Architectures multi-échelles de type encodeur-décodeur pour la stéréophotométrie

Clément Hardy

Normandie Université

UNIVERSITÉ CAEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Caen Normandie**

Architectures multi-échelles de type encodeur-décodeur pour la stéréophotométrie

Présentée et soutenue par

CLEMENT HARDY

Thèse soutenue le 18/11/2024

devant le jury composé de :

M. DAVID TSCHUMPERLE	Directeur de recherche au CNRS - CNRS	Directeur de thèse
MME JULIE DIGNE	Directeur de recherche au CNRS - CNRS	Président du jury
M. BENJAMIN BRINGIER	Maître de conférences - Université de Limoges	Membre du jury
M. YVAIN QUEAU	Chargé de recherche - CNRS	Membre du jury
M. ANDRÈS ALMANSA	Directeur de recherche - UNIVERSITE PARIS 5 UNIVERSITE PARIS DESCARTES	Rapporteur du jury
M. JEAN-DENIS DUROU	Maître de conférences HDR - Université de Toulouse 3 - Paul Sabatier	Rapporteur du jury

Thèse dirigée par **DAVID TSCHUMPERLE** (Groupe de recherche en informatique, image et instrumentation de Caen)

MIIΣ

GREYC
Laboratoire de recherche en sciences du numérique

Titre — Architectures multi-échelles de type encodeur-décodeur pour le problème de stéréophotométrie

Résumé — La stéréophotométrie est une technique de reconstruction 3D de la surface d'un objet. De plus en plus de recherches s'intéressent à ce problème qui se veut prometteur dans le monde industriel. En effet, la stéréophotométrie peut être utilisée pour détecter les défauts d'usinage de pièces mécaniques ou pour de la reconnaissance faciale par exemple. Cette thèse explore les méthodes d'apprentissage profond pour la stéréophotométrie, notamment les différents aspects liés aux bases de données d'entraînement et aux architectures considérées.

De manière générale, la sur-paramétrisation d'un réseau de neurones est souvent suffisante pour supporter la diversité des problèmes rencontrés. La base de données d'entraînement est alors considérée comme le point clé permettant de conditionner le réseau au problème traité. Par conséquent, pour répondre à ce besoin, nous proposons une nouvelle base de données d'entraînement synthétique. Cette base de données considère une très grande variété de géométries, de textures, de directions ou conditions lumineuses mais également d'environnements, permettant donc de générer un nombre de situation quasiment infini.

Le second point décisif d'une bonne reconstruction concerne le choix de l'architecture. L'architecture d'un réseau doit assurer une bonne capacité de généralisation sur de nouvelles données pour générer de très bons résultats sur des données inédites. Et ce, quelle que soit l'application. En particulier, pour la stéréophotométrie, l'enjeu est d'être capable de reconstruire des images très haute résolution afin de ne pas perdre de détails. Nous proposons alors une architecture multi-échelles de type encodeur-décodeur afin de répondre à ce problème.

Dans un premier temps, nous proposons une architecture fondée sur les réseaux convolutionnels pour répondre au problème de stéréophotométrie calibrée, i.e. quand la direction lumineuse est connue. Dans un second temps, nous proposons une version fondé sur les *Transformers* afin de répondre au problème de stéréophotométrie universelle. C'est-à-dire que nous sommes en capacité de gérer n'importe quel environnement, direction lumineuse, etc., sans aucune information préalable. Finalement, pour améliorer les reconstructions sur des matériaux difficiles (translucides ou brillants par exemple), nous proposons une nouvelle approche que nous appelons "faiblement calibrée" pour la stéréophotométrie. Dans ce contexte, nous n'avons qu'une connaissance approximative de la direction d'éclairage.

L'ensemble des pistes que nous avons explorées a conduit à des résultats convaincants, à la fois quantitatifs et visuels sur l'ensemble des bases de données de l'état-de-l'art. En effet, nous avons pu observer une amélioration notable de la précision de reconstruction des cartes de normales, contribuant ainsi à avancer l'état de l'art dans ce domaine.

Mots clés — Stéréophotométrie, *CNN*, *Transformers*, architecture multi-échelles, encodeur-décodeur, faiblement calibré.

Laboratoires d'accueils — Laboratoire GREYC, UMR CNRS 6072, Université de Caen Normandie, ENSICAEN, 6 Boulevard du Maréchal Juin, 14000 Caen, France.

Title — Multi-scale encoder-decoder architectures for the photometric stereo problem

Abstract — Photometric stereo is a technique for 3D surface reconstruction of objects. This field has seen a surge in research interest due to its potential applications in industry. Specifically, photometric stereo can be employed for tasks such as detecting machining defects in mechanical components or facial recognition. This thesis delves into deep learning methods for photometry stero, with a particular focus on training data and network architectures.

While neural network over-parameterization is often adequate, the training dataset plays a pivotal role in task adaptation. To generate a highly diverse and extensible training set, we propose a new synthetic dataset. This dataset incorporates a broad spectrum of geometric, textural, lighting, and environmental variations, allowing for the creation of nearly infinite training instances.

The second decisive point of a good reconstruction concerns the choice of architecture. The architecture of a network must ensure a good generalization capacity on new data to generate very good results on unseen data. And this, regardless of the application. In particular, for the photometric stereo problem, the challenge is to be able to reconstruct very high-resolution images in order not to lose any details. We therefore propose a multi-scale encoder-decoder architecture to address this problem.

We first introduce a convolutional neural network architecture for calibrated photometric stereo, where the lighting direction is known. To handle unconstrained environments, we propose a Transformers-based approach for universal photometric stereo. Lastly, for challenging materials shiny like translucent or shiny surfaces, we introduce a "weakly calibrated" approach that assumes only approximate knowledge of the lighting direction.

The approaches we have investigated have consistently demonstrated strong performance on standard benchmarks, as evidenced by both quantitative metrics and visual assessments. Our results, particularly the improved accuracy of reconstructed normal maps, represent a significant advancement in photometric stereo.

Keywords — Photometric stereo, CNN, Transformers, multi-scale architecture, encoder-decoder, weakly calibrated.

Institutes — Laboratory GREYC, UMR CNRS 6072, Université de Caen Normandie, ENSICAEN, 6 Boulevard du Maréchal Juin, 14000 Caen, France.

Remerciements

Je souhaite remercier David et Yvain d'avoir encadré ma thèse et d'avoir soutenu mes idées et mes travaux. Ensuite, je tiens à remercier les rapporteurs M. ALmansa et M. Durou d'avoir accepté de relire mon manuscrit, de m'avoir fait part de leurs remarques et conseils afin d'améliorer mon travail. Je remercie également les membres du jury, Mme. Digne et M. Bringier d'avoir accepté d'être présent pour la soutenance.

J'ai également eu la chance de rencontrer des chercheurs et des doctorants de tous horizons et ainsi d'avoir des discussions instructives avec plusieurs d'entre eux. Mes co-bureaux, Thibault et Sidney, ont toujours été d'un très grand soutient. La très bonne ambiance qui régnait dans notre bureau n'a pas toujours été propice au travail mais elle a permis de rendre le travail plus agréable tout au long de ces années.

Finalement, je remercie mes amis et ma famille. Plus particulièrement mes parents et ma sœur pour leur soutien tout au long de mes études, de m'avoir encouragé dans mes choix et pour avoir pris de mes nouvelles quand j'étais trop occupé par la thèse.

Merci à tous.

Table des matières

CHAPITRE 1

Introduction

Nous vivons dans un monde 3D, notre cerveau est donc capable d'interpoler les informations manquantes sur une image 2D afin d'analyser la surface d'un objet ou d'estimer la profondeur dans une scène. Les chercheurs s'intéressent depuis des années à reproduire cette capacité cérébrale, mais cela représente un problème complexe à modéliser mathématiquement et à automatiser. La reconstruction de surfaces 3D est utilisée dans de nombreux domaines d'application, tels que la réalité augmentée, le cinéma, l'imagerie biomédicale, la reconnaissance faciale, etc. Ces techniques sont devenues de plus en plus populaires avec l'évolution de la technologie. La troisième dimension permet non seulement la visualisation 3D, mais également l'extraction des caractéristiques de la surface de l'objet.

En pratique, la 3D est utilisée dans de nombreux domaines, par exemple la géomatique [29], c'est-à-dire le traitement des données géographiques. Dans un tel cas, la reconstruction 3D d'une ville peut être utilisée pour modéliser l'aménagement urbain, créer une maquette 3D avec de nouveaux bâtiments, réaliser des cartes précises pour les randonneurs ou alpinistes, ou encore, par exemple, pour la planification des opérations militaires. La géomatique peut notamment exploiter la stéréo-vision et les images satellites [21]. La reconstruction de notre environnement quotidien a ainsi beaucoup d'applications. La précision de la reconstruction 3D devient même fondamentale dans des domaines critiques, comme la détection de fissures dans les barrages hydrauliques [84, 107] et la numérisation d'objets archéologiques [73]. La 3D a également été exploitée dans le domaine maritime pour la cartographie des habitats, pour le tourisme sous-marin ou pour l'industrie avec le contrôle des canalisations pétrolières, gazoducs, etc. [11, 39, 71]. Ainsi, le champ d'application de la 3D est très large et important.

Pour répondre à ces demandes, de nombreuses approches de reconstruction 3D de surface ont déjà été proposées, comme la triangulation par laser [24, 25, 85] qui consiste à projeter un rayon laser et à étudier sa déviation pour en déduire la distance qui sépare la surface au scanner ainsi que l'orientation de celle-ci. L'impulsion laser [34, 78, 94] (ou plus couramment appelé LiDAR) est aussi fondée sur l'utilisation d'un laser mais mesure le "temps de vol" du faisceau pour réaliser l'aller-retour entre l'appareil et la surface de l'objet. On peut notamment retrouver ces méthodes par laser sur des chaînes de production pour le contrôle qualité. Les scanners "grand public" sont, quant à eux, généralement fondés sur la technique de lumière structurée [26]. Ce type de scanner projette sur l'objet un motif et la déformation du motif sur la surface de l'objet permet de déduire la forme de l'objet. Un autre type de scanner est le scanner par contact [36, 92]. Il utilise une pointe très fine qui vient au contact de l'objet et en déduit sa forme. Cela est possible car le scanner

connaît la position exacte de la pointe. L'avantage de ce système est sa très grande précision mais également la possibilité de scanner des matériaux très réfléchissants ou transparents. Cependant, il ne permet pas de scanner des objets aux formes très complexes et est également très lent.

Dans ce manuscrit, nous nous concentrons sur la stéréophotométrie. Ainsi, nous exploitons des photographies prises avec le même point de vue mais sous différentes conditions lumineuses. Le principal avantage de cette méthode est sa capacité à retrouver les détails les plus fins de la surface. Cependant, le type de matériau influe énormément sur la qualité de reconstruction. Nous verrons que le recours à des techniques fondées sur l'apprentissage supervisé permet de s'affranchir de ces deux difficultés.

1.1 Contexte : la reconstruction 3D à l'aide de photographies

La méthode de reconstruction 3D à base de photographies la plus connue est la photogrammétrie [49, 53, 54, 69], pour laquelle des solutions clé en main d'oeuvre comme Meshroom [30] existent déjà. Elle consiste à prendre en photo un même objet ou une même scène sous différents angles de vue. Des algorithmes sont ensuite employés afin de retrouver la géométrie, i.e. la forme 3D de l'objet ou de la scène en question. Ce type de technique permet d'obtenir la forme "grossière" de l'objet mais omet généralement les petits détails comme les fissures, les rayures, les légères variations dans le relief, etc.

À l'inverse, les méthodes fondées sur la photométrie utilisent la relation qui existe entre la surface d'un objet et le niveau de gris perçu par l'appareil photo. Ainsi, à l'aide d'une seule photographie il est possible de retrouver les petits détails d'un objet. Cette technique appelée le *"shape-from-shading"* a été mise au point par Horn [38].

Cependant, cette méthode n'est pas très robuste en pratique et ne permet pas de traiter des cas avec des réflectances complexes ou de résoudre l'ambiguïté concave-convexe. Les performances globales d'une telle méthode ne sont donc pas optimales du point de vue de la reconstruction.

Une extension possible pour pallier à ces inconvénients consiste non pas à prendre une seule photographie, mais plusieurs avec des conditions d'éclairages (i.e. direction lumineuses) différentes. Cette approche est la stéréophotométrie. Elle a été initialement proposée par Woodham [101], et constitue l'approche développée dans cette thèse.

1.2 La stéréophotométrie

La stéréophotométrie est une technique de reconstruction 3D qui, par l'analyse des photographies prises depuis le même point de vues mais sous différents éclairages, permet de reconstruire la surface d'un objet en 3D. De manière plus détaillée, cette technique fonctionne en analysant l'intensité lumineuse mesurée par le récepteur photographique sous différentes conditions lumineuses, comme illustré sur la figure 1.1, afin d'estimer la forme, et potentiellement la réflectance de l'objet. Le relief est typiquement reconstruit sous la forme d'une carte de normales, qui sont ensuite intégrées pour reconstruire la forme 3D [77] (voir la figure 1.2).

Photographie 1. Photographie 2. Photographie 3.

FIGURE 1.1 – Exemples d'acquisitions d'images de stéréophotométrie. Dans cet exemple, une statuette de Bouddha a été photographiée sous trois éclairages différents : la lumière vient de la droite sur la première photographie, de la gauche sur la deuxième et de face sur la dernière.

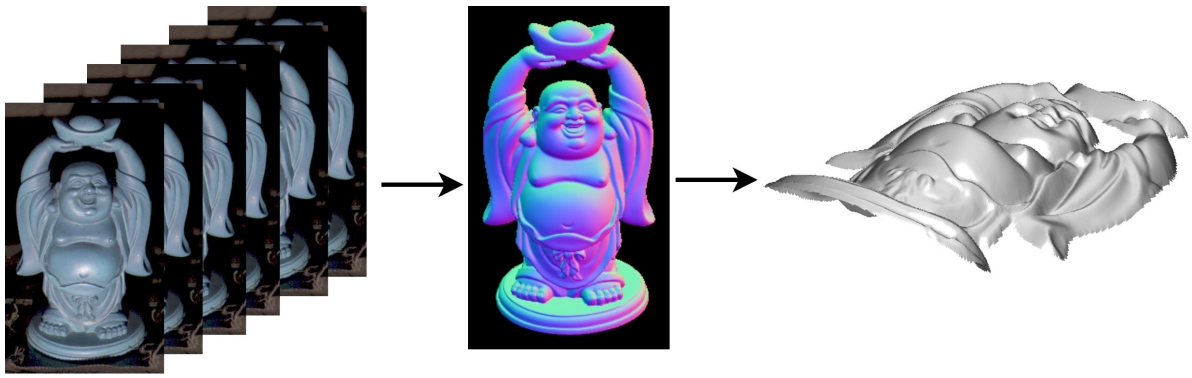


FIGURE 1.2 – Principe d'un algorithme de stéréophotométrie : la statuette est prise en photo sous différentes directions lumineuses. À l'aide de ces photos une carte de normales est estimée et par la suite, elle est intégrée pour obtenir un objet en 3D.

La stéréophotométrie peut être divisé en plusieurs sous-catégories correspondant à différents contextes d'applications. Les deux sous-catégories les plus utilisées sont la stéréophotométrie calibrée et non calibrée. Il existe également la stéréophotométrie universelle, qui a émergé ces dernières années, et qui montre un fort potentiel applicatif.

- La stéréophotométrie calibrée suppose les directions lumineuses ainsi que l'intensité des sources lumineuses connues. De plus, l'environnement est également contrôlé, c'est-à-dire que les prises de vues s'effectuent généralement dans le noir, sans lumière ambiante ou extérieure et sans réflexion sur d'autres objets ou surfaces. Dans ce cas, seule la lumière de la source lumineuse voulue est présente dans la scène. En revanche, cette méthode est contraignante à mettre en œuvre en pratique. En effet, pour beaucoup d'application, l'environnement n'est pas ou difficilement contrôlable et les directions lumineuses exactes sont difficiles à acquérir.

- Acquérir les directions et les intensités lumineuses est très complexe et contraignant car du matériel spécifique est nécessaire, ainsi qu'un algorithme d'étalonnage robuste. Ce procédé n'est donc pas simple à mettre en place. Pour contourner cette problématique, des techniques de stéréophotométrie *non calibrée* ont été développées pour travailler sans aucune information sur les sources lumineuses. L'environnement ambiant d'acquisition reste cependant contrôlé, i.e. les acquisitions sont prises dans le noir sans réflexion sur une autre surface que l'objet.

- La stéréophotométrie *universelle* va encore plus loin que la non calibrée car aucune information sur les sources lumineuses n'est connue mais surtout l'environnement ambiant lors de l'acquisition n'est plus du tout contrôlé. Cela signifie que les directions lumineuses et l'environnement ne sont pas connus, et tous les types de faisceaux lumineux peuvent être considérés (faisceaux parallèles et non parallèles). Ces approches sont donc très versatiles d'un point de vue applicatif car elles permettent de s'adapter à n'importe quel situation et sont utilisables facilement. En revanche, ce type d'approches est beaucoup plus complexes à mettre en place lors de l'apprentissage des réseaux de neurones. En effet, la base de données d'entraînement doit être complète et représentative de l'ensemble des situations possibles afin d'obtenir la meilleure paramétrisation du réseau pour une bonne généralisation.

1.3 Motivations et contributions

Pour aborder la stéréophotométrie, on propose et développe des méthodes d'apprentissage par réseaux de neurones. En effet, depuis quelques années, les réseaux de neurones ont démontré leur efficacité dans beaucoup de domaines et particulièrement pour le traitement d'images en général.

En traitement d'images, ce type d'approches constitue aujourd'hui l'état-de-l'art pour la majorité des applications car elles ont l'avantage de modéliser très efficacement la relation entre les variables d'entrée et de sortie par l'apprentissage des poids de différents filtres, par exemples convolutifs. Cependant, au début de la thèse, nous avons remarqué que beaucoup de ces méthodes innovent sur les architectures mais prennent rarement en compte la résolution native des images d'entrée. Par conséquent, nous pouvons observer une perte d'information importante dans les détails des cartes de normales reconstruites. Par ailleurs, les bases de données d'entraînement restent limitées, ce qui a des répercussions sur la généralisation des résultats en phase de test.

Dans cette perspective, cette thèse a pour objectif d'améliorer les précisions de reconstruction des cartes de normales, tout en conservant la résolution des images très haute résolution à l'aide de nouvelles architectures. Parmi tous les types de réseaux de neurones qui existent, nous avons choisi de nous placer dans un cadre de réseaux multi-échelles de type encodeur-décodeur. De plus, nous proposons une base de données d'entraînement permettant une meilleure généralisation et de meilleures performances sur tous les types de matériaux. Ces contributions ont été validées par les publications suivantes :

- Construction d'un jeu de données d'apprentissage adapté pour la reconstruction 3D par stéréophotométrie (GRETSI) [31] en 2022,
- MS-PS : A Multi-Scale Network for Photometric Stereo With a New Comprehensive Training Dataset (World Society for Computer Graphics) [33] en 2023,
- Uni MS-PS : A multi-scale encoder-decoder transformer for universal photometric stereo (Computer Vision and Image Understanding) [32] en 2024.

1.4 Plan du manuscrit

Cette thèse est ainsi orientée sur la stéréophotométrie à l'aide de méthodes fondées sur l'apprentissage. Nos contributions ont pour objectif d'améliorer les résultats selon deux axes principaux : la création de bases de données d'entraînement et la mise en place d'architectures multi-échelles de type encodeur-décodeur afin d'améliorer les résultats de l'état-de-l'art.

Le chapitre 2 est consacré à un état-de-l'art général sur les méthodes existantes pour aborder et résoudre la stéréophotométrie. Il existe une large variété de méthodes répondant à ce problème. En

effet, il existe des méthodes de traitement d'images classiques, comme les méthodes variationnelles, et plus récemment, les méthodes fondées sur l'apprentissage profond ont démontré une meilleure capacité de reconstruction d'un point de vue quantitatif et qualitatif. Ce chapitre a donc pour objectif d'exposer les points forts et les points faibles de ce type d'approche.

Le chapitre 3 présente l'ensemble des bases de données existantes pour répondre au problème de stéréophotométrie. Dans ce chapitre, nous étudions et comparons chacune des bases de données publiques pour l'entraînement et le test. De plus, nous présentons les bases de données synthétiques d'entraînement que nous avons générées, qui sont beaucoup plus diversifiées d'un point de vue des géométries, des matériaux, des directions lumineuses et des environnements que celles couramment utilisées dans la littérature. L'intérêt de générer une base de données plus diversifiée et complexe que celle existante a été démontré par une première publication au colloque GRETSI [31].

Le chapitre 4 se focalise sur la résolution de la stéréophotométrie calibrée. Nous proposons une architecture multi-échelles de type encodeur-décodeur utilisant un réseau de neurones convolutif (*CNN*) pour extraire les caractéristiques et reconstruire les cartes des normales. Nous montrons que notre architecture, couplée avec nos nouvelles bases de données d'entraînement, donne les meilleurs résultats de l'état-de-l'art pour le problème de stéréophotométrie calibrée. Ce chapitre a fait l'objet d'une contribution à la conférence WSCG [33].

Ensuite, le chapitre 5 introduit une architecture multi-échelles de type encodeur-décodeur fondée sur les *Transformers* pour répondre résoudre la stéréophotométrie universelle. Notre méthode se généralise sur des images en très grande dimension tout en conservant de très bons résultats, notamment sur des matériaux spéculaires réputés comme très difficiles, tels que l'aluminium. Cette contribution a été publiée dans la revue CVIU [32].

Finalement, le chapitre 6 présente une toute nouvelle approche pour résoudre la stéréophotométrie. En effet, les méthodes non-calibrées et universelle ont généralement du mal à prédire la direction lumineuse avec précision sur des matériaux difficiles comme les matériaux translucides, par exemple le plexiglas ou le verre. Ainsi, nous proposons de travailler dans un cadre que nous appelons "faiblement calibré". Cela signifie que nous ne donnons pas au réseau la direction lumineuse exacte mais seulement une indication globale, i.e. "haut", "bas", "droite", "gauche". Nous testons cette nouvelle approche faiblement calibrée sur nos architectures multi-échelles.

Un dernier chapitre conclut cette thèse et propose des pistes d'amélioration et des perspectives pour la stéréophotométrie. F

CHAPITRE 2

État-de-l'art général

La résolution du problème de la stéréophotométrie se fait traditionnellement par une approche de type problème inverse. Cette approche consiste à inverser un modèle physique dans le but de trouver le relief qui maximise la vraisemblance des observations. La seconde grande famille d'approches rassemble les méthodes fondées sur l'apprentissage profond. Ce type de méthodes permet de modéliser la relation entre les variables d'entrée et de sortie à l'aide de couches convolutives, de filtres linéaires, de fonctions d'activations, etc. Elles ont pour avantage d'être très performantes en termes d'erreur de reconstruction. Cependant, l'apprentissage en lui-même peut être long ou difficile.

Dans ce chapitre, nous commençons par présenter très succinctement les méthodes de type problème inverse, qui ne font pas partie des méthodes développées durant cette thèse. Ensuite, nous proposons une introduction à la théorie des réseaux de neurones. Nous poursuivons avec une présentation des méthodes d'apprentissage permettant de résoudre la stéréophotométrie au sens large dans la section 2.3, i.e. méthodes calibrées, non calibrées et universelles. Pour finir, nous présentons dans la section 2.4 les bases de données de stéréophotométrie disponibles dans la littérature pour l'entraînement et la validation.

2.1 Approches de type problème inverse

La première approche, développée par Woodham *et al.* en 1979 [100], considère le cas idéal d'une surface ayant une réflectance parfaitement lambertienne. C'est-à-dire que la luminance est proportionnelle au cosinus de l'angle entre l'éclairage incident et la normale à la surface, et le coefficient de proportionnalité est l'albédo (la couleur de la surface). En pratique, peu de matériaux respectent la loi de Lambert. Par exemple, certains ont une composante spéculaire comme les métaux, donnant des tâches très brillantes à la surface et d'autres sont translucides comme le verre ou le plexiglas.

Pour le cas des matériaux spéculaires, des solutions ont été proposées, par exemple les méthodes [9, 19, 103] qui détectent les tâches spéculaires et les ignorent durant l'optimisation. D'autres méthodes, telles que [27, 28, 90, 109], cherchent quant à elles à gérer ces zones de haute spécularité en utilisant un modèle physique plus poussé comme par exemple Torrance-Sparrow [27] ou Ward [28]. Enfin, une alternative est de changer la méthode d'optimisation et de remplacer la méthode des moindres carrés qui est la solution d'optimisation classique par une méthode plus robuste qui traite les tâches spéculaires comme des valeurs aberrantes [44, 76, 102]. Toutes ces méthodes restent cependant limitées au modèle Lambertien. Traiter le cas de réflectance plus complexes nécessite de

recourir à des approches neuronales.

2.2 Réseaux de neurones

L'essor des réseaux de neurones et leurs performances remarquables ont récemment révolutionné le domaine de la stéréophotométrie. Deux familles de réseaux de neurones sont exploitées dans cette thèse, les réseaux de neurones convolutifs et les *Transformers*. Dans cette partie, nous allons présenter globalement le fonctionnement général de ce type d'architecture.

Un réseau de neurones convolutifs (*CNN*) est une catégorie de modèle d'apprentissage automatique, à savoir un type d'algorithme d'apprentissage profond bien adapté pour l'analyse ou la reconstruction de données visuelles. Les *CNN* utilisent les principes de l'algèbre linéaire, en particulier les opérations de convolution, pour extraire des caractéristiques et identifier des modèles/statistiques dans les images. Bien que les *CNN* soient principalement utilisés pour traiter des images, ils peuvent également être adaptés pour fonctionner avec tout autre type de signal comme le langage ou l'audio par exemple.

Un des premiers réseaux convolutif est le modèle LeNet, proposé par LeCun *et al.* en 1998 [56], mais l'engouement autour des réseaux convolutifs est apparu à la suite du réseau AlexNet [55], proposé par Krizhevsky *et al.* et fortement inspiré du réseau LeNet, pour la classification.

2.2.1 Couches d'un réseau de neurones convolutif

Le principe de base des *CNN* est d'extraire et d'apprendre des caractéristiques au sein d'une base de données afin de classifier, reconstruire ou bien générer à partir des connaissances acquises sur cette base. Ainsi, les *CNN* sont composés de plusieurs niveaux permettant de modéliser efficacement la relation entre les variables d'entrée et de sortie. Par exemple, via l'utilisation de couches convolutives ou denses, de couches de pooling ou encore de fonctions d'activation, comme cela est détaillé ci-dessous.

• Les *couches convolutives* sont le cœur d'un *CNN* car elles permettent d'extraire des caractéristiques de complexité variable au sein d'une image à l'aide d'un filtrage convolutionnel. Ainsi, l'apprentissage des noyaux de convolution permettant l'extraction de ces caractéristiques s'adapte au problème considéré. C'est l'apprentissage automatique des poids qui rend les réseaux convolutifs très performants. La carte de caractéristiques résultante peut être donc vue comme un filtre qui indique où se situent les caractéristiques d'intérêt dans l'image après application de ce filtre.

• Les *fonctions d'activation* permettent d'introduire de la non-linéarité au réseau. Cela permet donc d'augmenter la capacité du réseau à modéliser des données complexes. Ces fonctions sont inspirées du potentiel d'activation du cerveau humain et reproduisent son comportement en permettant ou non le passage de l'information dans la suite du réseau. Elles vont donc décider si la réponse d'un neurone doit être activée ou non.

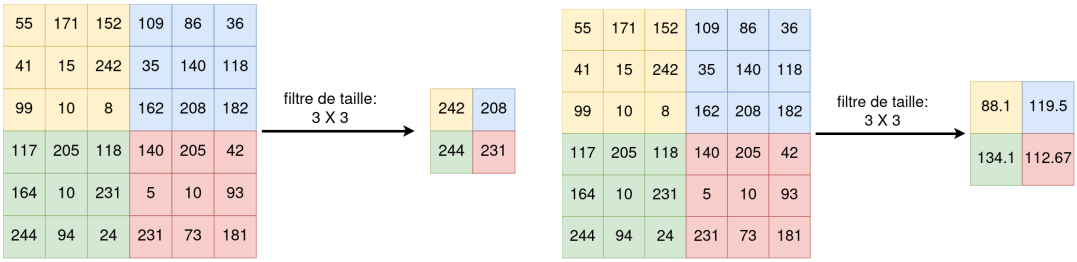
Les deux principales fonctions d'activations sont la fonction sigmoïde et la fonction *ReLU*. La fonction sigmoïde écrase les valeurs d'entrée en condensant la sortie entre 0 et 1. Cependant, elle peut être la cause du problème du "*vanishing gradient*" [106] vers 0 si le réseau est trop profond. Ainsi, elle est de moins en moins utilisée car elle perd l'information dûe à la saturation et donc le gradient a de très forte chance d'arriver à 0 lorsque les valeurs sont très grandes ou très petites.

La fonction *ReLU* (*Rectified Linear Unit*) laisse toutes les valeurs supérieures à 0 inchangées et attribue 0 aux valeurs négatives. La plupart des réseaux utilisent cette fonction ou l'une de ses variantes telles que *leaky ReLU* ou *ELU* [8]. Les variantes ont généralement pour objectif de nuancer les valeurs négatives afin que le gradient ne soit pas toujours égal à 0 dans cet intervalle.

Le principal avantage de ces fonctions est qu'elles sont parfaitement adaptées à l'apprentissage

automatique par rétro-propagation car elles sont différentiables ou sous-différentiables.

• La couche de pooling permet de sous-échantillonner une carte de caractéristiques données en sortie d'une couche convolutive. On retrouve principalement cette opération dans un encodeur ou un réseau de classification. Les deux types de *pooling* les plus répandus sont le *pooling* maximal et moyennant. Le premier garde les caractéristiques dominantes en sélectionnant la valeur maximale dans le voisinage de chaque pixel. Le second garde l'ensemble des informations extraites en moyennant le voisinage pour chaque pixel, comme cela est illustré sur la figure 2.1.



Pooling maximal. *Pooling* moyennant.

FIGURE 2.1 – Schéma représentatif d'une couche de *pooling* maximal et moyennant.

2.2.2 Les *Transformers*

Ces dernières années, les *Transformers* ont amélioré les performances de l'état-de-l'art pour l'ensemble des applications en intelligence artificielle. Initialement proposés pour le traitement du langage (*NLP*) [91], ce type d'architecture s'est propagé à la vision par ordinateur, le traitement du son ou encore les graphes par exemple.

Les *Transformers* ont initialement été proposés afin de résoudre le manque de parallélisation des réseaux récurrents (*RNN*) [68]. De plus, ils ont également été conçus afin de modéliser les relations/dépendances longues distances entre les mots d'une phrase [91]. Pour ce faire, le principe de base d'un *Transformers* est l'utilisation de mécanismes d'attention.

L'idée générale des *Transformers* pour le texte a été transposée à l'image dans une méthode appelée *ViT* (*Vision Transfomers*) [52]. Ainsi, de même que pour le langage, les mécanismes d'attention ont pour objectif de mettre en corrélation les caractéristiques intra-image afin d'en conserver toute l'information. Par conséquent, au lieu de traiter des mots dans une phrase, nous traitons des patchs dans une image. En effet, celle-ci est d'abord découpée en patchs avec ou sans chevauchement puis ces patchs sont mis sous forme de vecteur, comme cela est illustré sur la figure 2.2.

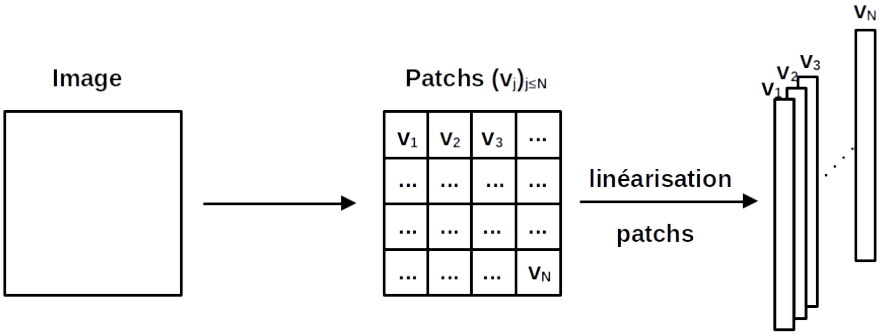


FIGURE 2.2 – Préparation d'une image en entrée d'un *Transfomer*.

Une fois que nous avons nos patchs, l'architecture d'un bloc se présente de la façon suivante :

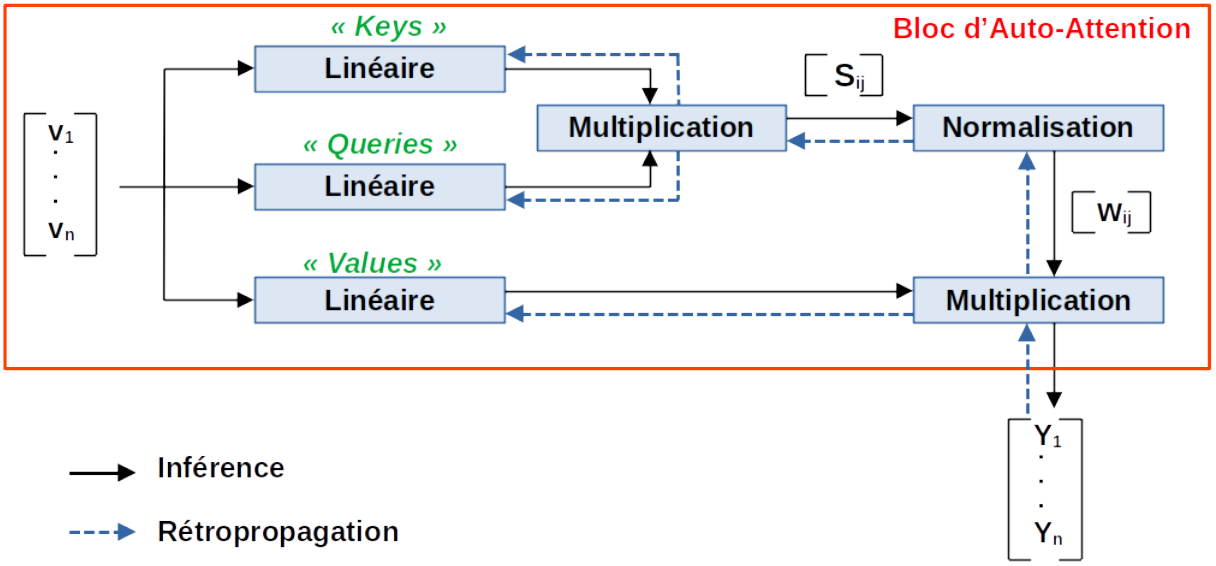


FIGURE 2.3 – Schéma global d'un bloc d'auto-attention.

Dans le schéma global d'un bloc d'auto-attention, on remarque que chaque patch passe trois fois dans une couche linéaire. Par ailleurs, une couche linéaire est équivalente à une multiplication matricielle. Cela est donc très coûteux en terme de nombre de paramètres et en temps de calcul. Ainsi, il n'y a que trois couches de paramètres entraînables par bloc et les blocs sont généralement empilés pour former un ViT [52] standard.

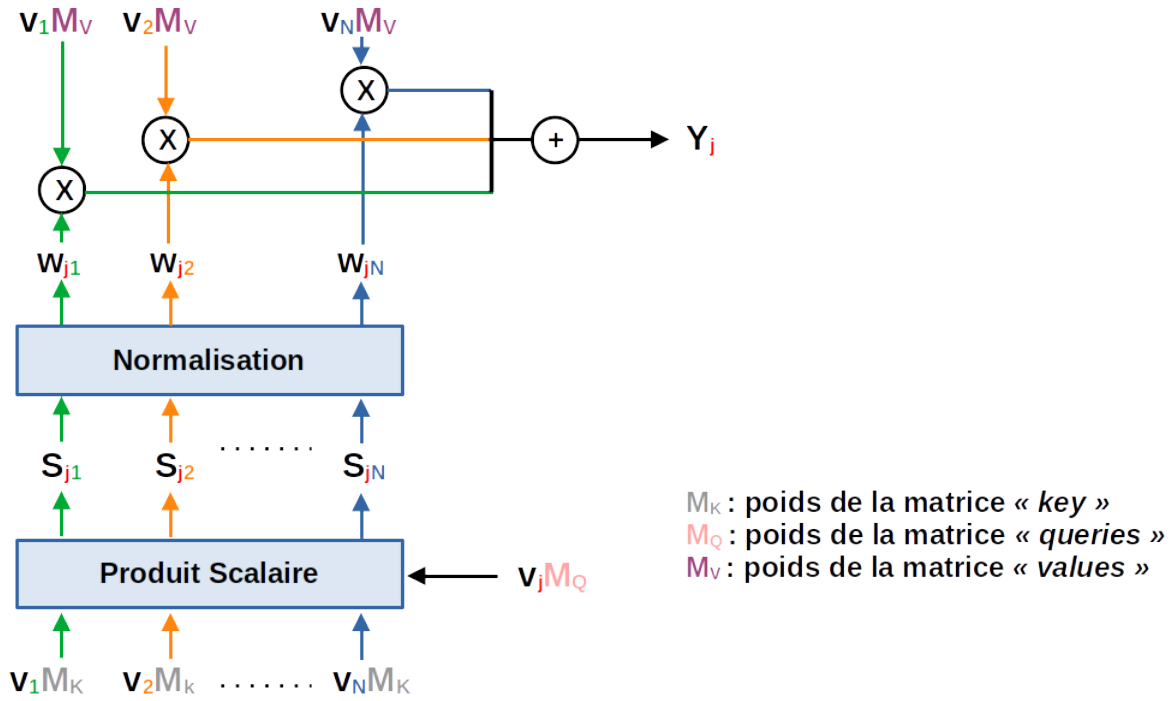


FIGURE 2.4 – Schéma détaillé de l'attention par rapport au patch v_j.

De façon plus détaillée, si l'on se concentre sur un patch v_j, les détails sont illustrés en figure 2.4. On peut donc ainsi dire que chaque patch v_j, $j \leq N$, est calculé de la façon suivante :

$$Y_j = \sum_{i \leq N} \frac{v_i M_K \cdot v_j M_Q}{||v_i M_K \cdot v_j M_Q||} v_i M_V, \tag{2.1}$$

où Y_i est la sortie du mécanisme d'attention pour le patch v_i, $(v_i)_{i \leq N}$ et M_Q, M_K et M_V les poids des couches linéaires à optimiser.

À l'aide de cette équation, on voit bien que toutes les informations extraites par chacune des couches linéaires pour l'ensemble des patchs sont synthétisées dans la sortie du mécanisme d'attention.

2.2.3 Optimisation des poids d'un réseau de neurones

L'optimisation d'un réseau F consiste à ajuster ses poids θ_F, par exemple à l'aide d'un algorithme de descente stochastique [81] pouvant être décrit comme une succession d'itérations. Une itération basique peut être résumée ainsi :

- ▷ Prendre m échantillons $\{x_1, ..., x_m\}$ à partir de p_{data}.
- ▷ Étape de SGD (Descente de Gradient Stochastique) :

$$\theta_F \leftarrow \theta_F + \varepsilon \nabla_{\theta_F} \frac{1}{m} \sum_{i \leq m} \mathcal{L}(y_i, F(x_i)) \tag{2.2}$$

où $\mathcal{L}$ représente la fonction de perte utilisée pour optimiser les poids du réseau, $\{x_1, ..., x_m\}$ des échantillons de notre base de données p_{data} et $\{y_1, ..., y_m\}$ les vérités terrains associées.

Le paramètre m correspond à la taille du batch. Une taille de batch trop petite ne permet pas au réseau de capter pleinement la distribution des données. En effet, lors de l'entraînement, les poids sont optimisés en faisant la moyenne des gradients de chaque échantillon du batch. Il faut qu'il y ait donc suffisamment d'échantillons pour avoir une estimation correcte. S'il y a trop peu d'échantillons par batch, les poids du réseau peuvent alors changer drastiquement d'une itération à l'autre et donc impacter les performances, voire ne pas converger dans un cas extrême. Au contraire, une taille de batch trop grande peut ralentir fortement l'entraînement car beaucoup d'échantillons doivent être prises en compte à chaque passage.

En pratique, les optimiseurs les plus utilisés sont Adam [51], SGD [88] et RMSprop [35].

Maintenant que les deux grands types de familles de réseau de neurones ont été présentées, nous allons présenter les méthodes de stéréophotométrie de l'état-de-l'art fondées sur l'apprentissage profond.

2.3 Méthodes neuronales pour la stéréophotométrie

2.3.1 Perceptron multi-couches

En 2017, avec la démocratisation des réseaux de neurones dans la recherche en traitement d'images, les premières méthodes fondées sur ceux-ci apparaissent pour répondre à la stéréophotométrie, notamment la méthode DPSN proposée par Santo *et al.* [83].

DPSN est une méthode pixel à pixel dans laquelle les auteurs ont utilisé un perceptron multi-couche, comme cela est illustré sur la figure 2.5, qui prend en entrée un vecteur dont la valeur correspond à la valeur perçue en un pixel sous différentes directions lumineuses. Les canaux des images RGB sont traités indépendamment et sont donc considérés comme 3 images en niveau de gris.

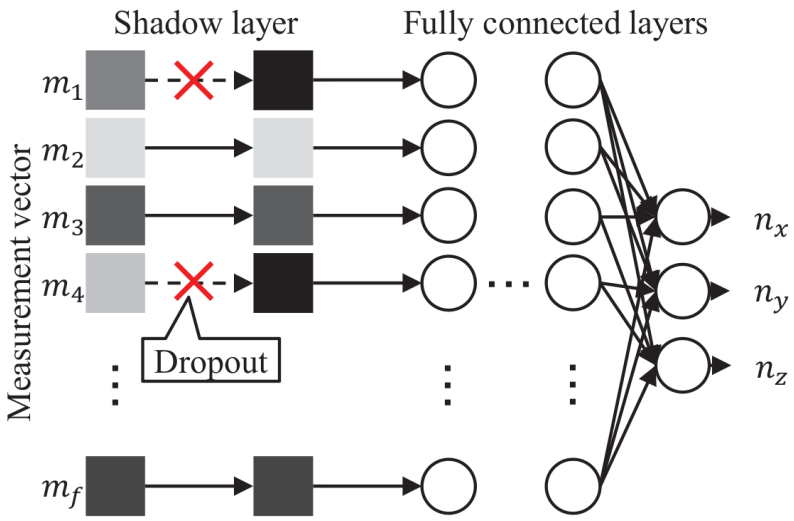
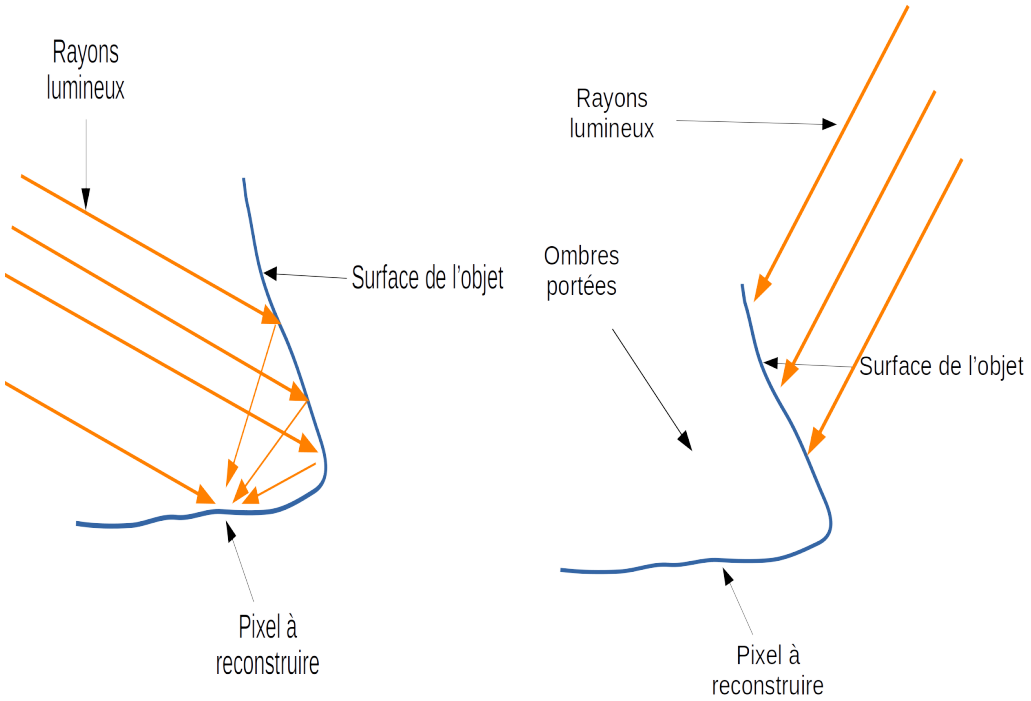


FIGURE 2.5 – Architecture DPSN proposée par Santo *et al.* [83] en 2017.

Le réseau de neurones prédit alors trois valeurs qui correspondent à la valeur de la normale au pixel traité. Dans le cas d'une image RGB, le réseau prédit ainsi trois normales par pixel qui sont ensuite moyennées pour obtenir la normale finale. Ce type d'approche pixel à pixel, où le modèle n'a pas accès à l'information des autres pixels, peut poser plusieurs difficultés comme par exemple les ombres portées ou encore les réflexions sur les autres parois de l'objet. Pour les réflexions, elles constituent un problème dans le sens où tous les rayons lumineux ne proviennent pas directement de la source lumineuse, comme cela est schématisé sur la figure 2.6(a). L'intensité lumineuse relevée en ce pixel n'est donc pas forcément la bonne.



(a) Inter-réflexion. (b) Ombres portées.

FIGURE 2.6 – Schéma de réflexion sur une autre paroi de l'objet (inter-réflexion) et phénomène d'ombres portées.

De même, une partie de l'objet peut créer une ombre sur une autre partie de celui-ci et ainsi empêcher les rayons lumineux d'éclairer les points qui devraient l'être, biaisant l'estimation de la normale. Ce phénomène, appelé "ombre portée", est illustré en figure 2.6(b).

Pour améliorer la gestion des ombres portées, les auteurs ont ainsi proposé un module appelé *"shadow layer"*. Son objectif est de mettre à zéro aléatoirement des valeurs du vecteur d'entrée à la manière d'un *dropout*. La mise à zéro de certains pixels simule les ombres portées en ces pixels.

Cependant, l'utilisation d'un perceptron multi-couche a une contrainte importante dûe à son architecture. En effet, le nombre d'images en entrée est fixe. D'un point applicatif, le fait de toujours avoir le même nombre d'images est très contraignant. Pour résoudre cette problématique, plusieurs approches ont été proposées, notamment celles fondées sur les couches de *pooling* ou les cartes d'observations.

2.3.2 Méthodes fondées sur le *pooling*

La première utilisation d'un réseau de neurones convolutifs pour la stéréophotométrie calibrée a été proposé par Chen *et al.* en 2018 [18]. Ce réseau, PS-FCN, fait passer chaque image indépendamment dans le réseau et fusionne les cartes de caractéristiques obtenues à l'aide d'une couche de *pooling*. Dans ce travail, le réseau de neurones a été conçu pour fonctionner dans le contexte de la stéréophotométrie calibrée en prenant en entrée les directions lumineuses. Ainsi, celles-ci sont concaténées avec leur image respective en entrée. D'un point de vue architectural, le réseau de neurones utilise un extracteur de caractéristiques composé de convolution. Celui-ci traite chaque couple image/direction lumineuse indépendamment. Le *max-pooling* fusionne ensuite les caractéristiques pixel à pixel. Ces caractéristiques fusionnées sont finalement utilisées par un décodeur composé de convolutions et de convolutions transposées pour générer la carte des normales. L'architecture de ce réseau est illustré sur la figure 2.7.

