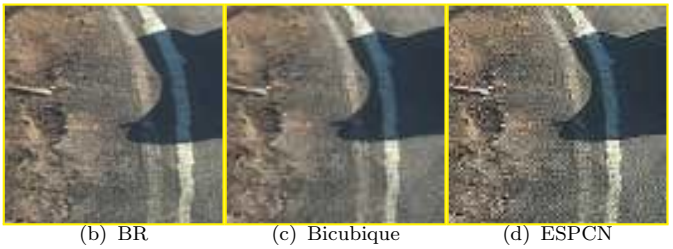
(e) SRGAN (f) ZSSR (g) M-VAE



(b) BR (c) Bicubique (d) ESPCN



(e) SRGAN (f) ZSSR (g) M-VAE



(b) BR (c) Bicubique (d) ESPCN

(e) SRGAN (f) ZSSR (g) M-VAE

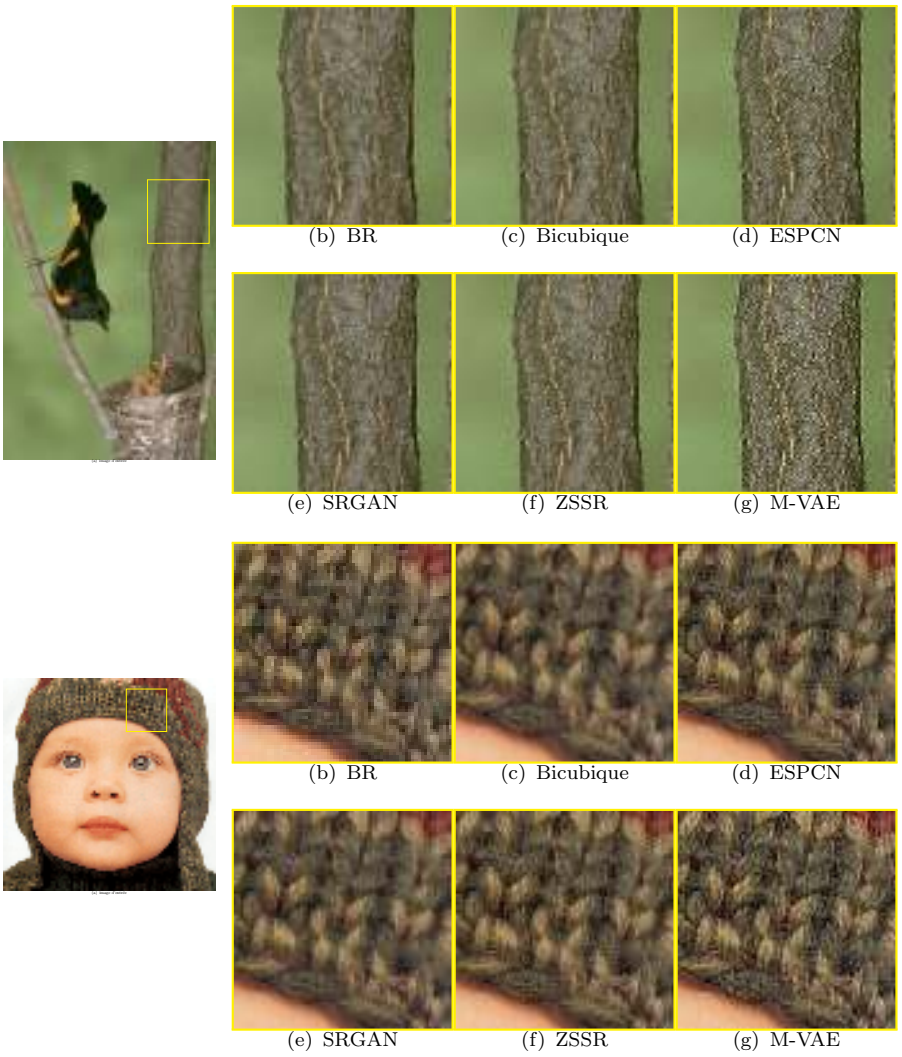


FIGURE 6.8 – Comparaison qualitative des performances des réseaux du benchmark de la reconstruction SISR sur des images réelles.

6.3.4 Expériences sur des photographies : images JPEG

Dans de nombreuses applications, les données réelles BR sont non seulement limitées par leur résolution spatiale, mais également par le processus de compression des données JPEG, tandis que la plupart des réseaux de reconstruction existants du benchmark n'impliquent aucune notion de compression dans les modèles analytiques de dégradation d'image, ni dans leur formule de régularisation. Un résultat très intéressant obtenu par notre méthode, comme complément des objectifs de la super-résolution appliquée aux images réelles, est la réduction des artéfacts de type blocking de la compression JPEG généralement perceptible par l'œil humain. Les

FIGURE 6.9 – Comparaison visuelle entre les résultats de reconstruction SISR d'une image compressé JPEG. (a) Résultat du réseau ESPCN entrainé sur ImageNet sur une base bicubique. (b) Résultat du réseau M-VAE.

résultats de reconstruction d'une image compressée par le modèle ESPCN (Figure 6.9(a)) et le modèle M-VAE (Figure 6.9(b)) sont confrontés. Nous pouvons noter que la présence des artéfacts de blocking se trouve être plus perceptible dans le résultat de reconstruction du modèle ESPCN. Un tel phénomène se produit lorsque la régularisation de la reconstruction est formulée autour d'un a priori peu fiable, et que le modèle de dégradation apprit ne prend pas en compte les images compressés en JPEG. Par conséquent, les effets de blocking sont considérés par le modèle comme étant des structures inhérentes à l'image étant donné que celui-ci n'a pas traité des images compressés durant son apprentissage. Ce résultat indique que les modèles de reconstruction basés sur la modélisation de la dégradation augmentent la résolution spatiale de l'image sans accentuer le bruit présent dans celle-ci.

6.4 Conclusion

Nous avons pu renforcer la validité de l'hypothèse autours du pouvoir de la redondance des patches par la reconstruction SISR. Nous avons exploité cette propriété dans l'apprentissage profond à travers l'implémentation d'un modèle génératif VAE à entrées multiples. Ce modèle, étant capable de dégrader les images en

préservant leurs distributions, capture avec efficacité l'a priori qui régularise au mieux la reconstruction mal-posée de la SISR. Cette manière s'avère plus efficace qu'une régularisation implicite à partir de donnée externes, ou explicite à travers des fonctions de coût idéales généralement inappropriées pour des cas pratiques complexes. Contrairement au KernelGAN, notre modèle proposé n'estime pas explicitement le noyau (matrice) de dégradation. Toutefois, le réseau $\boldsymbol{D_w}$ réducteur d'échelle conjointement entrainé avec le modèle M-VAE a permis de mieux modéliser la dégradation à partir d'une image unique, ce qui a abouti à de très bons résultats dans la reconstruction SISR, plus particulièrement la SISR semi-aveugle.

Chapitre 7

Conclusion générale et perspectives

7.1 Bilan général

L'objectif fixé pour ce travail de thèse est de contribuer à l'amélioration du conditionnement du problème mal-posé de la reconstruction SR en utilisant des méthodes du Deep Learning. Compte tenu du fait que la majorité des méthodes du benchmark actuel résolvent ce problème à travers des contraintes généralisées et prédéterminées, notre travail s'est focalisé sur la question de la connaissance du facteur du flou spécifique à chaque image de test.

Au chapitre 2, nous avons d'abord formalisé la reconstruction SR d'images comme étant un problème d'optimisation numérique mal posé et mal conditionné. Ceci rend la qualité de l'image reconstruite très sensible à la modélisation des paramètres de dégradation. Un petit nombre d'observations (réduit à une seule image dans le cas de la SISR) contraint la résolution de ce problème à s'appuyer particulièrement sur une hypothèse de régularité exprimée par l'a priori. Le gain en stabilité qu'offre la régularisation limite cependant le domaine d'application de la SR car un a priori non-adéquat à la scène d'acquisition peut générer des résultats moins bons.