Les auteurs introduisent également une version de ce réseau prévue pour la stéréophotométrie non calibrée. La seule modification apportée est le fait de ne prendre que les images en entrée.

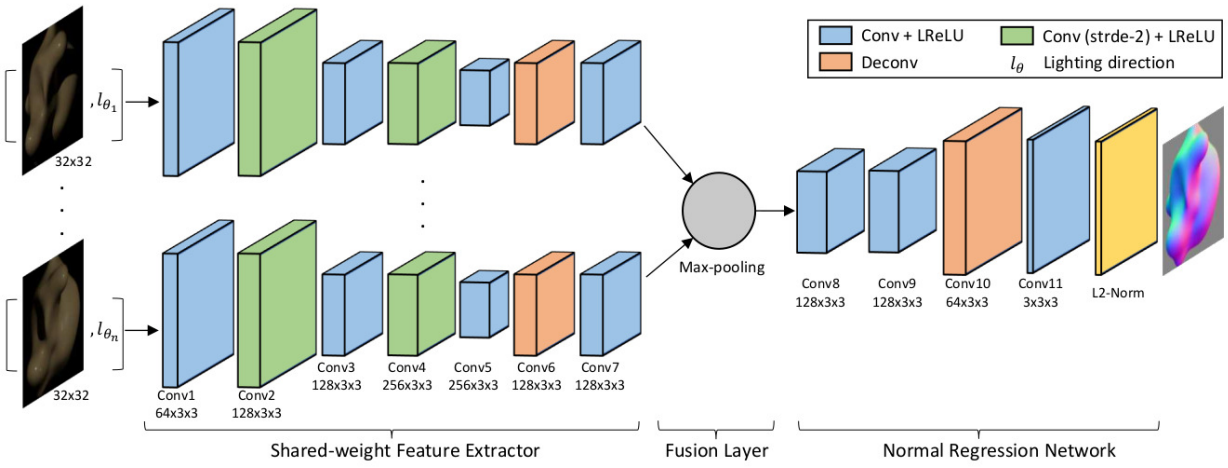


FIGURE 2.7 – Architecture PS-FCN proposée par Chen *et al.* [18] en 2018.

L'intérêt de cette méthode réside dans l'utilisation du *pooling*. En effet, celui-ci est capable de synthétiser l'information de multiples images, en prenant la valeur maximale dans le sens image à image de chaque pixel dans le cas d'un *max-pooling* ou la valeur moyenne dans le cas d'un *mean-pooling*. Cette force permet ainsi de pouvoir prendre en entrée un nombre d'images qui varie. Jusqu'à présent, les méthodes préexistantes ne pouvaient prendre en entrée qu'un nombre fixe d'images.

Les auteurs justifient le choix du *max-pooling* contre le *mean-pooling* par le fait que, pour la stéréophotométrie, les régions des images avec des tâches spéculaires ou autres fortes intensités lu-

mineuses donnent une information très pertinente pour reconstruire la normale. Or, le *max-pooling* extrait naturellement les caractéristiques de ces zones. À l'inverse, les zones des images qui n'ont pas ou peu d'intérêt sous certaines directions lumineuses pour la reconstruction de la normale peuvent être ignorées. Avec une opération de *mean-pooling*, les zones sans importance impactent les valeurs de sortie de la couche de *pooling*, ce qui n'est pas le cas avec un *max-pooling*.

L'architecture PS-FCN a ensuite été modifiée pour gérer la stéréophotométrie non calibrée par Chen *et al.* dans [17] en 2019. Pour cela, les auteurs introduisent un réseau de neurones qui prédit les directions lumineuses, ainsi que l'intensité lumineuse pour chaque image. Celles-ci sont ensuite utilisées par le réseau de neurones PS-FCN qui prend en entrée les images et les directions lumineuses associées et prédites par le premier réseau, comme dans le cadre calibré. Les intensités lumineuses prédites servent aussi à normaliser les images à l'entrée du second réseau.

De plus, les directions lumineuses prises en entrée sont discrétisées afin de simplifier le travail du premier réseau de neurones, comme cela est illustré sur la figure 2.8.

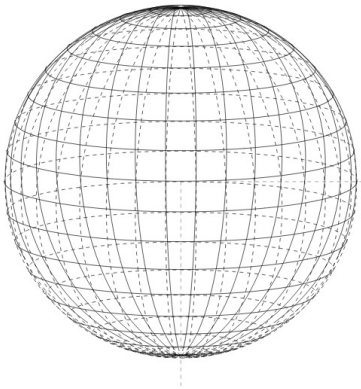


FIGURE 2.8 – Discrétisation régulière de l'hémisphère pour les directions lumineuses.

L'avantage de la discrétisation, comparativement à la régression naïve mentionnée plus hauts a été montré par les auteurs au niveau des performances. Remarquons que les directions lumineuses ne sont pas les seules à être discrétisées, l'intensité lumineuse est elle aussi discrétisée.

D'un point de vue architectural, le réseau en charge de prédire les directions/intensités lumineuses se compose d'un extracteur de caractéristiques traitant chaque image indépendamment. Ensuite, un *max-pooling* vient fusionner les caractéristiques de chaque image pixel à pixel. Cependant, à la différence du réseau PS-FCN, les caractéristiques fusionnées sont concaténées avec celles extraites (celles avant la fusion). Cela permet d'intégrer l'information globale à l'information locale pour chaque image. Ces caractéristiques concaténées sont ensuite utilisées par un classifieur qui prédit à quelle sous-zone de l'hémisphère appartiennent les directions lumineuses. L'architecture proposée par Chen *et al.* est illustrée sur la figure 2.9.

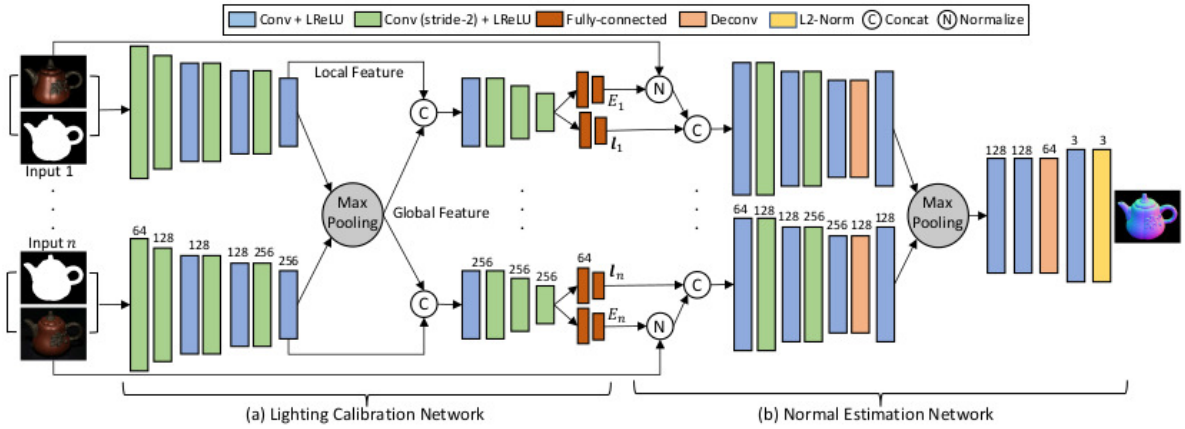


FIGURE 2.9 – Architecture SDPS-Net [17] proposée par Chen *et al.* en 2019.

Une autre variante de PS-FCN intitulée DSMF [64] a été proposée par Liu *et al.* en 2022. La méthode DSMF se base sur un double *max-pooling*, illustré sur la figure 2.10, et a été proposée pour répondre au problème de la Stéréophotométrie calibrée.

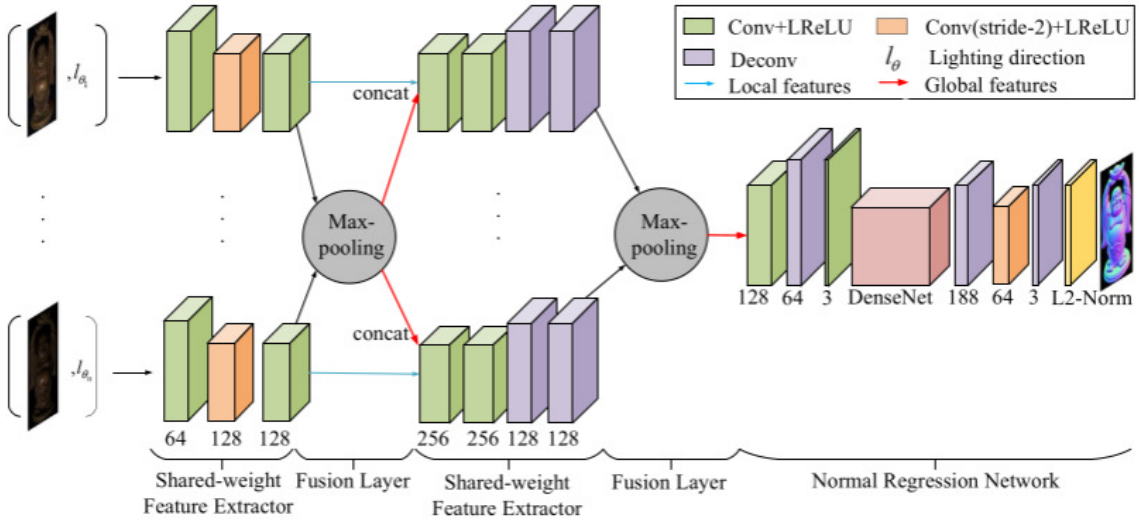


FIGURE 2.10 – Architecture DSMF [64] proposée par Liu *et al.* en 2022.

Le premier *max-pooling* est utilisé pour extraire à la fois des caractéristiques locales et globales inter-images. La concaténation du *max-pooling* et de la convolution précédente permet à la fois de conserver les caractéristiques locales tout en intégrant des caractéristiques globales image à image. Les auteurs argumentent que conserver uniquement le premier *max-pooling* réduit le nombre de caractéristiques utiles et par conséquent, réduit les performances des algorithmes. Au niveau des performances, intégrer un premier *max-pooling* au milieu de l'extracteur des caractéristiques permet d'obtenir des performances supérieures que ce soit avec peu d'images ou toutes les images disponibles.

Le deuxième *max-pooling*, quant à lui, ne conserve que des caractéristiques globales. Les auteurs indiquent que suffisamment de caractéristiques ont été synthétisées pour se permettre de ne conserver que les caractéristiques globales. Le simple ajout d'un deuxième *max-pooling* permet d'obtenir des gains de performances significatifs comparativement à l'architecture de base PS-FCN voir la figure 2.11. En effet, en supprimant le premier *max-pooling*, l'architecture n'est qu'autre que celle de PS-FCN.

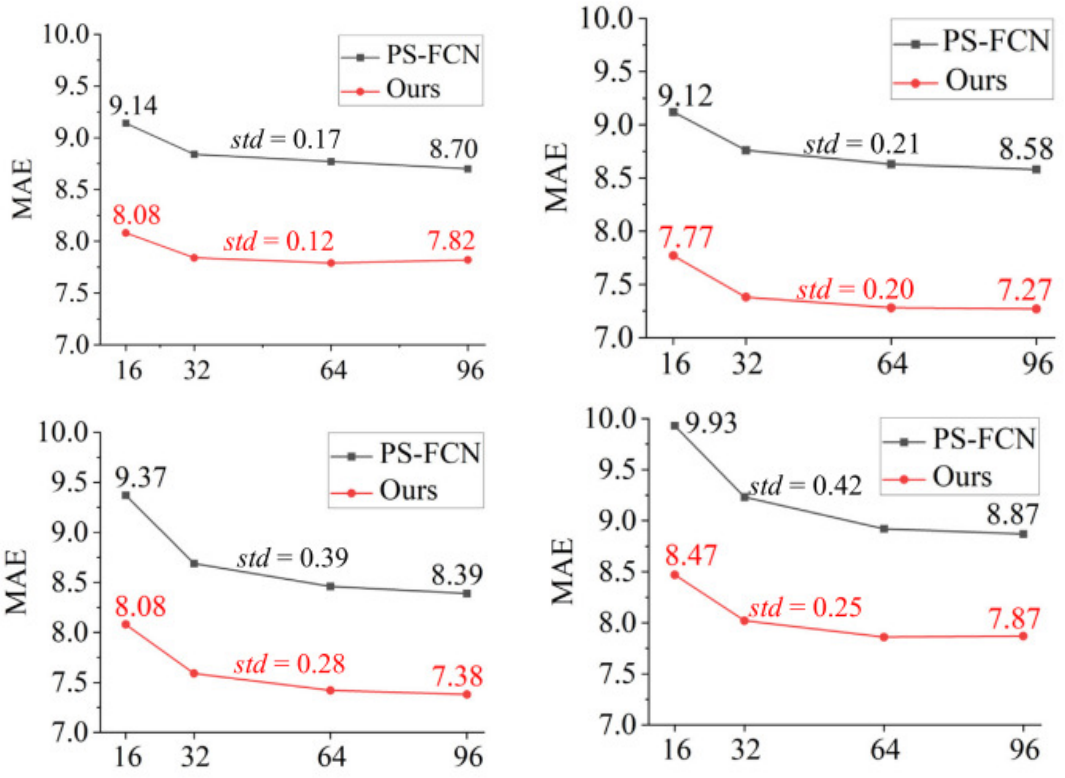


FIGURE 2.11 – Performance de l'architecture DSMF [64] proposée par Liu *et al.* proposé en 2022.

Pour conclure, l'utilisation du *max-pooling* est une méthode appropriée pour permettre de prendre un nombre variable d'images en entrée. Celle-ci a été exploitée dans plusieurs travaux pour résoudre la stéréophotométrie calibrée ou non calibrée. Cependant, le *max-pooling* ne permet pas de facilement traiter l'information image à image, contrairement aux les cartes d'observations.

2.3.3 Méthodes fondées sur les cartes d'observations

Pour pallier à la problématique du nombre variable d'images en entrée, une seconde approche possible est l'utilisation des cartes d'observations. Cette technique a été proposée pour la première fois en stéréophotométrie en 2018 dans CNN-PS [45] par Ikehata. La méthode CNN-PS repose sur l'exploitation de cartes d'observations. Il s'agit d'une méthode pixel à pixel, i.e. chaque pixel de l'image est traité indépendamment. D'un point de vue technique, une carte d'observation est une projection de chaque pixel sur un hémisphère.

Dans CNN-PS, Ikehata utilise une projection des vecteurs des directions lumineuses (l_x, l_y, l_z) dans un système de coordonnées cartésiennes (x,y) de taille fixe.

Remarquons qu'une carte d'observation est créée pour chaque pixel de l'image, d'où le traitement pixel à pixel. Les cartes d'observations sont, pour la plupart, des prises de vues parcimonieuses. En effet, avec une carte d'observations de taille 32×32 pixels et 100 directions lumineuses différentes (i.e. 100 images en entrées), il n'y a que 10% des valeurs de la carte d'observations qui sont différentes de 0, comme cela est illustré sur la figure 2.12.

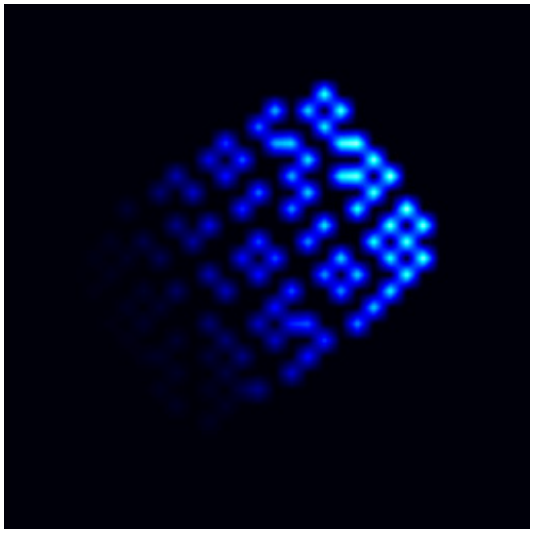


FIGURE 2.12 – Exemple d'une carte d'observation parcimonieuse (Ikehata, CNNPS [45] en 2018).

Une difficulté relevée par l'auteur est la gestion des matériaux ayant de fortes tâches spéculaires ou encore des effets d'anisotropie. Pour mieux gérer ce type de matériaux, l'auteur a intégré une méthode de rotation autour de l'axe de vision de la caméra. Chaque direction lumineuse est ainsi tournée d'un angle θ autour de cette axe. Pour chaque rotation, des nouvelles cartes d'observations sont générées et le réseau de neurones traite chaque rotation indépendamment. Au final, la fusion des prédictions pour chaque rotation se fait en moyennant les prédictions, remises dans le même plan en faisant la rotation inverse sur les prédictions.

L'intérêt principal des cartes d'observations est que, quel que soit le nombre d'images/directions lumineuses disponibles, la taille en entrée du réseau de neurones est toujours identique. Le réseau de neurones n'a donc pas besoin de passer par un système de fusion des caractéristiques ou des données, comme un *max-pooling* par exemple. En revanche, l'inconvénient principal réside dans l'approche pixel à pixel qui, comme pour le perceptron multi-couches, ne permet pas de générer les effets des autres parois de l'objet.

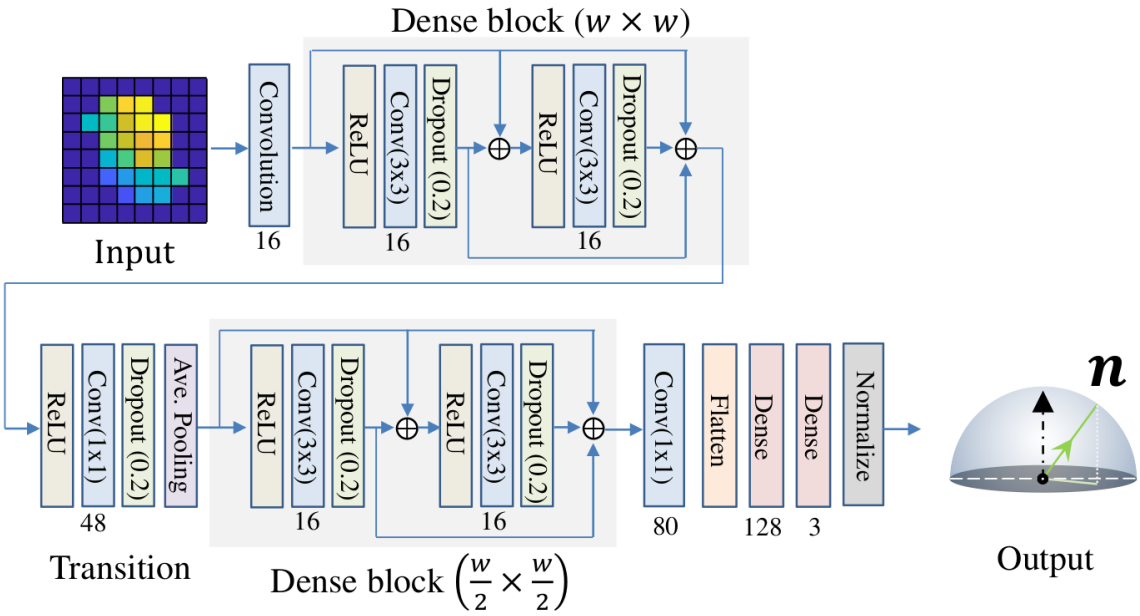


FIGURE 2.13 – Architecture CNN-PS [45] proposée par Ikehata en 2018.

L'architecture de CNN-PS est illustrée sur la figure 2.13. Elle est constituée d'un extracteur de caractéristiques puis d'un module de régression qui sort un vecteur de 3 composantes correspondant à la normale. Un schéma récapitulatif de la méthode est présenté sur la figure 2.14.

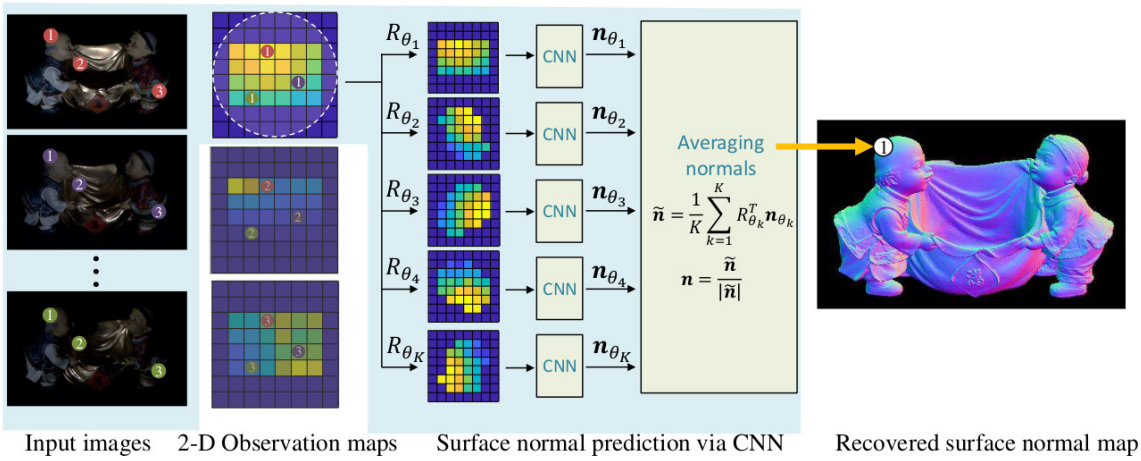


FIGURE 2.14 – Méthode CNN-PS [45] proposée par Ikehata en 2018.

Afin de résoudre le problème des ombres portées pour les cartes d'observations, deux autres méthodes ont été développées, LMPS [59] et PX-Net [65].

La méthode LMPS, proposée par Li *et al.* en 2019, introduit une couche d'occultation qui simule des ombres portées lors de l'entraînement. En effet, les auteurs ont remarqué que, dans une carte d'observations, les ombres portées créent souvent des frontières sous forme de ligne. D'un côté de cette frontière les valeurs sont toutes à zéro, de l'autre les valeurs sont non-nulles, comme cela est illustré sur la figure 2.15. Leur simulateur sépare ainsi en deux la carte d'observations et met à zéro la plus petite zone dans le but de simuler une ombre portée.

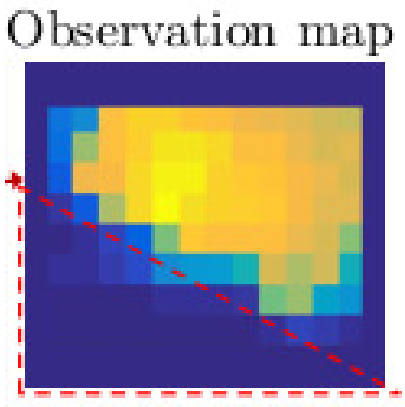


FIGURE 2.15 – Carte d'observations d'une couche d'occultation de la méthode LMPS [59] proposée par Li *et al.* en 2019.

Les auteurs se sont également intéressés à la sélection des images avec les directions lumineuses les plus pertinentes. Pour ce faire, un module de sélection des directions lumineuses est également proposé afin de ne prendre en entrée du réseau que les directions lumineuses les plus pertinentes. Ce module de sélection consiste en une table de connexions de la même taille que les cartes d'observations. Dans cette table de connexions tous les éléments sont supérieurs ou égaux à zéros. Durant l'inférence, les k plus grandes valeurs de cette table de connexions, correspondant au k directions lumineuses les plus pertinentes, sont conservées, les autres sont rejetées.

Un schéma récapitulatif de la méthode LMPS est illustré sur la figure 2.16. Bien que LMPS utilise les cartes d'observations proposées initialement par la méthode CNN-PS [45], la projection utilise des coordonnées polaires et non cartésiennes contrairement à la méthode CNN-PS. Ce type de coordonnées permet d'obtenir de meilleurs performances.

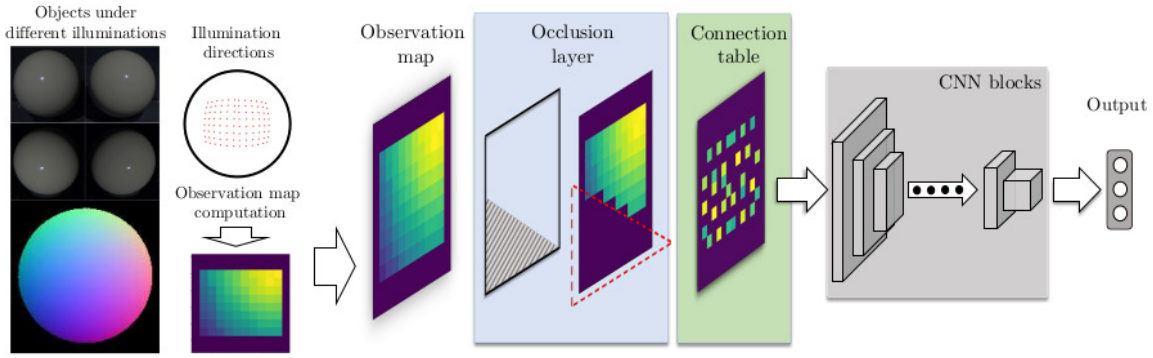


FIGURE 2.16 – Schéma de la couche d'occultation de la méthode LMPS [59] proposée par Li *et al.* en 2019.

PX-Net [65] est une méthode proposée par Logothetis *et al.* en 2021. Elle reprend l'idée des cartes d'observations introduite dans CNN-PS [45]. Cependant, au lieu de générer une base de données d'entraînement synthétique avec des rendus d'objets, les auteurs génèrent une base de données pixel à pixel et ainsi il n'y a qu'une seule carte d'observation à calculer. Cela évite d'avoir d'importants coûts de calcul pour la génération des rendus (i.e. la génération des images).

Cette méthode de génération pixel à pixel prend en entrée 9 paramètres qui représentent les caractéristiques du matériau voulu et les paramètres relatifs à la position dans l'objet du pixel, la normale en ce pixel, la direction lumineuse, etc.

Pour gérer le cas des ombre portées, les auteurs simulent un "mur" autour du pixel, comme illustré sur la figure 2.17. Cette simulation se fait en tirant de façon aléatoire une hauteur de mur en 20 points sur le cercle autour du pixel et ils mettent à zéro 25% des valeurs. Ainsi, toutes les directions lumineuses ne peuvent atteindre le pixel en question.

En plus des ombres portées, une autre problématique posée par une approche pixel à pixel comme PX-Net est la non prise en compte de phénomènes spatiaux comme les inter-réflexions (réflexions sur une autre paroi de l'objet). PX-Net tente d'intégrer cet aspect en générant des inter-réflexion durant l'entraînement. Pour cela, 5 points différents de l'objet de celui à estimer sont tirés aléatoirement. Ces 5 points et uniquement ces 5 points sont considérés comme influant via la réflexion sur le point/pixel à traiter. La réflexion d'un des 5 points (ici noté R) sur le pixel en question (ici noté 0) est donnée par :

$$r_r\left(L,\{L_R\}\right) = \sum_{L_R, L_R \neq L} B\left(N_R, L, L_R, ...\right) B\left(N, L_R, V_0 ...\right), \tag{2.3}$$

où $B\left(N, L_j, V_0, \rho, M\right)$ est la réflectance au point 0 avec une direction lumineuse L_j, N est la normale au point 0, ρ l'albédo et M les paramètres du matériau.

PX-Net, en utilisant toutes les techniques citées ci-dessus lors de la génération de la base d'entraînement, améliore grandement les résultats obtenus avec la méthode CNN-PS (PX-Net ayant conservé la même architecture que la méthode d'origine CNN-PS, illustrée sur la figure 2.13).

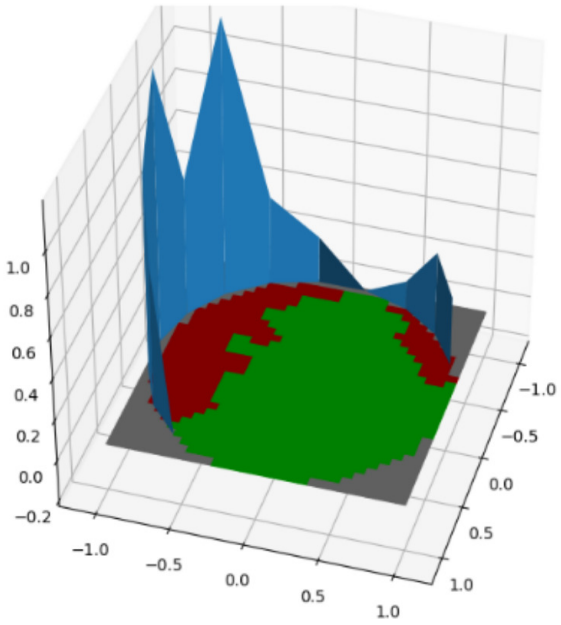


FIGURE 2.17 – Simulation d'un "mur" autour du pixel, comme cela est proposé dans le méthode PX-Net [65] par Logothetis *et al.* en 2021.

Une dernière méthode utilisant les cartes d'observations existe, il s'agit de la méthode SPLINE-Net [108] proposée par Zheng *et al.* en 2019. Celle-ci cherche à résoudre le problème de la parcimonie des cartes d'observations que les auteurs de CNN-PS avaient tenté de résoudre via une interpolation des valeurs non-nulles. L'approche proposée utilise un premier réseau de neurones générateur de type encodeur-décodeur.

Une fois la carte d'observations dense générée, celle-ci est ensuite injectée dans le second réseau qui est en charge de prédire la normale. Au niveau architectural, ce réseau est le même que celui de CNN-PS [45] qui est lui même une variation du réseau DenseNet [40]. L'architecture proposée est affichée sur la figure 2.18.

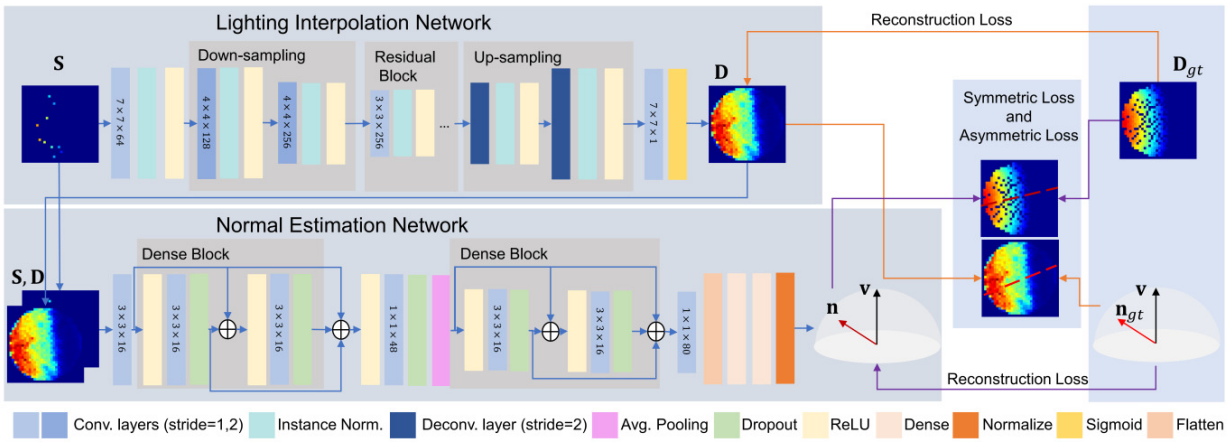


FIGURE 2.18 – Architecture de la méthode SPLINE-Net proposée par Zheng *et al.* [108] en 2019.

Il est important de noter que ce réseau génère un vecteur et non une carte de normales étant donné que, comme toutes les méthodes fondées sur les cartes d'observations, il s'agit d'une méthode pixel à pixel.

Un exemple de cartes d'observations générées à l'aide de cette méthode est affiché sur la figure 2.19. Comme on peut le voir, la carte générée par la méthode SPLINE-Net est beaucoup plus dense que celle générée par CNN-PS.

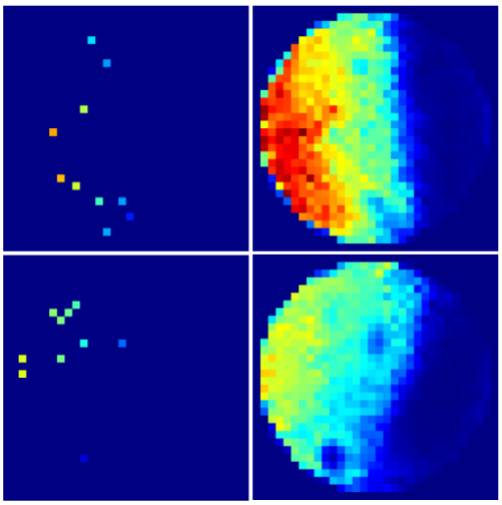


FIGURE 2.19 – Exemple de cartes d'observations générées par la méthode SPLINE-Net proposée par Zheng *et al.* [108] en 2019. À gauche, on peut voir des cartes d'observations avant l'utilisation du réseau générateur et à droite ces mêmes cartes après son utilisation.

La méthode Continuous Material Reflectance Map for Deep Photometric Stereo [74] proposée par Prouteau et al. en 2023 repose sur une technique de représentation différente des cartes d'observation. Les auteurs introduisent une version continue des cartes de réflectance, basée sur l'équation de rendu. Pour chaque pixel de l'image d'origine, une carte de réflectance est générée, suivant une approche pixel-par-pixel. Ces cartes de réflectance sont ensuite converties en images, qui servent d'entrée à un réseau de neurones convolutionnel chargé de prédire la normale associée au pixel concerné (voir figure 2.20).

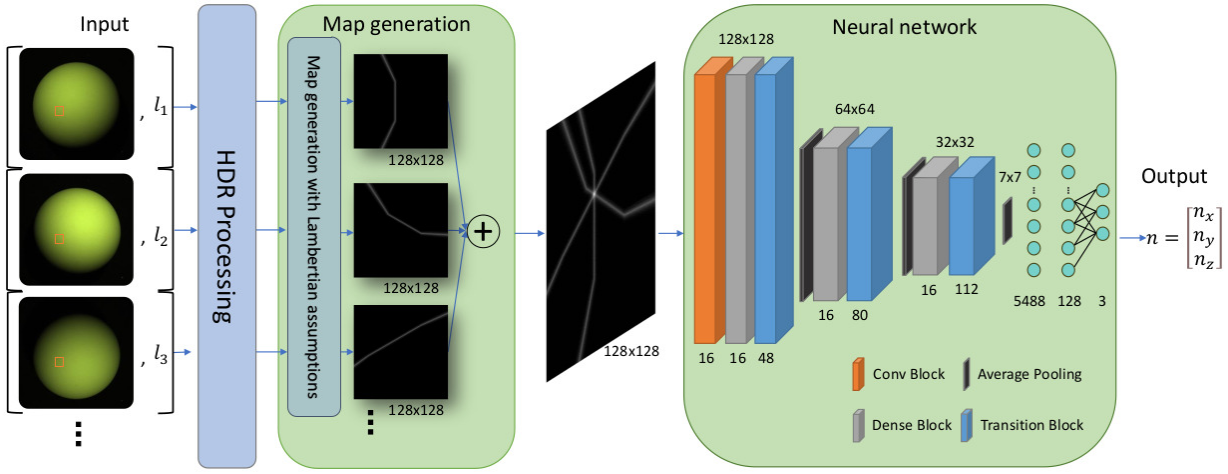


FIGURE 2.20 – Architecture de la méthode Continuous Material Reflectance Map for Deep Photometric Stereo proposée par Prouteau *et al.* [74] en 2023.

Pour résumer, il existe deux grands principes pour traiter un nombre variables d'images, le *max-pooling* et les cartes d'observations. Le *pooling* permet de traiter l'information spatiale via les convolutions et ainsi est plus robuste aux problématiques d'inter-réflexions et d'ombres portées. Cependant, une méthode convolutive combinée à un *pooling* ne permet que difficilement de traiter

l'information image à image lors de l'extraction des caractéristiques. À l'inverse, les cartes d'observations ont l'avantage de toujours avoir la même taille d'entrée quel que soit le nombre d'images et la taille des images et de "voir" l'information image à image. Par contre, la sensibilité aux problèmes spatiaux augmente car il s'agit d'une méthode pixel à pixel. De plus, les cartes d'observations sont une technique répondant uniquement au problème de la stéréophotométrie calibrée.

2.3.4 Méthodes mixtes fondées sur l'extraction spatiale et pixel à pixel

Des méthodes ont été introduites pour essayer de répondre cette problématique d'extraction des caractéristiques soit pixel à pixel ou spatiales. Nous pouvons notamment mentionner les méthodes LSPC [37], GPS-Net [104] ou encore PS-Transformer [41]. Ces trois méthodes se placent dans le contexte de la stéréophotométrie calibrée.

La méthode LSPC [37], proposée par Honzátko *et al.* en 2021, est illustrée sur la figure 2.21. Cette méthode reprend l'idée des cartes d'observations qui permettent d'extraire l'information pixel à pixel. Cependant, l'architecture est composée de convolutions 4D qui permettent d'extraire l'information spatiale. Ainsi, elle prend en entrée un tenseur I, en 4D, de taille $H \times W \times o \times o$ où o est la taille des cartes d'observations, H la hauteur des images et W la largeur. De cette manière, toute l'information est accessible directement par le réseau de neurones et celui-ci peut reconstruire la carte des normales sans avoir une approche pixel à pixel.

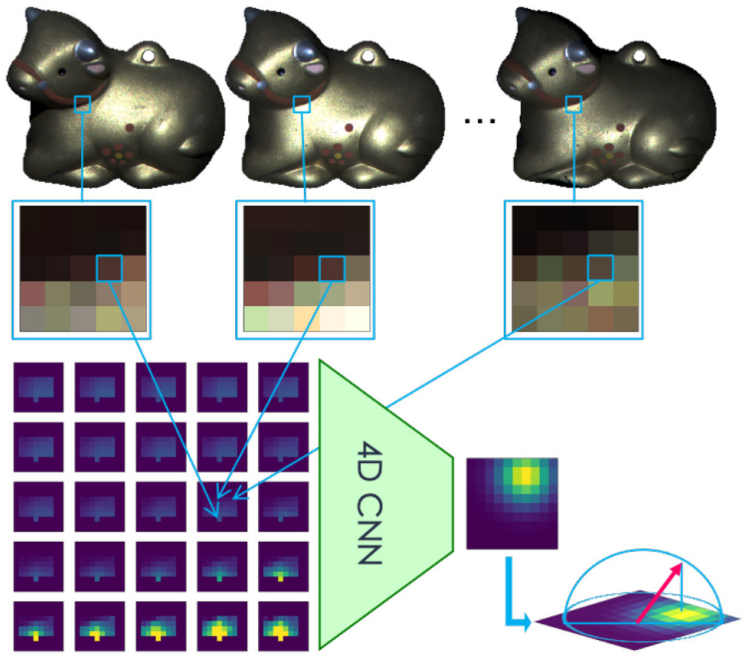


FIGURE 2.21 – Architecture LSPC proposée par Honzátko *et al.* [37] en 2021.

GPS-Net [104] est une méthode proposée par Yao *et al.* en 2020. Elle se distingue de l'état-de-l'art par l'utilisation des graphes pour répondre au problème de stéréophotométrie calibrée. Le réseau de neurones est décomposé en deux parties, comme illustré en figure 2.22.

La première partie traite l'information inter-images en mettant sous forme de graphe chaque pixel des différentes images. Les nœuds de ces graphes n'ont en réalité qu'une seule arête, excepté le nœud central. En effet, chaque nœud représente un pixel éclairé avec une direction lumineuse différente et il est relié uniquement au nœud central qui ne contient comme information que l'indice du pixel. Le nombre de nœuds est ainsi égal au nombre de directions lumineuses et peut varier en fonction du nombre de directions lumineuses disponibles. Les auteurs ont implémenté un module/filtre (SGC filter) qui permet de réaliser une convolution avec les graphes. Cette étape

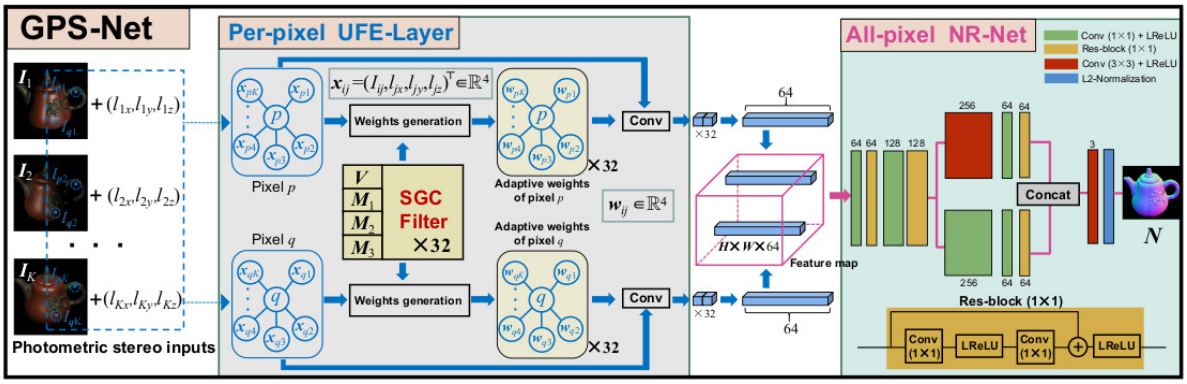


FIGURE 2.22 – Architecture GPS-Net [104] proposée par Yao *et al.* en 2020.

inter-images génère une carte de caractéristiques de taille $H \times W \times 64$ où H et W sont la hauteur et la largeur de l'image. Ainsi durant l'étape d'extraction des caractéristiques inter-images, un réseau de neurones convolutifs est utilisé pour prédire la carte des normales.

Enfin, PS-Transformer [41] proposé par Ikehata en 2021, fut la première méthode à utiliser les *Transformers* pour la stéréophotométrie. L'architecture du réseau de neurones est décomposée en deux branches, une pour l'information spatiale, l'autre pour l'information pixel à pixel. L'architecture proposée par Ikehata est affichée sur la figure 2.23.

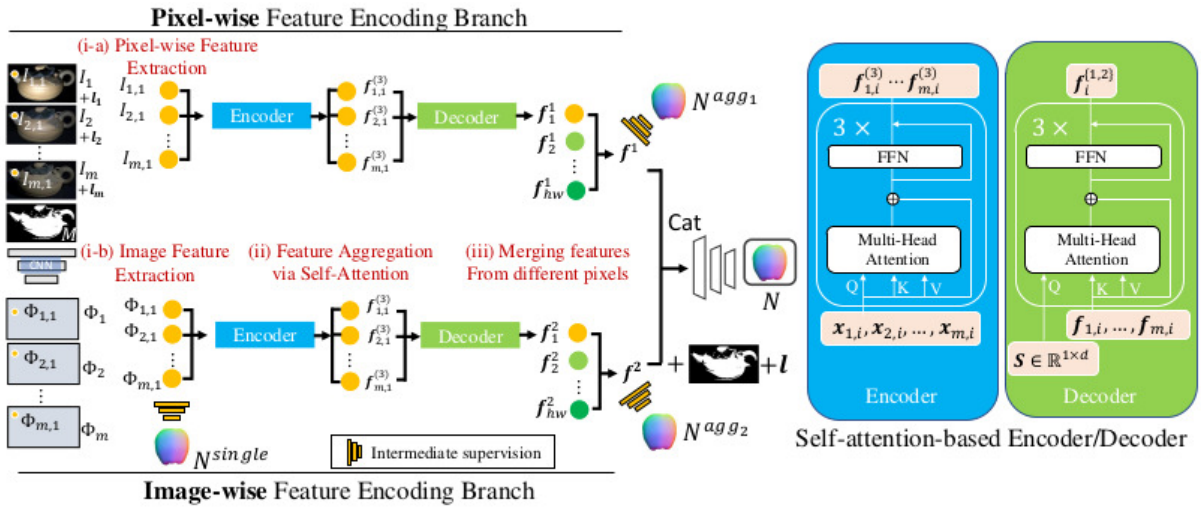


FIGURE 2.23 – Architecture PS-Transformer proposée par Ikehata [41] en 2021.

La branche pixel à pixel est constituée d'une succession de modules de type *multi-head self attention*. La branche spatiale, quant à elle, est d'abord constituée de blocs de convolutions puis de *multi-head self attention*. Pour les deux branches, la dernière étape consiste à effectuer l'agrégation des caractéristiques extraites pour obtenir des cartes de caractéristiques globales à l'ensemble des images. Généralement, cette étape de fusion se fait via des couches de *pooling*. PS-Transformer utilise aussi un module de *pooling*, le module *PMA* (i.e. *Pooling by Multi-head Attention*). Les caractéristiques finales extraites sont ensuite concaténées et des modules de convolutions prédisent la carte des normales.

À noter que cette architecture a été créée pour un faible nombre d'images en entrée. L'une des raisons évoquées réside dans la problématique de l'importante place mémoire requise pour les

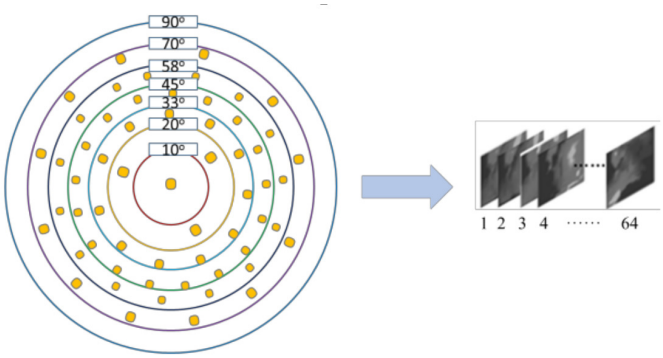


FIGURE 2.24 – Echantillonage polaire des directions lumineuses.

Transformers. L'utilisation des *Transformers* et des deux branches permet d'obtenir des performances intéressantes malgré cet inconvénient.

D'autres types d'architectures ont également été proposées, notamment les architectures multi-échelles qui présentent l'avantage d'être plus robuste au bruit comme expliqué dans [79] et d'être moins sensibles aux ombres portées ainsi qu'aux reflets spéculaires [63].

2.3.5 Méthodes multi-échelles

Trois grandes approches multi-échelles ont été développées pour la stéréophotométrie. La première consiste à agréger les caractéristiques en sortie de différentes couches d'un réseau de neurones. C'est le choix qui a été fait par Yu *et al.* [105]. Ceux-ci ont créé deux variantes de leur méthode, une pour le contexte calibré, l'autre pour le contexte non calibré.

La première se base sur l'architecture proposée dans PS-FCN [18], en ajoutant un module de fusion des caractéristiques. La seconde réutilise le même réseau mais ajoute le modèle d'estimation de la direction/intensité lumineuse introduit dans SDPS-Net [17]. Généralement, les méthodes de fusion multi-échelles prennent les caractéristiques extraites à différentes résolutions et sont ensuite remises à la même résolution à l'aide de méthodes de sous-échantillonnage ou de sur-échantillonnage. Or, dans cette méthode, pour "économiser" de la puissance de calcul, les auteurs prennent les caractéristiques extraites à la même résolution (résolution originale divisée par 2) mais pas au même niveau du réseau de neurones. La concaténation des caractéristiques ne requiert ainsi aucun traitement. Un *max-pooling* est ensuite appliqué sur cette fusion pixel à pixel pour obtenir une carte de caractéristique globale.

Une autre possibilité d'architecture multi-échelles est la prédiction d'une carte de normales à différentes échelles/résolutions, comme proposé par Ren *et al.* [79] en 2020, pour la stéréophotométrie calibrée. Leur méthode multi-échelles génère une carte de normales à trois échelles différentes : résolution originale, à la moitié de la résolution originale et au quart de la résolution originale. Les auteurs avancent comme argument que prédire la carte de normales à différentes résolutions permet d'améliorer la convergence et la stabilité lors de l'entraînement.

Une nouvelle méthode de discrétisation de l'hémisphère est également utilisée : l'espace cartésien communément utilisé est remplacé par l'espace polaire comme illustré sur la figure 2.24.