Ainsi, au chapitre 3, nous avons détaillé l'état de l'art en matière de reconstruction MISR et SISR en ciblant les prédéfinitions que font les méthodes sur le modèle d'acquisition et l'a priori de la régularisation. Nous avons noté que ces paramètres sont supposés connus pour la majorité des approches et les connaissances a priori sont extraites à partir de plusieurs observations de la même scène (pour le cas de la MISR), ou à partir d'un ensemble très large de données appelé dictionnaire (pour le cas de la SISR). Un intérêt particulier est ensuite accordé aux réseaux CNN au chapitre 4 étant données leur faculté à apprendre les paramètres des algorithmes de résolution à partir de données d'entraînement, ce qui diffère des algorithmes linéaires traditionnels à paramétrage prédéterminé. En plus de leur efficacité et efficience supérieure dans la reconstruction SISR en grande partie grâce au mappage de bout en bout des réseaux profonds. Nous avons passé en revue les architectures des modèles ainsi que leurs techniques d'apprentissage, et nous avons constaté que les résultats de telles méthodes sont supérieurs aux méthodes classiques de l'état de l'art. Néanmoins, ces approches échouent dans des cas plus complexes et manquent généralement de flexibilité. La plupart de ces méthodes ne tiennent pas compte du fait que le paramètre de dégradation est mal-connu, et modélisent cette méconnaissance en supposant que celui-ci est un noyau bicubique idéal. Plus précisément, les bases de données d'entraînement sont modifiées à cet effet, où les images de basse-résolution sont synthétisées avec ce noyau. Nous avons constaté à travers l'étude de l'état de l'art que cette stratégie de simulation est employée dans la majorité des méthodes SISR-DL, néanmoins, elle présente de nombreux inconvénients :

— Le modèle de dégradation prédéfini n'est pas le même que celui de la caméra ;

— Malgré leur diversité, les images naturelles du jeu de données peuvent ne pas correspondre aux structures internes de l'image d'entrée ;

— Réduire le résidu (erreur d'apprentissage) ayant une image de référence prédisposée conduit la reconstruction à une simple opération d'inversion du modèle vu en simulation, ce qui ne correspond pas aux applications réelles.

— Bien que la validation quantitative soit possible dans ce cas (vu la prédisposition de l'image de référence), la métrique d'évaluation, à savoir le PSNR basé sur

une norme L2, n'est pas fiable dans la mesure où elle ne correspond pas au système visuel humain.

Partant de ce constat, nous avons proposé de modéliser la connaissance imprécise du noyau de dégradation qui est défini, dans le contexte particulier de la SISR, par le flou.

Compte tenu de cette estimation, il faut noter que l'on dispose que d'une seule image ou d'une seule représentation, ce qui rend le processus d'estimation lui-même mal-posé. Ainsi, notre première contribution traitée dans le chapitre 5 était d'abord de révéler la propriété de l'autosimilarité dont dispose les images naturelles comme un a priori puissant de régularité. L'exploitation de cette dernière dans l'apprentissage de modèles profonds apporte un gain important dans la reconstruction notamment celle des textures. Les résultats d'évaluation d'une telle régularisation ont également montré l'inefficacité des paramétrages standards des réseaux CNN appliqués à la SISR, à savoir une fonction objective (i.e. MSE) et sa corrélation au critère d'évaluation (i.e. PSNR), une architecture profonde sophistiquée, et une large base de données externe.

Une fois l'intérêt du noyau estimé mis en avant, nous avons proposé dans le chapitre 6 une nouvelle méthode d'estimation de la dégradation inhérente à l'image basée un modèle générateur à apprentissage non-supervisé, qui représente notre seconde contribution. Nous avons présenté en détail l'implémentation de cette méthode fondée sur la théorie de l'autosimilarité de l'image dans une structure multi-échelle où l'apprentissage du modèle est réalisé conjointement par un réseau CNN réducteur d'échelle et un réseau VAE de modélisation des densités des patches de l'image.