Remarquons également qu'ils ordonnent les images en utilisant cette division de l'espace. Les images les plus au centre sont mises en premières, celles les plus éloignées sont mises en dernières. En effet, leur méthode a besoin d'avoir des images triées car toutes les images en entrées sont concaténées en une seule et unique "image" de $3n$ canaux, où n est le nombre d'images. Le fait

d'ordonner les images et de les concaténer permet une meilleure généralisation du modèle.

D'un point de vue architectural, le réseau de neurones reprend le principe d'une architecture de type U-Net [82] avec des *skip connection*, illustré sur la figure 2.25.

Cette méthode a cependant un inconvénient ; les images étant concaténées en entrée, il est impossible d'avoir un nombre variable d'images. Par exemple, si le réseau a été entraîné avec 64 images, l'inférence devra toujours se faire avec 64 images.

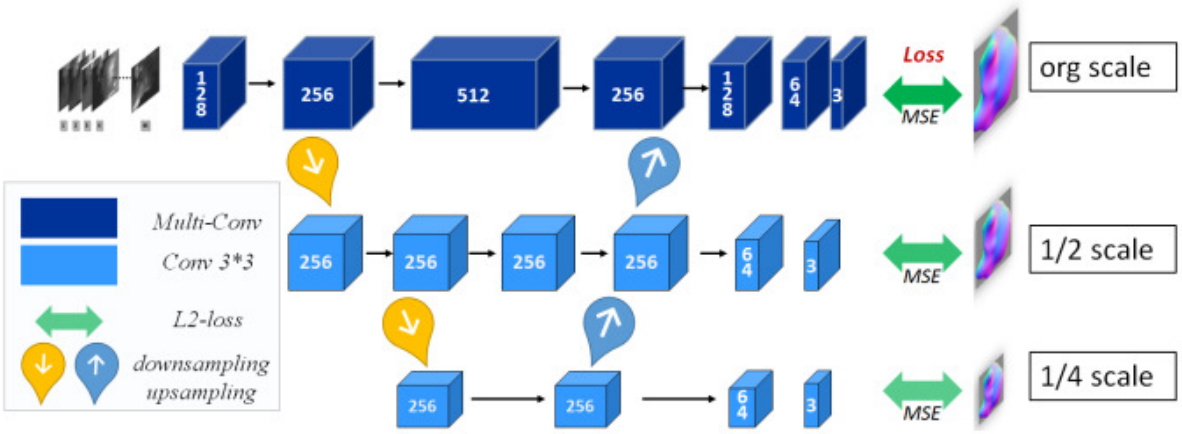


FIGURE 2.25 – Architecture multi-échelles avec prédiction à plusieurs résolutions proposée par Ren *et al.* [79] en 2020.

Enfin, une dernière méthode multi-échelles, développée dans [63] par Lichy *et al.* en 2021, consiste à créer une architecture récurrente qui augmente progressivement la résolution de la carte de normales générée. Une telle approche permet de pouvoir traiter n'importe quelle taille d'image. L'architecture neuronale employée est identique à chaque échelle/résolution de traitement. Le principe est que chaque étage prédit une carte de normale qui est ensuite utilisée à l'étage suivante et ainsi de suite jusqu'à atteindre la résolution souhaitée. Cette méthode est schématisée sur la figure 2.26.

Cependant, cette méthode a deux contraintes majeures. La première est que le nombre d'images en entrée doit être fixe. De plus, la source lumineuse doit être toujours positionnée au même endroit : droite, mi-droite, centre, mi-gauche et gauche de l'appareil photo. Ces aspects sont très contraignants pour l'utilisateur qui ne peut pas mettre les lampes où il veut et ne peut pas prendre autant d'images qu'il le souhaite.

Bien que l'ensemble des méthodes présentées précédemment, qu'elles soient fondées sur le *pooling*, sur les cartes d'observations ou sur l'aspect basées multi-échelles, possèdent des avantages d'un point de vue architectural, elles restent limitées dans un cadre applicatif. En effet, les méthodes calibrées doivent avoir en entrée les directions lumineuses ainsi que leur intensité, ce qui n'est pas forcément simple à acquérir ou estimer. La stéréophotométrie non calibrée tend à résoudre cette problématique en laissant l'estimation des directions lumineuses aux algorithmes. Néanmoins, dans les deux cas, les prises de vues doivent encore être réalisées dans une pièce obscure sans lumière réfléchissante. Cela constitue une contrainte importante pour le déploiement de la stéréophotométrie en dehors d'un laboratoire ou d'une entreprise.

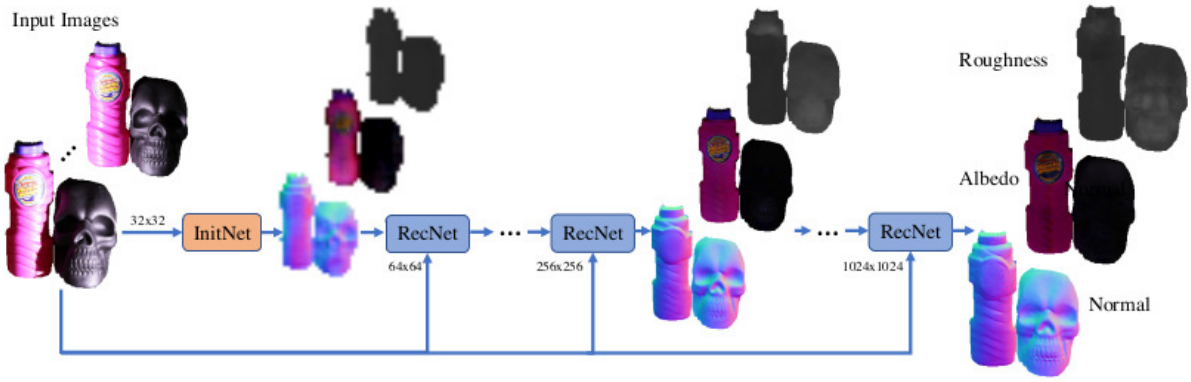


FIGURE 2.26 – Architecture récurrente multi-échelles proposée par Lichy *et al.* [63] en 2021.

Cependant, la stéréophotométrie sans aucune contrainte sur l'environnement ambiant de l'objet, ni sur la source éclairante reste possible via la stéréophotométrie dite *universelle*.

2.3.6 Méthodes universelles

La première méthode universelle avec de l'apprentissage profond a été développée par Ikehata en 2022 et se nomme UniPS [42]. UniPS est une méthode complètement fondée sur les données, c'est-à-dire qu'il n'est pas nécessaire d'apprendre un quelconque paramètre sur la lumière. En effet, le modèle gère de lui-même l'extraction de toutes les informations nécessaires à la prédiction des cartes de normales. Ainsi, cette approche repose totalement sur les données afin de se détacher des contraintes de directions lumineuses ou d'environnement présentes en stéréophotométrie calibrée ou non calibrée. Pour répondre à ce nouveau problème, l'auteur propose une architecture de type encodeur-décodeur. L'encodeur extrait les contextes d'éclairage globaux à partir des images, qui sont une représentation d'éclairage générique qui correspond aux paramètres d'éclairage physiques (par exemple la direction de la lumière). Le décodeur prend toutes les valeurs brutes de l'image et le contexte d'éclairage global, extrait par l'encodeur, interpolé à chaque pixel et prédit ensuite les normales à la surface. La méthode proposée par Ikehata est illustrée sur la figure 2.27.

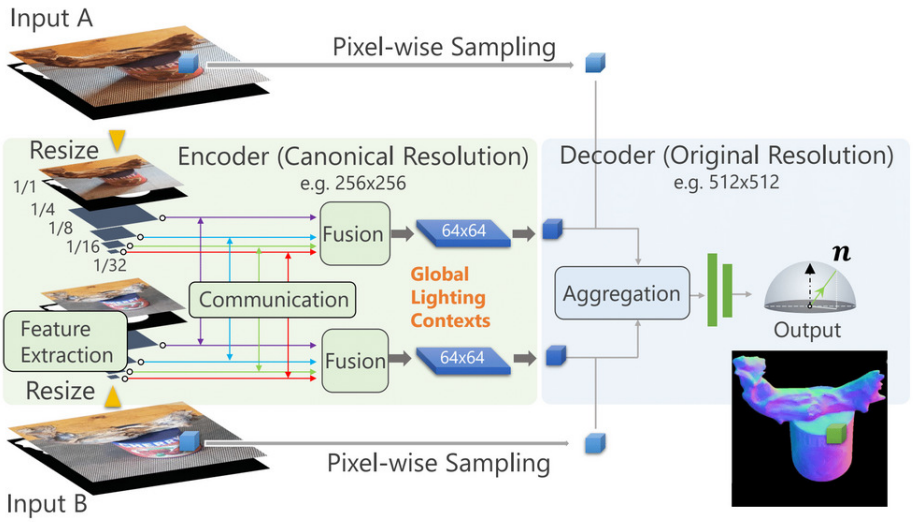


FIGURE 2.27 – Architecture UniPS proposée par Ikehata [42] en 2022.

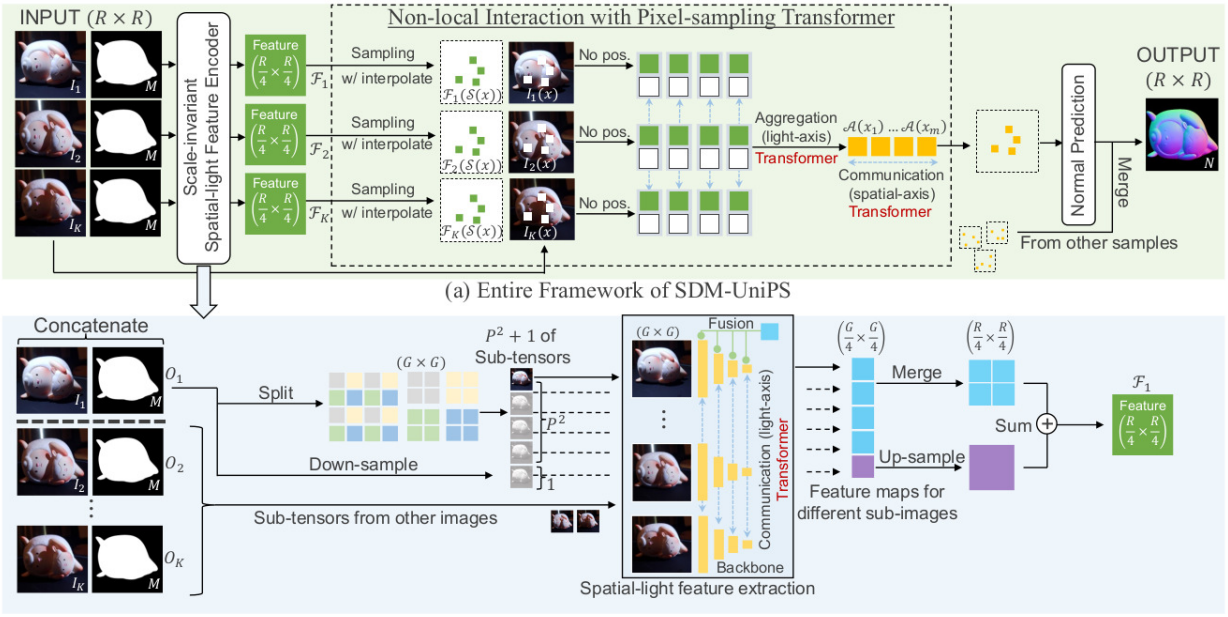


FIGURE 2.28 – Architecture SDM-UniPS [43] proposée par Ikehata en 2023.

Cette architecture présente 2 points clés :

— Le réseau prend plusieurs images en entrée donc l'ensemble des caractéristiques sont intégrées dans l'espace latent en tenant compte de leur interaction. Donc la mise en relation des caractéristiques se fait dans l'encodeur et l'agrégation dans le décodeur.

— Contrairement aux architectures classiques où les caractéristiques sont directement transmises au décodeur, différentes résolutions de travail pour l'encodeur et le décodeur sont considérées dans cette approche. La résolution de travail du décodeur est la même que la résolution des images d'origines, mais l'encodeur prend en entrée des images qui ont été redimensionnées à une résolution canonique prédéfinie, qui est fondamentalement plus petite que la résolution d'origine, sa sortie est transmise au décodeur après conversion inverse pour le travail du décodeur. Cela permet d'assurer l'évolutivité à la taille de l'image car les besoins en mémoire de l'encodeur dépendent uniquement de la résolution canonique, et non de celle d'origine, tandis que le décodeur traite chaque pixel un par un. Cela permet de garder le champ récepteur de l'encodeur invariant par rapport à la taille de l'image d'entrée. Sans cela, le champ de réception des réseaux risque de ne pas couvrir l'intégralité de l'objet dans les images de test à très haute résolution.

Ikehata a développé une seconde méthode universelle en 2023 appelée SDM-UniPS [43], illustrée sur la figure 2.28, dans le but d'améliorer sa première méthode en permettant notamment de traiter des images de très hautes résolutions. De même que dans la première méthode, l'auteur a utilisé une méthode purement fondée sur les données. L'extraction des différentes caractéristiques, de la lumière ambiante (e.g. direction lumineuse, intensité, etc.) ou paramètres nécessaires à la stéréophotométrie dans un tel environnement est entièrement confiée à un extracteur de caractéristiques. Dans ce travail, l'extracteur de caractéristiques est fondé sur un squelette composé de couches ConvNeXt [110] et de couches *Transformers*.

Comme évoqué précédemment, la principale caractéristique de cette contribution est sa capacité à gérer des images de tailles diverses et variées. Pour ce faire, un module appelé *Scale-invariant Spatial-light* a été proposé. Celui-ci prend des images et les divises en P^2 sous-images. Cette division

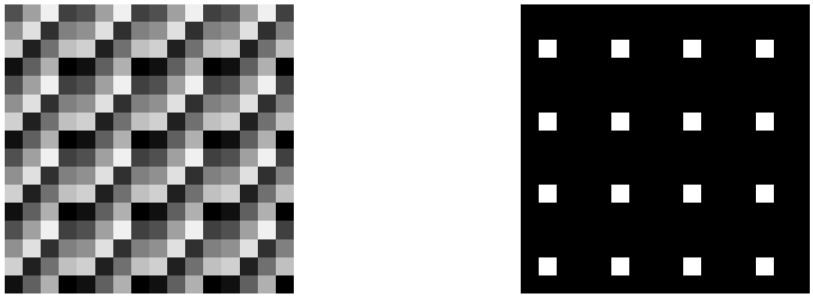


FIGURE 2.29 – Exemples de décomposition d'une image proposée par Ikehata [43] pour générer P^2 sous-images. À gauche, chaque couleur représente les pixels sélectionnées pour créer une sous-images. À droite, les pixels d'une sous-image sont indiqués en blanc, et les autres pixels de l'image d'origine sont mis à zéro.

s'effectue en prenant un pixel tous les P pixels dans le sens de la hauteur mais aussi dans le sens de la largeur. Cette astuce forme alors P^2 sous-images de taille fixe quelle que soit la résolution initiale de l'image.

Ces sous-images sont alors traitées de façon indépendantes par le même extracteur de caractéristiques. Celui-ci est composé de blocs ConvNeXt [110] pour l'extraction de caractéristiques spatiales et de blocs *Transformers* pour l'extraction des caractéristiques pixel à pixel. Une fois l'extraction des caractéristiques des sous-images faite, celles-ci sont restructurées pour former une seule et unique carte de caractéristiques.

2.3.7 Méthodes de rendu inverse

Seules les méthodes d'apprentissage profond entraînées de manière complètement supervisée ont été abordées jusqu'ici. Cependant, il existe des méthodes de rendu inverse qui n'ont pas besoin d'avoir les cartes de normales lors de leurs entraînements.

Nous pouvons notamment mentionner la méthode *Neural Inverse Rendering for General Reflectance Photometric Stereo* [89] qui est proposée en 2018 par Taniai *et al.*, pour résoudre le problème de stéréophotométrie calibrée. Cette méthode est décomposée en deux sous-réseaux. Le premier prédit la carte des normales. Cette carte des normales est ensuite utilisée pour générer une carte de spécularité via l'équation suivante :

$$S_{ip} = v^T \left[2 \left(l_i^T N_p \right) N_p - l_i \right], \tag{2.4}$$

où v est la direction de visée, N_p la normal prédite et l_i la direction associée à l'image i.

Enfin, cette carte de spécularité est concaténée aux images RGB et cette concaténation est utilisée en entrée d'un second réseau de neurones qui génère une carte de réflectance.

Finalement, le moteur de rendu génère les images RGB à l'aide des sorties de deux réseaux. Ces images générées sont ensuite utilisées dans la fonction de perte de reconstruction, en comparant les images générées aux images RGB d'origine, pour permettre l'optimisation des deux réseaux de neurones. L'architecture globale de cette approche est illustrée sur la figure 2.30.

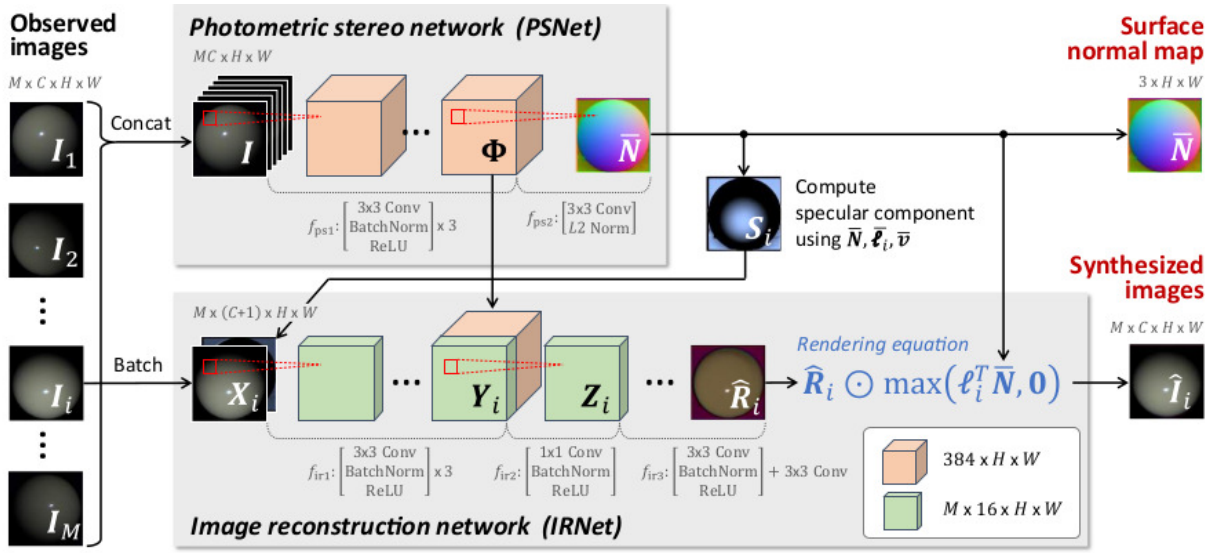


FIGURE 2.30 – Architecture *Neural Inverse Rendering* [89] proposée par Taniai *et al.* en 2018.

Ce type d'approche a été modifié en 2021 dans [50] par Kaya *et al.* pour répondre au problème de stéréophotométrie non calibrée. Le principe adopté est le plus communément utilisé dans la littérature, i.e. un premier réseau prédit les directions/intensités lumineuses puis celles-ci servent par la suite à traiter la problématique comme si le contexte était calibré.

Au niveau architectural, le premier réseau de cette méthode est le réseau classique introduit dans SDPS-Net [17], les deux autres sont des réseaux convolutifs, tout comme dans la version pour le contexte calibré présenté précédemment sur la figure 2.30. Cependant, il est important de remarquer que les images en entrée des deux derniers réseaux sont concaténées, il n'est donc pas possible d'avoir un nombre variable d'images en entrée, ce qui limite son applicabilité en pratique.

Pour résumer, les méthodes de rendu inverse présentent l'avantage de ne pas avoir besoin d'un jeu de données avec les cartes de normales (i.e. vérité terrain). En revanche, deux inconvénients sont notables :

— le temps d'inférence excessivement long,
— et le nombre fixe d'images en entrée.

Pour resynthétiser cette présentation de l'état-de-l'art, il existe une grande variété de méthodes pour résoudre le problème de la stéréophotométrie. L'introduction des réseaux de neurones a permis de grandement améliorer les performances. Plusieurs solutions ont été apportées pour traiter un nombre variable d'images en entrée (couche de *pooling* et cartes d'observations).

Cependant, la plupart des méthodes actuelles ne permettent pas de gérer des images de différentes résolutions. Cet aspect est pourtant primordiale dans le cadre de la stéréophotométrie. En effet, l'intérêt principal de la stéréophotométrie est sa capacité à reconstruire chaque petit détail. Il semble ainsi pertinent de traiter les images à leur résolution native pour éviter toute perte d'information. De plus, un manque de solution pour reconstruire des matériaux difficiles (e.g. fortement spéculaire) persiste.

Cette thèse se concentre principalement sur ces deux aspects : gérer les images de résolutions différentes, y compris de très haute résolution, ainsi que traiter tous les types de surfaces. Par conséquent, pour permettre l'entraînement des réseaux de neurones comme ceux présentés ci-dessus, il est nécessaire d'avoir des données d'entraînement. Cette thématique a été abordée et différentes bases de données ont été publiées dans la littérature.

2.4 Bases de données existantes

Plusieurs bases de données ont été proposées dans la littérature pour répondre au problème de la stéréophotométrie. Nous allons distinguer deux catégories : les bases de données avec et sans vérités terrains. Par vérité terrain, on parle ici des normales réelles pour chaque objet de la base de données.

2.4.1 Bases de données avec vérités terrains

Pour le problème de la stéréophotométrie, la création d'une base de données réelles avec vérités terrains est un travail difficile. En effet, il faut acquérir non seulement les images RGB mais également les normales de l'objet. Généralement, la création de ce type de bases se décompose en 3 étapes :

— Acquisition des images RGB : pour un objet, chaque image RGB doit avoir une direction lumineuse différente et la direction précise de la lumière doit être sauvegardée pour un usage dans un cadre de stéréophotométrie calibrée.
— Acquisition des normales : cette étape est généralement effectuée à l'aide d'un scanner 3D ou de modèles numériques 3D.
— Recalage des images et des normales : cette étape primordiale nécessite d'aligner les normales avec les images RGB au pixel près. Cela est difficile car, pour juger la qualité de reconstruction, il faut que l'alignement soit parfait. Bien que des algorithmes de recalage puissent être utilisés, la vérification doit être faite à la main.

La création de bases de données synthétiques est moins contraignante. En effet, l'ensemble des images RGB et des normales générées se fait principalement à l'aide d'un pipeline via un logiciel adapté, tel que *Blender* [20].

À ce jour, nous comptons 8 bases de données pour le problème de la stéréophotométrie. Tout d'abord, les bases de données d'images réelles *DiLiGenT* [87], *DiLiGenT10*^{2} [80], *DiLiGenT-Pi* [95] et *Luces* [72]. Ces bases de données sont des bases de données de tests. Elles permettent de comparer les différentes méthodes de l'état-de-l'art sur un pied d'égalité. On retrouve également les bases de données synthétiques *Blobbly* [18], *Sculpture* [18], *CyclePS* [45] et *CyclePS+* [41], utilisées pour l'entraînement.

DiLiGenT

La base de données *DiLiGenT* [87] a été proposée par Shi *et al.* en 2016. Cette base de données est pionnière dans ce domaine. Il s'agit d'une base de données réelles contenant 10 objets avec 96 images du même point de vue pour chacun des 10 objets. De plus, les conditions lumineuses sont connues. C'est-à-dire que l'on connaît à la fois la direction et l'intensité lumineuse. Les normales servant de vérités terrains ont été acquises à l'aide d'un scanner 3D.

Cette base de données a été la première base de données permettant d'évaluer les performances des méthodes de stéréophotométrie. Bien que composée de forme géométriques relativement peu complexes, des surfaces lisses sans grand détail, comme par exemple le chat illustré sur la figure 2.31(a), elle contient malgré tout des objets réputés difficiles en stéréophotométrie. En effet, ceux-ci présentent deux difficultés courantes en stéréophotométrie : les ombres portées, illustrées en figure 2.31(c), et les formes concaves, illustrées sur la figure 2.31(b). De plus, d'un point de vue des matériaux, cette base contient des objets avec des surfaces diffuses, des surfaces métalliques mais également des surfaces qui changent spatialement de matériau. Cela permet d'obtenir une première estimation pertinente des performances. La diversité des matériaux et des formes géométriques reste cependant beaucoup trop faible pour permettre de tester les performances sur des objets avec une surface plus complexe. Par exemple, les surfaces transparentes, spéculaires, ou encore très réfléchissantes comme pourrait des pièces en aluminium.

(a) (b) (c)

FIGURE 2.31 – Exemples d'images du jeu de données *DiLiGenT* [87] proposé par Shi *et al.* en 2016.

De plus, les images ont une résolution très restreinte, la partie de l'image contenant les objets ayant une résolution comprise entre 141 et 410 pixels au maximum. Aujourd'hui, la quasi totalité des capteurs ont une meilleure résolution que celle fournie. Cette faible résolution constitue un inconvénient pour tester la capacité des méthodes à travailler sur des scènes ou des objets très bien résolus, où beaucoup de détails sont visibles. Or, comme un des intérêts principaux de la stéréophotométrie est sa capacité à détecter et à reconstruire les moindres détails de la surface d'un objet, la capacité des méthodes à gérer des images très haute résolution est donc fondamentale.

Par ailleur, ce jeu de données est orienté pour tester les méthodes de stéréophotométrie prises sans lumière extérieure et sans lumière provenant de rebonds de la source lumineuse sur un mur ou sur un autre objet. Ainsi, les acquisitions ont été réalisées dans une pièce recouverte de noir, excepté les objets, les faisceaux lumineux sont considérés parallèles. La technique employée permettant de simuler des faisceaux lumineux parallèles consiste à éloigner le plus possible les LED de l'objet. En effet, plus les LED sont éloignées et plus la différence angulaire entre les différentes faisceaux lumineux arrivant sur l'objet est faible. À noter également que la distribution des positions des sources lumineuses est toujours la même, grâce à l'utilisation d'un panneau de LED comme source d'éclairage. Cette distribution est illustrée à la figure 3.16. Cette absence de diversité dans les distributions lumineuses est aussi un inconvénient étant donné qu'il est impossible de tester la capacité des méthodes à gérer d'autres types de distributions.

FIGURE 2.32 – Exemples d'images du jeu de données *Luces* [72] proposé par Mecca *et al.* en 2021.

FIGURE 2.33 – Exemple d'images du jeu de données *DiLiGenT10*^{2} [80] proposé par Ren *et al.* en 2022.

Non translucide : ABS | Translucide : Acrylic | Translucide : Acrylic (correction gamma)

FIGURE 2.34 – Exemple de l'impact d'un matériau translucide sur l'apparence de l'objet. Dans cette figure, la position de la source lumineuse est identique pour les trois images. Pourtant, dans le cas de l'acrylique, la lumière est visible du côté opposé de l'objet. La lumière a en réalité traversé l'objet.

DiLiGenT10^2

Pour répondre au manque de diversité des matériaux de la base de données *DiLiGenT* [87], la nouvelle base de données *DiLiGenT10*^{2} [80] a été proposée par Ren *et al.* en 2022. Cette base contient 10 objets, chacun manufacturé dans 10 matériaux différents. La diversité de matériaux dans cette base de données est donc importante. En effet, des matériaux diffus et modérément spéculaires sont présents mais également des matériaux métalliques à réflectance anisotrope ou encore des matériaux translucides. Cette diversité offre la possibilité de tester les performances des méthodes de stéréophotométrie sur des matériaux difficiles. Des exemples d'objets de cette base sont affichés sur la figure 2.33. La vérité terrain est également disponible mais elle a été obtenue en utilisant les modèles numériques 3D ayant servis à l'usinage et non pas un scanner 3D comme dans le jeu de données *DiLiGenT*.

De même que pour *DiLiGenT*, les images ont été acquises dans une salle complètement noire et les faisceaux lumineux sont considérés parallèles. Cependant, contrairement à la base de données *DiLiGenT*, les faisceaux parallèles ne sont pas obtenus grâce à l'éloignement de la source lumineuse mais en mettant une lentille convexe devant celle-ci. Les sources lumineuses quant à elles, sont réparties uniformément sur l'hémisphère comme illustré sur la figure 2.37b.

FIGURE 2.36 – Exemples d'images du jeu de données *DiLiGenT-Pi* [95] proposé par Wang *et al.* en 2023.

L'intérêt principal de cette base de données d'évaluation réside dans le type de matériaux présents. Il s'agit notamment de la seule base de données, toutes bases confondues, qui contient un matériau translucide, l'acrylique. Sur ce type de matériau, les méthodes doivent "comprendre" le trajet lumineux à travers l'objet pour prédire une carte de normales cohérente. En effet, comme nous pouvons le voir sur la figure 2.34, la lumière traverse le petit "dôme", le côté opposé à la source lumineuse étant ainsi illuminé. Par conséquent, pour obtenir une carte de normales cohérente sur un tel objet, les méthodes doivent comprendre que la lumière peut traverser l'objet, i.e. de quel côté celle-ci provient ainsi que son trajet. Sans ces connaissances, les méthodes pourraient supposer que le "dôme" est quasiment plat, ce qui permettrait à la lumière de passer par dessus.

Spéculaire : aluminium | Spéculaire : acier | Transluminescence : Nylon

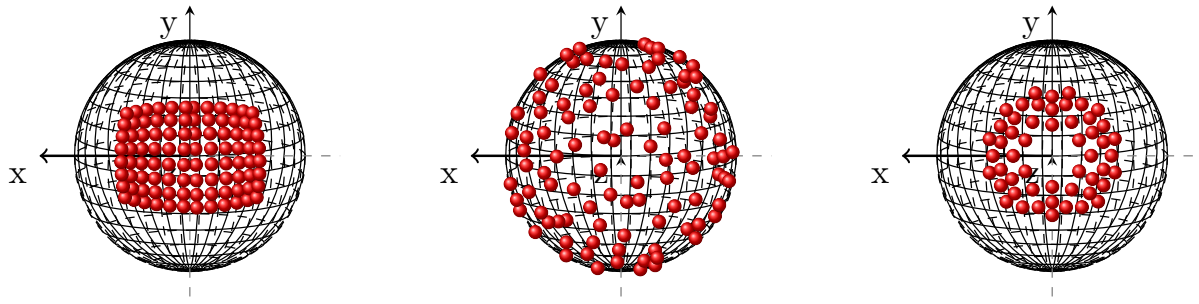
FIGURE 2.35 – Exemple d'objets avec une surface spéculaire et d'un objet avec une transluminescence.

Cette base de données contient d'autres matériaux intéressants comme des matériaux avec une réflectance très spéculaire comme l'aluminium, l'acier et des matériaux avec une transluminescence tel que le nylon, comme illustré sur la figure 2.35.

Cependant, la complexité des objets, d'un point vu géométrique, reste faible. En effet, la quasi totalité des objets sont de formes convexes, comme visible sur la figure 2.33.

DiLiGenT-Pi

Le jeu de données *DiLiGenT-Pi* [95] a été proposé par Wang *et al.* en 2023. Ce jeu de données a été créé pour tester la stéréophotométrie sur des surfaces quasi-planaires avec des surfaces



(a) *DiLiGenT* [87] (b) *DiLiGenT*^{2} [80] (c) *Luces* [72]

FIGURE 2.37 – Distributions des sources lumineuses dans les bases de données réelles *DiLiGenT*, *DiLiGenT10*^{2} et *Luces*. L'axe z correspond à l'axe optique de la caméra en considérant que l'objet se situe aux coordonnées $(0, 0, 0)$.

riches en détails comme des pièces de monnaie, des badges, etc. Des exemples d'objets sont illustrés sur la figure 2.36. Ce type de surface n'étant pas présent dans les bases de données *DiLiGenT* et *DiLiGenT10*^{2}, cette base de données permet ainsi de combler les manques de celles-ci. Quatre groupes de matériaux sont représentés dans cette base de données : surfaces métalliques, spéculaires, rugueuses et translucides. De plus, 30 objets sont présents et chaque objet dispose de 100 images RGB. Ensuite, comme pour le jeu de données *DiLiGenT*, la vérité terrain a été réalisée par un scanner 3D, mais adapté à ce type de surface avec de micro détails (scanner *Bruker Alicona Infinite Focus*).

Le système d'acquisition est équivalent à celui de *DiLiGenT10*^{2}. Les faisceaux parallèles sont créés par une lentille convexe et les distributions lumineuses sont réparties uniformément sur l'hémisphère. Tout comme les autres bases de données *DiLiGenT*, les images ont été acquises dans l'obscurité.

Luces

La base de données *Luces* [72], proposée par Mecca *et al.* en 2021, se distingue des base de données précédentes par les caractéristiques lumineuses d'acquisition. En effet, toutes les bases de données *DiLiGenT* sont des bases de données où la lumière est directionnelle avec des faisceaux lumineux parallèles. En revanche, la base de données *Luces* est une base de données avec des sources lumineuses ponctuelles très proches de l'objet induisant des faisceaux lumineux non parallèles. Le fait d'avoir des faisceaux non parallèles ajoute de la complexité car la direction lumineuse en chaque pixel n'est en réalité pas la même. Une estimation globale de la direction lumineuse n'est donc plus suffisante.

Cette base de données contient 14 objets et 52 images par vue. Les positions des sources lumineuses, leurs intensités ainsi que les paramètres intrinsèques de la caméra sont également connus. Comme pour les autres jeux de données, les cartes de normales servant de vérité terrain sont fournies. Celles-ci ont été acquises pas un scanner 3D.

Cette base de données contient quelques matériaux spéculaires mais la particularité et l'intérêt de cette base réside dans les faisceaux non parallèles et non dans le type de matériaux ou le type de géométrie présents. Cependant, les géométries présentes vont d'objets avec peu de détails et facile comme une balle à des objets beaucoup plus détaillés comme la maison ou la statue de la reine Elizabeth, présentés sur la figure 2.32.

Pour résumer, les différentes orientations des sources lumineuses sur les bases de données présentées précédemment sont affichées en figure 3.16.

FIGURE 2.38 – Exemple d'images du jeu de données *Skoltech3D* [93] introduit par Voynov *et al.* en 2023.

2.4.2 Bases de données sans vérité terrain

Quatre bases de données d'images réelles sans vérité terrain peuvent être utilisées pour tester les performances des méthodes de stéréophotométrie : *Skoltech3D* [93], *Shape and Material* [63], Uni-PS [42] et SDM-UniPS [43].

Skoltech3D [93] introduit par Voynov *et al.* en 2023 est un jeu de données pour la reconstruction 3D multi-vues. Il est constitué de 107 objets photographiés sous 100 angles de vue différents avec 14 conditions d'éclairage différentes pour chaque angle. Les objets photographiés sont divers et variés, certains avec beaucoup de détails comme le château, illustré sur la figure 2.38. De plus, les images sont biens résolues.
Cette base de données est ainsi un atout pour tester les méthodes de stéréophotométrie sur des objets avec des géométries très complexes. Un autre intérêt de cette base est que 7 types de capteurs ont été utilisés pour chaque angle de vue comme par exemple un Huawei Mate 30 Pro (smartphone), la caméra RGB de la Kinect v2 de Microsoft, une camera industrielle (DFK 33UX250), etc. Il est donc possible de tester l'impact du capteur photo avec des résultats qualitatifs. Des exemples d'images de cette base sont affichés sur la figure 2.38.

La base *Shape and Material* [63] créée par Lichy *et al.* en 2021, a été acquise à l'aide d'un appareil photo CANON 2000D DSLR. Elle est composée de 111 objets avec 6 sources lumineuses par vue (1 seule vue par objet). Ces 6 sources lumineuses sont toujours approximativement positionnées au même endroit, c'est-à-dire à 90° et 45° à gauche et à droite de l'appareil photo, à côté de la caméra et au dessus de l'objet avec une marge d'erreur sur l'angle d'environ 15°. Les objets de cette base sont principalement des objets du quotidien ou de décoration tels que des tasses, des peluches ou encore un casque de vélo (figure 2.39). Bien que les objets ne soient pas très complexes, les images ont été acquises dans une pièce avec une lumière ambiante naturelle. Cet aspect, lors de la prise de photo, permet de tester la capacité des méthodes à généraliser sur des photos prises par une personne "lambda", sans processus d'acquisition particulier.

Ce procédé d'acquisition très simple est également utilisé pour les jeux de données Uni-PS [42] et SDM-UniPS [43] qui sont des jeux de données de test pour la stéréophotométrie universelle. Tout comme pour *Shape and Material*, les photos ont été prises à l'intérieur d'une maison. La différence est que les sources lumineuses ne sont pas "fixes" mais sont aléatoirement positionnées et à une distance d'environ 30 cm de l'objet pour permettre d'avoir une variation spatiale de la direction lumineuse. Le fait de rapprocher les sources lumineuses permet d'amplifier ce phénomène et d'ainsi tester la robustesse des méthodes de stéréophotométrie dans un environnement lumineux complexe. Un autre avantage de Uni-PS et SDM-UniPS est que les images composant ces bases sont des images à haute résolution ce qui permet de tester les performances des méthodes de stéréophotométrie sur des images très bien résolues et prises dans un environnement non contraint. Sur les figures 2.40 et 2.41, nous pouvons voir des exemples d'images de ces deux bases de données.

FIGURE 2.39 – Exemples d'images du jeu de données *Shape and Material* [63] proposé par Lichy *et al.* en 2021.

FIGURE 2.40 – Exemples d'images du jeu de données Uni-PS [42] proposé par Ikehata en 2022.

FIGURE 2.41 – Exemples d'images du jeu de données SDM-UniPS [43] proposé par Ikehata en 2023.

2.4.3 Bases de données synthétiques

Les bases de données vues précédemment sont utilisées pour tester et comparer les performances des méthodes de stéréophotométrie car elles sont trop petites pour être utilisées pour l'entraînement d'un réseau. En pratique, il est très compliqué d'acquérir une base de données d'images réelles suffisamment importante et variée en terme d'objets, avec des matériaux représentant l'ensemble des possibilités et combinaisons possible entre matériaux, formes, directions lumineuses et environnements. À cela, il faut ajouter la difficulté en terme de temps nécessaire pour obtenir les vérités terrains, i.e. les cartes de normales, associées à chaque objet.

Pour résoudre cette problématique, les bases de données d'images synthétiques sont couramment utilisées et pertinentes pour entraîner un réseau de neurones pour la stéréophotométrie. Dans la littérature, plusieurs d'entre elles ont été créés pour répondre à ce besoin, notamment *Blobby* [18] introduit par Chen *et al.* en 2018, *Sculpture* [18] également introduit par Chen *et al.* en 2018, *CyclePS* [45] introduit par Ikehata en 2018, *CyclePS+* [41] introduit par Ikehata en 2021 et enfin *PS-Wild* [42] lui aussi introduit par Ikehata en 2022.

Blobby et *Sculpture* [18] ont été générées en utilisant le jeu de données MERL [70] qui contient 100 *BRDFs* (i.e réflectances bidirectionnelles) de matériaux "réels". La réflectance bidirectionnelle correspond à la fonction qui décrit la quantité de lumière réfléchie par une surface en fonction de l'angle d'incidence et de l'angle de vision.

Dans le cas de *Blobby*, ces matériaux ont permis de faire le rendu de 10 "blobs" sous 1 296 angles différents. Pour chaque angle de vue, deux matériaux sont appliqués tour à tour pour réaliser deux prises d'images. Au total, la base de données *Blobby* est constituée de 25 920 échantillons. Chaque vue est composée de 64 images d'une résolution de 128×128 pixels. La distribution des 64 sources lumineuses est illustrée sur la figure 2.45b : celles-ci sont réparties aléatoirement sur l'hémisphère entourant la caméra. Dans le cas de *Sculpture*, le même procédé de rendu est appliqué, la seule différence, comparativement à *Blobby*, réside dans le type de géométrie utilisé. Des formes lisses et sans détails sont utilisées dans *Blobby*. À l'inverse, *Sculpture* est composée de formes plus complexes, de formes beaucoup plus détaillées. Des exemples d'images de *Blobby* et *Sculpture* sont affichées sur la figure 2.42.

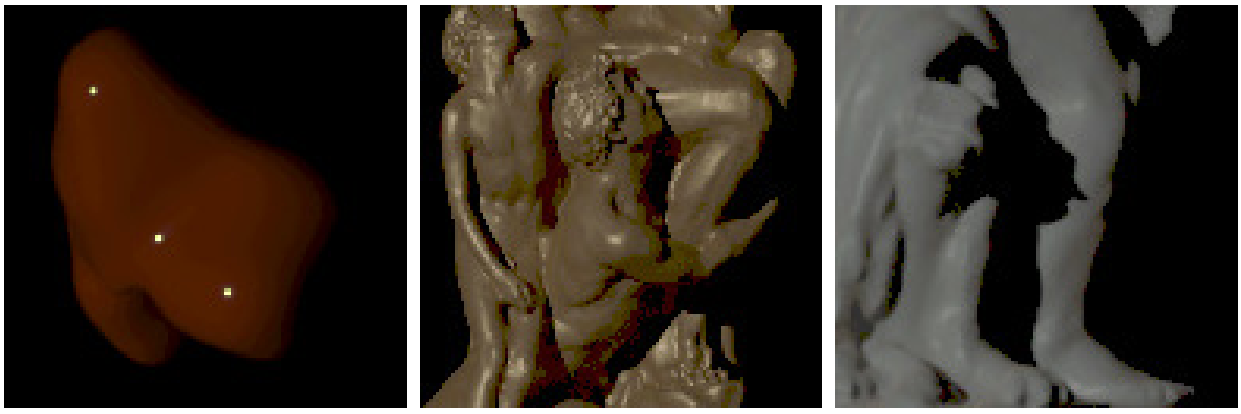


FIGURE 2.42 – Exemples d'images des jeux de données *Blobby* [18] et *Sculpture* [18].

Ces deux jeux de données ont relativement peu de géométries et peu de matériaux pour couvrir l'ensemble du spectre des objets observables dans le monde réels. De plus, les matériaux sont spatialement uniformes sur l'ensemble des objets, ce qui n'est pas réaliste.

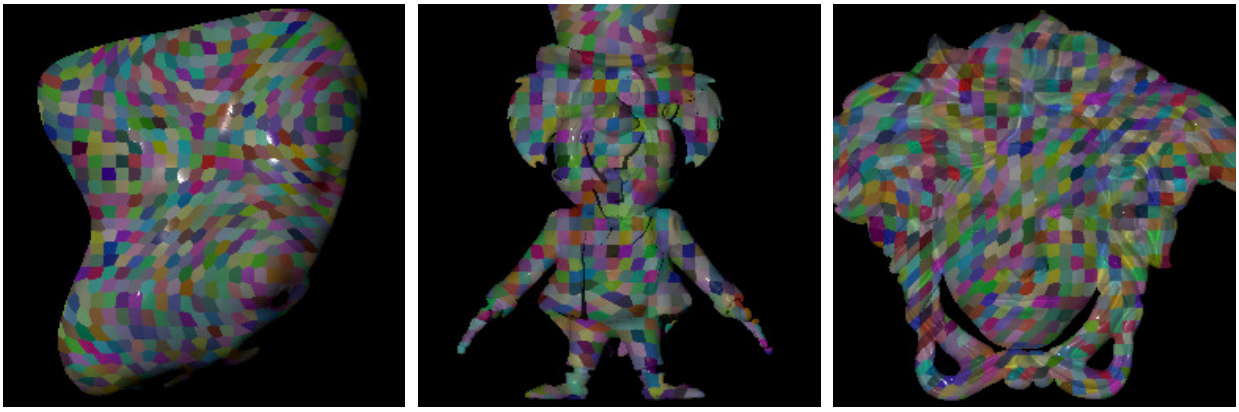


FIGURE 2.43 – Exemples d'images des jeux de données *CyclePS* [45] et *CyclePS+* [41].

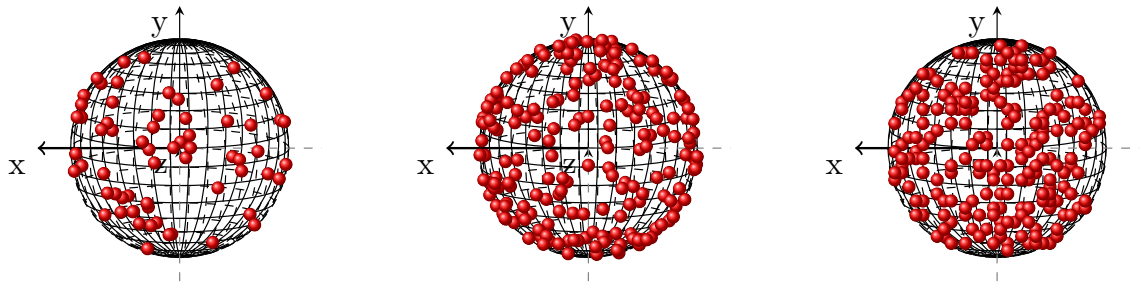
CyclePS [45] tente de résoudre ces problèmes. En effet, pour augmenter le nombre de matériaux, les auteurs se sont servis du *Disney's principled BSDF* [12]. Comparativement à la *BRDF*, la *BSDF* traite le cas de la transmittance et non pas uniquement le cas de la réflectance d'un matériau, ce qui permet de générer artificiellement l'ensemble du spectre des matériaux. Pour compenser la problématique d'uniformité des matériaux, la surface des objets a été divisée en 5 000 zones indépendantes. Dans chacune d'elle, un matériau différent est appliqué. Finalement, pour le nombre de géométries, les auteurs ont trouvé 15 modèles 3D libres de droits avec des surfaces suffisamment complexes et variées pour représenter un grand nombre de surfaces et de détails. Pour chaque échantillon, 1000 images d'une résolution de 256×256 pixels ont été synthétisées. Les directions lumineuses associées à chaque image sont réparties uniformément sur l'hémisphère autour de l'appareil photo, voir la figure 2.45b. Bien que le nombre d'images soit très conséquent par échantillon, *CyclePS* ne contient qu'une vue par objet. Ainsi, au final il n'y a dans ce jeu de données que 15 échantillons. *CyclePS+* [41] est une amélioration de *CyclePS* en augmentant le nombre d'objets et de géométries qui passe de 15 à 25. La méthode de génération des images restent inchangée mis à part cette augmentation de géométries.

FIGURE 2.44 – Exemples d'images du jeu de données *PS-Wild* [42].

La dernière base de données utilisée dans l'état-de-l'art est *PS-Wild* [42]. Celle-ci est différente des autres par son cadre d'application. En effet, elle est destinée à la stéréophotométrie universelle. La stéréophotométrie universelle est applicable avec un environnement ambiant ce qui n'est pas possible avec la stéréophotométrie classique qui est restreinte à un environnement noir. Pour permettre aux réseaux de neurones d'apprendre l'impact de la luminosité ambiante, il est nécessaire que la base d'entraînement prennent en compte cette nouvelle caractéristique. Pour simuler la luminosité ambiante, comme celle d'une pièce, de la nature ou du soleil par exemple, des images

360° *HDRI* sont utilisées dans les moteurs de rendu. Le nombre d'objets et géométries présent a considérablement augmenté comparativement aux autres bases avec pas moins de 410 objets. Le nombre de matériaux employés a aussi augmenté avec 926 matériaux. Pour finir, 31 images *HDRI* ont été utilisées. Les objets 3D, les matériaux et les images *HDRI* proviennent du site web *Adobe Stock* [1]. Pour chaque vue, 16 images ont été rendues avec une résolution de 512×512 pixels, voir la figure 2.44. Les informations sur les directions lumineuses ne sont pas fournies étant donné que cette base de données est orientée pour la stéréophotométrie universelle donc sans information sur les sources lumineuses.

Finalement, les positions des sources lumineuses des bases de données d'entraînements sont illustrées en figure 2.45.