Les évaluations des deux contributions sont présentées en deux temps. D'abord, les expérimentations axées sur l'exploitation de l'autosimilarité interne de l'image dans un apprentissage profond nous ont permis de balayer les éléments standardisés par la majorité des modèle DL de la SISR en les remplaçant par un a priori robuste. En amont, nous avons testé notre méthode proposée destinée à estimer cet a priori. Si l'on s'intéresse simplement à la qualité visuelle des images reconstruites, il ressort de nos expérimentations que la méthode que nous proposons est

au rang des plus efficaces tant en termes de qualité subjective des détails reconstruits (mesure de qualité visuelle) qu'en termes de tests objectifs par l'utilisation de métriques de qualité sans référence. Les résultats expérimentaux montrent que l'image de la reconstruction obtenue par notre méthode offre un bon compromis entre la netteté des contours, le rehaussement des textures, et l'atténuation des artéfacts de reconstruction comme le lissage des hautes fréquences, l'effet de ringing, et le blocking des algorithmes de compression. Les métriques d'évaluation basées sur une comparaison pixel à pixel tels le PSNR et SSIM enregistrent des scores relativement inférieurs quant à notre méthode proposée, néanmoins, elles s'avèrent également être de mauvaise corrélation avec le jugement humain de certaines images comme il a été démontré à l'issus des résultats de nos deux derniers chapitres.

7.2 Perspectives

Les travaux de recherche réalisés dans le cadre de cette thèse mettent en exergue de nombreuses perspectives à explorer. Comme tout travail appliqué par des méthodes du DL, il est toujours envisageable d'apporter à ces systèmes des améliorations techniques afin d'améliorer leur performances, on peut noter :

— L'initialisation des poids des réseaux CNN à l'aide de la technique d'apprentissage par transfert qui permet de profiter d'un apprentissage acquis précédemment pour accélérer la convergence du modèle ;

— L'implémentation d'une fonction objective basée sur la régularisation de variation totale (TV) ;

— L'extraction de la matrice représentative du noyau flou à partir du réseau $\boldsymbol{D_w}$ et l'introduire dans la fonction objective globale afin de privilégier la modélisation matricielle de la dégradation ;

— L'étude et la conception de critères d'évaluation plus pertinentes sur le plan perceptif pour les applications réelles.

— Une estimation conjointe entre le flou et le bruit suggère également une amélioration des performances du modèle à travers un meilleur conditionnement du problème inverse de la reconstruction

Une première perspective de notre méthode serait de l'adapter aux noyaux de flous cinétiques (ou flous de mouvement) généralement causé par une lente cadence des caméras classiques qui va jusqu'à capturer une animation rapide à 30 images par seconde, ou par l'instabilité de l'appareil d'acquisition pendant la prise de vue. Il est également important pour les applications émergentes de développer des systèmes SR capables de gérer des reconstructions de plus grands facteurs d'agrandissement allant au $\times 4$ et $\times 8$, où la préservation d'une haute qualité perceptive des détails locaux constitue un défi difficile à relever. Une seconde perspective consiste à proposer une base de données standardisée contenant des noyaux de flou de différents appareils d'acquisition destinée à l'apprentissage profond des réseaux étant donnée qu'un tel paramètre représente mieux l'information de la scène qu'une image photographique de type RVB. Ceci va permettre également d'éliminer l'usage de la bicubique retrouvé dans la totalité des bases de données conçues pour la SISR.

Bibliographie

[1] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2) :295–307, 2015.

[4] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[5] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[6] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[7] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[8] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[9] Maria Zontak and Michal Irani. Internal statistics of a single natural image. In *CVPR 2011*, pages 977–984. IEEE, 2011.

[10] Kawther Aarizou and Abdelhamid Loukil. Perceptual-based super-resolution reconstruction using image-specific degradation estimation. *Journal of Electronic Imaging*, 30(3) :033007, 2021.

[11] Joseph B Keller. The american mathematical monthly. *Inverse Problems*, 83 :107–118, 1976.

[12] Jacques Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. Yale university press, 1923.

[13] Nhat Nguyen, Peyman Milanfar, and Gene Golub. A computationally efficient superresolution image reconstruction algorithm. *IEEE transactions on image processing*, 10(4) :573–583, 2001.

[14] Andrey N Tikhonov and Vasiliy Y Arsenin. Solutions of ill-posed problems. *New York*, 1(30) :487, 1977.

[15] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2) :56–65, 2002.

[16] Kevin Su, Qi Tian, Qing Xue, Nicu Sebe, and Jingsheng Ma. Neighborhood issue in single-frame image super-resolution. In *2005 IEEE international conference on multimedia and expo*, pages 4–pp. IEEE, 2005.

[17] Mei Gong, Kun He, Jiliu Zhou, and Jian Zhang. Single color image super-resolution through neighbor embedding. *Journal of computational information systems*, 7(1) :49–56, 2011.

[18] Weisheng Dong, Guangming Shi, Lei Zhang, and Xiaolin Wu. Super-resolution with nonlocal regularized sparse representation. In *Visual Communications and Image Processing 2010*, volume 7744, pages 152–161. SPIE, 2010.

[19] Guangming Shi, Weisheng Dong, Xiaolin Wu, and Lei Zhang. Context-based adaptive image resolution upconversion. *Journal of Electronic Imaging*, 19 (1) :013008, 2010.

[20] Xiaogang Wang and Xiaoou Tang. Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(3) :425–434, 2005.

[21] Xiang Ma, Junping Zhang, and Chun Qi. Hallucinating face by position-patch. *Pattern Recognition*, 43(6) :2224–2236, 2010.

[22] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, 2001.

[23] Karl S Ni, Sanjeev Kumar, Nuno Vasconcelos, and Truong Q Nguyen. Single image superresolution based on support vector regression. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II. IEEE, 2006.

[24] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on non-negative neighbor embedding. 2012.

[25] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, pages 349–356. IEEE, 2009.

[26] Samuel Schulter, Christian Leistner, and Horst Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3791–3799, 2015.

[27] Shengyang Dai, Mei Han, Wei Xu, Ying Wu, and Yihong Gong. Soft edge smoothness prior for alpha channel super resolution. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[28] Raanan Fattal. Image upsampling via imposed edge statistics. In *ACM SIGGRAPH 2007 papers*, pages 95–es. 2007.

[29] Bryan S Morse and Duane Schwartzwald. Image magnification using level-set reconstruction. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[30] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Gradient profile prior and its applications in image super-resolution and enhancement. *IEEE Transactions on Image Processing*, 20(6) :1529–1542, 2010.

[31] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[32] Yu-Wing Tai, Shuaicheng Liu, Michael S Brown, and Stephen Lin. Super resolution using edge prior and single image detail synthesis. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2400–2407. IEEE, 2010.

[33] Xin Li and Michael T Orchard. New edge-directed interpolation. *IEEE transactions on image processing*, 10(10) :1521–1527, 2001.

[34] Seung-Jun Lee, Mun-Cheon Kang, Kwang-Hyun Uhm, and Sung-Jea Ko. An edge-guided image interpolation method using taylor series approximation. *IEEE Transactions on Consumer Electronics*, 62(2) :159–165, 2016.

[35] Jian Sun, Nan-Ning Zheng, Hai Tao, and Heung-Yeung Shum. Image hallucination with primal sketch priors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–729. IEEE, 2003.

[36] Wei Fan and Dit-Yan Yeung. Image hallucination using neighbor embedding over visual primitive manifolds. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–7. IEEE, 2007.

[37] Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Robust web image/video super-resolution. *IEEE transactions on image processing*, 19(8) :2017–2028, 2010.