(a) *Sculpture*, *Blobby* [18] (b) *CyclePS* [45] (c) *CyclePS+* [41]

FIGURE 2.45 – Distributions des sources lumineuses dans les bases de données synthétiques d'entraînement *Sculpture* [18], *Blobby* [18], *CyclePS* [45] et *CyclePS+* [41]. Dans le cas de *CyclePS* et *CyclePS+*, seule une partie des sources lumineuse est affichée. L'axe z correspond à l'axe optique de la caméra en considérant que l'objet se situe aux coordonnées $(0, 0, 0)$.

2.4.4 Récapitulatif des bases de données existantes

Dans le tableau 2.1, un récapitulatif des caractéristiques des bases de données existantes a été effectué. En ce qui concerne les bases de données d'entraînement, la plupart des bases de données ont soit très peu de géométries, soit très peu de matériaux ou très peu des deux. *PS-Wild* [42] est celle qui se distingue le plus par un nombre beaucoup plus important de géométries/objets et de matériaux. Cependant, le nombre d'objets (410 objets) reste encore trop faible pour synthétiser l'ensemble des formes géométriques possibles dans la réalité. De même, elle ne présente que 926 matériaux. Cela représente déjà une certaine diversité mais ce n'est pas encore suffisant pour permettre à un réseau de neurones d'apprendre à gérer l'ensemble du spectre des matériaux.

Du point de vue des bases de données de test existantes, celles-ci sont récapitulées également dans le tableau 2.1. Le principal inconvénient de l'ensemble de ces bases est que chacune d'entre elle a été acquise dans un environnement contrôlé. Par conséquent, ces bases ne sont pas totalement adaptées pour tester les performances des méthodes universelles par exemple. De plus, la diversité des objets, positions lumineuses ou matériaux n'est pas très grande pour chacune des bases individuellement. En revanche, ces bases se complètent dans le sens où elles présentent chacune des caractéristiques différentes d'un point de vue des matériaux ou des géométries.

Les bases de données sans vérité terrain sont récapitulées dans le tableau 2.2. *Skoltech3D* [93] est celle qui se distingue le plus par un grand nombre d'objets, de vues, d'échantillons et de matériaux. En revanche, elle souffre d'un manque de diversité au niveau de l'environnement. A contrario, les autres bases sont moins diversifiées mais ont l'avantage de prendre en compte différents types d'environnement. Ainsi, même sans vérité terrain, on peut évaluer qualitativement les reconstructions pour les méthodes de stéréophotométrie universelle.

		# objets	# vues	# échantillons	# lumières	# matériaux	# environnements
Entraînement	*Blobby* [18]	10	1 296	25 920	64	100	0
	Sculpture [18]	8	**1 387-6 874**	**59 292**	64	100	0
	CyclePS [45]	15	10	180	**1 300**	**90 000**	0
	CyclePS+ [41]	25	10	180	**1 300**	**90 000**	0
	PS-Wild [42]	**410**	≈ 24	10 099	10	926	**31**
Test	*DiLiGenT* [87]	10	1	10	96	10	0
	DiLiGenT10^2 [80]	10	10	**100**	**100**	10	0
	DiLiGenT-Pi [95]	**30**	**1**	30	**100**	30	0
	Luces [72]	14	1	14	52	14	0

TABLE 2.1 – Tableau comparatif des caractéristiques des différentes bases de données avec vérité terrain disponibles dans la littérature. La meilleure diversité dans chaque catégorie est mise en gras.

	# objets	# vues	# échantillons	# lumières	# matériaux	# environnements
Skoltech3D [93]	107	100	10700	11	107	0
Shape and Material [63]	111	1	111	6	111	1
Uni-PS [42]	6	1	6	13-21	6	6
SDM-UniPS [42]	13	1	13	5-13	13	13

TABLE 2.2 – Tableau comparatif des caractéristiques des différentes bases de données n'ayant pas de vérité terrain.

Le prochain chapitre s'intéresse ainsi à la création d'une base de données d'entraînement permettant un apprentissage plus performant pour les matériaux et les géométries complexes.

CHAPITRE 3

Création d'une nouvelle base de données synthétique

La base de données est un élément décisif dans l'apprentissage d'un réseau de neurones, quelle que soit l'application. En effet, une base de données d'entraînement se doit d'être la plus diversifiée possible afin de répondre à la problématique souhaitée.

Dans le cadre de la stéréophotométrie, la base de données "idéale" prend en compte les quatre points clefs suivants : les matériaux, la géométrie des objets, les directions lumineuses et l'environnement. Les matériaux sont très importants car ceux-ci réagissent différemment en fonction de l'exposition lumineuse. Certains vont réfléchir, d'autres absorber la lumière. De même, la géométrie d'un objet change l'angle de réflexion de la lumière. Finalement, l'environnement et la source lumineuse permettent au réseau de s'adapter à n'importe quelle situation d'acquisition.

Dans ce chapitre, nous élaborons un nouveau jeu de données synthétiques d'entraînement adapté au problème de stéréophotométrie. Cette nouvelle base de données a pour objectif de diversifier les matériaux, les géométries des objets, les environnement et les directions lumineuses afin d'améliorer les performances de reconstruction des normales.

Comme nous avons pu le voir précédemment dans le paragraphe 2.4, les bases de données d'entraînement disponibles n'ont pas une grande diversité de matériaux et de géométries. De plus, la diversité d'environnement ou de directions lumineuses est également très limitée. En effet, ces bases de données sont restreintes à un contexte dans un milieu contrôlé, sans aucun éclairage externe. En exploitant uniquement ces bases de données d'entraînement, il est donc impossible d'entraîner des réseaux de neurones pour des contextes spécifiques. Par exemple, dans un contexte où la lumière ambiante d'une pièce est présente, il est impossible d'obtenir de bonnes reconstructions avec celles-ci.

Il est également important de noter que l'utilisation de ces bases de données pour entraîner ne permet pas de s'adapter à toutes les distributions lumineuses sur l'hémisphère. En effet, pour obtenir les meilleurs résultats possibles pour une distribution lumineuse précise, l'idéal est d'entraîner sur celle-ci en particulier. Par exemple, un réseau de neurones entraîné suivant la distribution de *DiLiGenT* aura de meilleurs résultats sur *DiLiGenT* que le même réseau entraîné sur une autre distribution lumineuse. Il est donc important, pour avoir une grande flexibilité, de pouvoir générer n'importe quelle distribution lumineuse, type de matériaux, etc. Donc, pour s'adapter au mieux à l'objectif de la méthode développée et obtenir les meilleurs résultats possibles, la diversité est décisive.

Pour résoudre toutes ces problématiques et contraintes, la création d'un nouveau jeu de données est pertinente. Dans un cadre d'entraînement, la création d'un jeu de données réelles avec la vérité terrain associée est irréalisable. Le nombre d'objets nécessaires et la difficulté d'obtenir une vérité terrain suffisamment précise pour permettre l'entraînement, rendent la tâche irréalisable. En effet, pour entraîner correctement un réseau de neurones, le nombre de matériaux et de géométrie doit être très conséquent. Ainsi, l'option la plus adaptée est de générer des bases de données d'entraînement synthétiques les plus réalistes et diversifiées possible. Pour ce faire, la solution adoptée est d'utiliser un moteur de rendu réaliste tel que *Cycle* [23] de *Blender* [20]. Il s'agit d'un moteur de rendu non biaisé, de type *"path tracing"*, générant des images photo-réalistes. La facilité à gérer le type de matériaux et même à générer de nouveaux types de matériaux font de ce moteur de rendu un candidat idéal.

Pour résumer, plusieurs améliorations sont à réaliser :
— augmentation du nombre et de la diversité des géométries des objets,
— augmentation du nombre et du type de matériaux, notamment du verre, des métaux ou encore du plastique,
— ajout d'environnements ambiants,
— possibilité de créer des distributions lumineuses bien spécifiques.

Ces différents points d'amélioration sont détaillés ci-dessous, et ces contributions apparaissent en partie dans les publications [31, 32, 33].

3.1 Augmentation du nombre et de la diversité des géométries

Les objets contenus dans le jeu de données d'entraînement doivent non seulement contenir des géométries avec des angles très saillants, mais également des géométries avec des formes lisses pour permettre au réseau de neurones de voir le maximum de formes possibles. Pour ce faire, deux stratégies ont été employées. La première consiste à générer des objets avec des sommes de potentiels gaussiens et ensuite d'extraire les iso-surfaces à l'aide de l'algorithme *Marching Cubes* [67]. Des exemples d'objets générés par cette méthode sont affichés sur la figure 3.1.

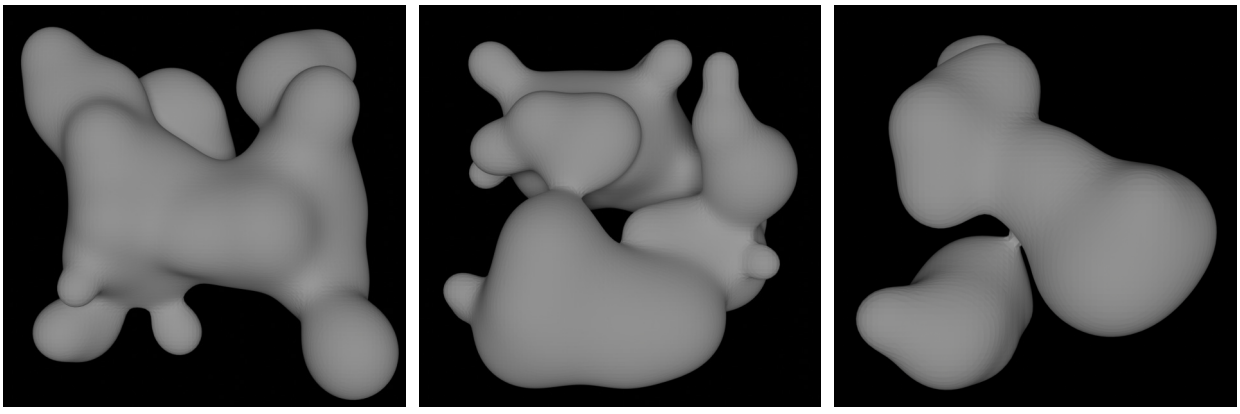


FIGURE 3.1 – Exemples d'objets *"Blobby"* générés par une somme de potentiels gaussiens suivi de l'algorithme *Marching Cubes* [67].

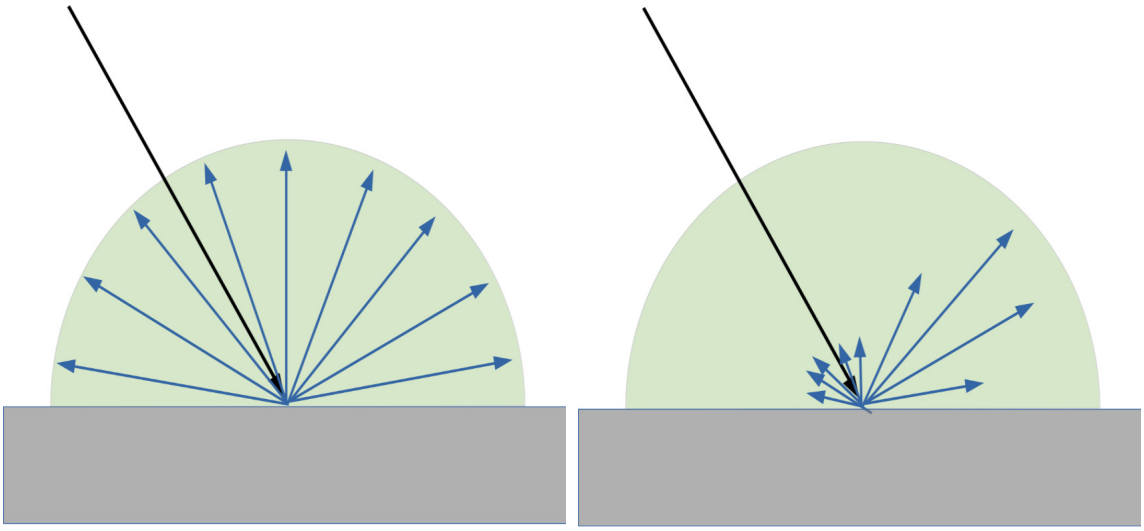
La seconde méthode consiste à utiliser des objets 3D disponibles sur différents sites internet, notamment des sites d'impressions 3D. Ainsi, environ 11 000 objets représentant des gravures, statues, jouets ou autres ont été téléchargés à partir des sites *Scan the World* [6], *Sketchfab* [7] et *Minifactory* [4]. Des exemples sont affichés sur la figure 3.2.

FIGURE 3.2 – Exemples d'objets 3D disponibles sur *Scan the World* [6], *Sketchfab* [7] et *Minifactory* [4].

À l'aide de ces deux méthodes, nous pouvons obtenir à la fois des formes simples et complexes. Ainsi, beaucoup plus de géométries sont présentes dans notre base de données d'entraînement.

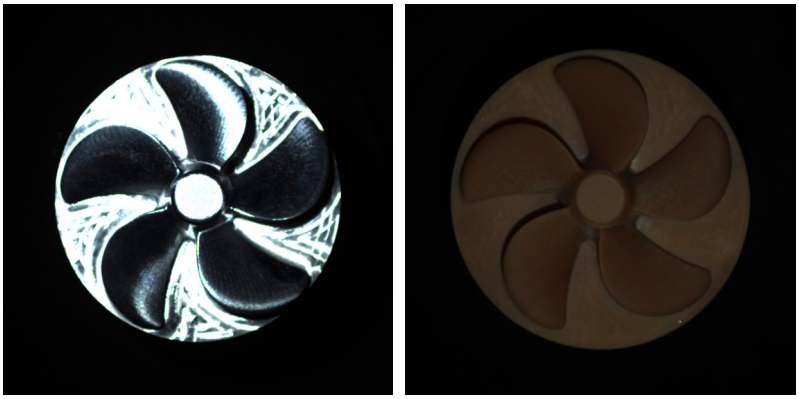
3.2 Augmentation du nombre et du type de matériaux

De même que pour la géométrie des objets, la base de données d'entraînement doit contenir la plus grande variété possible de matériaux afin que l'ensemble des matériaux du monde réel soit représenté. Ainsi, les réseaux pourront apprendre les caractéristiques physiques et les effets lumineux de chaque matériau. En effet, il n'est pas possible pour un réseau de neurones entraîné uniquement sur des matériaux avec une réflectance diffuse par exemple, de généraliser sur des matériaux avec une réflectance spéculaire. Les effets lumineux sur ces deux types de matériaux sont beaucoup trop éloignés l'un de l'autre. Nous pouvons voir sur la figure 3.3 la différence du trajet lumineux entre un matériau avec une réflectance diffuse et un matériau avec une réflectance spéculaire. Cette différence sur la perception d'un objet est illustrée sur la figure 3.4 sur un objet réel. Il est donc primordial que l'ensemble du spectre des matériaux soit représenté pour permettre la généralisation.



Réflectance diffuse. Réflectance spéculaire.

FIGURE 3.3 – Schémas théoriques d'une réflectance diffuse à gauche et spéculaire à droite.



Réflectance spéculaire. Réflectance diffuse.

FIGURE 3.4 – Images d'une surface réelles présentant une réflectance spéculaire (à gauche) et diffuse (à droite).

Pour obtenir une telle diversité dans les matériaux, une première possibilité est d'utiliser des cartes de matériaux réels créées pour des moteurs de rendu comme ceux disponibles sur le site internet *AmbientCG* [3]. Au total, seulement 1 100 matériaux étaient disponibles, répartis en une vingtaine de type de surfaces d'objets (bois, métaux, tissus, tapis, etc.). Cela est malheureusement trop peu. Cependant, pour mettre au point un algorithme de scannage des matériaux [22], une équipe de chercheurs a généré une grande quantité de matériaux synthétiques (*SVBRDF*). Ce jeu de données contient environ 200 000 matériaux. L'exploitation de ces deux sources de matériaux permet de générer une base de données d'entraînement solide d'un point de vue des matériaux.

Finalement, pour parfaire cette base de données en terme de matériaux, et être en capacité de générer un type bien particulier de matériaux, comme par exemple des matériaux transparents tels que l'acrylique ou le diamant, une chaîne de génération de matériaux synthétiques a été créée. Cette chaîne de génération de matériaux se concentre principalement sur les matériaux compliqués à gérer pour un algorithme de stéréophotométrie : les métaux, les matériaux transparents ou encore les matériaux avec une dispersion surfacique. Ainsi, les couches "*BSDF* guidée", "*BSDF* verre", "*BSDF* brillante", "dispersion sous-surfacique" intégrées à *Blender* sont utilisées. Les couches "verre", "brillante" et "dispersion sous-surfacique" sont spécialisées dans un type de matériaux bien spécifique tandis que la couche "*BSDF* guidée" est globale et en capacité de générer une grande diversité de matériaux, y compris ceux des couches spécialisées. En revanche, les couches spécialisées sont conservées afin d'être certain que l'ensemble des types de matériaux sont bien générés et que chaque possibilité pour chaque type de matériaux est bien prise en compte, i.e. que toutes les possibilités ont bien été générées. La couche globale permet, quant à elle, de générer des matériaux un peu plus atypiques en mélangeant facilement des caractéristiques de plusieurs matériaux très différents.

Un résumé des paramètres des couches *BSDF* utilisées du logiciel *Blender* pour la génération de notre base de données est présenté sur le tableau 3.1.

	Couleur	Subsurface	Métallique	Spéculaire	Rugosité	Lustre	Vernis	IOR	Transmission	Normale	Anisotropie
BSDF verre	√				√			√		√	
BSDF brillante	√				√					√	√
Dispersion sous-surfacique	√							√		√	√
BSDF guidée	√	√	√	√	√	√	√	√	√	√	√

TABLE 3.1 – Les paramètres des couches *BSDF* de *Blender* utilisés pour générer les bases de données d'entraînement.

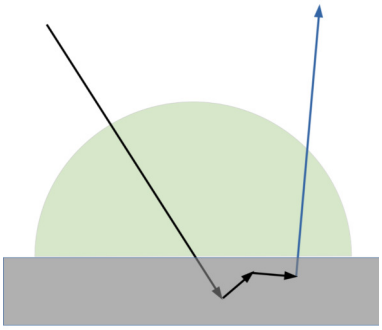


FIGURE 3.5 – Illustration d'une dispersion sous-surfacique. Les rayons pénètrent dans l'objet et suivent un trajet "aléatoire" à l'intérieur du matériau avant de ressortir en un autre point que le point de pénétration.

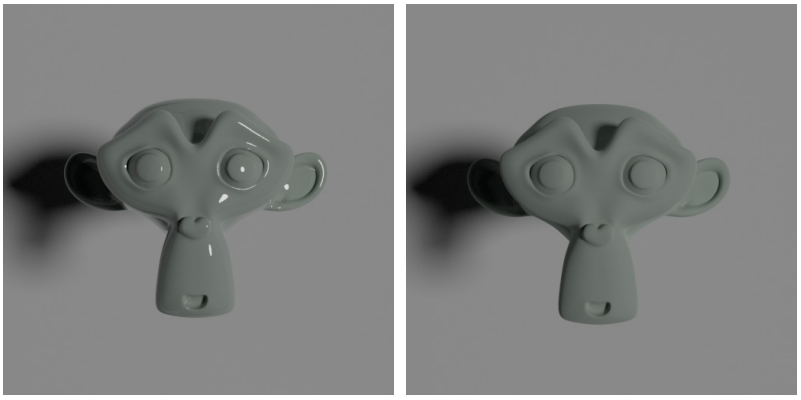
Ces couches prennent en entrée des paramètres numériques qui représentent les caractéristiques du matériau généré. Ainsi, le choix de ces paramètres est primordial pour obtenir une base de données diversifiée et pertinente. Dans le cas de la couche "*BSDF* guidée", tous les paramètres ne sont pas utilisés en même temps. Pour avoir le maximum de diversité le choix des paramètres à utiliser se fait de façon aléatoire.

Avec la dispersion sous-surfacique, la lumière, au lieu d'être réfléchie sur la surface de l'objet, pénètre dans l'objet et va être réfléchie à l'intérieur de celui-ci. Le point de sortie de la lumière n'est ainsi pas le même que le point d'impact de la lumière à la surface de l'objet. Un schéma de ce phénomène est illustré à la figure 3.5. Ce phénomène se présente pour des matériaux tels que la peau, les bougies ou certaines lampes d'ambiance par exemple. Ce type de réflexion peut aussi être géré par le paramètre *"subsurface"* de la couche "*BSDF* guidée".

Pour rentrer un peu plus dans les détails des différents paramètres utilisés :

— Couleur : couleur de la surface de l'objet.

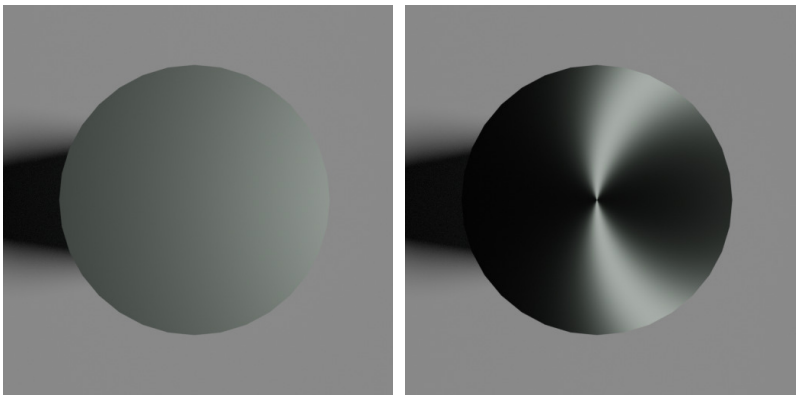
— Rugosité : la rugosité d'une surface correspond à des irrégularités dues à de légères différences de niveau. D'un point de vue de la lumière, une surface avec peu de rugosité va renvoyer beaucoup de reflets. À l'inverse, une surface avec beaucoup de rugosité va "étaler" les reflets pour devenir une surface visuellement mate, comme illustré sur la figure 3.6.



Rugosité=0. Rugosité=1.

FIGURE 3.6 – Comparaison du rendu entre un matériau avec et sans rugosité.

— Anisotropie : un matériau anisotrope est l'inverse d'un matériau isotrope, cela signifie que les propriétés physiques de celui-ci varient en fonction de la direction radiale. Par exemple, lors de l'observation d'une coupe transversale d'un tronc d'arbre, les anneaux de croissance sont visibles dans le sens radial mais pas dans le sens axial. Celle-ci est également visible dans le dos d'une poêle, la réflexion de la lumière varie en fonction de la direction radiale. L'anisotropie est aussi caractérisée dans *Blender* par le paramètre "rotation" qui représente l'angle de rotation de la direction de la tangente anisotrope ou en d'autre termes la direction de l'allongement des reflets anisotropes. Ce phénomène est illustré sur la figure 3.7.



Anisotropie=0. Anisotropie=1.

FIGURE 3.7 – Comparaison du rendu entre un matériau avec et sans anisotropie.

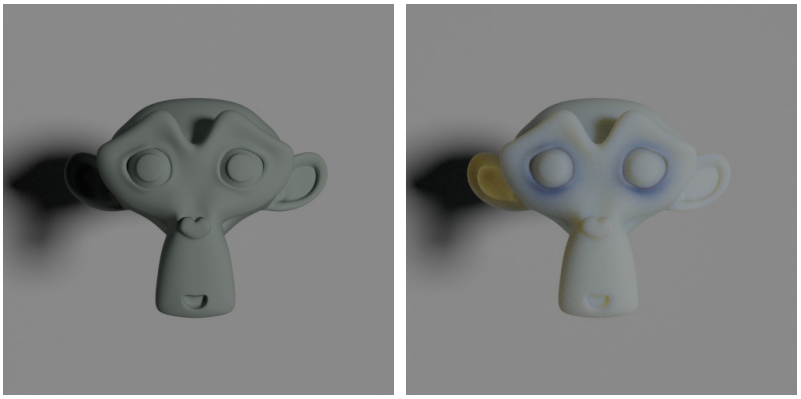
— *IOR* (indice de réfraction) : l'*IOR* caractérise le changement de vitesse de la lumière lors d'un changement de milieu. Plus le changement de vitesse est important, plus l'*IOR* est élevé. En pratique, cela implique un changement dans la direction de la lumière. Plus l'indice est important, plus la lumière va être déviée, comme dans l'exemple de la figure 3.8.

IOR=1. *IOR*=1.33 (eau). *IOR*=2.417 (diamant).

FIGURE 3.8 – Comparaison du rendu entre un matériau avec un *IOR* plus ou moins élevé.

— Rayon (dispersion sous-surfacique) : distance de la surface de l'objet à laquelle la lumière se réfléchit, se disperse à l'intérieur de celui-ci (si la matière est sous-surfacique).

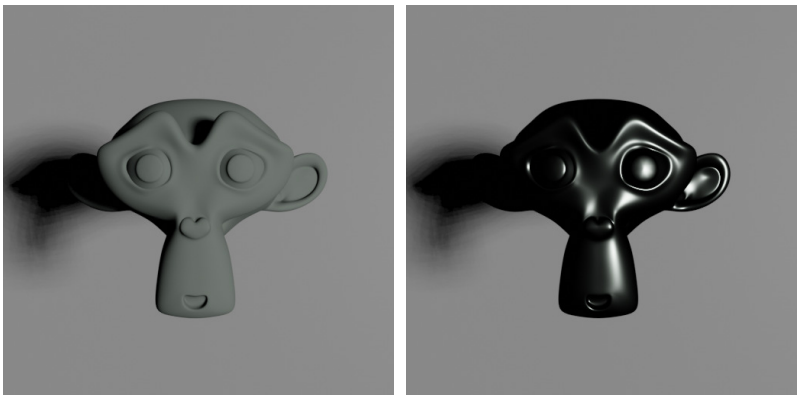
— Sous-surface : correspond à la dispersion sous-surfacique décrite précédemment. Un exemple de ce phénomène est affiché sur la figure 3.11. Plus la valeur est proche de 1 plus la matière présente les caractéristiques sous-surfaciques tandis qu'une valeur de 0 indique que la lumière ne "rebondit" pas du tout à l'intérieur de la matière.



Sous-surface=0. Sous-surface=1.

FIGURE 3.9 – Comparaison du rendu entre un matériau avec ou sans dispersion sous-surfacique.

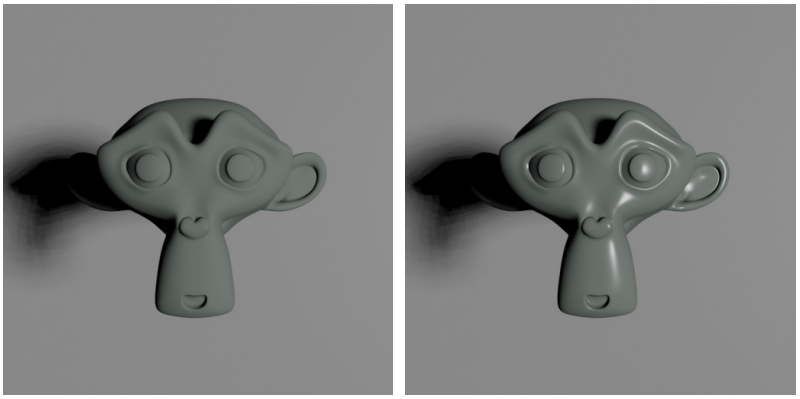
— Métallique : comme son nom l'indique, cette valeur contrôle le pourcentage métallique d'un objet, plus cette valeur est élevée plus l'objet sera métallique et plus il reflètera la lumière. À l'inverse, une valeur faible correspondra à un matériau avec une réflectance diffuse.



Métallique=0. Métallique=1.

FIGURE 3.10 – Comparaison du rendu entre un matériau non métallique et un métallique.

— Spéculaire : contrôle le taux de spécularité de la surface. Plus la valeur sera proche de 1 plus la surface aura des reflets, à l'inverse une valeur proche de 0 représente un matériau qui "absorbe" ces reflets voir figure 3.11.



Spéculaire=0. Spéculaire=1.

FIGURE 3.11 – Comparaison du rendu entre un matériau non spéculaire et un spéculaire.

— Lustre : ce paramètre permet de simuler une réflexion similaire à celle des tissus et velours en ajoutant au-dessus de la surface une légère couche qui absorbe et réfléchit la lumière de manière diffuse. La teinte contrôle la couleur de la réflexion diffuse.

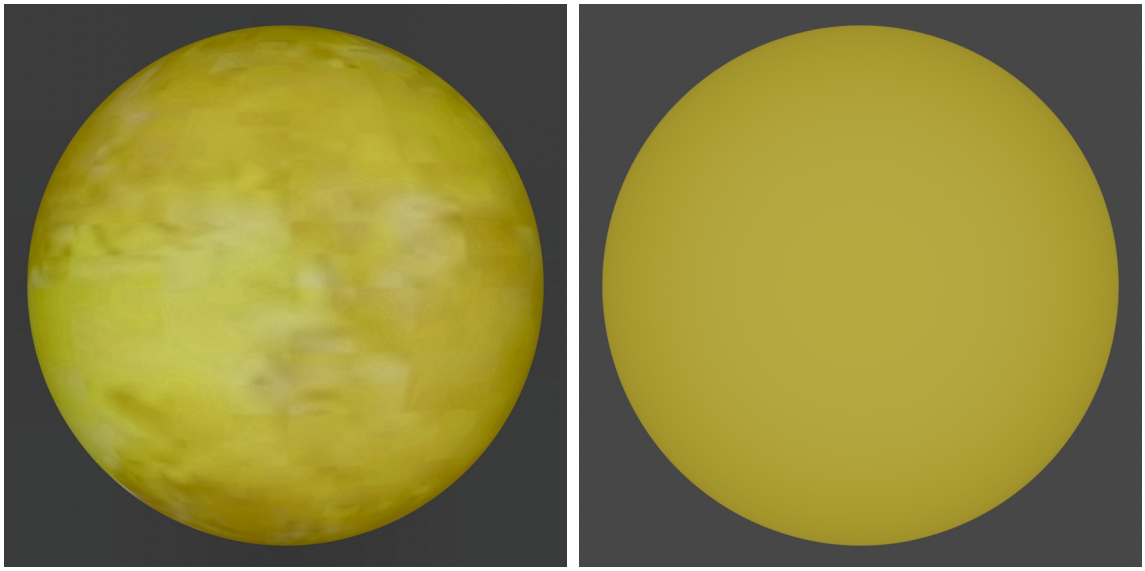
— Vernis : comme son nom l'indique, ce paramètre sert à simuler du vernis comme sur les voitures ou sur les meubles en ajoutant une couche blanche au-dessus de la surface pour obtenir un matériau à l'aspect lisse et brillant. Cet effet est également contrôlé par un paramètre de rugosité.

Le choix de ces valeurs de paramètres en entrée peut se faire de deux manières. Soit une unique valeur par paramètre est utilisée, soit une matrice de valeurs est utilisée pour simuler une variation spatiale du matériau. Dans le cas d'une valeur unique, le choix des valeurs des paramètres se fait de façon aléatoire dans des intervalles de valeurs fixés au préalable pour chaque *BSDF*.

Dans le cas d'une matrice de valeurs, une image quelconque (image RGB, carte de normale, carte de profondeur ou autre) est utilisée en normalisant ses valeurs dans un intervalle de valeurs fixé au préalable, le même que celui dans le cas d'une valeur unique. L'intérêt d'utiliser une matrice de valeurs est illustrée sur la figure 3.12. Le fait de mettre une matrice de valeurs permet d'avoir une variation spatiale. Les réseaux de neurones sont ainsi capables de gérer des matériaux uniformes mais également des objets avec une réflectance non uniforme.

3.3 Environnement ambiant

L'ajout d'un environnement ambiant autour de l'objet pour simuler la lumière ambiante, telle qu'elle pourrait être dans une pièce en pleine journée, se fait via l'utilisation d'une image 360°. Pour ce faire, 1 100 images ont été téléchargées provenant de diverses sources telles que *Poly Haven* [5], *AmbientCG* [3] et *Alexandre Duret-Lutz* [2]. Comme nous pouvons le voir sur la figure 3.13, l'utilisation d'une image 360° permet d'obtenir un arrière plan et de prendre en compte son impact sur l'objet. Illustré aux figures 3.14 et 3.15, l'impact qu'a cet environnement ambiant sur la réflexion de la lumière est visible sur la surface de l'objet. Cela ajoute du réalisme et une difficulté car l'intensité lumineuse perçue en chaque pixel de l'objet n'est alors plus forcement liée à la lumière utilisée pour la stéréophotométrie mais également à celle de l'environnement.



Matrice de valeurs. Valeur unique.

FIGURE 3.12 – Comparaison de l'utilisation d'une matrice de valeurs en entrée des paramètres des *BSDF* contre une valeur unique. Une matrice de valeurs permet d'obtenir une variation spatiale, apportant de la robustesse à notre base de données.

FIGURE 3.13 – Exemples d'environnements ambiants possibles.

3.4 Distributions lumineuses

Les jeux de données d'entraînement et de test existants ont des distributions lumineuses spécifiques. Par exemple, les jeux de données d'entraînement *Blobby* et *Sculpture* ont une distribution lumineuse complètement aléatoire sur l'hémisphère (voir figure 2.45a). Le jeu de données réelles de test *DiLiGenT* [87] adopte une distribution lumineuse sous forme de rectangle (voir figure 2.37a). *DiLiGenT-Pi* [95] et *DiLiGenT*10^2 [80] adoptent quant à eux une distribution uniforme des directions lumineuses sur l'hémisphère (voir figure 2.37b). Pour finir, *Luces* [72] a une distribution en forme de disque autour de la caméra (voir figure 2.37c). En effet, les auteurs ont utilisé un disque sur lequel ils ont positionné plusieurs rangées de LED sous forme de cercle.

Les distributions lumineuses ne sont ainsi pas toutes les mêmes. Il est important de restreindre les contraintes expérimentales pour les utilisateurs souhaitant utiliser la stéréophotométrie. Par conséquent, les méthodes doivent être capable de généraliser sur n'importe quelle type de distri-

FIGURE 3.14 – Comparaison de l'effet de l'ajout d'un environnement sur un objet avec un matériau très réfléchissant.

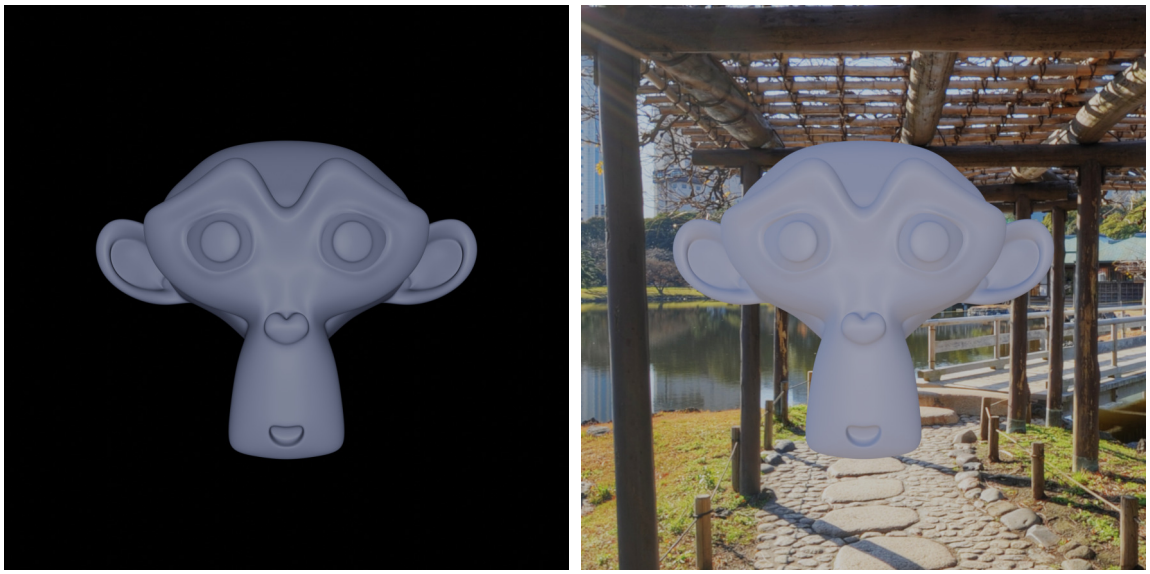
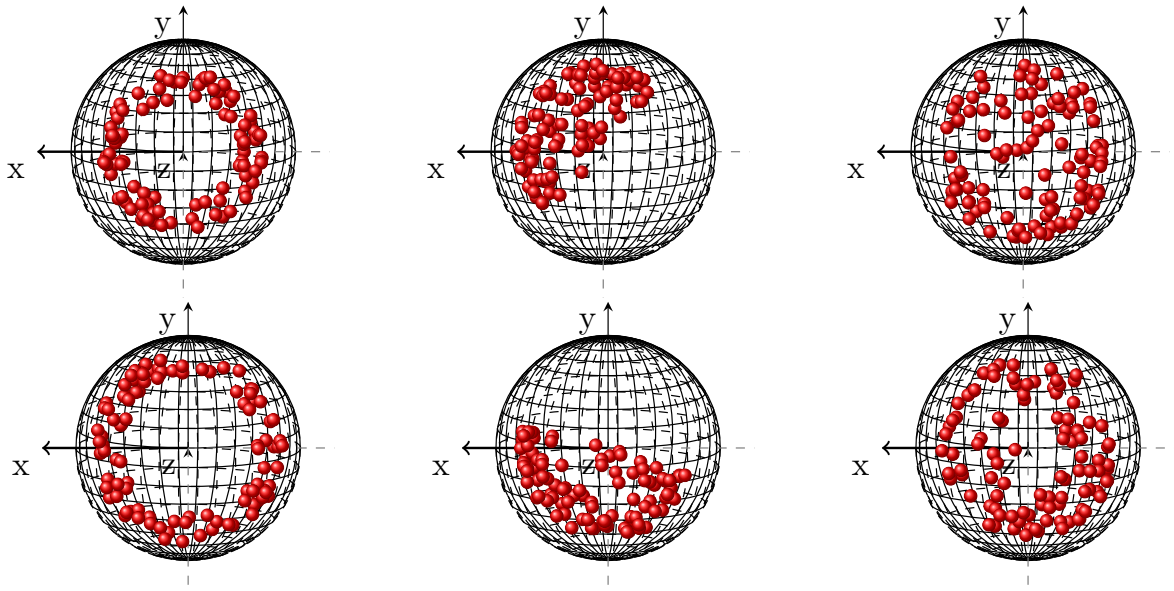


FIGURE 3.15 – Comparaison de l'effet de l'ajout d'un environnement sur un objet présentant une très grande rugosité.



Anneau. Demi-hémisphère. Hémisphère.

FIGURE 3.16 – Exemples de distributions lumineuses utilisées dans les jeux d'entraînement générés. L'axe z correspond à l'axe de vue de la caméra avec l'objet positionné à la position $(0, 0, 0)$.

bution lumineuse. Trois types de distributions lumineuses ont ainsi été implémentées : l'anneau, le demi-hémisphère et l'hémisphère, illustrées sur la figure 3.16. Le type "hémisphère" permet aux réseaux de gérer les lampes positionnées aléatoirement autour de la caméra. Le type "demi-hémisphère" quant à lui sert dans le cas ou l'utilisateur ne peut positionner la lampe que sur un coté de la caméra. Par exemple, dans le contexte où l'objet et la caméra sont positionnés sur une table, il est alors impossible d'avoir une direction lumineuse provenant du dessous. À noter que le demi-hémisphère n'est pas toujours orienté de la même manière. De plus, pour ne mettre aucune contrainte sur la taille des hémisphères et demi-hémisphères autour de la caméra, la distance maximale d'une lampe par rapport à la caméra n'est pas toujours la même dans les bases de données générées. L'anneau, quant à lui, permet de gérer le contexte où toutes les lampes sont plus ou moins à la même distance de la caméra.

3.5 Chaîne de génération et exemples de rendus

Finalement, une fois les géométries, les matériaux, les environnements et les positions lumineuses définis, un algorithme a été mis en place afin de générer un ensemble d'images pour un objet. La figure 3.17 montre le déroulement de la chaîne de génération de données dans *Blender*.

Dans un premier temps, le choix entre couche *BSDF* et matériau réel est fait afin de sélectionner le type du matériau. Ensuite, dans le cas d'une couche *BSDF*, les paramètres de cette couche sont fixés aléatoirement afin de personnaliser le matériau choisi. Le matériau est finalisé en choisissant soit une valeur fixe, soit une matrice en entrée pour créer un matériau unique ou un *patchwork* de matériaux sur l'objet. Dans un second temps, les choix aléatoires de la géométrie de l'objet, de la distribution lumineuse et de l'environnement sont effectués. Finalement, le processus de *Blender* construit plusieurs images de rendu pour un objet.

Voici quelques exemples de rendus des différents types de matériaux.

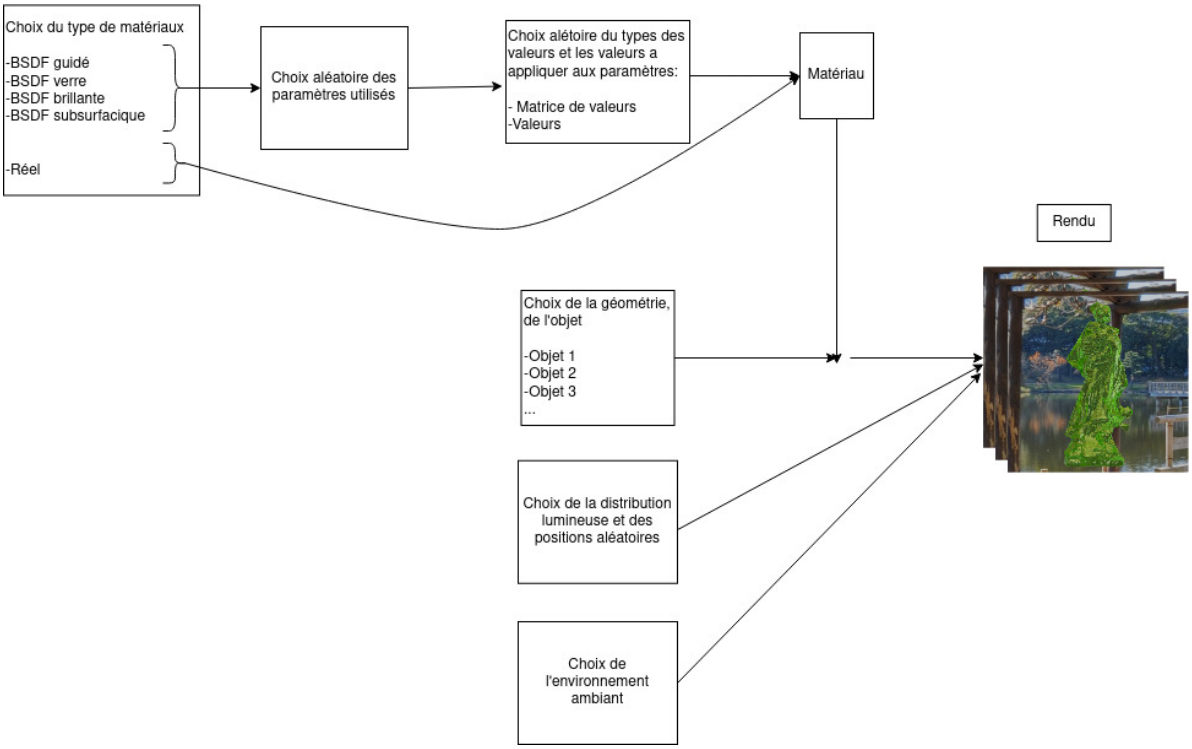


FIGURE 3.17 – Schéma récapitulatif de la chaîne de génération complète.

FIGURE 3.18 – Exemples de rendus d'objets translucides par la chaîne de génération.

FIGURE 3.19 – Exemples de rendus d'objets en métal par la chaîne de génération.

FIGURE 3.20 – Exemples de rendus d'objets avec une dispersion surfacique par la chaîne de génération.

FIGURE 3.21 – Exemples de rendus d'objets en matériaux "réels", issus du site *AmbientCG* [3], par la chaîne de génération.

3.6 Synthèse des bases de données d'entraînement

Trois bases de données ont été générées au cours de cette thèse pour répondre à différentes problématiques et questionnements. Notamment la comparaison des performances entre un apprentissage sur des matériaux "réels", issus du site *AmbientCG* [3], contre l'apprentissage avec une plus grande diversité de matériaux incluant des matériaux non réalistes ou encore pour permettre l'apprentissage de modèle pour la stéréophotométrie universelle. Par conséquent, une première base de données avec uniquement des matériaux "réels" a été générée avec *AmbientCG* [3] contenant 1 100 matériaux ainsi que 56 objets et 3 000 formes blobs. Pour cette première base de données aucun environnement, ni lumière ambiante n'a été ajouté et seul le cas des faisceaux parallèles est traité. Au total, 30 000 prises de vues ont été générées. Cette première base de données a permis de valider l'intérêt d'ajouter des formes géométriques ainsi que des matériaux dans la base d'entraînement.

Le concept étant validé, une seconde base de données a été générée, cette fois-ci avec non seulement les 1 1000 matériaux réels mais également avec la méthode de génération de matériaux synthétiques décrite précédemment. De même que pour la précédente base de données, aucun envi-

ronnement ambiant n'a été ajouté et les faisceaux lumineux sont considérés parallèles. Cette base de données a permis de grandement améliorer les performances sur la base de données $DiLiGent10^2$ [80] notamment sur les matériaux spéculaires et translucides dans le cadre de la stéréophotométrie calibrée. Ainsi, 60 000 vues sont présentes dans cette base de données.

Une dernière base de données a été générée pour résoudre la problématique de la stéréophotométrie universelle. Pour ce faire, l'ensemble des formes géométriques disponibles ont été utilisées (11 000 objets + 3 000 formes blobs), de même l'ensemble des possibilités des matériaux ont été générées (réels, synthétiques). De plus, tous les types de faisceaux lumineux sont également considérés. Finalement, un environnement ambiant est ajouté à chaque prise de vue. Cette base de données générée est la plus importante par sa taille, avec pas moins de 100 000 vues. Un grand nombre de vues est en effet nécessaire pour que l'ensemble des caractéristiques ait pu être générées.

Par la suite de ce manuscrit, la concaténation des bases de données préexistantes *Blobby* et *Sculpture* portera le nom de DB1. Notre première base de données portera le nom de DB2, la seconde de DB3 et la dernière de DB4. Un récapitulatif des différentes caractéristiques des bases de données générées est disponibles dans le tableau 3.2.

	# objets	# vues	# échantillons	# lumières	# materiaux	# environnements
DB1 (Sculpture+Blobby)	18	1296-6874	85212	64	100	0
DB2	56+3000	≈10	30000	100	1100	0
DB3	56+3000	≈20	60000	100	1100+infini	0
DB4	11000+3000	≈7	100000	50	1000+infini	1100

TABLE 3.2 – Tableau comparatif des caractéristiques des différentes bases de données générées.

3.7 Conclusion

Dans ce chapitre, nous avons mis en évidence les points forts et les points faibles des bases de données de l'état-de-l'art. Par exemple, cela peut être le nombre d'objets, la diversité des matériaux, des environnements ou des sources lumineuses. En revanche, un manque persistant de matériaux spéculaires constitue le point faible dans la littérature. Les bases de données d'entraînement, quant à elles, manquent de diversité sur l'ensemble des points pour répondre à l'ensemble des possibilités présentes dans la nature.

Ainsi, nous avons proposé une nouvelle chaîne de génération d'images synthétiques d'entraînement pour la stéréophotométrie. Nos bases de données générées sont beaucoup plus diversifiées en termes de matériaux. En effet, nous offrons la possibilité de générer des objets à partir de plusieurs matériaux, nous proposons également plus de géométries, réelles ou non, ce qui permet ainsi de répondre au plus grand nombre de cas possibles. De plus, nous proposons un grand nombre d'environnements, ce qui a pour avantage de prendre en compte les différents reflets de la lumière en milieu naturel. Par ailleurs, nous proposons plusieurs schémas lumineux pour prendre en compte le plus grand nombre de configurations possibles. Cela rend donc nos bases de données d'entraînement très diversifiées et en capacité de répondre aux problèmes de stéréophotométrie calibrée, non calibrée et universelle.