[38] Jan P Allebach. 7.1 image scanning, sampling, and interpolation. *Handbook of Image and Video Processing (Second Edition), Communications, Networking and Multimedia*, 2005.

[39] Alberto Biancardi, Luigi Cinque, and Luca Lombardi. Improvements to image magnification. *Pattern Recognition*, 35(3) :677–687, 2002.

[40] Sergio Carrato, Giovanni Ramponi, and Stefano Marsi. A simple edge-sensitive image interpolation filter. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, pages 711–714. IEEE, 1996.

[41] Qing Wang and Rabab Ward. A new edge-directed image expansion scheme. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 3, pages 899–902. IEEE, 2001.

[42] Lingfeng Wang, Shiming Xiang, Gaofeng Meng, Huaiyu Wu, and Chunhong Pan. Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(8) :1289–1299, 2013.

[43] H Chang, D Y Yeung, and Y Xiong. Super-resolution through neighbor embedding. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1 :I–I, 2004.

[44] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[45] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11) :2861–2873, 2010.

[46] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[47] Noriaki Suetake, Morihiko Sakano, and Eiji Uchino. Image super-resolution based on local self-similarity. *Optical review*, 15(1) :26–30, 2008.

[48] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Single-image super-resolution via linear mapping of interpolated self-examples. *IEEE Transactions on image processing*, 23(12) :5334–5347, 2014.

[49] Chanzi Liu, Qingchun Chen, and Hengchao Li. Single image super-resolution reconstruction technique based on a single hybrid dictionary. *Multimedia Tools and Applications*, 76(13) :14759–14779, 2017.

[50] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[51] Xiaoguang Li, Kin Man Lam, Guoping Qiu, Lansun Shen, and Suyu Wang. Example-based image super-resolution with class-specific predictors. *Journal of Visual Communication and Image Representation*, 20(5) :312–322, 2009.

[52] Congyong Su, Yueting Zhuang, Li Huang, and Fei Wu. Steerable pyramid-based face hallucination. *Pattern Recognition*, 38(6) :813–824, 2005.

[53] Simon Baker and Takeo Kanade. Hallucinating faces. In *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, pages 83–88. IEEE, 2000.

[54] Simon Baker and Takeo Kanade. Super-resolution : Reconstruction or recognition. In *Proc. of IEEE-Eurasip Workshop on Nonlinear Signal and Image Processing*, pages 349–385, 2001.

[55] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9) :1167–1183, 2002.

[56] Simon Baker and Takeo Kanade. Super-resolution : Limits and beyond. In *Super-resolution imaging*, pages 243–276. Springer, 2002.

[57] Shu-Fan Lui, Jin-Yi Wu, Hsi-Shu Mao, and Jenn-Jier James Lien. Learning-based super-resolution system using single facial image and multi-resolution wavelet synthesis. In *Asian conference on computer vision*, pages 96–105. Springer, 2007.

[58] CV Jiji, Manjunath V Joshi, and Subhasis Chaudhuri. Single-frame image super-resolution using learned wavelet coefficients. *International journal of Imaging systems and Technology*, 14(3) :105–112, 2004.

[59] Karl S Ni and Truong Q Nguyen. Image superresolution using support vector regression. *IEEE Transactions on Image Processing*, 16(6) :1596–1610, 2007.

[60] Seung P Kim and W-Y Su. Recursive high-resolution reconstruction of blurred multiframe images. *IEEE Transactions on Image Processing*, 2(4) :534–539, 1993.

[61] Russell C Hardie, Kenneth J Barnard, John G Bognar, Ernest E Armstrong, and Edward A Watson. High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering*, 37(1) :247–260, 1998.

[62] Jin Chen, Jose Nunez-Yanez, and Alin Achim. Video super-resolution using generalized gaussian markov random fields. *IEEE Signal Processing Letters*, 19(2) :63–66, 2011.