Ces bases de données générées seront utilisées dans la suite de ce manuscrit pour entraîner nos réseaux de neurones. Dans le chapitre suivant, un premier réseau de neurones, fondé sur les CNN, et une architecture multi-échelles, sera introduit pour la stéréophotométriee calibrée.

CHAPITRE 4

MS-PS : Une architecture multi-échelles pour la stéréophotométrie calibrée

L'intérêt principal des méthodes de stéréophotométrie comparativement aux autres méthodes de reconstruction 3D est leur capacité et à reconstruire les détails les plus fins de la surface des objets. Il est donc très important qu'elles puissent gérer les résolutions natives des images, sans avoir besoin de réaliser la moindre déformation, compression ou réduction de taille. En effet, il serait dommageable de réduire la taille des images et d'en perdre des détails précieux. Ainsi, considérer des méthodes capables de prendre en compte cet aspect est très important.

Du point de vue de l'apprentissage profond, l'architecture est un deuxième point décisif, après la base de données d'entraînement, pour obtenir les meilleurs résultats possibles. En effet, l'architecture d'un réseau doit assurer une bonne capacité de généralisation sur de nouvelles données. La plupart des réseaux de neurones ont des difficultés à gérer des résolutions d'images différentes. Généralement, il s'agit d'une contrainte bloquante dans leurs architectures. Par exemple, il peut s'agir d'une couche dans l'architecture prenant une taille fixe en entrée. Et même dans certains cas où le réseau est, d'un point architectural, apte à prendre en entrée toutes tailles d'images, on se rend compte qu'en pratique, les performances des réseaux décroissent rapidement dès que la résolution est très différente de celle utilisée pour l'entraînement. En effet, le nombre de couches de convolution nécessaires pour synthétiser l'information globale n'est pas du tout le même pour une image de 200×200 pixels ou une image de 8000×8000 pixels.

Pour faire face à ce problème, deux méthodes principales ont été proposées dans la littérature. La première est la séparation en patchs des images en taille fixe, avec ou sans chevauchement. La seconde est un sous-échantillonnage particulier consistant à prendre un pixel tous les N pixels, où N est la taille de l'image divisée par le nombre d'images souhaitées. Cette dernière méthode a notamment été utilisée par Ikehata [43]. Ces méthodes ont pour défaut de ne pas permettre au réseau d'avoir accès à l'ensemble de l'information contenue dans les images pour prédire la carte des normales. En effet, les patchs côte-à-côte ne permettent qu'une analyse locale de la géométrie de l'objet.

Par conséquent, dans ce chapitre, nous présentons l'architecture générale qui est l'une de nos contributions et qui servira de fil rouge à travers l'ensemble des chapitres suivants. Elle consiste en une architecture multi-échelles originale de type encodeur-décodeur pour résoudre le problème de la stéréophotométrie calibrée. Cette architecture permet de gérer des images, quelle que soit la résolution de l'image d'entrée, grâce à son caractère multi-échelles. De plus, malgré ces aspects multi-échelles et encodeur-décodeur, nous restons dans une architecture de taille moyenne mais qui reste performante. Les résultats de ce chapitre ont été présentés lors de la conférence WSCG [33].

4.1 Architecture proposée

La reconstruction des normales de matériaux translucides ou réflectifs est l'un des problèmes clés en stéréophotométrie. C'est l'un des principaux défis pour l'ensemble des méthodes de l'état-de-l'art. Ici, nous présentons une méthode, inspirée par Chen *et al.* [18], fondée sur une architecture multi-échelles de type encodeur-décodeur pour résoudre le problème de la stéréophotométrie calibrée, comme décrite dans l'état-de-l'art au chapitre 2. Notre architecture est décomposée en deux sous-réseaux indépendants :

— Le premier réseau, qui correspond à la première échelle, est en charge de générer une première estimation à basse résolution de la carte de normales en prenant uniquement les images à basse résolution en entrées.

— Le second réseau, utilisé pour toutes les autres échelles, est, quant à lui, en charge de raffiner l'estimation à l'échelle précédente en doublant la résolution à chaque fois.

Ce raffinement s'effectue en prenant en entrée non seulement les images redimensionnées à la résolution de traitement actuelle (sous-échantillonnée), mais également une estimation de la normale redimensionnée (sur-échantillonnée). Le processus de raffinement avec la deuxième échelle se termine lorsque la résolution atteinte est celle souhaitée. La première échelle de traitement se fait typiquement à l'échelle 16×16 pixels. Chacune des échelles suivantes double la résolution, jusqu'à ce que l'on retrouve la résolution des images d'origine. Lorsque la taille de l'image n'est pas un multiple de 2, l'image est sur-échantillonnée pour permettre l'inférence. Par la suite, un sous-échantillonnage est appliqué à la carte des normales afin de retrouver la taille d'origine.

D'un point de vue architectural, le traitement d'une échelle est décomposé en trois blocs distincts : un bloc d'extraction des caractéristiques, un bloc de fusion des caractéristiques des différentes images et un bloc décodeur qui prédit la carte de normales. Le bloc d'extraction des caractéristiques est constitué de sept blocs de convolutions et le bloc de fusion est un simple *max-pooling* le long de l'axe inter-images qui garde l'information la plus importante pour chaque pixel. Ensuite, une série de couches convolutives transposées sont mises en place pour le décodeur afin de générer la carte de normale. Remarquons que l'extracteur des caractéristiques est appliqué sur chaque image de façon indépendante. Le schèma global de l'architecture est présenté sur la figure

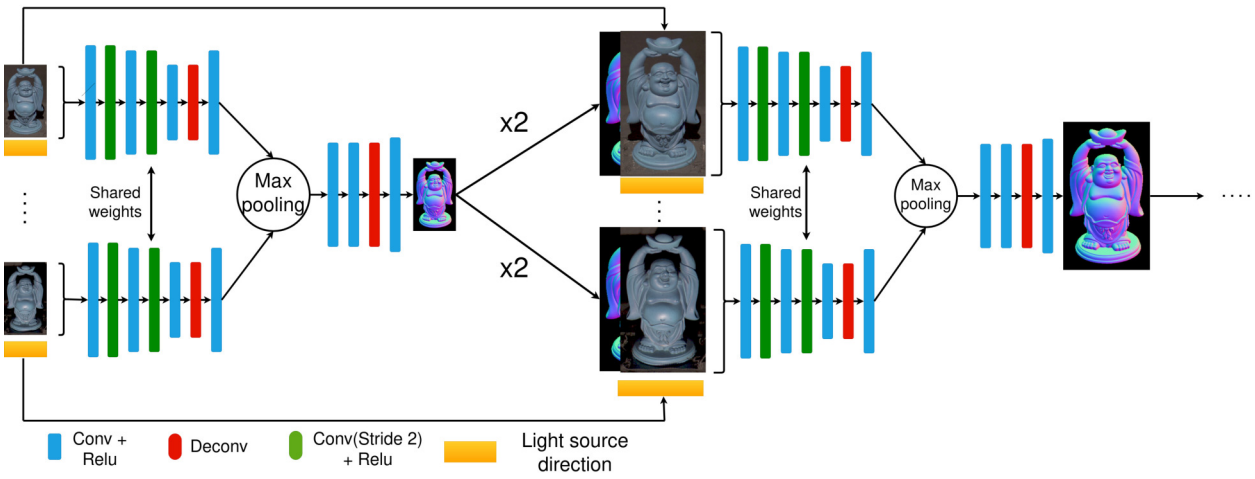


FIGURE 4.1 – Architecture de notre méthode multi-échelles MS-PS.

4.1.1 Processus de raffinage

Afin d'illustrer le processus de raffinage mentionné précédemment, nous présentons sur la figure 4.2 un exemple de reconstruction des normales au fur et à mesure des échelles.

Comme nous pouvons le voir les basses fréquences sont reconstruites avec les premières échelles de traitement et les détails hautes fréquences apparaissent au fur et à mesure du traitement des échelles. Les détails sur la surface du lapin commencent à être visibles sur la deuxième échelle présentée sur la figure 4.2 mais il faut attendre la dernière échelle pour que ceux-ci soit clairement visibles. Ainsi, notre architecture multi-échelles permet d'analyser des détails de finesse quelconque sur la surface quelle que soit la résolution des images.

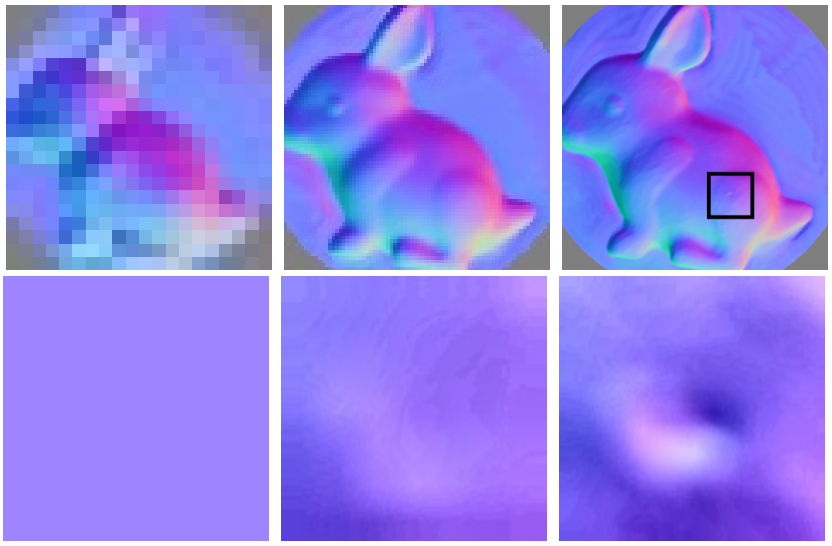


FIGURE 4.2 – Illustration du processus de raffinement de la carte des normales au fur et à mesure des échelles. On peut voir que le caractère multi-échelles de notre réseau permet de capturer l'ensemble des détails de l'objet. L'image provient de la base de données *DiLiGenT* [80].

Pour résumer, nous apprenons deux réseaux, un pour la première échelle et un second pour les autres échelles. Ces deux réseaux ont donc la même architecture mais des poids différents. Par conséquent, dans notre cas, le nombre d'échelles appliquées augmente en fonction de la résolution de l'image d'entrée mais la taille du réseau global est la même, quelle que soit la résolution. En effet, nous apprenons le passage d'une résolution à celle du-dessus indépendamment de la résolution des images d'entraînement. Par conséquent, les poids appris sont utilisables quelle que soit la résolution d'entrée.

D'un point de vue de la taille de l'architecture globale, celle-ci reste légère ne comptant que 4.4 millions de paramètres, malgré l'aspect multi-échelles et encodeur-décodeur qui peuvent être gourmands en paramètres.

4.1.2 Pré-traitement des données

Étant donné que nous travaillons dans un contexte de stéréophotométrie calibrée, nous avons l'information de la direction lumineuse pour chaque image d'entrée donnée au réseau.

Cette direction lumineuse est mise elle-même sous forme d'une image concaténée aux canaux des images RGB en entrée dans la première échelle du réseau. Pour les échelles suivantes, qui prennent en plus une estimation de la carte des normales, nous concaténons donc les images RGB, l'image formée avec la direction lumineuse et l'estimation de la carte des normales, pour définir l'entrée des modules de raffinement.

De plus, les images RGB sont normalisées par les intensités lumineuses connues afin que seuls la direction lumineuse impacte les valeurs RGB des pixels et que l'intensité de la lumière n'ait pas d'incidence sur celles-ci.

4.1.3 Entraînement

L'entraînement d'un réseau multi-échelles comme le nôtre est relativement simple étant donné qu'il est possible d'entraîner à basse résolution et ensuite généraliser sur de très hautes résolutions sans perte de performance car le réseau apprend le passage d'une échelle à une autre et non pas une résolution fixe. Le système informatique nécessaire est ainsi relativement peu puissant.

En pratique, nous avons entraîné l'architecture multi-échelles proposée en ne prenant que 32 images (i.e. 32 directions lumineuses) de taille 128×128 pixels par objet. L'entraînement a pris quelques jours sur une seule carte graphique Nvidia GeForce GTX 1080 Ti avec une taille de batch fixée à 3 (i.e. le maximum que nous pouvons utiliser sur notre carte graphique). Le temps d'inférence quant à lui dépend du nombre d'images en entrée et de leur résolution. Par exemple, en prenant 100 images 256×256 pixels, cela prend environ 1,6 secondes pour notre méthode multi-échelles sur une GeForce GTX 1080 Ti. Ce temps d'inférence évolue de manière linéaire avec le nombre d'images en entrée et est multiplié par un facteur quatre lorsque la résolution des images est multipliée par deux. Lors de l'inférence, le nombre d'échelles est ajusté pour obtenir la résolution voulue comme par exemple sept pour des images de taille 1024×1024.

Finalement, l'entraînement se fait en utilisant la fonction de perte de similarité cosinus qui correspond à la mesure angulaire entre la normale vérité terrain et la normale générée en chaque point. Cette métrique est définie de la façon suivante :

$$l_{stage} = 1 - \sum_{ij} N_{ij}^{\top} \hat{N}_{ij}, \tag{4.1}$$

où $\hat{N}_{ij}$ est la normale estimée au pixel (i, j) et N_{ij} celle de la vérité terrain.

Pour permettre d'entraîner chaque sous réseau la fonction de perte globale est donnée par :

$$l_{normal} = \sum_{i=0}^{N} l_{stage_i} \tag{4.2}$$

où N est le nombre d'échelles utilisé pendant l'entraînement, dans notre cas quatre, et $i = \{1, ..., N\}$ est une échelle du réseau.

4.2 Résultats

Dans cette section, nous allons comparer à la fois quantitativement et qualitativement notre méthode avec les méthodes de l'état-de-l'art dans un cadre calibré. De plus, afin de justifier notre choix d'une architecture multi-échelles, nous allons comparer l'approche multi-échelles et l'approche mono-échelle équivalente.

Pour les bases de données, nous allons reprendre les notations données dans le chapitre 1 : DB1 correspond à *Blobby + Sculpture*, DB2 à notre première base de données générée (1 100 matériaux réels, 56 objets, 3 000 formes blob, faisceaux parallèles, environnement ambiant noir).

4.2.1 Mono vs multi-échelles

Dans un premier temps, nous comparons quantitativement l'impact de l'architecture multi-échelles avec la même architecture mono-échelle sur $DiLiGenT10^2$ [80]. Nous traitons les images à leur résolution d'origine (1024×1024 pixels), ce qui nécessite 7 échelles dans l'architecture multi-échelles. Pour l'architecture mono-échelle, le réseau prend directement en entrée les images avec leur résolution initiale sans passer par une méthode par patch. Celle-ci sera abordée dans la partie 4.2.4. Nous montrons dans le tableau 4.1 la différence entre les approches mono-échelle et multi-échelles, lorsqu'elles sont toutes deux entraînées sur le jeu de données DB1 (*Sculpture* + *Blobby*). Comme nous pouvons le voir, un gain significatif de $9,3\%$ est observé avec l'architecture multi-échelles. Le gain le plus visible est sur les objets de forme sphérique construits dans des matériaux anisotropes (coin supérieur droit dans le tableau 4.1c). Un exemple qualitatif sur un matériau anisotrope est également affiché sur la figure 4.3 sur l'objet "Golf" en cuivre. Les objets en acrylique (matériau translucide) sont également bien mieux gérés par l'architecture multi-échelles.
De même, l'architecture multi-échelles permet de sensiblement améliorer les résultats sur la base de données *DiLiGenT* [87] (voir le tableau 4.2).

mean: 17.77 median: 17.33

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	16.0	6.1	12.67	9.7	14.33	10.67	24.33	20.67	24.0	29.67
GOLF	15.0	14.0	14.0	9.9	14.0	12.0	22.67	17.67	21.0	31.33
SPIKE	16.33	14.67	11.33	7.53	9.97	14.0	26.33	13.67	26.33	35.67
NUT	19.0	14.0	19.33	6.73	21.0	11.67	23.67	18.0	20.0	26.0
SQUARE	21.67	20.33	21.0	18.0	23.0	11.0	23.0	15.67	15.33	25.0
PENTAGON	20.67	11.33	19.0	9.1	21.0	14.0	21.67	17.33	18.67	19.0
HEXAGON	18.67	12.67	16.33	8.3	20.67	11.33	24.0	21.33	24.33	26.33
PROPELLER	19.33	15.0	21.67	9.17	19.67	13.0	15.0	13.0	13.0	10.67
TURBINE	30.33	17.33	33.0	14.67	32.33	25.0	28.33	25.67	25.0	24.67
BUNNY	15.0	10.33	17.33	7.6	17.33	11.0	13.0	10.33	11.67	12.67

(a) *Mono* (DB1)

mean: 16.27 median: 16.0

POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
15.33	4.83	9.83	9.4	11.0	7.43	22.0	18.67	21.67	25.67
14.67	11.67	12.67	8.7	13.0	9.53	18.67	14.33	17.67	28.67
15.0	12.67	10.33	7.03	9.83	11.0	23.33	11.67	23.67	31.33
18.0	11.33	18.33	5.83	17.33	8.93	21.67	16.0	18.0	23.0
20.67	19.0	20.67	17.33	21.0	8.63	20.33	11.0	12.67	21.33
21.33	10.33	20.67	9.97	20.0	12.33	20.0	15.33	17.33	17.33
18.67	11.0	16.67	8.13	18.33	9.27	22.0	17.67	22.33	22.0
19.33	13.0	22.33	8.3	18.33	10.33	14.67	12.33	12.33	11.0
32.33	14.0	37.0	11.0	32.0	20.33	27.33	24.67	24.67	24.67
16.0	9.57	18.0	6.97	17.0	9.5	13.0	9.77	11.0	13.0

(b) *Multi* (DB1)

mean: -1.5 median: -1.33

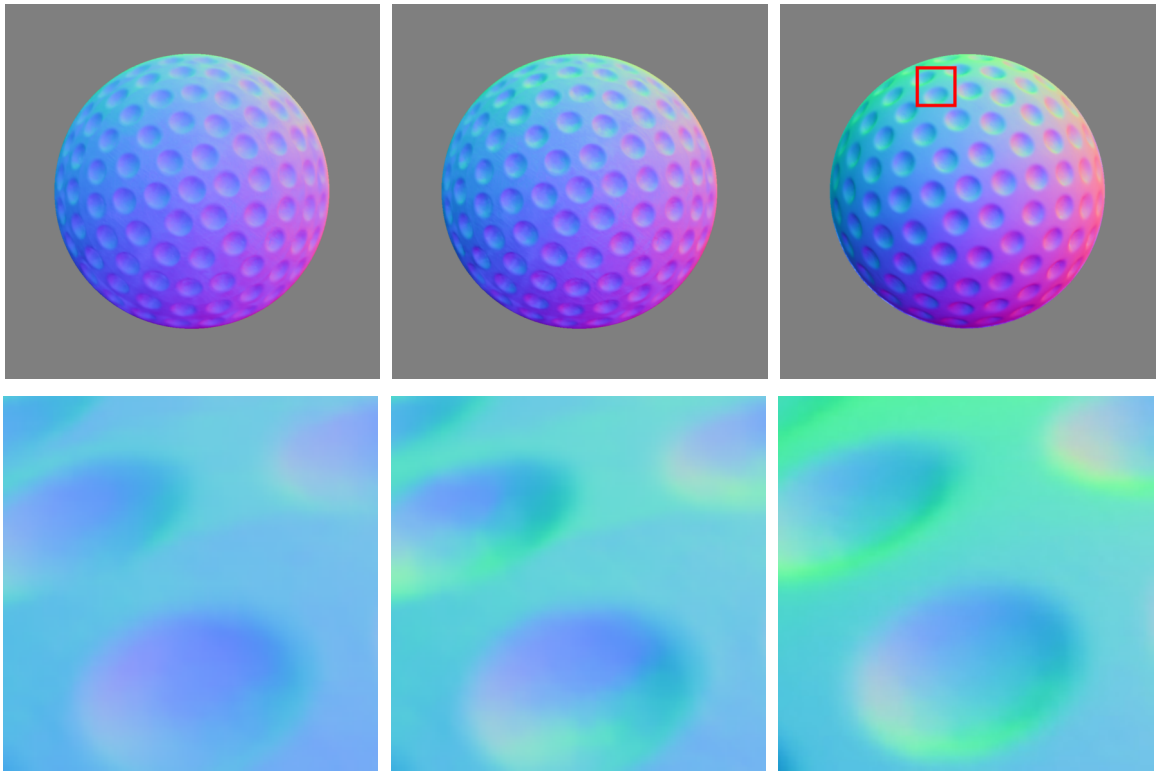
	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	-0.67	-1.27	-2.84	-0.3	-3.33	-3.24	-2.33	-2.0	-2.33	-4.0
GOLF	-0.33	-2.33	-1.33	-1.2	-1.0	-2.47	-4.0	-3.34	-3.33	-2.66
SPIKE	-1.33	-2.0	-1.0	-0.5	-0.14	-3.0	-3.0	-2.0	-2.66	-4.34
NUT	-1.0	-2.67	-1.0	-0.9	-3.67	-2.74	-2.0	-2.0	-2.0	-3.0
SQUARE	-1.0	-1.33	-0.33	-0.67	-2.0	-2.37	-2.67	-4.67	-2.66	-3.67
PENTAGON	0.66	-1.0	1.67	0.87	-1.0	-1.67	-1.67	-2.0	-1.34	-1.67
HEXAGON	0.0	-1.67	0.34	-0.17	-2.34	-2.06	-2.0	-3.66	-2.0	-4.33
PROPELLER	0.0	-2.0	0.66	-0.87	-1.34	-2.67	-0.33	-0.67	-0.67	0.33
TURBINE	2.0	-3.33	4.0	-3.67	-0.33	-4.67	-1.0	-1.0	-0.33	0.0
BUNNY	1.0	-0.76	0.67	-0.63	-0.33	-1.5	0.0	-0.56	-0.67	0.33

(c) *Multi* (DB1) - *Mono* (DB1)

TABLE 4.1 – Erreurs angulaires moyennes (en degrés) sur la base de données $DiLiGenT10^2$ [80]. L'architecture mono-échelle ainsi que la multi-échelles ont été entraînées sur la base de données DB1 (*Sculpture* + *Blobby*). L'architecture multi-échelles améliore d'environ 9% les performances comparativement à l'architecture mono-échelle. Ce gain est principalement visible dans le coin supérieur droit du tableau (objets sphériques avec une réflectance anisotrope).

	ball	bear	buddha	cat	cow	goblet	harvest	pot1	pot2	reading	moyenne
Mono (DB1)	2.63	6.66	8.27	4.47	4.77	8.24	12.78	6.00	5.38	9.68	6.88
Multi (DB1)	1.60	7.82	7.55	4.33	4.18	7.85	12.36	5.22	5.36	9.04	6.54

TABLE 4.2 – Erreurs angulaires en degrés sur la base de données *DiLiGenT* [87]. L'architecture mono-échelle ainsi que la multi-échelles ont été entraînées sur la base de données DB1 (*Sculpture* + *Blobby*).



Mono-échelle (DS1). Multi-échelles (DS1). Vérité terrain.

FIGURE 4.3 – Résultats qualitatifs des architectures mono-échelle et multi-échelles (toutes les deux entraînées sur la base de données DB1) sur la balle de golf en cuivre de *DiLiGenT10*^{2} [80]. L'architecture multi-échelles donne un résultat bien plus net, notamment autour des enfoncements.

4.2.2 Performances de notre nouvelle base de données d'entraînement

Dans cette section, la base de données avec un nombre beaucoup plus important de matériaux et de géométries (DB2) a été ajoutée lors de l'entraînement des réseaux de neurones mono-échelle et multi-échelles dans le but de tester son impact sur les performances.

Comme nous pouvons le voir sur le tableau 4.3, l'ajout de cette base de données aux bases de données préexistantes (DB1) améliore significativement les performances sur *DiLiGenT*10^2 pour les deux architectures. L'amélioration est particulièrement visible sur les matériaux les plus complexes à traiter, tels que les métaux (matériau spéculaire) ou l'acrylique (matériaux translucides). L'augmentation des performances est beaucoup plus importante dans le cas de l'architecture multi-échelles avec un gain de près de 30% pour le multi-échelles contre 13.5% pour le mono-échelle.

Pour *DiLiGenT*, seule l'architecture multi-échelles améliore ses performances (voir le tableau 4.4). Les matériaux présents dans la base de données *DiLiGenT* sont pour la majorité diffus. Par conséquent, l'apport d'une grande diversité de matériaux lors de l'entraînement est donc moins perceptible.

mean: 17.77 median: 17.33

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	16.0	6.1	12.67	9.7	14.33	10.67	24.33	20.67	24.0	29.67
GOLF	15.0	14.0	14.0	9.9	14.0	12.0	22.67	17.67	21.0	31.33
SPIKE	16.33	14.67	11.33	7.53	9.97	14.0	26.33	13.67	26.33	35.67
NUT	19.0	14.0	19.33	6.73	21.0	11.67	23.67	18.0	20.0	26.0
SQUARE	21.67	20.33	21.0	18.0	23.0	11.0	23.0	15.67	15.33	25.0
PENTAGON	20.67	11.33	19.0	9.1	21.0	14.0	21.67	17.33	18.67	19.0
HEXAGON	18.67	12.67	16.33	8.3	20.67	11.33	24.0	21.33	24.33	26.33
PROPELLER	19.33	15.0	21.67	9.17	19.67	13.0	15.0	13.0	13.0	10.67
TURBINE	30.33	17.33	33.0	14.67	32.33	25.0	28.33	25.67	25.0	24.67
BUNNY	15.0	10.33	17.33	7.6	17.33	11.0	13.0	10.33	11.67	12.67

(a) Mono-échelle (DB1)

mean: 16.27 median: 16.0

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	15.33	4.83	9.83	9.4	11.0	7.43	22.0	18.67	21.67	25.67
GOLF	14.67	11.67	12.67	8.7	13.0	9.53	18.67	14.33	17.67	28.67
SPIKE	15.0	12.67	10.33	7.03	9.83	11.0	23.33	11.67	23.67	31.33
NUT	18.0	11.33	18.33	5.83	17.33	8.93	21.67	16.0	18.0	23.0
SQUARE	20.67	19.0	20.67	17.33	21.0	8.63	20.33	11.0	12.67	21.33
PENTAGON	21.33	10.33	20.67	9.97	20.0	12.33	20.0	15.33	17.33	17.33
HEXAGON	18.67	11.0	16.67	8.13	18.33	9.27	22.0	17.67	22.33	22.0
PROPELLER	19.33	13.0	22.33	8.3	18.33	10.33	14.67	12.33	12.33	11.0
TURBINE	32.33	14.0	37.0	11.0	32.0	20.33	27.33	24.67	24.67	24.67
BUNNY	16.0	9.57	18.0	6.97	17.0	9.5	13.0	9.77	11.0	13.0

(b) *Multi-échelles* (DB1)

mean: 15.35 median: 14.05

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	5.0	8.4	7.6	9.7	6.0	18.0	16.0	22.0	22.0
GOLF	11.0	7.1	10.0	6.1	10.0	6.9	13.0	9.8	14.0	21.0
SPIKE	11.0	7.8	10.0	7.2	8.6	8.0	20.0	11.0	20.0	30.0
NUT	19.0	11.0	18.0	7.7	15.0	11.0	19.0	14.0	17.0	26.0
SQUARE	19.0	10.0	19.0	11.0	15.0	8.7	17.0	9.5	13.0	18.0
PENTAGON	22.0	12.0	21.0	10.0	18.0	13.0	17.0	14.0	16.0	22.0
HEXAGON	18.0	9.8	17.0	8.8	14.0	8.8	18.0	12.0	18.0	23.0
PROPELLER	23.0	12.0	24.0	9.6	19.0	12.0	14.0	12.0	13.0	14.0
TURBINE	36.0	18.0	38.0	14.0	33.0	22.0	29.0	25.0	27.0	26.0
BUNNY	18.0	11.0	19.0	9.2	15.0	11.0	14.0	12.0	12.0	14.0

(c) Mono-échelle (DB1+DB2)

mean: 11.33 median: 9.98

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	3.4	8.7	5.2	8.4	4.3	8.5	12.0	14.0	8.6
GOLF	10.0	7.3	9.8	5.8	10.0	6.87	7.9	7.7	9.8	12.0
SPIKE	12.0	8.8	9.9	6.3	8.5	7.9	12.0	7.6	12.0	17.0
NUT	14.0	8.9	15.0	5.8	10.0	5.8	9.2	7.6	8.2	16.0
SQUARE	18.0	11.0	17.0	8.2	14.0	5.5	12.0	7.2	7.9	11.0
PENTAGON	18.0	8.4	17.0	8.0	17.0	9.4	11.0	9.4	13.0	20.0
HEXAGON	16.0	7.5	15.0	6.1	13.0	7.1	11.0	8.1	11.0	20.0
PROPELLER	13.0	8.9	11.0	7.9	16.0	9.7	11.0	8.4	8.2	19.0
TURBINE	21.0	12.0	24.0	11.0	18.0	15.0	23.0	16.0	18.0	22.0
BUNNY	17.0	8.2	16.0	6.5	12.0	8.3	8.6	7.3	8.0	18.0

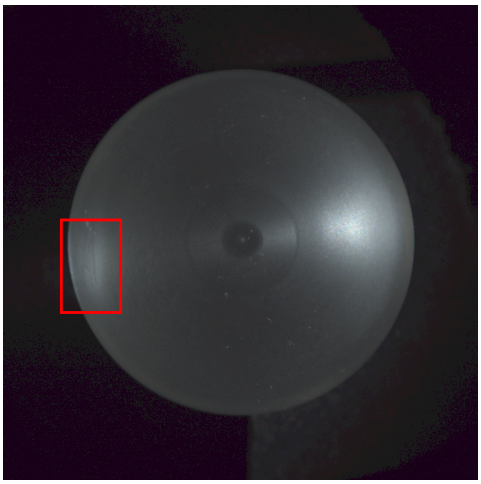
(d) Multi-échelles (DB1+DB2)

TABLE 4.3 – Comparaison des performances sur la base de données $DiLiGenT10^2$ [80] en erreur angulaire en (degrés) avec ou sans l'ajout de la base de données DB2 lors de l'entraînement des architectures mono-échelle et multi-échelles. Pour les deux architectures, l'ajout de DB2 permet un gain significatif des résultats.

L'intérêt principal d'une architecture multi-échelles tient dans sa capacité à extraire des caractéristiques géométriques locales mais aussi globales des objets imagés. Dans le contexte de matériaux complexes tels que les métaux ou les matériaux translucides, cette capacité est primordiale. Par exemple, dans le cas d'une balle en acrylique, illustré sur la figure 4.4, la lumière peut venir d'un côté de l'objet et ressortir de l'autre. Un réseau de reconstruction de normales qui ne verrait l'information que localement ne pourrait donc pas déterminer que la lumière a traversé la matière. Il lui est alors impossible de reconstruire convenablement une carte de normales cohérente. À l'inverse, notre réseau multi-échelles bénéficie de l'analyse des premières échelles pour synthétiser l'information globale et la transmettre aux échelles suivantes. Cette problématique de matériaux translucide est illustrée sur les figures 4.4 et 4.5. Notre architecture multi-échelles est ainsi plus apte à gérer des matériaux complexes, ce qui explique une telle différence de performances entre les deux types d'architectures.

	ball	bear	buddha	cat	cow	goblet	harvest	pot1	pot2	reading	average
Mono (DB1)	2.63	6.66	8.27	4.47	4.77	8.24	12.78	6.00	5.38	9.68	6.88
Multi (DB1)	1.60	7.82	7.55	4.33	4.18	7.85	12.36	5.22	5.36	9.04	6.54
Multi (DB1+DB2)	2.05	4.24	7.03	3.9	4.00	7.57	11.01	4.94	5.22	8.47	5.84

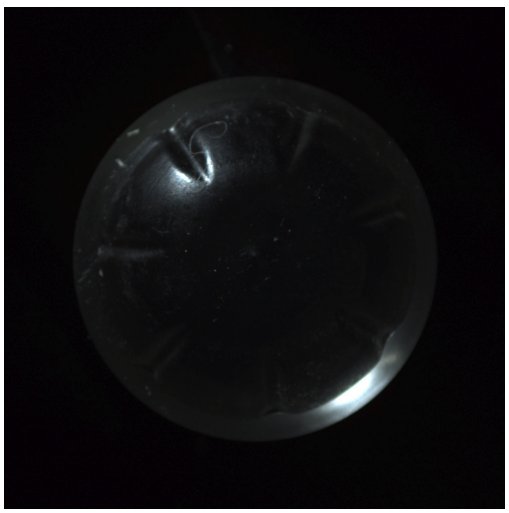
TABLE 4.4 – Comparaison des performances sur la base de données *DiLiGenT* [87] en erreur angulaire (en degrés) moyenne avec ou sans l'ajout de la base de données DB2 lors de l'entraînement des architectures mono-échelle et multi-échelles.



Balle en acrylique.

FIGURE 4.4 – Image d'une balle en acrylique éclairée par la droite. La lumière passe à travers la balle pour ressortir du côté opposé (rectangle rouge).

Plastique.



Acrylique.

FIGURE 4.5 – Le même objet, fabriqué avec un matériau diffus (ABS) et un matériau translucide (acrylique), est éclairé depuis la même direction venant d'en haut à gauche. La zone en bas à droite, qui est dans l'ombre dans le cas diffus, paraît beaucoup plus brillante sur l'objet translucide car la lumière a "traversé" l'objet. Ces cas sont extrêmement difficiles à analyser pour un réseau de reconstruction de normales.

4.2.3 Comparaison à l'état-de-l'art

Afin d'évaluer les performances de notre modèle, nous nous sommes comparé à l'état-de-l'art des méthodes de stéréophotométrie calibrée. Ainsi, nous avons comparé notre méthode aux méthodes L2 [100], GPS-NET [104], CHR-PSN [46], PS-Transformer [41], CNN-PS [45], MT-PS-CNN [14], OB-CNN [37] et PX-NET [65].

Comme on peut le voir dans le tableau 4.5, nous améliorons les résultats en moyenne de 0.5 degrés sur la base de données *DiLiGenT*. Quand on regarde avec plus de détails, les résultats par objet, on voit que l'on obtient les meilleurs ou les second meilleurs pour 7 objets sur 10 de la base de données. On voit notamment une grande amélioration sur les objets *"reading"* où nous gagnons quasiment 2 degrés et l'objet *"cat"* avec une erreur de 3.9 degrés contre 4.1 degrés pour le second. Cependant, pour les autres objets, nous obtenons malgré tout des résultats très proches de l'état-de-l'art.

	ball	bear	buddha	cat	cow	goblet	harvest	pot1	pot2	reading	average
L2 (Baseline)[100]	4.10	8.39	14.92	8.41	25.60	18.5	30.62	8.89	14.65	19.80	15.39
GPS-NET [104]	2.92	5.07	7.77	5.42	6.14	9.00	15.14	6.04	7.01	13.58	7.81
CHR-PSN [46]	2.26	6.35	7.15	5.97	6.05	8.32	15.32	7.04	6.76	12.52	7.77
PS-transformer (10 images) [41]	3.27	4.88	8.65	5.34	6.54	9.28	14.41	6.06	6.97	11.24	7.66
MT-PS-CNN [14]	2.29	5.87	**6.92**	5.79	6.89	**6.85**	**7.88**	11.94	7.48	13.71	7.56
PS-FCN [15]	2.67	7.72	7.52	4.75	6.72	7.84	12.39	6.17	7.15	10.92	7.39
CNN-PS [45]	2.2	4.6	7.9	4.1	8.0	7.3	14.0	5.4	6.0	12.6	7.2
OB-Cnn [37]	2.49	3.59	7.23	4.69	4.89	6.89	12.79	5.10	**4.98**	11.08	6.37
PX-NET [65]	2.03	**3.58**	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.28
Notre méthode	2.05	4.24	7.03	**3.9**	**4.00**	7.57	11.01	**4.94**	5.22	**8.47**	**5.84**

TABLE 4.5 – Erreurs angulaires en degrés sur la base de données *DiLiGenT* [87]. Les meilleurs résultats sont en gras et les second meilleurs sont soulignés. La ligne en bleue est notre méthode multi-échelles. On peut voir que nous obtenons les meilleurs résultats en moyenne.

De plus, la combinaison de la nouvelle architecture multi-échelles et de la nouvelle base de données d'entraînement permet d'atteindre une erreur moyenne de 11.33 degrés sur *DiLiGenT10*^{2}. Il faut comparer ce résultat à celui de 15.78 degrés obtenu par CNN-PS [45] et aux 16.21 degrés de PS-FCN [18] (tableau 4.9), qui étaient les deux méthodes les plus performantes jusqu'à présent sur *DiLiGenT10*^{2}. De plus, en comparant nos résultats avec toutes les méthodes de l'art disponibles [17, 18, 45, 86, 89, 102, 104, 108], nous avons constaté que notre méthode est la plus performante sur 73% des objets de ce benchmark, comme indiqué aux tableau 4.7.

De nouveau, notre méthode multi-échelles est plus performante que les méthodes de l'état-de-l'art existantes sur *DiLiGenT-Pi* [95]. Nous obtenons une amélioration significative de près de 5% (voir le tableau 4.8). De plus, elle permet d'obtenir les meilleurs résultats sur 45% des objets de cette base de données. Pour rappel, la particularité de cette base de données est qu'elle est principalement composée d'objets plats avec une surface fortement spéculaire. Nous pouvons donc en conclure que notre méthode reste robuste sur ce type d'objets.

Pour résumer, notre méthode multi-échelles, entraînée sur notre nouvelle base de données permet d'obtenir les meilleurs performances comparativement à toutes les autres méthodes calibrées de l'état-de-l'art disponible et sur les trois bases de données de test.

mean: 16.21 median: 15.1

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	13.0	4.3	12.0	8.6	11.0	5.3	19.0	17.0	21.0	24.0
GOLF	15.0	9.7	15.0	8.7	12.0	9.0	16.0	12.0	15.0	27.0
SPIKE	14.0	10.0	11.0	7.3	9.9	8.9	22.0	10.0	22.0	31.0
NUT	19.0	12.0	21.0	10.0	17.0	9.8	20.0	15.0	17.0	24.0
SQUARE	19.0	14.0	20.0	13.0	18.0	10.0	19.0	9.0	11.0	19.0
PENTAGON	21.0	11.0	22.0	13.0	19.0	13.0	17.0	13.0	17.0	21.0
HEXAGON	19.0	11.0	19.0	11.0	17.0	9.5	21.0	14.0	20.0	22.0
PROPELLER	24.0	13.0	28.0	11.0	18.0	12.0	12.0	9.9	9.9	14.0
TURBINE	37.0	19.0	38.0	17.0	34.0	25.0	25.0	23.0	23.0	28.0
BUNNY	19.0	9.9	21.0	9.1	17.0	11.0	12.0	8.8	10.0	14.0

(a) PS-FCN [18] (DS1)

mean: 15.78 median: 13.99

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	5.1	6.4	4.2	4.5	6.9	7.3	16.0	14.0	16.0	19.0
GOLF	14.0	8.0	12.0	6.8	14.0	9.4	12.0	9.2	13.0	22.0
SPIKE	11.0	9.4	11.0	11.0	12.0	9.5	14.0	8.3	16.0	28.0
NUT	20.0	8.8	19.0	6.9	17.0	8.0	16.0	13.0	14.0	22.0
SQUARE	21.0	8.1	22.0	6.7	19.0	8.1	13.0	4.9	7.9	18.0
PENTAGON	26.0	9.5	26.0	9.8	22.0	9.6	15.0	13.0	15.0	23.0
HEXAGON	18.0	7.5	19.0	7.2	17.0	28.0	18.0	10.0	17.0	21.0
PROPELLER	28.0	12.0	35.0	8.4	23.0	11.0	16.0	9.6	9.8	17.0
TURBINE	54.0	20.0	51.0	16.0	39.0	21.0	25.0	22.0	21.0	32.0
BUNNY	24.0	11.0	27.0	7.8	21.0	9.1	12.0	7.7	12.0	14.0

(b) CNN-PS [45] (DS1)

mean: 11.33 median: 9.98

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	3.4	8.7	5.2	8.4	4.3	8.5	12.0	14.0	8.6
GOLF	10.0	7.3	9.8	5.8	10.0	6.87	7.9	7.7	9.8	12.0
SPIKE	12.0	8.8	9.9	6.3	8.5	7.9	12.0	7.6	12.0	17.0
NUT	14.0	8.9	15.0	5.8	10.0	5.8	9.2	7.6	8.2	16.0
SQUARE	18.0	11.0	17.0	8.2	14.0	5.5	12.0	7.2	7.9	11.0
PENTAGON	18.0	8.4	17.0	8.0	17.0	9.4	11.0	9.4	13.0	20.0
HEXAGON	16.0	7.5	15.0	6.1	13.0	7.1	11.0	8.1	11.0	20.0
PROPELLER	13.0	8.9	11.0	7.9	16.0	9.7	11.0	8.4	8.2	19.0
TURBINE	21.0	12.0	24.0	11.0	18.0	15.0	23.0	16.0	18.0	22.0
BUNNY	17.0	8.2	16.0	6.5	12.0	8.3	8.6	7.3	8.0	18.0

(c) *Multi-échelles* (DS1+DS2)

TABLE 4.6 – Erreurs angulaires moyennes en degrés sur la base de données *DiLiGenT10*^{2}, avec les résultats de CNN-PS [45] et de PS-FCN [18] pour la comparaison.

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	3.4	8.7	5.2	8.4	4.3	**8.5**	12.0	14.0	**8.6**
GOLF	**10.0**	7.3	**9.8**	5.8	**10.0**	**6.87**	**7.9**	**7.7**	**9.8**	**12.0**
SPIKE	12.0	8.8	**9.9**	**6.3**	**8.5**	**7.9**	12.0	**7.6**	12.0	**17.0**
NUT	**14.0**	8.9	15.0	**5.8**	**10.0**	**5.8**	**9.2**	**7.6**	**8.2**	**16.0**
SQUARE	18.0	11.0	**17.0**	8.2	**14.0**	**5.5**	12.0	7.2	7.9	**11.0**
PENTAGON	**18.0**	**8.4**	**17.0**	**8.0**	**17.0**	**9.4**	**11.0**	**9.4**	**13.0**	20.0
HEXAGON	**16.0**	7.5	**15.0**	**6.1**	**13.0**	**7.1**	**11.0**	**8.1**	**11.0**	20.0
PROPELLER	**13.0**	**8.9**	**11.0**	**7.9**	**16.0**	**9.7**	**11.0**	**8.4**	**8.2**	19.0
TURBINE	**21.0**	**12.0**	**24.0**	**11.0**	**18.0**	**15.0**	**23.0**	**16.0**	**18.0**	**22.0**
BUNNY	**17.0**	**8.2**	**16.0**	**6.5**	**12.0**	**8.3**	**8.6**	**7.3**	**8.0**	18.0

TABLE 4.7 – Erreur angulaire moyenne obtenue par le meilleur modèle parmi [17, 18, 45, 86, 89, 102, 104, 108] et l'architecture multi-échelles proposée, sur les 100 objets de *DiLiGenT10*^{2} [80]. Les cases en vert indiquent quand l'architecture proposée, combinée au nouvel ensemble de données, donne les meilleurs résultats.

	Astro	Bagua-R	Bagua-T	Bear	Bird	Cloud-R	Cloud-T	Crab	Fish	Flower	
	Lion-R	Lion-T	Lions	Lotus-R	Lotus-T	Lung	Ocean	Panda-R	Panda-T	Para	moyenne
	Queen	Rhino	Sail	Ship	Sun	TV	Taichi	Tree	Wave	Whale	
NormAttention-PSN [47]	7.2	12.0	16.5	7.4	6.9	13.4	17.3	**4.4**	4.4	**4.6**	
	16.4	21.0	**4.4**	**10.8**	13.7	7.8	5.8	**13.9**	16.6	4.2	9.2
	4.9	5.1	5.2	4.9	**5.6**	**7.6**	9.7	9.6	6.1	8.7	
PS-FCN [18]	7.2	13.0	16.8	7.4	7.2	14.3	17.8	5.3	4.6	**4.6**	
	18.4	21.2	4.5	11.8	13.6	9.7	5.8	14.8	17.2	4.7	9.85
	4.7	5.3	**5.1**	6.1	6.7	8.0	10.2	10.6	6.8	12.2	
CNN-PS [45]	6.0	12.2	16.4	7.4	**6.8**	14.6	17.2	4.5	**4.2**	4.7	
	15.8	20.3	4.7	10.9	13.5	**5.7**	**4.6**	14.2	16.6	3.9	9.16
	5.4	**4.9**	5.2	**4.9**	5.8	8.3	**7.8**	11.3	5.3	11.6	
Méthode proposée	**5.96**	**11.32**	**15.1**	**6.9**	7.69	**13.28**	**14.74**	4.58	4.68	5.43	
	14.37	**15.71**	5.5	11.92	**12.8**	7.51	4.97	14.75	**14.72**	**4.09**	8.78
	6.37	5.18	5.26	5.14	6.46	8.63	9.91	**8.22**	**5.29**	**7.09**	

TABLE 4.8 – Erreurs angulaires moyennes en degrés sur la base de données *DiLiGenT-Pi*. La méthode multi-échelles proposée donne les meilleurs résultats comparativement aux méthodes de l'état-de-l'art (CNN-PS [45], PS-FCN [18] et NormAttention-PSN [47] qui donnent les meilleurs performances de l'état-de-l'art). Les meilleurs résultats sur chaque objet sont indiqués en gras.

4.2.4 Étude d'ablation : inférence *full-scale* vs *patch-based*

L'objectif de cette section est de montrer l'intérêt de l'architecture multi-échelles à traiter les images à leur résolution initial originale sans devoir passer par une méthode par patch, contrairement à la partie 4.2.1 où le réseau mono-échelle prenait directement les images à leur résolution initiale. Ainsi, nous avons comparé ces deux techniques d'inférence avec notre architecture sur la base de données *DiLiGenT10*^{2}. Nous avons donc testé les deux méthodes suivantes :

— Méthode *full-scale* : nous utilisons notre réseau sans modification, avec autant d'échelles que nécessaire pour arriver à la résolution d'origine des images RGB. Par exemple, pour des images de taille 1001×1001 pixels, nous testons notre architecture avec sept échelles pour que l'image passe entièrement en inférence.

— Méthode *patch-based* : nous utilisons notre architecture d'entraînement avec trois échelles pour reconstruire des patchs de normales à la résolution utilisée pendant l'entraînement. Pour cela, les images RGB ont été découpées en patchs 128×128 pixels avec un chevauchement d'environ 50%, puis nous inférons avec le réseau d'entraînement. Avec le même exemple d'images de taille 1001×1001, les images sont découpées en patchs 128×128 pixels (i.e. la résolution d'entraînement) et inférées sur le réseau d'entraînement.

Nous avons comparé ces deux techniques sur la base de données *DiLiGenT10*^{2} [80]. Les résultats sont affichés dans le tableau 4.9.

Comme on peut le voir dans le tableau 4.9, l'ensemble des résultats a été amélioré avec la méthode *full-scale*. En effet, l'amélioration est d'autant plus marquée pour les objets (e.g. la turbine) et les matériaux difficiles tels que l'acrylique et le nylon par exemple. Pour l'acrylique, les erreurs angulaires sont entre 13 et 27 degrés en fonction de l'objet pour la méthode *patch-based* alors qu'on obtient des erreurs angulaires comprises entre 8 et 20 degrés avec notre méthode *full-scale* pour le même matériau. Nous obtenons donc une amélioration conséquente sur ce matériau difficile. De plus, en moyenne, sur la base de données, on améliore la reconstruction de quasiment 2 degrés, ce qui est non négligeable.