[63] Ruimin Pan and Stanley J Reeves. Efficient huber-markov edge-preserving image restoration. *IEEE Transactions on Image Processing*, 15(12) :3728–3735, 2006.

[64] Michael K Ng, Huanfeng Shen, Edmund Y Lam, and Liangpei Zhang. A total variation regularization based super-resolution reconstruction algorithm for digital video. *EURASIP Journal on Advances in Signal Processing*, 2007 : 1–16, 2007.

[65] José M Bioucas-Dias, Mario AT Figueiredo, and Joao Pedro Oliveira. Total variation-based image deconvolution : a majorization-minimization approach. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II. IEEE, 2006.

[66] Mingqiang Zhu and Tony Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA Cam Report*, 34 : 8–34, 2008.

[67] Yu-Mei Huang, Michael K Ng, and You-Wei Wen. A new total variation method for multiplicative noise removal. *SIAM Journal on imaging sciences*, 2(1) :20–40, 2009.

[68] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE transactions on Image Processing*, 7(3) :370–375, 1998.

[69] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10) :1327–1344, 2004.

[70] Heng Lian. Variational local structure estimation for image super-resolution. In *2006 International Conference on Image Processing*, pages 1721–1724. IEEE, 2006.

[71] Yu He, Kim-Hui Yap, Li Chen, and Lap-Pui Chau. A nonlinear least square technique for simultaneous image registration and super-resolution. *IEEE Transactions on Image Processing*, 16(11) :2830–2841, 2007.

[72] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos. Total variation super resolution using a variational approach. In *2008 15th IEEE International Conference on Image Processing*, pages 641–644. IEEE, 2008.

[73] Antonio Marquina and Stanley J Osher. Image super-resolution by tv-regularization and bregman iteration. *Journal of Scientific Computing*, 37 (3) :367–382, 2008.

[74] Dennis Mitzel, Thomas Pock, Thomas Schoenemann, and Daniel Cremers. Video super resolution using duality based tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 432–441. Springer, 2009.

[75] Akihiro Yoshikawa, Shotaro Suzuki, Tomio Goto, Satoshi Hirano, and Masaru Sakurai. Super resolution image reconstruction using total variation regularization and learning-based method. In *2010 IEEE International Conference on Image Processing*, pages 1993–1996. IEEE, 2010.

[76] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society : series B (statistical methodology)*, 67(2) :301–320, 2005.

[77] Xuelong Li, Yanting Hu, Xinbo Gao, Dacheng Tao, and Beijia Ning. A multi-frame image super-resolution method. *Signal Processing*, 90(2) :405–414, 2010.

[78] Qiang Chen, Philippe Montesinos, Quan Sen Sun, Peng Ann Heng, et al. Adaptive total variation denoising based on difference curvature. *Image and vision computing*, 28(3) :298–306, 2010.

[79] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Robust shift and add approach to superresolution. In *Applications of Digital Image Processing XXVI*, volume 5203, pages 121–130. SPIE, 2003.

[80] Vorapoj Patanavijit and Somchai Jitapunkul. An iterative super-resolution reconstruction of image sequences using fast affine block-based registration with btv regularization. In *APCCAS 2006-2006 IEEE Asia Pacific Conference on Circuits and Systems*, pages 1717–1720. IEEE, 2006.

[81] Adam WM van Eekeren, Klamer Schutte, and Lucas J van Vliet. Super-resolution on small moving objects. In *2008 15th IEEE International Conference on Image Processing*, pages 1248–1251. IEEE, 2008.

[82] Andrey S Krylov, Alexey S Lukin, and Andrey V Nasonov. Edge-preserving nonlinear iterative image resampling method. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 385–388. IEEE, 2009.

[83] Osama A Omer and Toshihisa Tanaka. Image superresolution based on locally adaptive mixed-norm. *Journal of Electrical and Computer Engineering*, 2010, 2010.