Des exemples visuels de reconstruction des cartes de normales sont affichés sur la figure 4.6 et sur la figure 4.8.

4.3 Limitations

Même si la combinaison de notre base de données DB2, présentée au chapitre 3 et de notre nouvelle architecture multi-échelles de type encodeur-décodeur basée sur un *CNN* améliore les résultats de l'état-de-l'art sur les matériaux non lambertiens, quelques lacunes persistent.

mean: 11.33

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	3.4	8.7	5.2	8.4	4.3	8.5	12.0	14.0	8.6
GOLF	10.0	7.3	9.8	5.8	10.0	6.87	7.9	7.7	9.8	12.0
SPIKE	12.0	8.8	9.9	6.3	8.5	7.9	12.0	7.6	12.0	17.0
NUT	14.0	8.9	15.0	5.8	10.0	5.8	9.2	7.6	8.2	16.0
SQUARE	18.0	11.0	17.0	8.2	14.0	5.5	12.0	7.2	7.9	11.0
PENTAGON	18.0	8.4	17.0	8.0	17.0	9.4	11.0	9.4	13.0	20.0
HEXAGON	16.0	7.5	15.0	6.1	13.0	7.1	11.0	8.1	11.0	20.0
PROPELLER	13.0	8.9	11.0	7.9	16.0	9.7	11.0	8.4	8.2	19.0
TURBINE	21.0	12.0	24.0	11.0	18.0	15.0	23.0	16.0	18.0	22.0
BUNNY	17.0	8.2	16.0	6.5	12.0	8.3	8.6	7.3	8.0	18.0

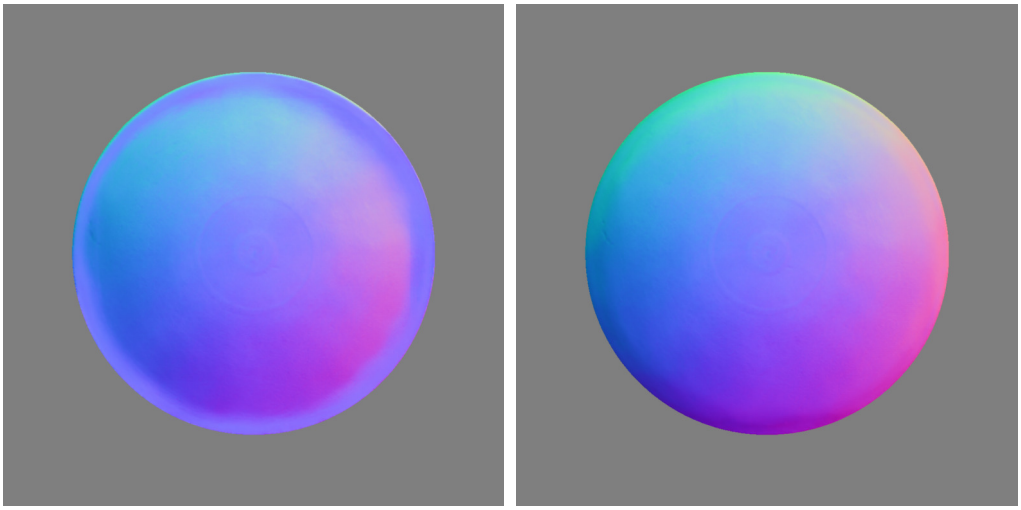
(a) Méthode *full-scale*.

mean: 13.16

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	10.0	3.3	8.8	5.3	8.6	4.5	12.0	13.0	16.0	20.0
GOLF	11.0	7.4	10.0	5.9	10.0	6.8	8.5	8.2	11.0	18.0
SPIKE	12.0	9.1	10.0	6.2	8.6	8.0	16.0	9.2	16.0	27.0
NUT	17.0	11.0	16.0	6.2	13.0	6.2	15.0	12.0	13.0	23.0
SQUARE	19.0	12.0	18.0	10.0	15.0	5.5	14.0	7.8	9.0	15.0
PENTAGON	20.0	8.7	20.0	8.1	18.0	9.4	13.0	10.0	15.0	19.0
HEXAGON	17.0	8.9	15.0	6.7	14.0	7.5	16.0	10.0	16.0	22.0
PROPELLER	16.0	9.9	14.0	8.3	19.0	10.0	12.0	9.9	9.6	13.0
TURBINE	30.0	13.0	34.0	11.0	29.0	17.0	25.0	22.0	22.0	24.0
BUNNY	18.0	8.4	15.0	6.9	14.0	8.4	9.5	7.6	8.7	15.0

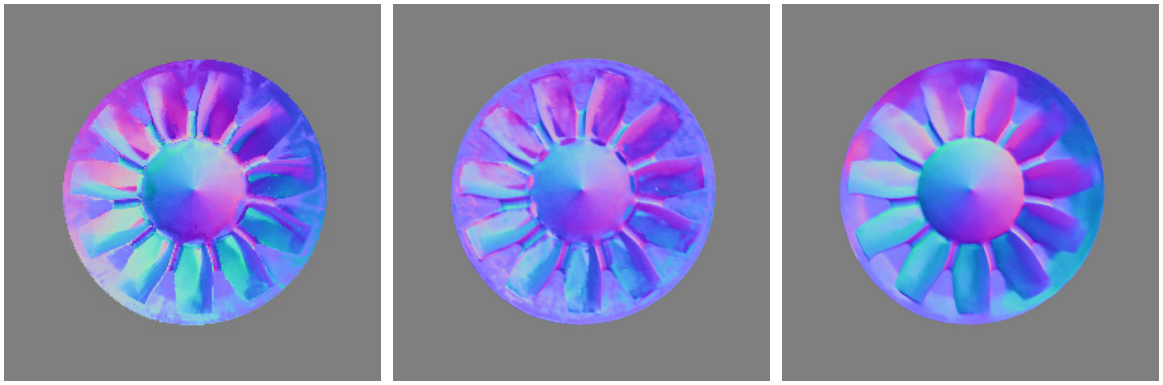
(b) Méthode *patch-based*.

TABLE 4.9 – Erreurs angulaires moyennes en degrés sur la base de données *DiLiGenT10*[2]. La méthode *full-scale* donne de meilleurs résultats.



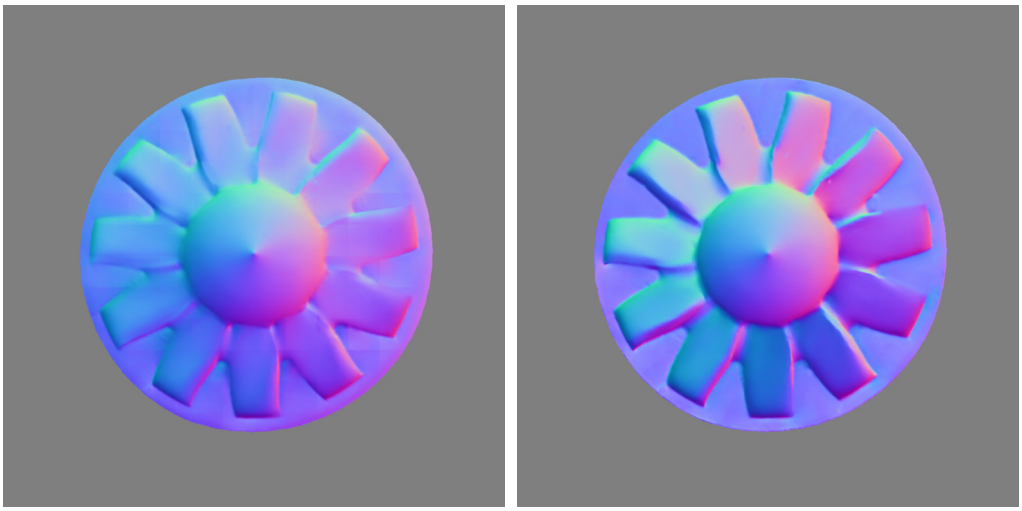
Méthode *patch-based*. Méthode *full-scale*.

FIGURE 4.6 – Comparaison entre la reconstruction de la carte de normale de la balle en acrylique de *DiLiGenT10*[2] avec une approche *patch-based* et avec une approche *full-scale*.



CNN-PS (DS1) Erreur : 32°. Mono (DS1) Erreur : 26°. Multi (DS1+DS2) Erreur : 22°.

FIGURE 4.7 – Résultats de CNN-PS, de notre architecture mono-échelle et de notre architecture multi-échelles sur la turbine en acrylique.



Méthode patch-based. Méthode full-scale.

FIGURE 4.8 – Comparaison entre la reconstruction de la carte de normale de la Turbine en ABS (Acrylonitrile butadiène styrène) de *DiLiGenT10*[2] avec une approche *patch-based* et avec une approche *full-scale*.

Par exemple, on remarque que les bords et contours des objets translucides ne sont pas forcément bien prédits, comme illustré sur la figure 4.7. En effet, comme déjà mentionné précédemment dans l'étude d'ablation sur la figure 4.4, la lumière vient du côté opposé de la partie brillante de la turbine en arcylique. La forme de l'objet est donc difficile à reconstruire. Bien que notre approche multi-échelles gère mieux cette anisotropie que l'approche mono-échelle ou les méthodes existantes telles que CNN-PS, elle montre ses limites lorsque l'anisotropie est très importante.

4.4 Conclusion

Dans ce chapitre, nous avons présenté notre architecture multi-échelles de type encodeur-décodeur basée sur les *CNN* pour résoudre le problème de la stéréophotométrie calibrée. La combinaison de notre architecture multi-échelles avec notre nouvelle base de données, présentée au chapitre 3, nous permet d'obtenir les meilleurs résultats de l'état-de-l'art. L'aspect multi-échelles nous permet de traiter, synthétiser et utiliser l'information globale contenue dans les images d'entrée, quelle que soit la résolution de celles-ci. De plus, notre base de données beaucoup plus diversifiée permet au réseau de gérer des exemples plus variés d'un point de vue des formes géométriques des objets, mais aussi des matériaux.

En revanche, quelques lacunes concernant les matériaux anisotropes persistent. En effet, la reconstruction des normales reste difficile dans les cas de matériaux très très réfléchissant ou translucides. De plus, les contraintes d'un point applicatif restent importantes, puisque les directions et les intensités lumineuses doivent être connues. L'environnement doit également être obscur. Pour résoudre cette problématique, dans le prochain chapitre, nous proposons une nouvelle architecture multi-échelles de type encodeur-décodeur basée sur les *Transformers* dans un contexte de stéréophotométrie universelle.

CHAPITRE 5

Uni-MS-PS : Une architecture multi-échelles de type encodeur-décodeur fondée sur les *Transformers* pour la stéréophotométrie universelle

Comme mentionné dans le chapitre précèdent, la stéréophotométrie calibrée a des contraintes très exigeantes sur les conditions d'acquisition. Ainsi, le passage à la stéréophotométrie universelle est avantageux d'un point de vue expérimental. Pour rappel, la stéréophotométrie universelle s'affranchit de toutes contraintes expérimentales sur l'environnement, les directions lumineuses, etc.

Dans ce chapitre, nous combinons l'avantage du multi-échelles, démontré dans le chapitre précédent avec les *Transformers* dans un contexte de stéréophotométrie universelle. En effet, le principal avantage du multi-échelles est qu'il permet de d'avoir de bonnes reconstructions sur les détails à la fois basse et haute fréquence. Ainsi, nous proposons une nouvelle méthode multi-échelles de type encodeur-décodeur fondée sur les *Transformers*. Notre choix est motivé par le fait que, dans la littérature, les *Transformers* donnent globalement de meilleures performances que les *CNN* dans un cadre de stéréophotométrie universelle [42, 43]. En effet, le problème de la stéréophotométrie universelle étant plus complexe, les *Transformers* sont plus aptes à capturer les inter-dépendances et les similarités au sein des images en entrée et ainsi de mieux gérer la lumière ambiante, contrairement aux réseaux de convolutions. Les résultats présentés ont été publiés dans la revue Computer Vision and Image Understanding(CVIU) [32].

5.1 Architecture multi-échelles fondée sur les *Transformers*

Nous reprenons l'architecture multi-échelle globale présentée dans le chapitre 4 et nous adaptons chaque échelle pour utiliser des Transformers. Notre architecture de type encodeur-décodeur composée de blocs *PVT* (*Pyramid Vision Transformer*) [97], de blocs *SAB* (*Self Attention Blocks*) [57] et de *PMA* (*Pooling by Multi-head Attention*) [57].

5.1.1 Architecture à une échelle donnée

Architecture de l'encodeur. Chaque échelle de notre modèle est composée d'un encodeur et d'un décodeur. La partie encodeur combine trois modules : le premier extrait l'information spatiale sur chaque image indépendamment, le second extrait les informations relatives à la lumière pour toutes les images et le dernier regroupe les informations obtenues pour les connexions entre l'encodeur et le décodeur (i.e. *skip connections*). La partie décodeur est composée de modules de

régression.

Le module d'extraction spatiale est fondé sur le *PVT* (*Pyramid Vision Transformer*). En effet, ce type d'architecture génère des caractéristiques haute-résolution permettant ainsi de considérer le problème au niveau du pixel (comme pour la segmentation). Le principal avantage du *PVT* est cette capacité à prendre en entrée des images de n'importe quelle résolution tout en gardant un temps calculatoire modéré. Ce dernier point est très important pour la stéréophotométrie car pour obtenir de les meilleurs résultats possibles pour la reconstruction des normales, il est nécessaire de conserver l'image à la résolution initiale.

Ensuite, l'extraction des informations relatives à la lumière se fait par le biais d'un module *SAB* (*Self Attention Block*). L'information est extraite à l'échelle du pixel. Les valeurs des pixels pour chaque image sont concaténées afin d'obtenir l'information à chaque localisation. Par ailleurs, le bloc *SAB* est appliqué à chaque localisation indépendamment.

Finalement, un module de *PMA* (*Pooling by Multi-head Attention*) est utilisé en parallèle du module *SAB* afin d'agréger les informations données par le bloc *PVT*. L'objectif est alors de créer une carte de caractéristiques et de l'utiliser pour les connexions entre l'encodeur et le décodeur (i.e. *"skip-connection"*).

Architecture du décodeur. Après l'exécution des blocs d'encodage, le décodeur est utilisé pour reconstruire les cartes de normales. Pour le décodeur, nous considérons trois couches convolutives transposées avec des *skip-connection* provenant de l'encodeur et plus exactement des sorties des blocs *PMA*. En effet, à chaque étape, on concatène la sortie *PMA* obtenue dans l'encodeur avec la sortie de la convolution transposée, jusqu'à ce que l'on retrouve la résolution de l'image en entrée de l'encodeur. Finalement, une dernière étape de convolution transposée a été ajoutée. Elle permet de garder la taille de la dernière carte de caractéristiques mais de changer le nombre de canaux en profondeur pour prédire la carte des normales finale.

La figure 5.1 présente l'architecture de notre méthode à une seule échelle.

5.1.2 Base de données d'entraînement

Afin d'obtenir les meilleures cartes de normales possibles, une base de données appropriée est nécessaire pour le processus d'entraînement. La plupart des bases de données d'entraînement existantes [18, 45] ont été acquises dans un environnement sombre avec des faisceaux lumineux parallèles. Pour la stéréophotométrie universelle, Ikehata a proposé la base de données *PS-Wild* [42]. Malheureusement, cette base de données souffre d'un manque de diversité d'un point de vue des géométries des objets proposés mais aussi d'un point de vue des matériaux et environnements. Comme on peut le voir dans le tableau 5.1, cette base de données est composée de 10099 échantillons comprenant 410 géométries, 926 matériaux et 31 environnements différents. En pratique, cela est clairement pas suffisant pour entraîner un réseau de neurones fondé sur les *Transformers*.

Base de données d'entraînement	nb géométries	nb matériaux	nb environnements	nb échantillons
PS-Wild	410	926	31	10 099
Notre base d'entraînement	11 000	100 000	1 100	100 000

TABLE 5.1 – Comparaison entre notre base de données d'entraînement et la base de données *PS-Wild* [42]. Notre base de données propose beaucoup plus de diversité, que cela soit en nombre d'échantillons, de géométries, de matériaux et d'environnements que *PS-Wild* [42].

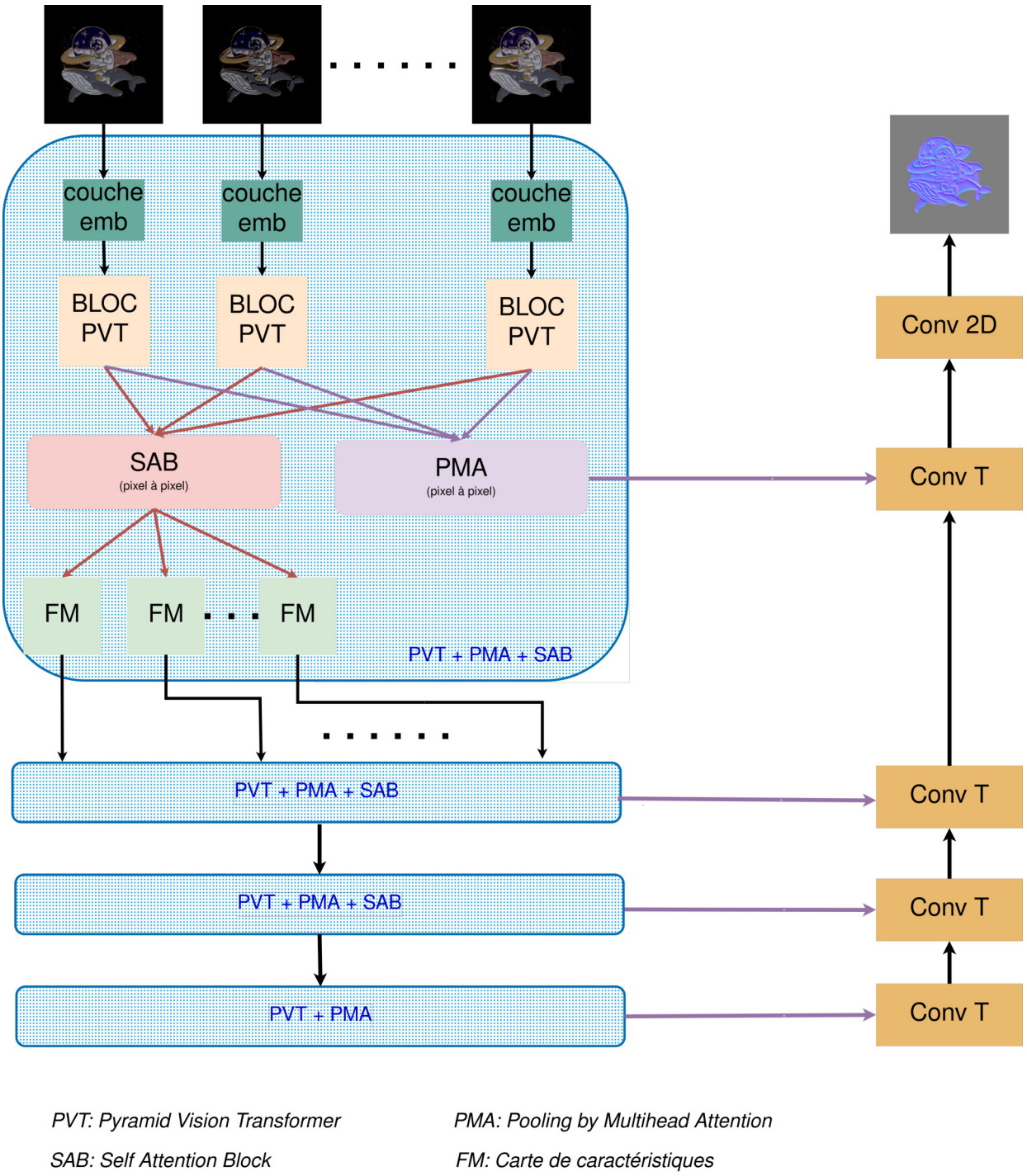
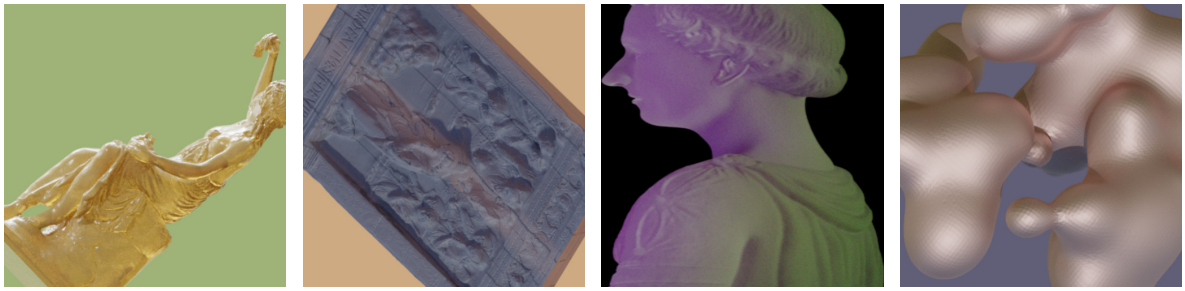


FIGURE 5.1 – Architecture détaillée d'une échelle de réseau. L'entrée peut être soit uniquement des images (pour la première échelle de réseau), soit des images concaténées avec la carte de normales précédente ré-échantillonnée (pour les autres échelles). Nous traitons les images au niveau du pixel afin de capturer tous les détails géométriques. *SAB* signifie *Self Attention Block* [57], *PMA* signifie *Pooling by Multi-head Attention* [57], *PVT* signifie *Pyramid Vision Transformers* [97] et *FM* signifie *Feature Map*.

Par conséquent, comme décrit au chapitre 3, nous avons proposé une base de données d'entraînement afin de prendre en compte le plus de matériaux et de géométries possibles dans une grande diversité d'environnements. Cette base de données a été créée à l'aide du logiciel *Blender* [20]. Au final, nous avons généré 100 000 échantillons avec 11 000 types de géométries, 100 000 matériaux et 1 100 environnements. Nous pouvons donc voir que chaque échantillon à un matériau différent.

Des exemples d'images générées par notre chaîne de génération sont affichées sur la figure 5.2. Comme nous pouvons le voir dans cet exemple, nous proposons une grande diversité en termes de géométries. En effet, nous proposons à la fois des objets "réels", c'est-à-dire des objets tels que des statues ou des sculptures (figures 5.2 a et c), mais également des objets abstraits sur la figure 5.2d. De plus, on peut voir que nous proposons des matériaux mats, par exemple aux figures 5.2 a et d, et des matériaux brillants aux figures 5.2 b et c. De même pour les environnements, notre base de données est composée d'environnements sombres (figures 5.2 a et b) et lumineux (figures 5.2 c et d).



(a) (b) (c) (d)

FIGURE 5.2 – Exemples d'images générées avec notre chaîne de génération sur *Blender*. Comme on peut le voir, nous proposons beaucoup de diversité en termes de géométrie, de matériaux et de environnements.

5.1.3 Processus d'entraînement

Pour entraîner notre architecture multi-échelles, nous utilisons des images 128×128 pixels. Pour atteindre cette résolution, trois échelles sont nécessaires : 32×32 pixels, 64×64 pixels et 128×128 pixels. Grâce à la basse résolution des images d'entraînement, nous sommes capables d'entraîner notre architecture *Transformer* avec 23 images par vue tout en n'utilisant qu'une seule et unique carte graphique A100. À titre de comparaison SDM-UniPS [43] est entraîné sur 4 cartes graphiques A100 avec au maximum 6 images par vue en entrée. Notre réseau est optimisé avec l'algorithme Adam [51], présenté au chapitre 4, avec un taux d'apprentissage fixé à 10^{-4}. Les trois échelles sont entraînées simultanément avec une fonction de perte cosinus.

5.1.4 Inférence sur des images très haute résolution

Comme mentionné plus haut, notre méthode peut être utilisée sur des images très haute résolution en inférence tout en gardant le maximum d'information. Cependant, le processus d'inférence sur des images très haute résolution est un défi. En effet, même en considérant une taille de *batch* égale à 1, l'image ne peut pas être contenue dans une seule carte graphique. Ainsi, afin de faire tourner notre réseau sur des images très haute résolution, on utilise une méthode fondée sur les patchs. Pour ce faire, nous considérons les images entières jusqu'à 256×256 pixels. Ensuite, une fois que nous avons reconstruit les normales jusqu'à cette taille, nous découpons les images et les cartes de normales prédites en patchs 256×256 pixels avec un chevauchement de 64 pixels. Nous pouvons ensuite inférer sur chaque patch indépendamment. Finalement, nous fusionnons tous les patchs en utilisant une moyenne pondérée. Pour deux patchs voisins, des poids gaussiens sont attribuées :

$$w(x, y) = e^{-\frac{||(x,y)-(x_c,y_c)||^2}{2\sigma^2}}, \tag{5.1}$$

où (x_c, y_c) est le centre du patch et l'écart type σ a été fixé à 25.

Cette méthode permet d'éviter le calcul de la carte d'attention sur toute l'image haute résolution en une fois. En effet, ce calcul de carte d'attention est très coûteux et peut augmenter drastiquement l'espace mémoire utilisé dans le *PVT* lorsque la taille de l'image augmente.

Cependant, les résultats, présentés dans le tableau 5.2, montrent que les performances peuvent être dégradées si la taille du patch est trop petite. En effet, nous avons testé notre réseau combiné avec notre méthode d'inférence sur images très haute résolution sur *DiLiGenT10*^{2} [80] avec 30 images par objet. Suite aux résultats obtenus, nous avons choisi une taille de patch de 256×256 pixels car elle offre un compromis raisonnable entre l'utilisation de la mémoire et la précision. En effet, augmenter la taille du patch de 256×256 à 512×512 pixels n'améliore les performances sur *DiLiGenT10*^{2} que de $1,7\%$, tout en nécessitant six fois plus de mémoire. De plus, les différences visuelles entre ces deux tailles de patch ne sont pas perceptibles, même sur un matériau hautement spéculaire qui nécessite le contexte complet de l'image pour comprendre les trajectoires des faisceaux lumineux (voir la figure 5.3). L'empreinte mémoire et le temps de traitement dépendent également du nombre d'images d'entrée. Le tableau 5.3 illustre cette dépendance pour une taille de patch de 256×256 pixels sur des images de 1000×1000 pixels, indiquant l'empreinte mémoire GPU et le temps requis pour l'inférence pour différents nombres d'images d'entrée.

Taille patch	**Chevauchement**	**Mémoire utilisée**	**Erreur (*degré*)**
128	32	3.5 Go	15.52
256	64	21 Go	13.19
512	128	130 Go	12.96

TABLE 5.2 – Précision de reconstruction et utilisation mémoire de la méthode proposée sur un exemple d'image très haute résolution de *DiLiGenT10*^{2}. On peut voir que le meilleur compromis mémoire/erreur est de choisir des patchs de taille 256×256 pixels.

Nombre d'images	**Utilisation mémoire (Go)**	**Temps de calcul (secondes)**
3	4.27	66
6	7.5	120
9	12.3	173
15	21	280
32	34	560

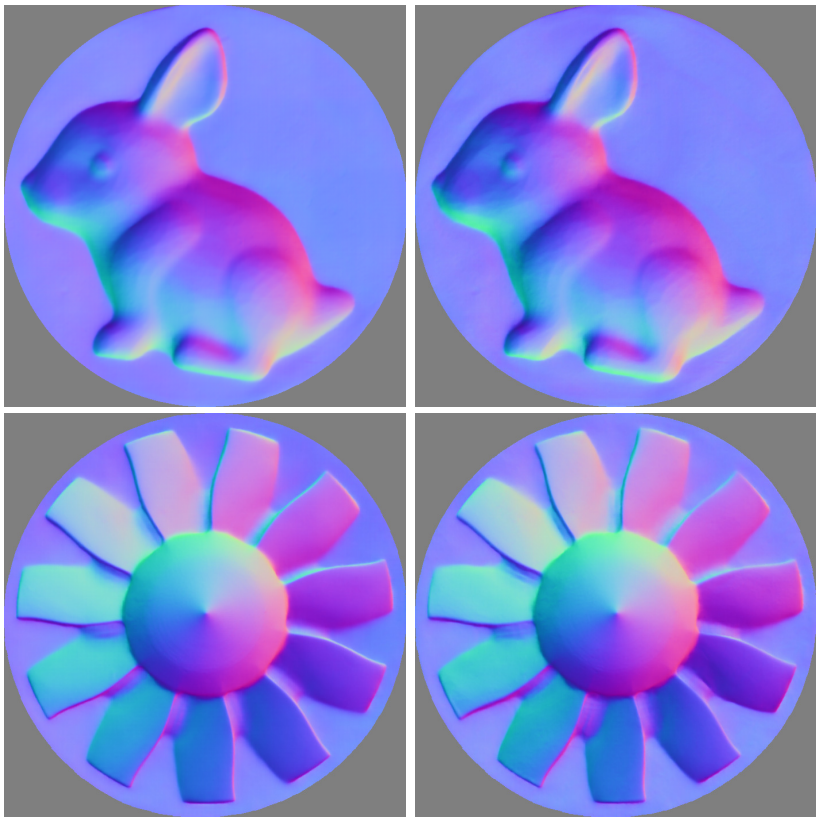
TABLE 5.3 – Utilisation de la mémoire et temps de calcul sur la résolution complète de l'image de *DiLiGenT10*^{2} avec une taille de patch de 256×256 pixels.

5.2 Résultats

D'un point de vue quantitatif, nous comparons notre approche à toutes les méthodes de l'état-de-l'art. En effet, nous considérons les méthodes calibrées [15, 33, 37, 45, 47, 62, 65, 66], non calibrées [17, 58, 62] et universelles [42, 43]. Comme mentionné précédemment, nous avons fixé notre taille de patch à 256×256 pixels pour inférer sur des images haute résolution, pour garder le meilleur compromis mémoire/performance.

5.2.1 Base de données de tests

Nous évaluons les différentes méthodes de l'état-de-l'art et notre méthode sur différentes bases de données publiques. Les bases de données *DiLiGenT* [87], *DiLiGenT10*^{2} [80] et *DiLiGenT-Pi* [95]



Taille de patch : 256×256 pixels. Taille de patch : 512×512 pixels.

FIGURE 5.3 – Comparaison de notre méthode universelle entre une taille de patch de 256×256 pixels et une taille de patch de 512×512 pixels. Visuellement, la différence entre les deux est minime. Cependant, l'inférence avec une taille de patch de 512×512 pixels utilise six fois plus de mémoire GPU. Le lapin et la turbine sont tous deux en matériau très spéculaire, l'aluminium pour le lapin et le laiton pour la turbine.

où la lumière est directionnelle et également la base de données *Luces* [72] pour laquelle les faisceaux sont non parallèles.

Ces méthodes nous offrent l'opportunité de tester notre méthode sur une grande variété d'objets (géométries, matériaux et environnements). Cependant, pour compléter notre évaluation et nous permettre de tester différents types de lumière, caméra ou environnement, nous avons testé notre méthode sur des bases de données où les normales réelles ne sont pas connues, *Skoltech3D* [93], *Shape and Material* [63], UNI-PS [42] et SDM-UniPS [43]. Ces bases de données permettent seulement de juger visuellement de la qualité de reconstruction des normales.

5.2.2 Comparaison quantitative

Dans un premier temps, nous comparons les résultats sur la base de données *DiLiGenT* [87] dans le tableau 5.4. Nous comparons les performances entre toutes les méthodes de l'état-de-l'art, y compris les méthodes calibrées, non calibrées et universelles. Nous avons également testé notre méthode en version calibrée afin de pouvoir comparer avec pertinence notre proposition avec les méthodes calibrées.

Comme on peut le voir dans le tableau 5.4, notre méthode donne les meilleurs résultats en moyenne, toutes méthodes confondues. En effet, si l'on compare l'ensemble des méthodes (calibrées, non calibrées et universelle), nous obtenons une erreur moyenne angulaire de 4.97 degrés alors que la plus petite erreur angulaire des autres méthodes de l'état-de-l'art est de 5.8 degrés, soit un gain de 16%.

	type	ball	bear	buddha	cat	cow	goblet	harvest	pot1	pot2	reading	moyenne
PS-FCN [15]	C	2.67	7.72	7.52	4.75	6.72	7.84	12.39	6.17	7.15	10.92	7.39
CNN-PS [45]	C	2.2	4.6	7.9	4.1	8.0	7.3	14.0	5.4	6.0	12.6	7.2
OB-Cnn [37]	C	2.49	3.59	7.23	4.69	4.89	6.89	12.79	5.10	4.98	11.08	6.37
PX-NET [65]	C	2.03	3.58	7.61	4.39	4.69	6.90	13.10	5.08	5.10	10.26	6.28
NormAttention-PSN [47]	C	2.93	5.48	7.12	4.65	5.99	7.49	12.28	5.96	6.42	9.93	6.83
MS-PS : [33] (cf. chapitre 4)	C	2.05	4.24	7.03	3.9	4.00	7.57	11.01	4.94	5.22	8.47	5.84
SDPS-Net [17]	UC	2.8	6.9	9.0	8.1	8.5	11.9	17.4	8.1	7.5	14.9	9.5
SCPS-NIR [58]	UC	1.24	3.82	9.28	4.72	5.53	7.12	14.96	6.73	6.50	10.54	7.05
UNI-PS [42]	UC/Uni	4.9	9.1	19.4	13.0	11.6	24.2	25.2	10.8	9.9	18.8	14.7
SDM-UniPS [43]	UC/Uni	**1.5**	3.6	7.5	5.4	4.5	8.5	10.2	4.7	4.1	8.2	5.8
Notre méthode (K=30)	C	1.93	2.64	5.88	3.05	3.76	6.40	10.44	3.85	4.32	7.31	4.96
Notre méthode (K=96)	UC/Uni	1.92	3.14	6.16	3.60	4.04	6.35	8.84	**4.08**	4.88	7.09	5.01
Notre méthode (K=30)	UC/Uni	1.84	3.14	**6.04**	**3.45**	**3.99**	6.49	8.9	4.12	**4.7**	**7.0**	**4.97**
Notre méthode (K=15)	UC/Uni	1.93	**3.05**	6.31	3.97	4.06	7.0	9.27	4.25	4.9	7.41	5.22
Notre méthode (K=6)	UC/Uni	2.4	3.7	7.14	4.52	4.7	8.06	12.43	5.32	5.84	9.4	6.35
Notre méthode (K=3)	UC/Uni	3.58	4.83	11.46	7.13	6.68	17.8	18.05	8.79	7.75	15.65	10.17

TABLE 5.4 – Erreurs angulaires en degrés pour chaque objet de la base de données *DiLiGenT* [87] pour les méthodes de l'état-de-l'art. Le type C indique les méthodes calibrées, UC les méthodes non calibrées et Uni les méthodes universelles. La méthode proposée donne les meilleurs résultats, toutes méthodes confondues.

Nous avons également affiché les résultats de notre méthode universelle avec $K = (3, 6, 15, 30, 96)$ images en entrée du réseau. On peut voir qu'à partir de 15 images, nous obtenons déjà les résultats de l'état-de-l'art. Et qu'à partir de 6 images, nous obtenons des résultats très proche de l'état-de-l'art. Cela signifie que notre méthode ne nécessite pas d'utiliser l'ensemble des 96 images pour être performante.

Ensuite, nous avons testé notre méthode universelle sur la base de données *DiLiGenT10*^{2} [80] qui est plus difficile. Les résultats sont présentés dans le tableau 5.5. Sur cette base de données, nous avons comparé notre méthode version calibrée avec la meilleure méthode de l'état-de-l'art, MS-PS [33] et nous avons comparé notre méthode universelle avec les deux meilleures de l'état-de-l'art, CNN-PS [45] et SDM-UniPS [43]. Dans le cas calibré, nous obtenons les meilleurs résultats en moyenne mais nous avons une faiblesse sur le matériau Nylon qui est en deçà de la méthode MS-PS [33]. Pour les méthodes non calibrées et universelles, nous obtenons également les meilleurs résultats en moyenne avec une amélioration nette d'environ 1.7 degrés, soit 13%. Cependant, nous pouvons observer une difficulté à reconstruire les cartes des normales pour le matériau Acrylique. Dans ce cas, il s'agit d'un matériau translucide qui est très difficile à reconstruire car il est difficile pour le réseau de savoir de quel coté vient la lumière comme nous l'avons expliqué précédemment (section 4.2.4). Sans aucun a priori sur la forme de l'objet, il est difficile de voir d'où vient la lumière. Cela a donc un impact dans les cas du non calibré et universel car la lumière pourrait venir de deux côtés opposés. Les méthodes calibrées sont moins impactées car la direction lumineuse est donnée en entrée du réseau.

La seconde base de données très difficile est la base de données *DiLiGenT-Pi* [95]. Les résultats obtenus sont présentés dans le tableau 5.6. Encore une fois, on remarque que la méthode proposée donne les meilleurs résultats de l'état-de-l'art. En effet, si l'on compare notre méthode version calibrée et les autres méthodes de l'état-de-l'art calibrées, on obtient une erreur moyenne de 7.75 degrés alors que la seconde meilleure donne 8.78 degrés d'erreur moyenne, soit un gain de 1 degré.

De plus, on remarque que l'erreur moyenne n'est pas nécessairement la meilleure métrique de comparaison entre les méthodes calibrées et non calibrées. En effet, pour certains objets reconstruits dans un cadre non calibré ou universel, toutes les méthodes de l'état-de-l'art prédisent des normales inversées par rapport aux normales réelles. Par exemple, en figure 5.4, on peut voir que notre méthode universelle reconstruit des normales inversées. Même à l'œil humain, on peut avoir l'impression que le poumon est bombé et non creusé. Cela peut être dû au fait que les méthodes non calibrées et universelles ne sont pas capables de prédire correctement d'où vient la lumière,

mean: 11.33

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	9.3	3.4	8.7	5.2	8.4	4.3	8.5	12.0	14.0	8.6
GOLF	10.0	7.3	9.8	5.8	10.0	6.87	7.9	7.7	9.8	12.0
SPIKE	12.0	8.8	9.9	6.3	8.5	7.9	12.0	7.6	12.0	17.0
NUT	14.0	8.9	15.0	5.8	10.0	5.8	9.2	7.6	8.2	16.0
SQUARE	18.0	11.0	17.0	8.2	14.0	5.5	12.0	7.2	7.9	11.0
PENTAGON	18.0	8.4	17.0	8.0	17.0	9.4	11.0	9.4	13.0	20.0
HEXAGON	16.0	7.5	15.0	6.1	13.0	7.1	11.0	8.1	11.0	20.0
PROPELLER	13.0	8.9	11.0	7.9	16.0	9.7	11.0	8.4	8.2	19.0
TURBINE	21.0	12.0	24.0	11.0	18.0	15.0	23.0	16.0	18.0	22.0
BUNNY	17.0	8.2	16.0	6.5	12.0	8.3	8.6	7.3	8.0	18.0

(a) MS-PS [33] (cf. chapitre 4) (C).

mean: 11.01

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	4.5	3.3	5.0	3.6	5.8	4.1	3.6	7.8	8.8	6.4
GOLF	13.0	6.4	14.0	5.1	11.0	6.8	7.0	6.4	8.1	9.3
SPIKE	10.0	7.3	11.0	7.5	9.1	8.3	8.5	8.3	9.0	11.0
NUT	11.0	5.0	19.0	4.5	8.1	5.0	6.6	6.8	6.4	24.0
SQUARE	18.0	8.5	23.0	7.7	13.0	7.1	7.9	5.0	7.6	19.0
PENTAGON	15.0	8.3	22.0	8.9	13.0	8.4	11.0	9.5	9.5	21.0
HEXAGON	16.0	5.8	20.0	5.9	12.0	6.4	7.1	5.2	6.7	20.0
PROPELLER	16.0	9.2	32.0	8.2	9.6	6.9	15.0	7.8	10.0	17.0
TURBINE	30.0	9.4	30.0	10.0	19.0	9.8	22.0	15.0	16.0	23.0
BUNNY	12.0	8.0	29.0	6.0	8.0	8.3	9.4	8.5	8.4	13.0

(b) Notre méthode (C).

mean: 15.78

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	5.1	6.4	4.2	4.5	6.9	7.3	16.0	14.0	16.0	19.0
GOLF	14.0	8.0	12.0	6.8	14.0	9.4	12.0	9.2	13.0	22.0
SPIKE	11.0	9.4	11.0	11.0	12.0	9.5	14.0	8.3	16.0	28.0
NUT	20.0	8.8	19.0	6.9	17.0	8.0	16.0	13.0	14.0	22.0
SQUARE	21.0	8.1	22.0	6.7	19.0	8.1	13.0	4.9	7.9	18.0
PENTAGON	26.0	9.5	26.0	9.8	22.0	9.6	15.0	13.0	15.0	23.0
HEXAGON	18.0	7.5	19.0	7.2	17.0	28.0	18.0	10.0	17.0	21.0
PROPELLER	28.0	12.0	35.0	8.4	23.0	11.0	16.0	9.6	9.8	17.0
TURBINE	54.0	20.0	51.0	16.0	39.0	21.0	25.0	22.0	21.0	32.0
BUNNY	24.0	11.0	27.0	7.8	21.0	9.1	12.0	7.7	12.0	14.0

(c) CNN-PS [45] (C).

mean: 14.96

POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
1.7	1.2	2.5	2.8	2.7	2.4	2.8	5.0	6.9	3.6
12.0	6.2	13.0	5.0	12.0	7.1	6.7	6.3	7.5	9.2
12.0	6.8	11.0	6.5	8.7	7.6	8.4	5.7	9.2	13.0
16.0	5.1	18.0	4.7	8.4	4.8	18.0	13.0	7.5	20.0
23.0	5.4	25.0	5.3	12.0	7.3	19.0	5.5	20.0	32.0
24.0	8.1	29.0	8.4	13.0	10.0	25.0	26.0	25.0	29.0
18.0	6.5	20.0	4.8	11.0	7.0	20.0	14.0	19.0	34.0
28.0	8.3	44.0	6.1	24.0	7.2	22.0	12.0	19.0	28.0
46.0	10.0	51.0	9.4	36.0	11.0	31.0	25.0	25.0	31.0
36.0	8.9	44.0	6.3	19.0	8.1	27.0	7.8	11.0	28.0

(d) SDM-UniPS [43] (UC).

mean: 13.19

POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
7.4	6.3	7.5	8.4	9.9	7.2	9.0	9.8	13.0	43.0
14.0	7.6	14.0	5.3	11.0	7.6	8.0	8.2	9.6	38.0
9.3	7.1	10.0	6.7	7.4	6.7	11.0	7.7	13.0	29.0
16.0	6.5	20.0	5.8	8.8	8.1	9.9	10.0	9.1	35.0
18.0	6.0	21.0	7.4	9.1	6.5	6.7	5.5	7.4	35.0
20.0	8.8	24.0	9.9	12.0	9.7	13.0	13.0	12.0	40.0
17.0	8.2	18.0	5.5	11.0	6.2	9.4	8.0	7.9	41.0
15.0	8.1	20.0	6.2	8.4	7.3	16.0	9.2	12.0	21.0
30.0	9.9	33.0	9.6	15.0	10.0	22.0	13.0	17.0	33.0
15.0	8.0	23.0	6.0	7.2	8.1	10.0	7.8	8.2	12.0

(e) Notre méthode (Uni).

TABLE 5.5 – Comparaison des erreurs angulaires moyennes sur la base de données *DiLiGenT10*[2] pour les méthodes CNN-PS [45], SDM-UniPS [43], MS-PS [33] et notre méthode en version calibrée et non calibrée/universelle. Notre méthode donne les meilleurs résultats.

	Type	Astro Lion-R Queen	Bagua-R Lion-T Rhino	Bagua-T Lions Sail	Bear Lotus-R Ship	Bird Lotus-T Sun	Cloud-R Lung TV	Cloud-T Ocean Taichi	Crab Panda-R Tree	Fish Panda-T Wave	Flower Para Whale	moyenne
NormAttention-PSN [[47]]	C	7.2 16.4 **4.9**	12.0 21.0 5.1	16.5 **4.4** 5.2	7.4 10.8 **4.9**	6.9 13.7 **5.6**	13.4 7.8 **7.6**	17.3 5.8 9.7	**4.4** 13.9 9.6	4.4 16.6 6.1	4.6 4.2 8.7	9.2
PS-FCN [[18]]	C	7.2 18.4 **4.7**	13.0 21.2 5.3	16.8 4.5 **5.1**	7.4 11.8 6.1	7.2 13.6 6.7	14.3 9.7 8.0	17.8 5.8 10.2	5.3 14.8 10.6	**4.6** 17.2 6.8	4.6 4.7 12.2	9.85
CNN-PS [[45]]	C	6.0 15.8 5.4	12.2 20.3 **4.9**	16.4 4.7 5.2	7.4 10.9 4.9	6.8 13.5 5.8	14.6 5.7 8.3	17.2 **4.6** 7.8	4.5 14.2 11.3	4.2 16.6 5.3	4.7 **3.9** 11.6	9.16
MS-PS [[33]] (cf. chapitre 4)	C	**5.96** 14.37 6.37	11.32 15.71 5.18	15.1 5.5 5.26	6.9 11.92 **5.14**	7.69 12.8 6.46	13.28 7.51 8.63	14.74 4.97 9.91	4.58 14.75 8.22	4.68 14.72 5.29	5.43 4.09 7.09	8.78
SDPS-Net [[17]]	UC	37.7 20.8 16.5	22.5 23.6 24.9	28.9 19.6 16.7	30.7 21.7 19.0	17.6 26.5 31.5	27.4 40.2 26.9	27.5 31.4 34.1	20.5 21.8 41.1	23.6 23.7 39.1	12.8 19.8 29.8	25.93
SDM-UniPS [[43]]	UC/Uni	37.8 15.9 10.6	14.6 16.2 17.0	17.1 9.2 10.5	23.8 11.8 22.0	26.5 13.6 26.2	17.1 46.6 36.6	19.2 34.6 47.2	25.4 17.1 34.4	24.5 17.6 34.9	15.2 23.2 33.8	23.34
Notre méthode (k=30)	C	6.03 13.12 5.69	**9.57** 11.43 5.22	11.75 5.37 6.66	**6.72** **10.17** 6.25	**6.55** **8.09** 5.9	12.61 **5.41** 10.24	**11.01** 5.44 **7.26**	5.75 **12.98** **6.08**	**4.11** **11.39** 5.48	4.85 4.73 **6.71**	**7.75**
Notre méthode (k=100)	UC/Uni	7.58 12.66 7.43	10.19 11.24 6.69	**11.12** 6.63 7.2	12.49 11.29 5.35	8.14 10.38 6.54	**12.45** 42.1 10.39	11.63 6.35 8.54	6.0 13.5 47.27	8.32 11.9 6.11	5.88 7.2 7.84	11.35
Notre méthode (k=30)	UC/Uni	7.14 **12.73** 9.54	10.43 **11.2** 6.68	11.69 6.16 6.62	14.09 11.51 5.65	7.35 10.39 6.05	13.08 41.98 11.5	11.92 5.91 8.95	5.32 13.28 47.15	5.96 12.22 5.93	5.14 7.13 8.77	11.38
Notre méthode (k=15)	UC/Uni	10.93 14.19 7.68	10.46 12.2 6.83	13.44 7.31 8.2	12.16 11.79 5.99	8.21 11.19 8.05	12.71 43.73 11.37	14.04 10.46 10.45	8.23 13.81 48.9	8.76 12.65 7.49	8.02 7.34 9.49	12.54

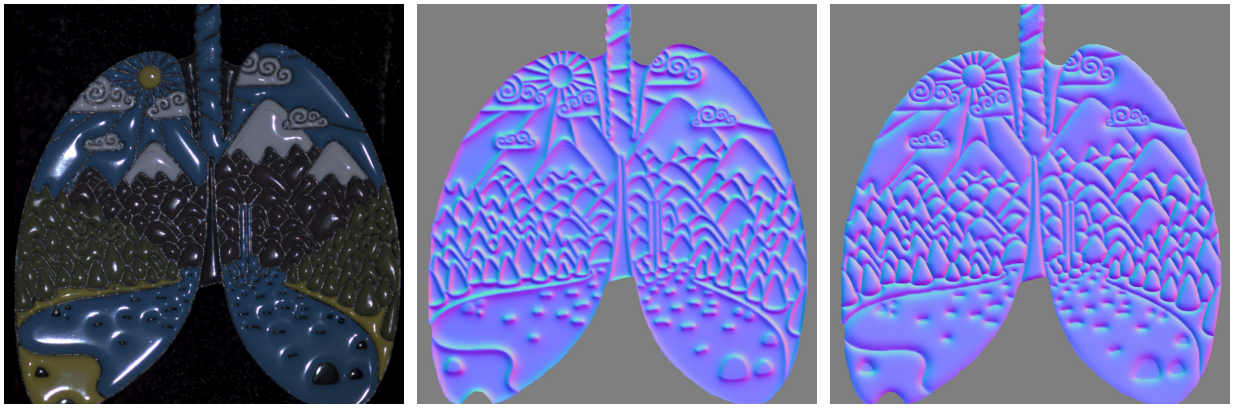
TABLE 5.6 – Erreur angulaire en degrés sur *DiLiGenT-Pi* [95]. Les meilleurs résultats sont en gras et les seconds sont soulignés. Le type C est pour les méthodes calibrées, UC les méthodes non calibrées et Uni pour les méthodes universelle. On peut voir que la méthode proposée donne les meilleurs résultats tous types de méthodes confondus.

contrairement aux méthodes calibrées où la direction lumineuse est donnée en entrée du réseau. Ainsi, le réseau prédit une direction lumineuse opposée à la réalité, d'où les normales inversées. Dans l'exemple de la figure 5.4, il est difficile de dire si la lumière est orientée vers le haut ou le bas. Les deux possibilités sont plausibles et cela peut donc amener à des normales inversées.

En revanche, notre méthode est beaucoup plus robuste à ce problème que les autres méthodes non calibrées ou universelles. En effet, notre méthode inverse les normales pour seulement deux objets de la base de données *DiLiGenT-Pi* contre 11 objets pour la méthode SDM-UniPS [43] et 8 pour la méthode SDPS-NET [17]. Donc notre méthode donne à nouveau les résultats de l'état-de-l'art avec une amélioration de 12% en moyenne.

Finalement, nous avons également testé notre méthode sur la base de données *Luces* [72] qui considère des faisceaux lumineux non parallèles. Les résultats sont affichés dans le tableau 5.7.

On peut voir que notre méthode universelle est capable de gérer également des faisceaux non parallèles. On obtient les meilleurs résultats en comparaison avec l'ensemble des méthodes universelles et non calibrées, même celles construites spécifiquement pour résoudre ce problème de faisceaux non parallèles (Fast-PS(v2) [62]). En effet, les méthodes non calibrées et universelles ont des erreurs moyennes comprises entre 13.5 et 23.77 degrés et nous sommes seulement à 11.10 degrés d'erreur avec notre méthode. Nous obtenons donc un gain significatif. Maintenant, si l'on compare notre méthode en considérant les méthodes calibrées également, nous n'obtenons pas les résultats de l'état-de-l'art mais nous en sommes très proches. Nous obtenons une erreur moyenne de 11.10 degrés contre 9.90 degrés pour la méthode calibrée L22 [66]. Il est cependant important de noter que dans le cas calibré avec des faisceaux non parallèles, non seulement la direction globale de la lumière et l'intensité doivent être étalonnés, mais également la dispersion de la lumière. Dans le cas non calibré ou universel, toutes ces informations ne sont pas nécessaires.



(a) Image gamma corrigée. (b) Carte des normales réelle. (c) Carte des normales générées.

FIGURE 5.4 – Résultats de notre architecture sur l'objet *Lung* de la base de données *Diligent-Pi* [95]. On peut voir que la carte prédite est inversée.

		Ball Hippo	Bell House	Bowl Jar	Buddha Owl	Bunny Queen	Cup Squirrel	Die Tool	average
Fast-PS (v1) [62]	C	8.55 8.01	6.20 29.00	7.0 **5.32**	12.69 12.32	8.63 12.90	17.28 13.00	**5.16** 12.33	11.32
L22 [66]	C	8.84 **5.60**	7.51 22.97	**5.95** 6.19	**11.59** 8.89	**7.06** **9.97**	15.35 11.77	5.19 **11.64**	9.90
Fast-PS (v2) [62]	UC	**6.59** 10.64	**7.17** 31.00	10.17 9.14	14.50 15.92	11.75 18.39	18.98 15.97	8.63 18.61	14.11
UNI-PS [42]	UC/Uni	11.012 21.41	24.12 35.93	23.84 14.53	27.90 32.87	23.51 28.36	28.64 25.36	16.24 19.03	23.77
SDM-UniPS [43]	UC/Uni	13.30 8.86	12.76 26.07	8.44 8.30	18.58 12.67	8.53 15.97	19.67 16.01	7.25 12.54	13.50
Notre méthode (K=52)	UC/Uni	10.20 8.33	10.52 25.29	6.98 6.30	12.83 11.47	9.60 12.45	13.68 **11.36**	6.19 11.79	11.21
Notre méthode (K=30)	UC/Uni	10.29 8.44	10.51 25.46	6.79 6.10	12.57 11.38	9.6 15.97	**13.35** 11.37	6.27 12.22	11.10
Notre méthode (K=15)	UC/Uni	10.47 8.54	10.8 25.30	7.91 6.49	13.14 11.82	9.90 12.49	13.96 11.64	6.52 11.89	11.50
Notre méthode (K=6)	UC/Uni	10.94 12.42	11.40 **9.41**	9.38 26.68	13.75 **7.37**	11.029 12.62	15.38 12.85	7.80 12.79	12.47
Notre méthode (K=3)	UC/Uni	10.93 11.06	15.95 31.61	12.07 10.49	16.78 15.73	14.53 14.99	16.09 15.67	9.09 15.69	15.05

TABLE 5.7 – Erreur angulaire en degrés sur la base de données *Luces* [72]. Les meilleurs résultats sont en gras et les seconds sont soulignés. Le type C est pour les méthode calibrées, UC pour les méthodes non calibrées et Uni pour les méthodes universelles. La méthode proposée donne les résultats de l'état-de-l'art.

5.2.3 Comparaison qualitative

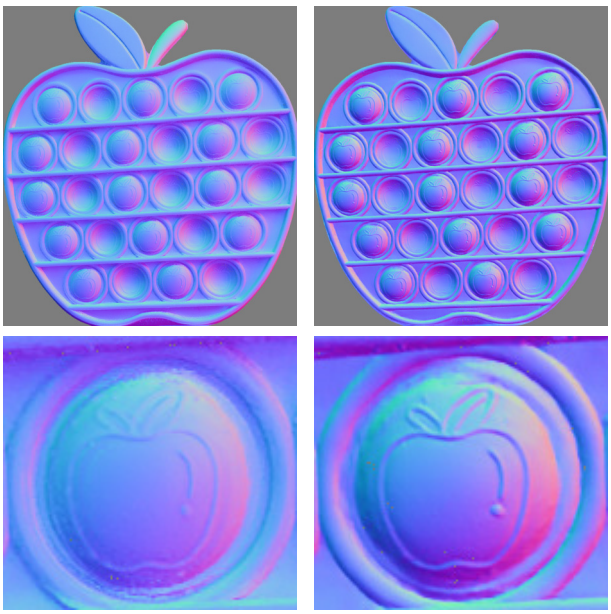
Afin de tester la robustesse de notre méthode universelle dans divers contextes (lumière, caméra, environnement, etc.), nous avons également évalué qualitativement notre réseau sur d'autres bases de données pour lesquelles nous n'avons pas les normales réelles (la vérité terrain). Nous avons choisi de nous comparer principalement à la méthode SDM-UniPS [43] car il s'agit de la méthode universelle la plus performante à ce jour.

Des exemples de reconstruction sont affichés sur les figures 5.5, 5.6, 5.7, 5.8, 5.9 et 5.10. Pour l'ensemble de ces méthodes, on peut voir que globalement les cartes de normales obtenues sont cohérentes. En revanche, lorsque l'on regarde en détails les parties zoomées, on peut voir que notre méthode donne de meilleurs résultats que SDM-UniPS [43]. En effet, les détails de textures et de géométries sont beaucoup mieux reconstruits. Cela est particulièrement visible par exemple sur les écailles de l'alligator à la figure 5.7, sur les détails architecturaux du château à la figure 5.10 ou sur les mailles du textile à la figure 5.9.

5.2.4 Inférence sans masque

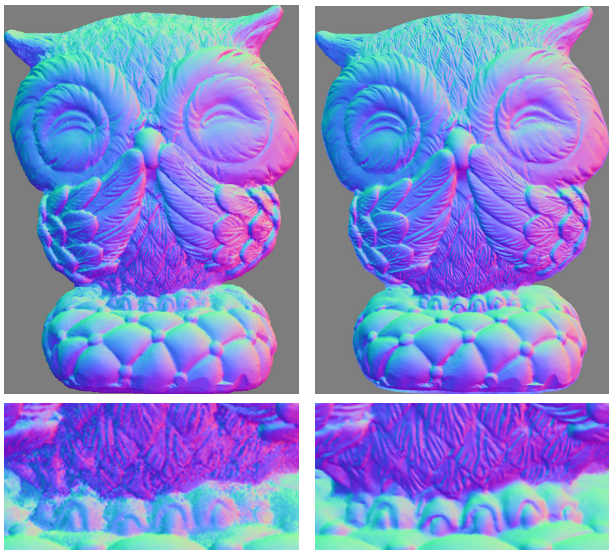
Un avantage de SDM-UniPS de [43] par rapport aux autres méthodes de stéréophotométrie est sa capacité à ne pas avoir besoin des masques des objets en entrée pour obtenir de bons résultats. Il s'agit d'une nouveauté pour les méthodes de stéréophotométrie fondées sur l'apprentissage profond. En effet, les autres méthodes nécessitent le plus souvent de masquer l'arrière-plan des images pour fonctionner correctement.

Seules les bases de données *DiLiGenT* [87] et *Luces* [72] sont intéressantes dans ce contexte car les autres bases de données sont prises dans le noir complet et ainsi l'arrière plan n'est pas visible. Par conséquent, pour tester notre réseau universel sur ces deux bases de données, nous avons inféré sans masquer l'environnement et avons calculé l'erreur seulement sur les normales de l'objet qui sont les seules connues. Cette technique nous permet alors de mesurer l'impact de l'environnent sur la reconstruction des normales d'un objet. Les résultats sont présentés dans le tableau 5.8.



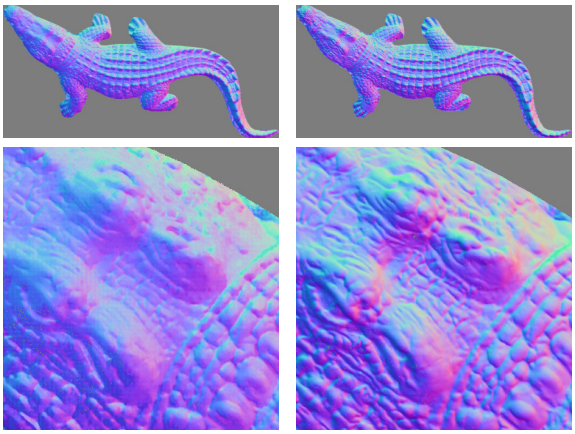
SDM-UniPS [43]. Notre méthode.

FIGURE 5.5 – Exemple de reconstruction d'une image proposée par Ikehata [42]. La deuxième ligne présente une partie zoomée.



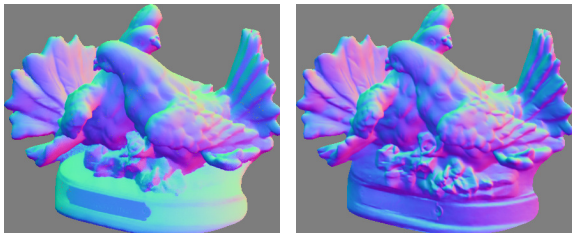
SDM-UniPS [43]. Notre méthode.

FIGURE 5.6 – Exemple de reconstruction d'une image proposée par Ikehata [43]. La deuxième ligne présente une partie zoomée.



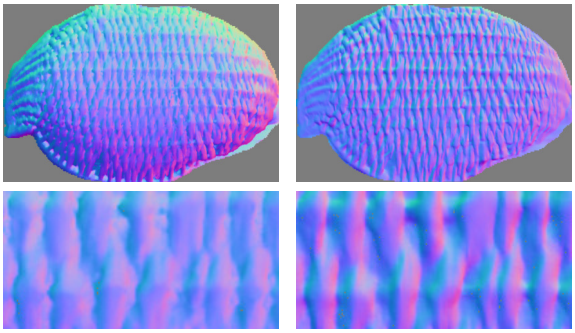
SDM-UniPS [43]. Notre méthode.

FIGURE 5.7 – Exemple de reconstruction d'une image proposée par Ikehata [43]. La deuxième ligne présente une partie zoomée.



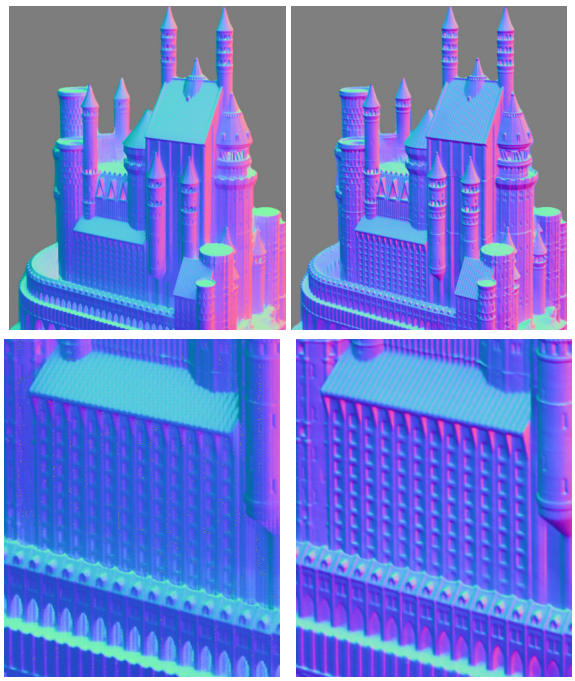
SDM-UniPS [43]. Notre méthode.

FIGURE 5.8 – Exemple de reconstruction d'une image proposée par Ikehata [43].



SDM-UniPS [43]. Notre méthode.

FIGURE 5.9 – Exemple de reconstruction d'une image de la base *Shape and Material* [63]. La deuxième ligne présente une partie zoomée.



SDM-UniPS [43]. Notre méthode.

FIGURE 5.10 – Exemple de reconstruction d'une image de la base de données *Skoltech3D* [93]. La deuxième ligne est une partie zoomée.

		ball	bear	buddha	cat	cow	goblet	harvest	pot1	pot2	reading	moyenne
Sans masque	SDM-UniPS [43]	4.42	4.21	8.54	5.59	7.24	10.37	14.92	5.44	6.72	12.97	8.04
	Notre méthode	11.46	4.64	7.46	4.11	7.80	7.14	10.34	5.27	5.59	7.93	7.17
Avec masque	SDM-UniPS [43]	1.5	3.6	7.5	5.4	4.5	8.5	10.2	4.7	4.1	8.2	5.8
	Notre méthode	1.84	3.14	6.04	3.45	3.99	6.49	8.9	4.12	4.7	7.0	4.97

TABLE 5.8 – Erreurs angulaires en degrés sur la base de données *DiLiGenT* [87] sans l'utilisation d'un masque pour inférer. On peut voir que l'arrière-plan a une influence négative sur les résultats, quelle que soit la méthode.

Afin de faciliter les comparaisons, nous nous sommes comparé à la meilleure méthode de l'état-de-l'art universelle, SDM-UniPS [43] pour constater l'influence de l'environnement sur la reconstruction des normales. Nous pouvons constater que quelle que soit la méthode, les performances décroissent sans utilisation de masques pour l'inférence. Pour les deux méthodes, les performances décroissent en moyenne de 2.5 degrés environ. Notre méthode maintient tout de même de bonnes performances et reste à l'état-de-l'art.

En revanche, même si les résultats quantitatifs ne démontrent pas d'amélioration significatifs de notre méthode, on peut voir sur la figure 5.11 que notre méthode reconstruit beaucoup mieux les détails de l'objet mais aussi de l'environnement. Là où l'environnement a l'air plat avec la méthode SDM-UniPS [43], notre méthode permet de distinguer la texture, notamment le tapis sur lequel est posé le crocodile en figure 5.11.

SDM-UniPS [43]. Notre méthode.

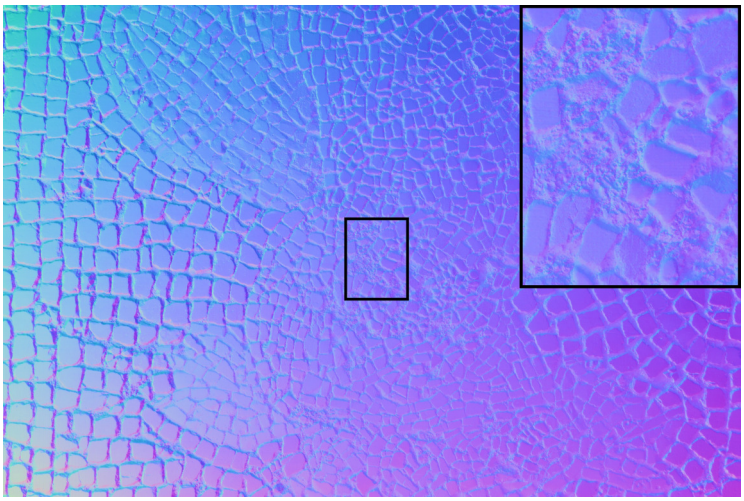
FIGURE 5.11 – Comparaison visuelle des cartes des normales pour la méthode SDM-UniPS [43] et notre méthode sur un objet de la base de données *DiLiGenT* [87] et un de la base SDM-UniPS [43]. On peut voir que notre méthode reconstruit beaucoup mieux les détails et textures.

5.2.5 Inférence sur images très haute résolution

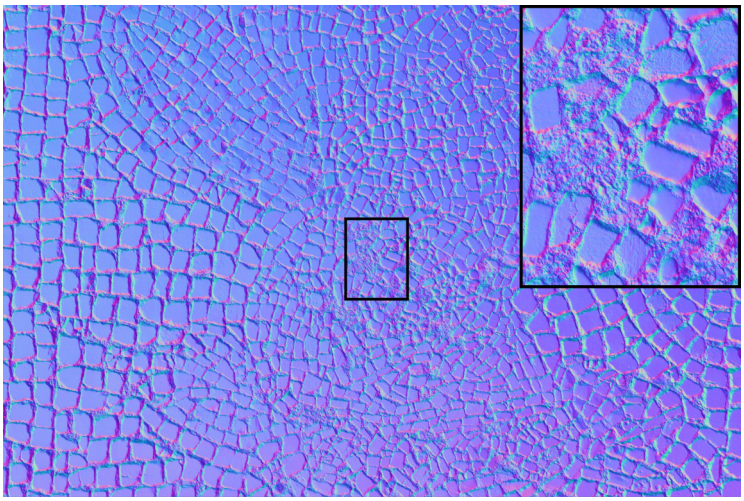
Dans cette section, nous souhaitons présenter quelques résultats sur des images très haute résolution.

Les résultats présentés sur les figure 5.12, 5.13 et 5.14 sont obtenus sur des images 6000×8000 pixels, 5500×8200 pixels et 4000×4000 pixels respectivement. Les résultats obtenus sont impressionnants, notre méthode est capable de gérer des images de telles résolution, en reconstruisant bien tous les détails. Notre méthode ne souffre donc pas du changement d'échelle et demeure performante.

Image RGB



SDM-UniPS [43].



Notre méthode.

FIGURE 5.12 – Comparaison visuelle entre SDM-UniPS [43] et notre méthode universelle sur la Mosaïque des "Saisons". La résolution des images est de 5500 × 8200 pixels. Nous pouvons voir que notre méthode gère la très haute résolution et surpasse SDM-UniPS [43] en terme de qualité de reconstruction de la carte de normales. Crédits des images : A. Laurent (INRT, UMR 5055 IRIT)/MAN.

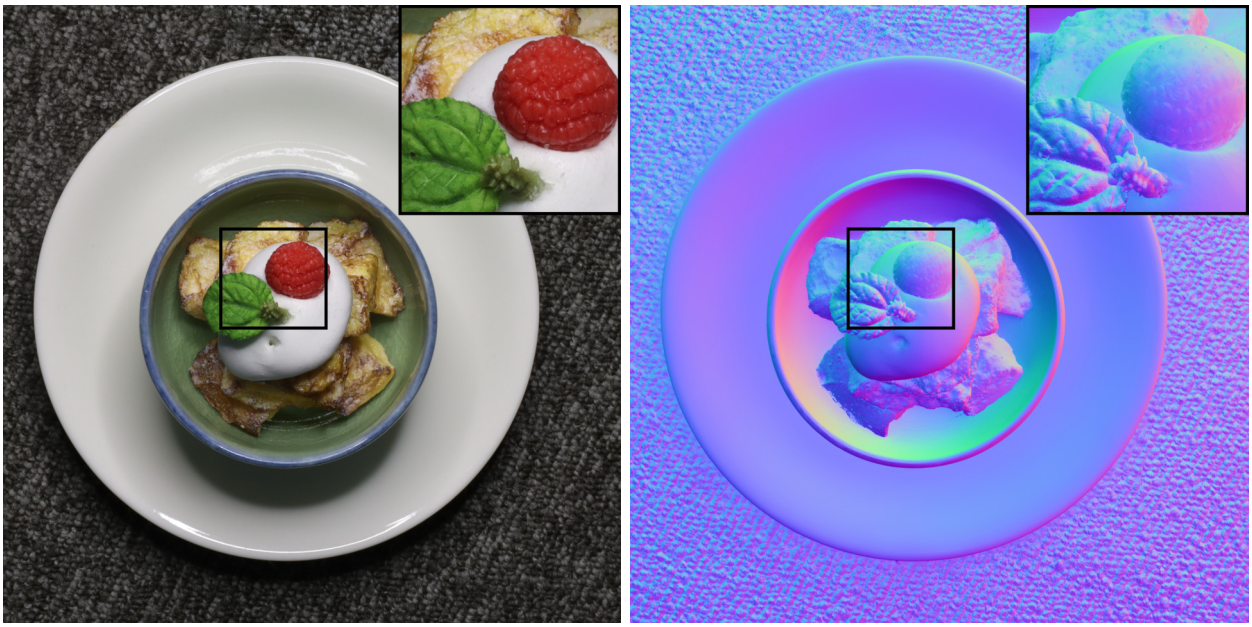
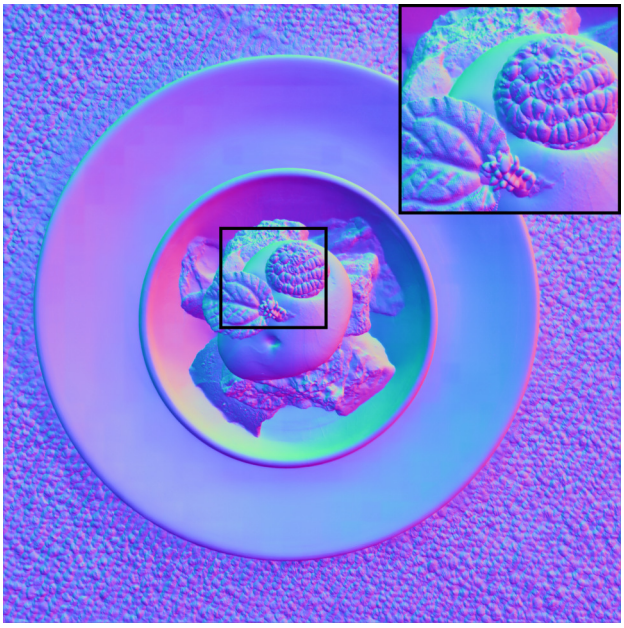


Image RGB

SDM-UniPS [43].



Notre méthode.

FIGURE 5.13 – Comparaison visuelle entre SDM-UniPS [43] et notre méthode universelle sur l'objet *"Sweet"* de [43]. La résolution des images est de 4000×4000 pixels. Nous pouvons voir que notre méthode gère la très haute résolution et surpasse SDM-UniPS [43] en terme de qualité de reconstruction de la carte de normales.

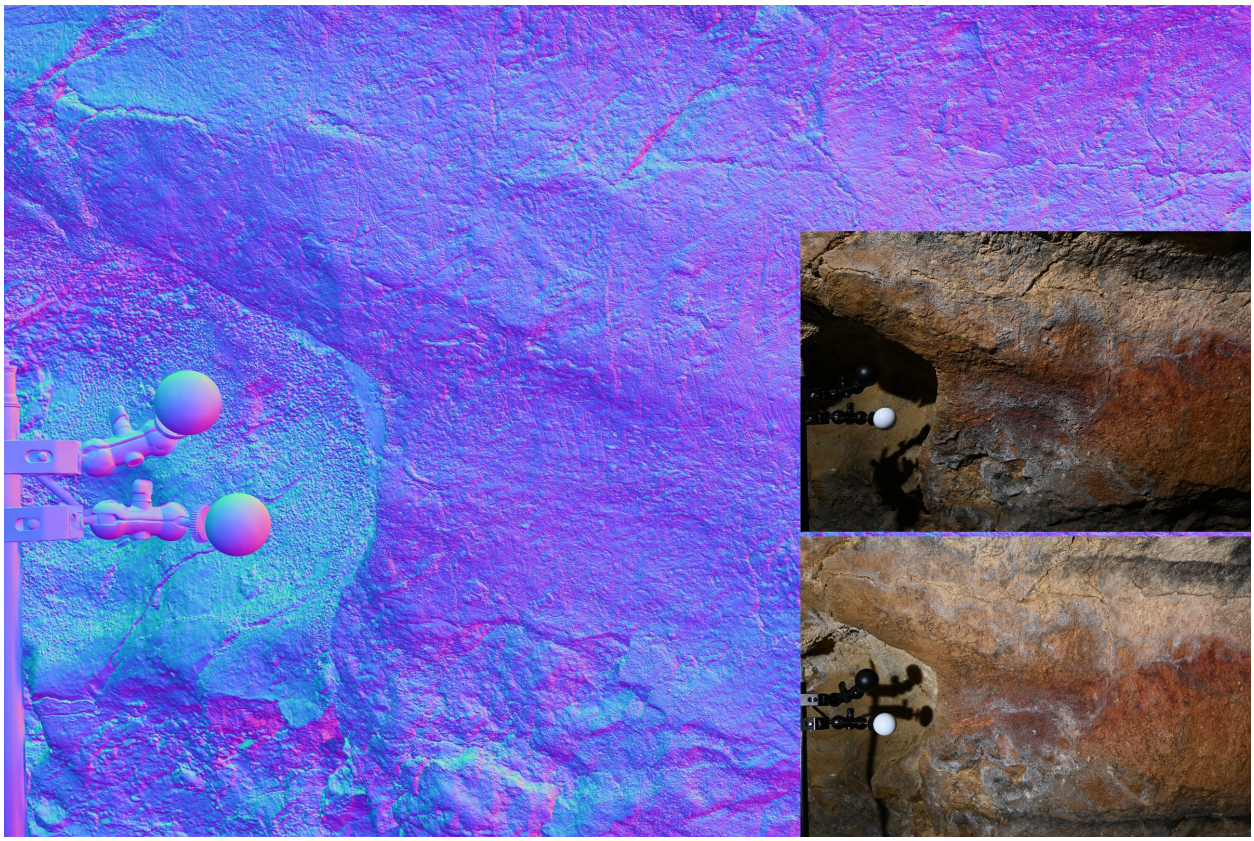


FIGURE 5.14 – Reconstruction de la carte des normales de taille 6000 × 8000 pixels de la cave de Marsoulas à l'aide de notre méthode universelle multi-échelles. Notre méthode permet à la fois d'avoir des reconstruction de la carte des normales très précises tout en gérant des images à très hautes résolutions. En effet, touts les détails de la roche apparaissent. Crédits des images : A. Laurent 2023 (INPT, UMR 5505 IRIT), C. Fritz and G. Tosello team (CREAP-E.Cartailhac), MSHS-T (UAR 3414).

5.3 Conclusion

Dans ce chapitre, nous avons présenté notre architecture multi-échelles de type encodeur-décodeur fondée sur les *Transformers* pour résoudre le problème de stéréophotométrie non calibrée et universelle. Les avantages du multi-échelles nous ont permis de démontrer l'efficacité de notre méthode pour reconstruire des cartes de normales d'images très haute résolution avec beaucoup de détails et de textures. De plus, la considération d'une architecture de type encodeur-décodeur fondée sur les *Transformers* nous permet de garder le maximum d'information tout au long du réseau et de capturer les dépendances concernant la lumière au sein des images afin de reconstruire avec précision tous les détails présent dans l'image. En effet, notre méthode donne les résultats de l'état-de-l'art pour l'ensemble des bases de données publiques. Par ailleurs, notre méthode ne souffre pas du changement de résolution des images en entrée et elle continue d'être performante sur des images très haute résolution.

Cependant, nous avons pu constater encore quelques difficultés de reconstruction dans le cas de matériaux translucides lorsque la direction lumineuse n'est pas connue. Ainsi, dans le prochain chapitre, nous proposons une application alternative entre le mode calibré et le mode non calibré pour améliorer la reconstruction des matériaux difficiles.

CHAPITRE 6

Stéréophotométrie faiblement calibrée

Comme nous l'avons vu dans le chapitre précédent, dans un cadre de stéréophotométrie non-calibrée ou universelle, les méthodes ont parfois du mal à définir la direction lumineuse sur certains objets. En effet, dans certains cas, la carte de normales générée est "l'opposé" de la vérité terrain. Cependant, dans un cadre calibré, nous voyons que cette inversion des normales n'apparaît pas, comme illustré en figure 6.1.

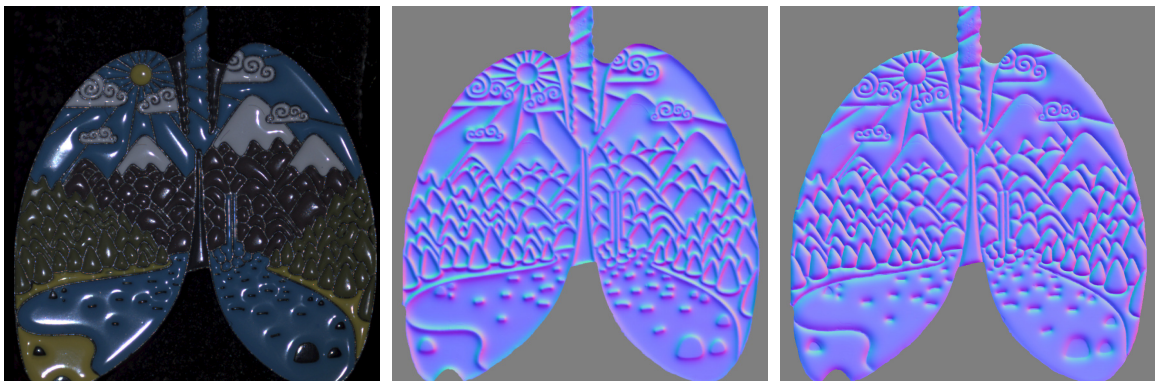


Image gamma corrigée | Calibrée | Universelle

FIGURE 6.1 – Exemple d'un cas d'inversion des normales sur l'objet "*Lung*" de la base de données *DiLiGenT-Pi* [95].

Notre hypothèse est que, dans le contexte non calibré ou universel, le réseau doit estimer tout seul au moins de manière implicite la direction lumineuse et que, sur certains objets, celui-ci se trompe et pense que la lumière vient de la direction opposée à la réalité. Cette hypothèse est démontrée dans la suite de ce chapitre. En effet, dans le cas où les estimations des directions lumineuses sont inversées, il est logique que les normales estimées soient elles aussi inversées car leurs estimations dépendent directement de la direction des sources lumineuses.

Pour résoudre cette problématique, nous proposons ici d'étudier un nouveau type de contexte pour la stéréophotométrie, que nous appelons la stéréophotométrie faiblement calibrée. Dans ce cadre, le réseau n'aura plus une direction lumineuse précise par image mais une approximation grossière de cette direction, par exemple "haut", "bas", "droite" ou "gauche" par exemple. D'un point de vue expérimental, l'estimation précise de la direction lumineuse n'est plus nécessaire.

De plus, aucune autre contrainte sur le contrôle de l'environnement n'est nécessaire. C'est-à-dire que tous types de faisceaux lumineux ou d'environnements peuvent être considérés. Le but de ce nouveau type de stéréophotométrie est de pouvoir lever l'ambiguïté de la direction lumineuse, illustré sur la figure 6.1 qui persiste dans certains cas mais sans pour autant mettre des contraintes expérimentales trop fortes sur l'utilisateur.

6.1 Estimation de la direction lumineuse

Généralement les méthodes non calibrées hormis les méthodes universels se décomposent en deux sous-réseaux indépendants. Le premier estime la direction lumineuse et le second prend cette estimation ainsi que les images pour prédire la carte des normales. Ainsi, pour vérifier l'hypothèse que les méthodes non calibrées se trompent dans les estimations des directions lumineuses sur plusieurs types d'objets ou matériaux, notamment avec une inversion totale des estimations, nous avons testé certains sous-réseaux permettant d'estimer la direction lumineuse. Si ces sous-réseaux se trompent dans leurs estimations alors notre hypothèse sur la mauvaise estimation de la direction lumineuse sera validée.

Dans la littérature, plusieurs méthodes pour estimer la direction des sources lumineuse ont été proposées, notamment UPS-GCNet [16] et DANI-Net [61]. Nous allons utiliser ces deux méthodes pour valider notre hypothèse. De plus en complément de cette méthode de test, nous avons testé deux modifications possibles de notre architecture Uni-MS-PS présentée au chapitre 5, afin que celle-ci prédisse également les directions lumineuses. Ces deux modifications permettent également de valider notre hypothèse selon laquelle la mauvaise estimation de la direction lumineuse est validée.

La première modification consiste à ajouter une branche de régression de la direction lumineuse entre l'encodeur et le décodeur (entre le dernier bloc *PMA* et la première couche de convolution transposée). Cette branche est composée d'un *max-pooling*, suivi d'une couche linéaire qui prédit la direction lumineuse (x, y, z) normalisée. Pour cette méthode, seul la partie régresseur est apprise, le reste des poids est équivalent à ceux de Uni-MS. Lors de l'inférence, seul le résultat du régresseur de la dernière échelle est conservé, comme illustré sur la figure 6.2. Dans la suite de ce chapitre, nous appellerons cette méthode Uni-MS-PS-Reg.

La seconde modification consiste simplement à utiliser l'architecture de la partie encodeur et à remplacer le décodeur par un régresseur qui doit non pas sortir une classe d'objet mais la direction (x, y, z) de la source lumineuse, comme illustré sur la figure 6.3. Cette fois ci, l'ensemble des poids soit appris. Dans la suite de ce chapitre, nous appellerons cette méthode Uni-MS-PS-Cls.

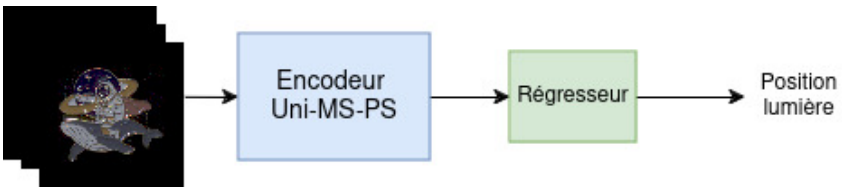


FIGURE 6.3 – Deuxième modification de Uni-MS-PS pour prédire la position de la source lumineuse. Seule l'architecture de l'encodeur est conservé et l'architecture complète est apprise depuis le début.

Sur la figure 6.4, nous présentons les résultats sur d'estimation des directions lumineuses obtenus sur l'objet *"Lung"* de la base de données *DiLiGenT-Pi* [95], connu pour être un objet compliqué à reconstruire. L'extrémité en forme d'étoile représente la position prédite et l'extrémité en forme de point la vérité terrain. Comme nous pouvons le voir avec les quatre méthodes employées (UPS-

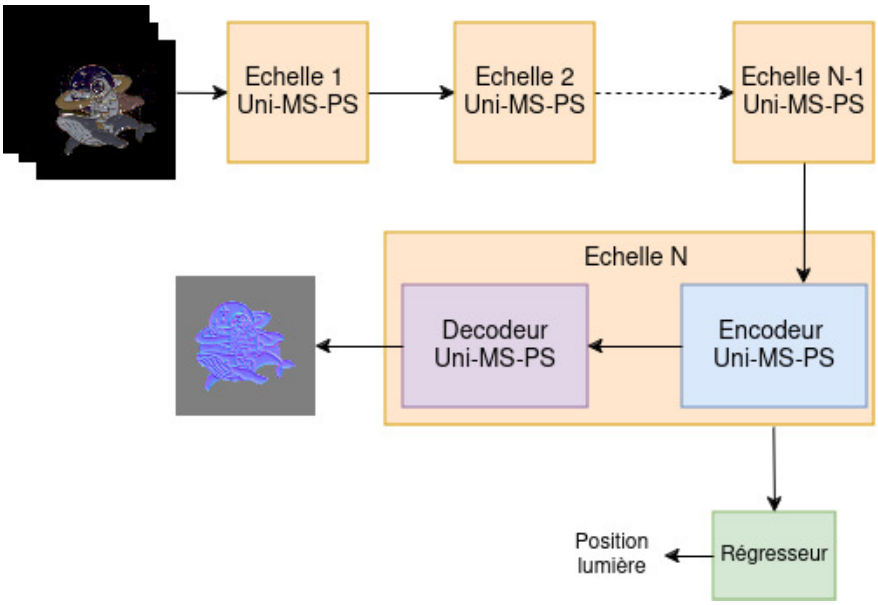
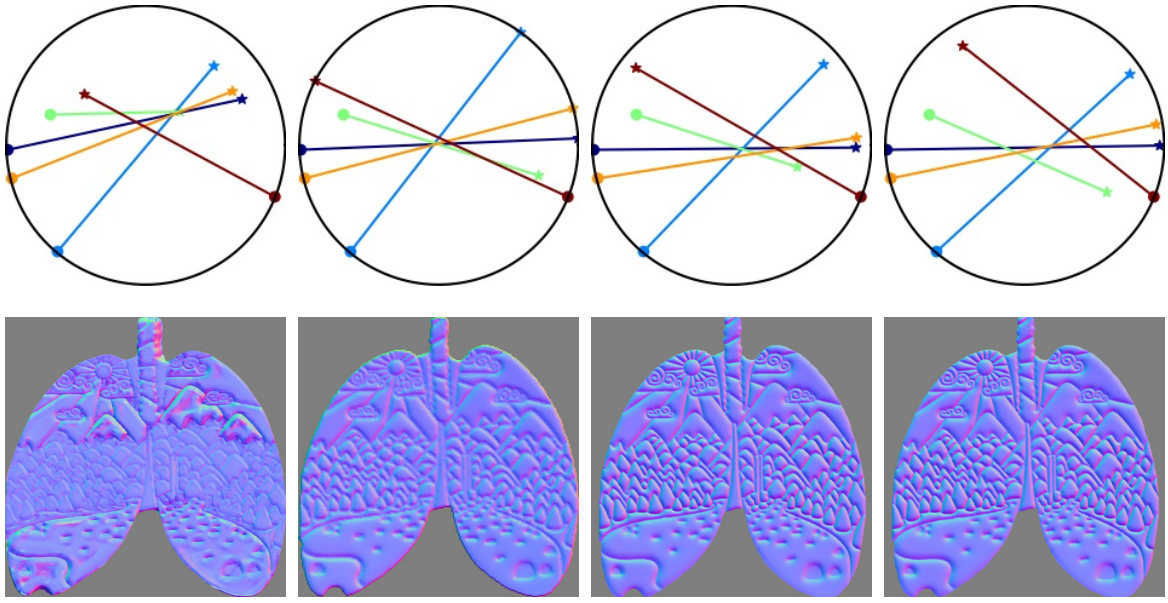


FIGURE 6.2 – Première modification de Uni-MS-PS pour prédire la direction lumineuse. Les poids des échelles 1 à N sont repris de Uni-MS-PS, seul le régresseur de l'échelle N est appris.

GCNet [16], DANI-Net [61] ainsi que nos deux modifications), les directions lumineuses estimées sont opposés à la vérité terrain.

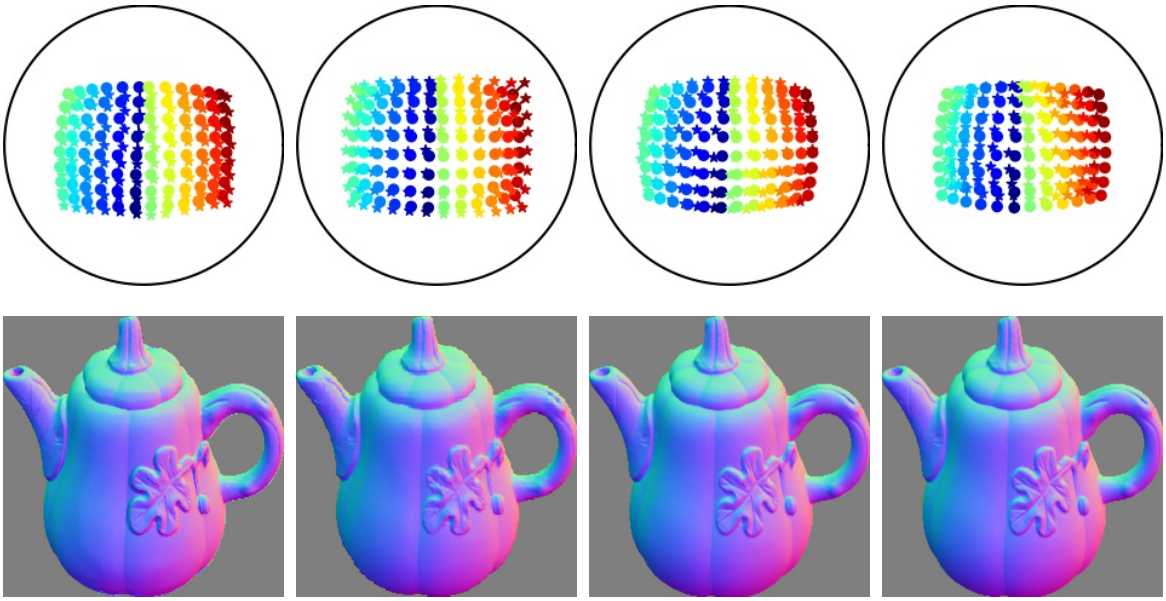


UPS-GCNet [16]. Dani-Net [61]. Uni-MS-PS-Reg. Uni-MS-PS-Cls.

FIGURE 6.4 – Sur les objets où la carte de normales prédite est inversée comme sur l'objet *"Lung"* de *DiLiGenT-Pi* [95], les estimations des directions lumineuses sont à l'opposé de la vérité terrain pour toutes les méthodes testées. Ici, les estimations sont sous formes d'étoile et la vérité terrain est sous forme de point. Pour plus de clarté, seulement 5 estimations ont été affichées.

A contrario, dans le cas d'un objet "facile" comme pour le *"pot2"* du jeu de données *DiLiGenT* [87] (voir figure 6.5), les directions lumineuses estimées sont très proches de la réalité, quelle que soit la méthode utilisée et en tout point du domaine.

L'ensemble de ces expériences valide l'hypothèse que les méthodes non calibrées et universelles



UPS-GCNet [16]. Dani-Net [61]. Uni-MS-PS-Reg. Uni-MS-PS-Cls.

FIGURE 6.5 – Sur l'objet *"pot2"* de *DiLiGenT* [87], réputé "facile", les méthodes d'estimation de la direction lumineuse fonctionnement très bien. Les estimations en forme d'étoile sont juste à côté de la vérité terrain, ici sous forme de point.

ont des difficultés à estimer les directions lumineuses et les inversent mêmes complètement sur certains objets ou matériaux. L'inversion de la carte des normales est dûe à cette inversion dans l'estimation des directions lumineuses.

6.2 Contribution : méthode faiblement calibrée

Pour résoudre cette problématique d'inversion des cartes des normales, il semblerait pertinent d'utiliser simplement une méthode calibrée. Effectivement, les méthodes calibrées n'ont pas ce problème. Par exemple, l'objet *"Lung"* est bien reconstruit avec la version calibrée de notre architecture Uni-MS-PS. Cependant, d'un point de vue expérimental, le contexte calibré requiert une installation et du matériel restreignant considérablement son applicabilité. En effet, les méthodes calibrées ont besoin des directions précises. Cependant, pour résoudre la problématique des inversions des normales, les directions précises ne sont pas nécessaires, une indication peu précise sur les directions lumineuses suffit. La solution apportée par notre méthode consiste ainsi à fixer quatre directions de références (ou plus si besoin) dans l'hémisphère et à indiquer au modèle de quel direction de référence la direction lumineuse est la plus proche. Pour une question de praticabilité, les quatre directions de références choisies correspondent à : "haut", "bas", "droite" et "gauche", comme illustré sur la figure 6.6.

L'intérêt du faiblement calibré ne se retreint pas à la problématique de l'inversion des normales. Celui-ci est également utile sur tous les types d'objets où l'estimation de la direction lumineuse est difficile.

D'un point de vue architectural, le faiblement calibré prend en entrée les images, chacune concaténée avec un indice indiquant la direction de référence mis sous forme d'images pour permettre la concaténation.

FIGURE 6.6 – Représentation de la séparation de l'hémisphère en 4 directions de références. Toutes les directions lumineuses dans la zone orange auront la même direction lumineuse en entrée du réseau de neurones, de même pour les autres zones de références.

6.3 Résultats quantitatifs

Dans un premier temps, nous avons comparé notre architecture Uni-MS-PS avec sa version calibrée et sa version faiblement calibrée sur les bases de données *DiLiGenT* [87], *DiLiGenT10*^{2} [80] et *DiLiGenT-Pi* [95].

Les résultats obtenus sur la base de données *DiLiGenT* sont affichés dans le tableau 6.1. Les performances sont comparables à la méthode calibrée et universelle. Il n'y a donc pas un énorme écart de performance entre les trois *Transformers* multi-échelles. En effet, sur cette base de données, les estimations implicites des directions lumineuses étaient déjà excellentes comme cela est visible sur la figure 6.5. Ces estimations sont précises car les matériaux utilisés et les géométrie des objets sont relativement faciles à reconstruire.

	ball	bear	buddha	cat	cow	goblet	harvest	pot1	pot2	reading	moyenne
Universelle	1.84	3.14	6.04	3.45	3.99	6.49	**8.9**	4.12	4.7	**7.0**	4.97
Faiblement calibré	**1.39**	**2.41**	**5.59**	3.66	**3.86**	6.54	9.15	**3.75**	**4.24**	7.63	**4.82**
Calibré	1.93	2.64	5.88	**3.05**	3.76	**6.40**	10.44	3.85	4.32	7.31	4.96

TABLE 6.1 – Erreur angulaire en degrés pour chaque objet de la base de données *DiLiGenT* [87]. Les trois méthodes présentées ont toutes les trois la même architecture (i.e. *Transformers* multi-échelles) présentée dans le chapitre 5 et modifiée si nécessaire pour le contexte calibré, universel et faiblement calibré.

Les résultats obtenus sur la base de données *DiLiGenT*10^2 sont affichés dans le tableau 6.2. Comme on peut le voir, on améliore globalement les résultats obtenus avec la version faiblement calibrée de notre architecture Uni-MS-PS. Les grandes difficultés rencontrées avec les matériaux translucides sont résolues. En effet, les résultats quantitatifs sont très proches de ceux obtenus avec les méthodes calibrées. Finalement, nous avons également comparé les résultats sur la base de données *DiLiGenT-Pi* [95] dans le tableau 6.3. Les résultats obtenus pour le faiblement calibré sont à mi-chemin entre ceux obtenus avec la méthode calibrée et la méthode universelle. De plus, les résultats obtenus sur les objets *"Lung"* et *"Tree"* qui sont très difficiles à reconstruire sont affichés en gras dans le tableau. Comme on peut le voir, les résultats ont été grandement améliorés sur ces objets difficiles.