[84] Huihui Song, Lei Zhang, Peikang Wang, Kaihua Zhang, and Xin Li. An adaptive l 1–l 2 hybrid error model to super-resolution. In *2010 IEEE International Conference on Image Processing*, pages 2821–2824. IEEE, 2010.

[85] Pulak Purkait and Bhabatosh Chanda. Super resolution image reconstruction through bregman iteration using morphologic regularization. *IEEE Transactions on Image Processing*, 21(9) :4029–4039, 2012.

[86] Gabriele Steidl, Stephan Didas, and Julia Neumann. Relations between higher order tv regularization and support vector regression. In *International*

Conference on Scale-Space Theories in Computer Vision, pages 515–527. Springer, 2005.

[87] Jonathan S Yedidia, William Freeman, and Yair Weiss. Generalized belief propagation. *Advances in neural information processing systems*, 13, 2000.

[88] Malcolm Acock. Vision : A computational investigation into the human representation and processing of visual information. by david marr. *The Modern Schoolman*, 62(2) :141–142, 1985.

[89] Michal Irani and Shmuel Peleg. Motion analysis for image enhancement : Resolution, occlusion, and transparency. *Journal of visual communication and image representation*, 4(4) :324–335, 1993.

[90] Qiang Wang, Xiaoou Tang, and Harry Shum. Patch based blind image super resolution. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 709–716. IEEE, 2005.

[91] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500) :2323–2326, 2000.

[92] Bo Li, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao. Hallucinating facial images and features. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[93] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Neighbor embedding based single-image super-resolution using semi-nonnegative matrix factorization. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1289–1292. IEEE, 2012.

[94] Jinjun Wang and Shenghuo Zhu. Resolution-invariant coding for continuous image super-resolution. *Neurocomputing*, 82 :21–28, 2012.

[95] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4) :1289–1306, 2006.

[96] W T Freeman, E C Pasztor, and O T Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1) :25–47, 2000.

[97] Jianchao Yang, Hao Tang, Yi Ma, and Thomas Huang. Face hallucination via sparse coding. In *2008 15th IEEE international conference on image processing*, pages 1264–1267. IEEE, 2008.

[98] Shuyuan Yang, Min Wang, Yiguang Chen, and Yaxin Sun. Single-image super-resolution reconstruction via learned geometric dictionaries and clustered sparse coding. *IEEE Transactions on Image Processing*, 21(9) :4016–4028, 2012.

[99] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12) :3736–3745, 2006.

[100] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[101] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[102] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment : from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4) :600–612, 2004.

[103] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[104] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

[105] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv :1508.06576*, 2015.

[106] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[107] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3) :211–252, 2015.

[108] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.

[109] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016.

[110] Kawther Aarizou and Abdelhamid Loukil. Self-similarity single image super-resolution based on blur kernel estimation for texture reconstruction. *International Journal of Computational Science and Engineering*, 25(1) :64–73, 2022.

[111] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018.

[112] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[113] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution : Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

[114] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5) :898–916, 2010.

[115] Mariusz Oszust. No-reference image quality assessment using image statistics and robust feature descriptors. *IEEE Signal Processing Letters*, 24(11) : 1656–1660, 2017.

[116] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1) :372–387, 2015.

[117] Diederik P Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

[118] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1) :145–151, 1999.

[119] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop : Divide the gradient by a running average of its recent magnitude. *COURSERA : Neural networks for machine learning*, 4(2) :26–31, 2012.

[120] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[121] Alex Krizhevsky and Geoffrey E Hinton. Using very deep autoencoders for content-based image retrieval. In *ESANN*, volume 1, page 2. Citeseer, 2011.

[122] Xue Feng, Yaodong Zhang, and James Glass. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1759–1763. IEEE, 2014.

[123] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders : Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[124] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[125] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 2013.

[126] Ilya Loshchilov and Frank Hutter. Sgdr : Stochastic gradient descent with warm restarts. *arXiv preprint arXiv :1608.03983*, 2016.

[127] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment : Unifying structure and texture similarity. *arXiv preprint arXiv :2004.07728*, 2020.