6.4 Résultats qualitatifs

Les résultats obtenus sur les objets *"Lung"* et *"Tree"* de la base de données *DiLiGenT-Pi* [95] sont affichés dans la figure 6.7. Comme on peut le voir, le faiblement calibré permet de résoudre

mean: 13.19

	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	7.4	6.3	7.5	8.4	9.9	7.2	9.0	9.8	13.0	43.0
GOLF	14.0	7.6	14.0	5.3	11.0	7.6	8.0	8.2	9.6	38.0
SPIKE	9.3	7.1	10.0	6.7	7.4	6.7	11.0	7.7	13.0	29.0
NUT	16.0	6.5	20.0	5.8	8.8	8.1	9.9	10.0	9.1	35.0
SQUARE	18.0	6.0	21.0	7.4	9.1	6.5	6.7	5.5	7.4	35.0
PENTAGON	20.0	8.8	24.0	9.9	12.0	9.7	13.0	13.0	12.0	40.0
HEXAGON	17.0	8.2	18.0	5.5	11.0	6.2	9.4	8.0	7.9	41.0
PROPELLER	15.0	8.1	20.0	6.2	8.4	7.3	16.0	9.2	12.0	21.0
TURBINE	30.0	9.9	33.0	9.6	15.0	10.0	22.0	13.0	17.0	33.0
BUNNY	15.0	8.0	23.0	6.0	7.2	8.1	10.0	7.8	8.2	12.0

(a) Notre méthode universelle.

mean: 11.78

POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
6.3	3.8	6.0	5.6	7.9	6.1	5.6	8.1	6.7	9.8
12.6	6.3	13.1	5.1	9.4	6.9	7.1	7.2	8.8	11.1
10.0	6.8	10.1	6.5	7.9	6.4	9.4	7.5	10.4	14.3
14.9	5.3	20.2	4.6	7.6	7.3	9.0	8.9	7.8	23.6
18.5	6.1	22.3	6.1	10.6	5.9	11.8	7.0	10.3	23.3
17.9	7.9	24.8	8.4	10.5	8.6	17.4	14.9	19.1	21.9
14.5	7.3	19.4	6.0	9.6	5.2	12.2	6.3	11.3	25.9
18.0	7.2	29.6	6.4	8.5	6.4	16.9	9.1	10.0	17.4
29.4	9.6	32.4	9.9	14.1	10.1	24.8	14.0	19.0	25.2
21.6	6.4	36.7	5.5	8.0	6.3	8.2	6.3	6.9	10.6

(b) Notre méthode faiblement calibrée .

mean: 11.01

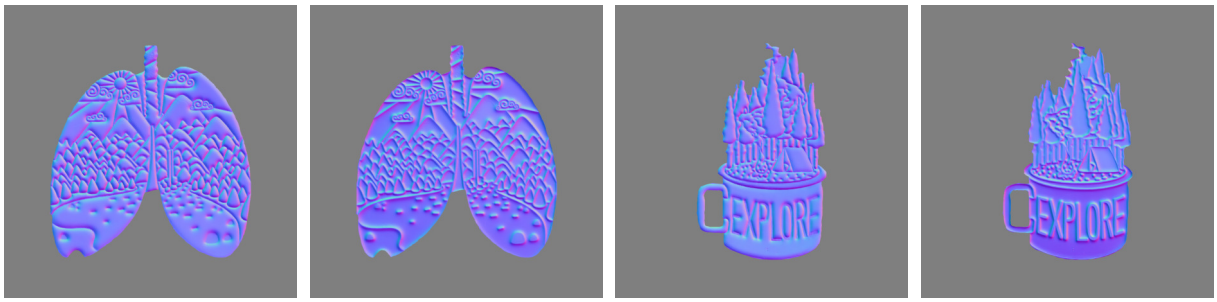
	POM	PP	NYLON	PVC	ABS	BAKELITE	AI	CU	STEEL	ACRYLIC
BALL	4.5	3.3	5.0	3.6	5.8	4.1	3.6	7.8	8.8	6.4
GOLF	13.0	6.4	14.0	5.1	11.0	6.8	7.0	6.4	8.1	9.3
SPIKE	10.0	7.3	11.0	7.5	9.1	8.3	8.5	8.3	9.0	11.0
NUT	11.0	5.0	19.0	4.5	8.1	5.0	6.6	6.8	6.4	24.0
SQUARE	18.0	8.5	23.0	7.7	13.0	7.1	7.9	5.0	7.6	19.0
PENTAGON	15.0	8.3	22.0	8.9	13.0	8.4	11.0	9.5	9.5	21.0
HEXAGON	16.0	5.8	20.0	5.9	12.0	6.4	7.1	5.2	6.7	20.0
PROPELLER	16.0	9.2	32.0	8.2	9.6	6.9	15.0	7.8	10.0	17.0
TURBINE	30.0	9.4	30.0	10.0	19.0	9.8	22.0	15.0	16.0	23.0
BUNNY	12.0	8.0	29.0	6.0	8.0	8.3	9.4	8.5	8.4	13.0

(c) Notre méthode calibrée.

TABLE 6.2 – Erreur angulaire en degrés sur la base de données *DiLiGenT10*[2]. Les trois méthodes présentées ont toutes les trois le même type d'architecture *Transformers* multi-échelles présentée dans le chapitre 5 et modifiée si nécessaire pour le contexte calibré, universel et faiblement calibrée.

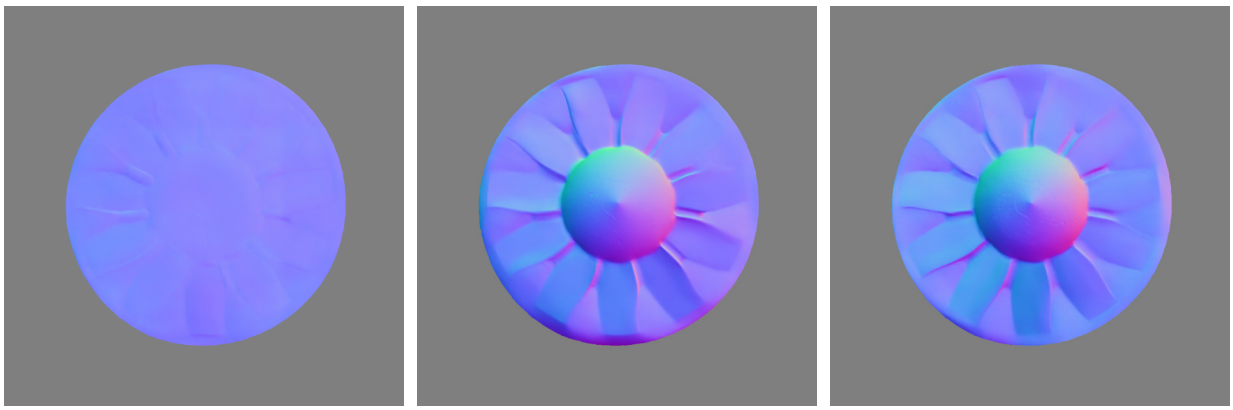
	Astro Lion-R Queen	Bagua-R Lion-T Rhino	Bagua-T Lions Sail	Bear Lotus-R Ship	Bird Lotus-T Sun	Cloud-R Lung TV	Cloud-T Ocean Taichi	Crab Panda-R Tree	Fish Panda-T Wave	Flower Para Whale	moyenne
Calibré	6.03	9.57	11.75	6.72	6.55	12.61	11.01	5.75	4.11	4.85	
	13.12	11.43	5.37	10.17	8.09	**5.41**	5.44	12.98	11.39	4.73	7.75
	5.69	5.22	6.66	6.25	5.9	10.24	7.26	**6.08**	5.48	6.71	
Semi calibré	7.98	9.17	12.71	8.47	6.51	12.46	12.25	7.56	7.18	7.27	
	12.64	11.76	5.98	9.98	9.68	**10.03**	6.42	12.6	11.67	5.9	9.12
	6.74	5.78	6.55	9.31	7.27	11.04	8.56	**8.59**	6.13	15.31	
Universelle	7.14	10.43	11.69	14.09	7.35	13.08	11.92	5.32	5.96	5.14	
	12.73	11.2	6.16	11.51	10.39	**41.98**	5.91	13.28	12.22	7.13	11.38
	9.54	6.68	6.62	5.65	6.05	11.5	8.95	**47.15**	5.93	8.77	

TABLE 6.3 – Erreur angulaire moyenne (en degrés) sur *DiLiGenT-Pi* [95]. Les trois méthodes présentées ont toutes les trois le même type d'architecture *Transformers* multi-échelles présentée dans le chapitre 5 et modifiée si nécessaire pour le contexte calibré, universel et faiblement calibrée. Les valeurs en gras représentent celles obtenues pour les objets *"Lung"* et *"Tree"*. On peut voir que le faiblement calibré a permis de retrouver des normales orientées dans le bon sens.



"Lung" Universel. | *"Lung"* faiblement calibré. | *"Tree"* Universel. | *"Tree"* faiblement calibré.

FIGURE 6.7 – Résultats obtenus sur les objets *"Lung"* et *"Tree"* de la base de données *DiLiGenT-Pi* [95]. On peut voir que le faiblement calibré résout la problématique de l'inversion des normales sur ces deux objets très difficiles.



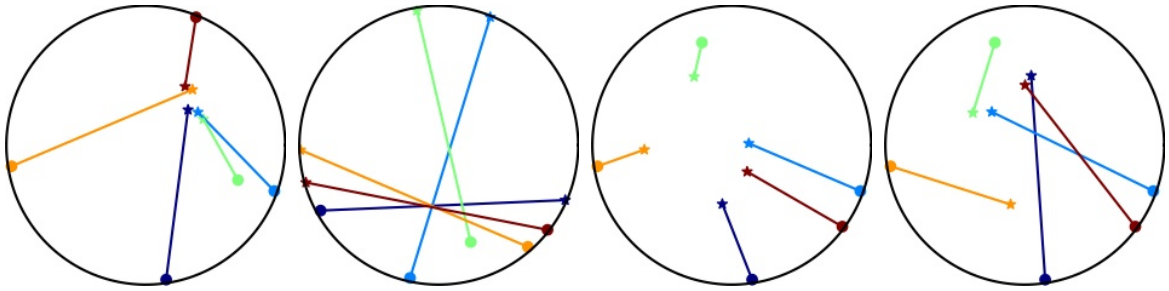
Universel. | Faiblement calibré. | Calibré.

FIGURE 6.8 – Résultats obtenus sur l'objet Turbine en plexiglas de la base de données *DiLiGenT-10*^{2} [80]. On peut voir que le faiblement calibré résout la problématique rencontrée par le non-calibré. La reconstruction est de bien meilleure qualité.

le problème d'inversion des normales obtenu avec notre méthode universelle Uni-MS-PS. Cette solution peu contraignante permet de résoudre l'inversion des normales.

Par ailleurs, nous avons également testé notre méthode sur un matériau translucide où la reconstruction de la carte des normales est difficile. Par exemple sur la *"Turbine"* en plexiglas de la base de données *DiLiGenT10*^{2} [80] affichée sur la figure 6.8.

Comme on peut le voir, notre méthode faiblement calibrée améliore grandement le résultat obtenus avec notre méthode universelle pour se rapprocher de la version calibrée de notre architecture Uni-MS-PS, mais les résultats restent perfectibles. En effet, si l'on regarde d'un peu plus près les estimations des directions lumineuses obtenues sur cet objet, présentées sur la figure figure 6.9, on remarque que l'estimation semble "aléatoire".



UPS-GCNet [16]. Dani-Net [61]. Uni-MS-PS-Reg. Uni-MS-PS-Cls.

FIGURE 6.9 – Sur les objets translucides comme ici sur l'objet *"Turbine"* de *DiLiGenT10*^{2} [80], l'estimation des directions est difficile. L'estimation est en effet bien différente en fonction de la méthode employée.

6.5 Conclusion

Motivés par la difficulté à estimer la direction lumineuse pour les matériaux réfléchissants ou translucides, nous avons proposé une nouvelle approche pour la stéréophotométrie, appelée stéréophotométrie faiblement calibrée. Cette nouvelle méthode a pour avantage de ne pas avoir besoin d'une direction précise, seulement d'une indication approximative de la lumière pour reconstruire la carte des normales, permettant ainsi de nous libérer des contraintes exigeantes du cadre calibré, en indiquant une direction lumineuse vague, sans contraintes d'environnement contrôlé.

Les résultats obtenus sont très prometteurs. En effet, lorsque l'on compare cette approche aux approches calibrées et non calibrées/universelles, on remarque que l'on améliore significativement les résultats obtenus dans un contexte universel en se rapprochant du cas calibré. Tout cela, sans les contraintes imposées dans le cas calibré.

CHAPITRE 7

Conclusion : bilan et perspectives

7.1 Bilan

Dans cette thèse, nous nous sommes intéressés au problème de reconstruction 3D de la surface d'un objet par stéréophotométrie en utilisant des méthodes d'apprentissage profond. Les deux principaux défis de ce problème sont d'obtenir des cartes de normales les plus précises possibles afin de retrouver l'ensemble des détails de l'objet mais également de pouvoir considérer des images de toutes résolutions pour conserver l'ensemble de ces détails. Ainsi, nos contributions concernent à la fois les aspects sur les bases de données et sur les architectures multi-échelles originales et performantes de type encodeur-décodeur.

Notre première contribution a été la création automatisée d'une nouvelle base de données d'entraînement synthétique et universelle. En effet, les bases de données d'entraînement pour la stéréophotométrie sont rares ou peu variées. Cette première contribution présente une très grande variété d'objets, de directions lumineuses, d'environnements et de matériaux. Les tests effectués sur l'état-de-l'art montrent l'importance de la base de données d'entraînement mais également l'apport de la diversité des matériaux, des sources lumineuses, des objets et des environnements pour l'amélioration des performances. Malgré le caractère synthétique de la base ainsi générée, nous avons pu montrer qu'elle permet d'améliorer significativement les performances des réseaux de l'état de l'art pour l'estimation des normales.

Suite à ces premiers résultats vis-à-vis de la base de données, notre seconde contribution s'est concentrée sur l'élaboration d'une architecture idéale afin de conserver l'ensemble des détails présents dans les images en conservant la résolution native de celles-ci. Ainsi, dans un contexte de stéréophotométrie calibrée, nous avons proposé une architecture multi-échelles de type encodeur-décodeur. Les avantages de cette architecture sont sa capacité à gérer les images très hautes résolution à l'aide du multi-échelles et de propager l'ensemble des détails présents grâce au caractère encodeur-décodeur de notre architecture. Combinée avec notre base de données, cette contribution répond aux deux principaux enjeux de la stéréophotométrie vue dans le premier paragraphe de ce chapitre. Cette méthode donne les résultats de l'état-de-l'art dans le domaine de la stéréophotométrie calibrée.

Afin de se libérer du cadre calibré et par conséquent des contraintes d'étalonnage de la lumière et de contrôle de l'environnement, nous avons élaboré une architecture pour la stéréophotométrie universelle. Cette troisième contribution est une extension de notre seconde contribution car elle

reprend la caractère multi-échelles de type encodeur-décodeur. Même si la base de l'achitecture choisi n'est plus convolutif mais de type *PVT* (*Pyramid Vision Transformer*).La où les réseaux convolutifs deviennent insuffisants par leur champ d'action local. Nous avons montré que les *Transformers* sont parfaits pour capturer les similarités inter et intra images pour mieux estimer de manière implicite l'environement lumineux. Notre méthode donne les résultats de l'état-de-l'art à la fois quantitatifs et visuels sur l'ensemble des bases de données publiques avec ou sans vérités terrains.

Finalement, afin de reconstruire le maximum de détails et trouver un compromis entre calibré et universel, nous avons proposé une nouvelle approche hybride, appelée "faiblement calibrée", où la connaissance de la position lumineuse est approximative. Nous avons testé cette approche sur nos deux architectures et nous avons montré que nous pouvons gagner en performance sur des objets avec une réflectance très complexe par rapport au cadre universel, tout en relaxant la contrainte de la position lumineuse.

7.2 Perspective court terme

Les méthodes de stéréophotométrie classiques ne permettent pas encore d'obtenir une reconstruction complète d'un objet en 3D (i.e. une reconstruction à 360°). En effet, aussi bonne la prédiction des normales soit elle, la stéréophotométrie n'exploite qu'une seule vue de l'objet. Il est ainsi impossible de reconstruire en 3D les zones invisibles de l'objet sur la vue en question. En revanche, il est toujours possible d'intégrer mathématiquement la carte de normales pour obtenir une partie de l'objet en 3D sous la forme d'une carte d'élévation, comme illustré en figure 7.1. Lorsque les performances de prédiction des normales sont bonnes, les résultats d'une intégration permet d'obtenir des résultats concluants, notamment en retrouvant de petits détails comme nous pouvons le voir sur les objets en 3D de la figure 7.2. Cependant, les méthodes d'intégration mono-vues ont également des inconvénients, autres que l'impossibilité de reconstruire l'objet en entier. En effet, elles sont très sensibles au bruit par exemples. et aux sauts de discontinuité de profondeur. De telles problématiques pourraient être résolues en utilisant plusieurs vues de l'objet, comme le fait traditionnellement la stéréovision.

Ainsi, une perspective intéressante à cette thèse serait d'utiliser les cartes de normales prédites par stéréophotométrie sur toutes les faces de l'objet indépendamment puis d'utiliser un algorithme de reconstruction 3D robuste multi-vues fondé sur les normales.

À des fins d'exploration, nous avons utilisé la méthode de reconstruction 3D RNb-NeuS [10], fondée sur NeuS [96]. L'objectif ici est de montrer l'intérêt d'une approche multi-vues utilisant des cartes de normales, au lieu d'une approche multi-vues exploitant directement les images RGB (comme la stéréovision classique), mais également de montrer le niveau de détails captés par une telle approche fondée sur les normales.
Dans notre cas, les cartes de normales seront générées par notre méthode de stéréophotométrie universelle décrite dans le chapitre 5. RNb-Neus a besoin pour fonctionner non seulement des cartes de normales mais également des positions des caméras, de leurs focales, etc. Pour obtenir ces paramètres nous utilisons le logiciel *Meshroom* [30]. Finalement, pour permettre une comparaison adéquate, les résultats obtenus sont comparés à la méthode NeuS [96].

Comme nous pouvons le voir dans les exemples de reconstruction affichés sur les figures 7.3,7.4,7.5, l'utilisation des cartes de normales (RNb-Neus) au lieu des images brutes (Neus) permet d'obtenir un niveau de détails dans la reconstruction 3D très satisfaisant. Les écailles de l'escargot (figure 7.3) sont bien plus visibles et détaillés avec la méthode exploitant les cartes de normales, de même pour les pattes de la grenouille (figure 7.4), ainsi que la devanture du jouet (figure 7.5).

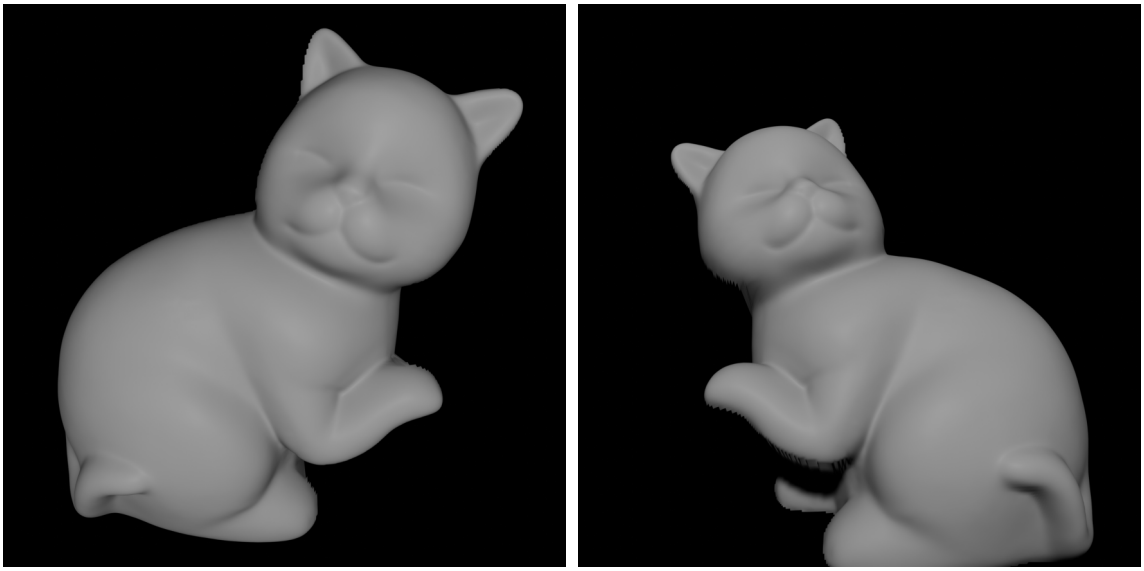


FIGURE 7.1 – Reconstruction 3D mono-vue de l'objet "*cat*" de *DiLiGenT* [87] à l'aide la méthode [13]. Bien que cette reconstruction ait été faite avec une seule et unique vue, la reconstruction 3D est de bonne qualité. Cependant, seul cette face du chat peut être reconstruite, son dos ne peut l'être.

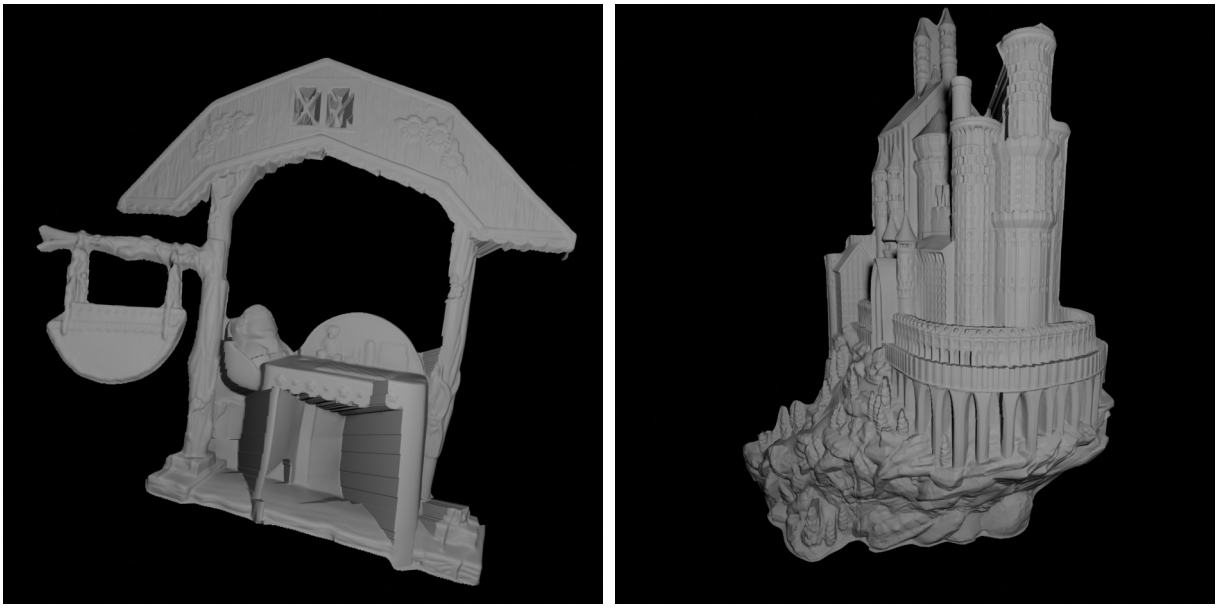
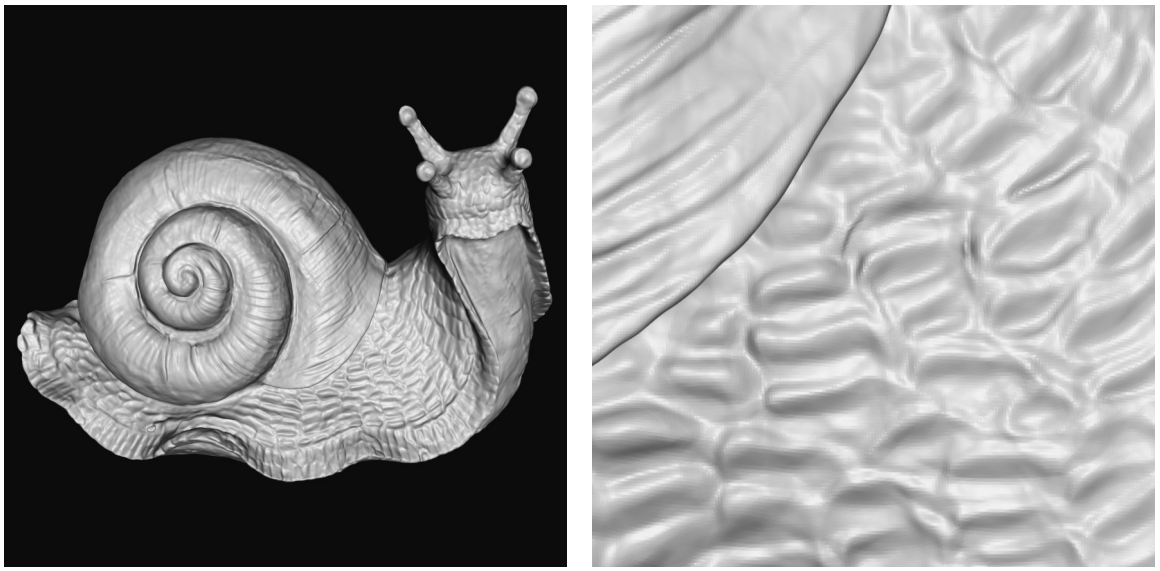


FIGURE 7.2 – Reconstruction 3D mono-vue de des objets "*pink-toy-park*" et "*white-castle-towers*" du jeu de données [93] à l'aide la méthode [13]. L'intégration de la carte de normales sur des formes plus complexe, permet d'obtenir une reconstruction détaillée des objets. Les zones non visibles depuis la vue de la caméra comme sous la table de l'objet de gauche ne peuvent être correctement reconstruit.

Une extension possible de notre travail pour la suite pourrait donc être d'essayer de combiner directement la stéréophotométrie et une méthode de type Neus pour obtenir directement un maillage 3D. Une telle méthode permettrait de combiner les avantages de deux approches en une méthode *end to end*. Par exemple, il pourrait être intéressant en premier test de se baser sur les travaux comme *Neural Target Object 3D Reconstruction with Segment Anything* [99] ou *MeshLRM* [98] en remplaçant les images en entrée par un réseau qui estime les cartes de normales.

Images RGB de l'objet.

Neus [96].



RnB Neus [10].

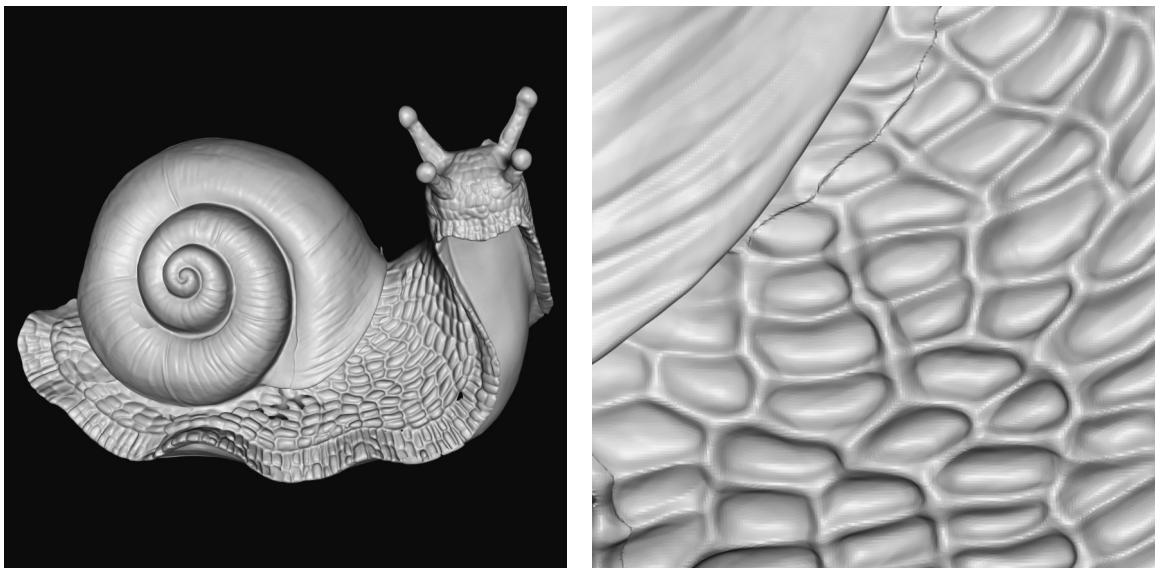
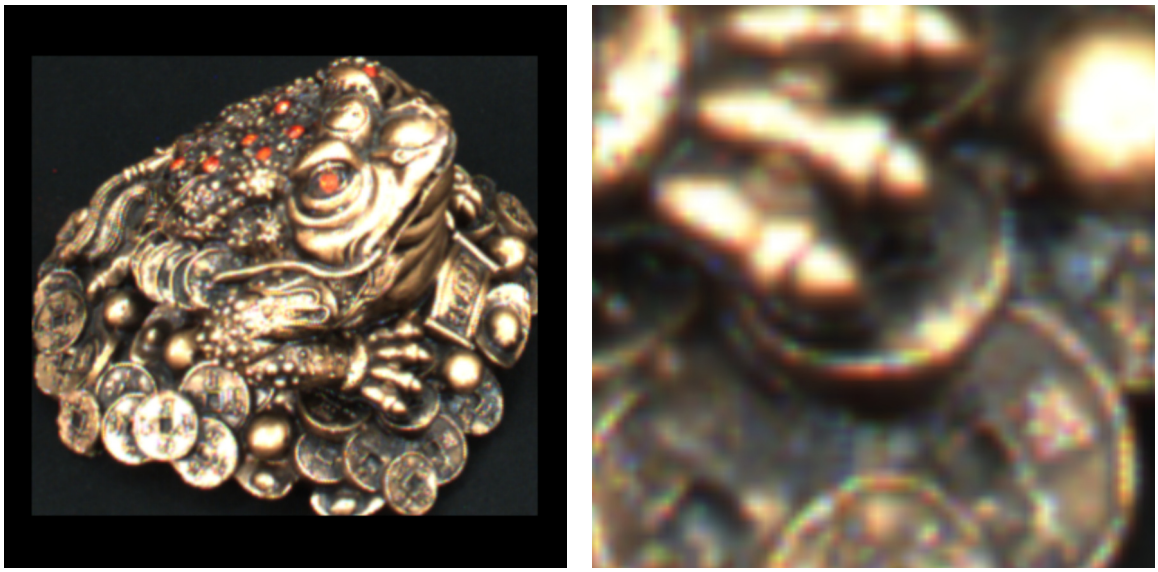
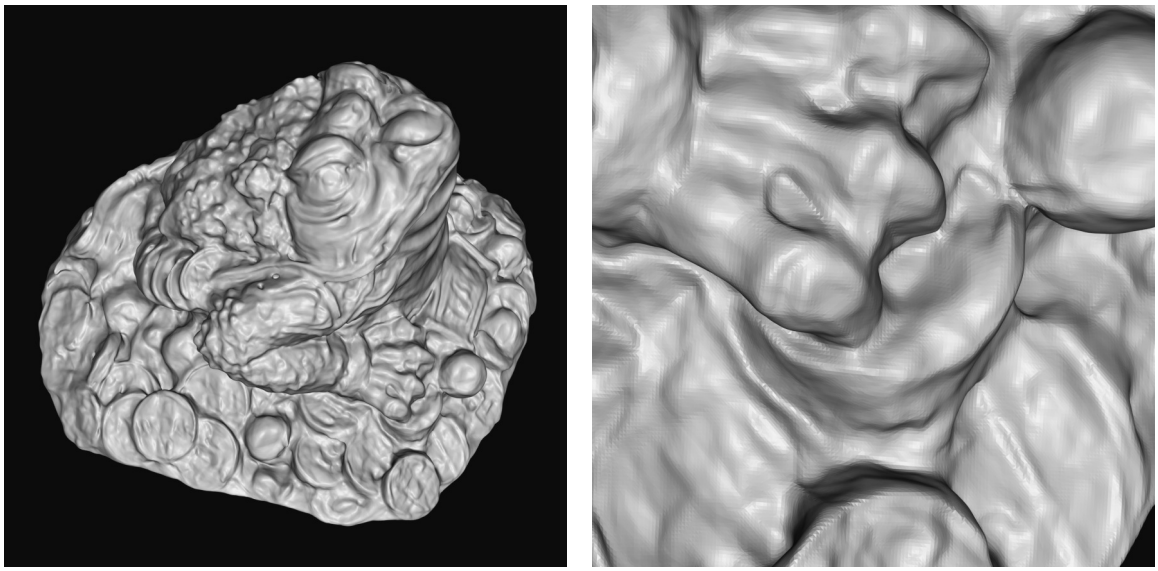


FIGURE 7.3 – Reconstruction 3D multi-vues de l'objet "*golden-snail*" du jeu de données [93]. RNb-Neus permet d'obtenir une reconstruction 3D beaucoup plus détaillée, précise et correcte des objets comparativement à Neus. La coquille et les écailles sont bien mieux reconstruites avec RNb-Neus.

Images RGB de l'objet.



Neus [96].



RnB Neus [10].

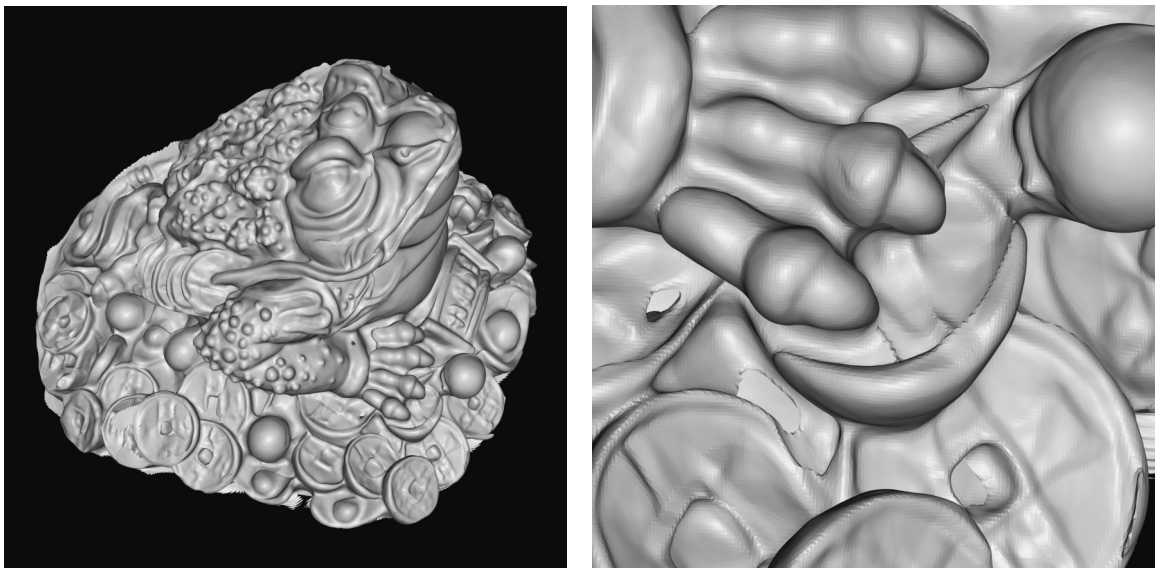


FIGURE 7.4 – Reconstruction 3D multi-vues de l'objet "*jin-chan*" du jeu de données [93]. RNb-Neus permet d'obtenir une reconstruction 3D beaucoup plus détaillée, précise et correcte des objets comparativement à Neus. Les pattes et pièces de monnaies sont notamment beaucoup mieux reconstruites.

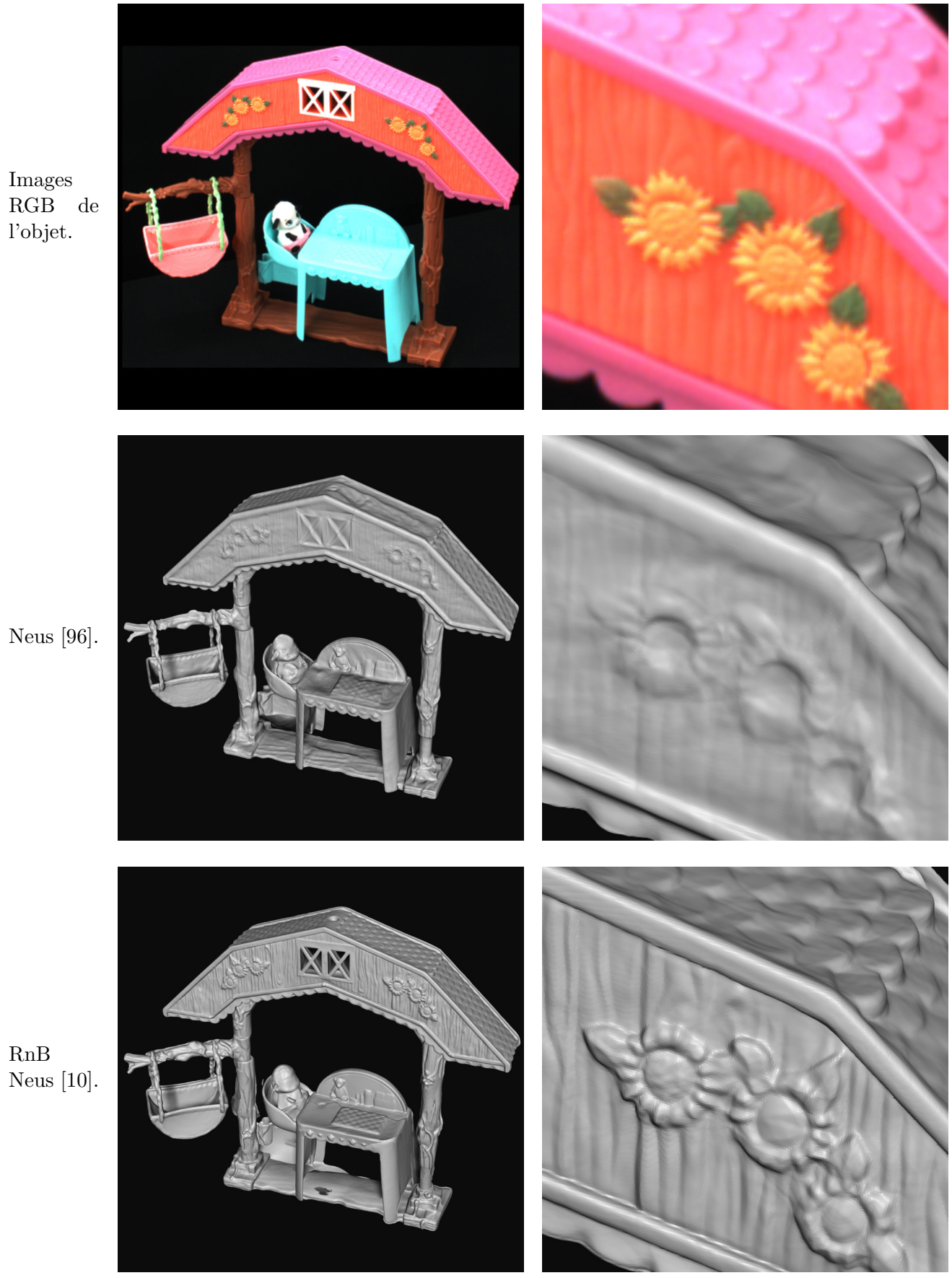


FIGURE 7.5 – Reconstruction 3D multi-vues de l'objet "*pink-toy-arc*" du jeu de données [93]. RNb-Neus permet d'obtenir une reconstruction 3D beaucoup plus détaillée, précise et correcte des objets comparativement à Neus.

7.3 Perspectives moyen et long terme

- Une seconde perspective peut être de combiner plusieurs modalités pour obtenir une reconstruction 3D plus précise et cohérente. En effet, l'intégration à partir d'une seule vue basée sur la carte des normales peut engendrer une propagation de l'erreur lors de l'estimation de celle-ci. Même une erreur infime sur une partie de la carte des normales peut se répercuter sur l'ensemble de la reconstruction. Pour illustrer ce phénomène, nous allons prendre en exemple une acquisition réalisée sur les hauts-reliefs de la caserne de Metz. Comme on peut le voir à la figure 7.6, la carte des normales est très détaillée et précise. À première vue, l'intégration semble également de très bonne qualité et sans défaut apparent (voir figure 7.7). Cependant, la courbure de l'intégration n'est pas correcte, comme on peut l'observer sur la figure 7.8. Pour résoudre cette problématique, une piste pourrait être d'intégrer une autre modalité, comme par exemple quelques points de référence acquis avec un LiDAR. Des méthodes utilisant une carte de profondeur comme a priori ont déjà été développées, notamment par Quéau *et al.*[75] ou encore par Cao *et al.*[13]. La combinaison de plusieurs approches a déjà fait ses preuves, comme démontré dans [48]. Une première possibilité de travail est de s'appuyer sur les travaux de Li *et al.* [60] en utilisant des modèles plus modernes, ainsi qu'en adaptant la méthode à un contexte universel.
- Pour finir, à plus long terme, il pourrait être intéressant de développer des méthodes permettant de reconstruire en 3D de grandes scènes, et non plus de se limiter à des objets comme cela peut être fait à l'aide de la photogrammétrie.

FIGURE 7.6 – Reconstruction de la carte des normales des hauts-reliefs de la caserne de Metz à l'aide de notre méthode universelle développée dans le chapitre 5.

FIGURE 7.7 – Intégration de la carte des normales de la figure 7.6 à l'aide de la méthode proposé par Cao *et al.* [13], sans aucun a priori sur la profondeur. La reconstruction 3D est globalement très détaillée.

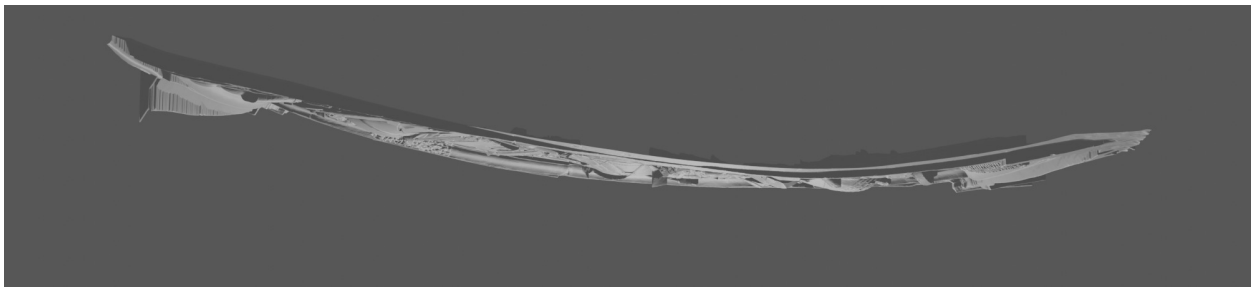


FIGURE 7.8 – Intégration de la carte des normales de la figure 7.6 à l'aide de la méthode proposée par Cao *et al.* [13], sans aucun a priori sur la profondeur. Nous pouvons observer que la reconstruction 3D est courbée (vue du dessus du haut-relief), ce qui ne devrait pas être le cas.

Bibliographie

[1] Adobe stock. `https://stock.adobe.com`.

[2] Alexandre Duret-Lutz. `https://www.flickr.com/people/gadl/`.

[3] AmbientCG. `https://ambientcg.com/`.

[4] MyMiniFactory. `https://www.myminifactory.com/fr`.

[5] Poly Haven. `https://polyhaven.com`.

[6] Scan the world. `https://www.myminifactory.com/scantheworld/`.

[7] Sketchfab. `https://sketchfab.com`.

[8] A. F. Agarap. Deep learning using rectified linear units (relu). *ArXiv*, 2018.

[9] S. Barsky and M. Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[10] Baptiste Brument, Robin Bruneau, Yvain Quéau, Jean Mélou, François Bernard Lauze, Jean-Denis Durou, and Lilian Calvet. Rnb-neus : Reflectance and normal-based multi-view 3d reconstruction. 2024.

[11] F. Bruno, A. Lagudi, L. Barbieri, M. Muzzupappa, M. Mangeruga, M. Cozza, A. Cozza, G. Ritacco, and R. Peluso. Virtual reality technologies for the exploitation of underwater cultural heritage. *Latest developments in reality-based 3D surveying and modelling*, 2018.

[12] B. Burley and W. D. Studios. Physically-based shading at disney. *ACM Transactions on Graphics Courses (SIGGRAPH)*, 2012.

[13] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[14] Y. Cao, B. Ding, Z. He, J. Yang, J. Chen, Y. Cao, and X. Li. Learning inter-and intraframe representations for non-lambertian photometric stereo. *Optics and Lasers in Engineering (OLEN)*, 2022.

[15] G. Chen, K. Han, B. Shi, Y. Matsushita, and K. K. Wong. Deep photometric stereo for non-lambertian surfaces. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[16] G. Chen, K. Han, Y. Shi, B. ANDMatsushita, and K. Wong. Deep photometric stereo for non-lambertian surfaces. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[17] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K. K. Wong. Self-calibrating deep photometric stereo networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Guanying Chen, Kai Han, Kwan-Yee K. Wong, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Ps-fcn : A flexible learning framework for photometric stereo. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2018.

[19] E. R. Coleman and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing.*

[20] Blender Online Community. *Blender - a 3D modelling and rendering package*, 2018.

[21] Julie Delon and Andrés Almansa. *Inverse Problems in Vision and 3D Tomography.*

[22] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (SIGGRAPH)*, 2018.

[23] Cycles Developers. *Cycles is a physically based production renderer developed by the Blender project*, 2022.

[24] D. Ding, W. Ding, R. Huang, Y. Fu, and F. Xu. Research progress of laser triangulation on-machine measurement technology for complex surface : A review. *Measurement*, 2023.

[25] J.G.D.M. Franca, M.A. Gazziro, A.N. Ide, and J.H. Saito. A 3d scanning system based on laser triangulation and variable field of view. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2005.

[26] J. Geng. *Advances in Optics and Photonics.*

[27] A. Georghiades. Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2008.

[28] D.B. Goldman, B. Curless, A. Hertzmann, and S.M. Seitz. Shape and spatially-varying BRDFs from photometric stereo. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2005.

[29] M.A. Gomarasca. Basics of geomatics. 2009.

[30] C. Griwodz, S. Gasparini, L. Calvet, P. Gurdjos, F. Castan, B. Maujean, G. De Lillo, and Y. Lanthony. Alicevision meshroom : An open-source 3d reconstruction pipeline. *Proceedings of the ACM Multimedia Systems Conference (MMSys)*, 2021.

[31] C. Hardy, Y. Quéau, and D. Tschumperlé. Construction d'un jeu de données d'apprentissage adapté pour la reconstruction 3d par stéophotométrie. *Groupe de Recherche en Traitement du Signal et des Images (GRETSI)*, 2022.

[32] C. Hardy, Y. Quéau, and D. Tschumperlé. Uni ms-ps : A multi-scale encoder-decoder transformer for universal photometric stereo. *Computer Vision and Image Understanding (CVIU)*, 2024.

[33] Clément Hardy, Yvain Quéau, and David Tschumperlé. MS-PS : A Multi-Scale Network for Photometric Stereo With a New Comprehensive Training Dataset. *Proceedings of the International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2023.

[34] R. Harrap and M. Lato. An overview of LIDAR : collection to application. *NGI publication*, 2010.

[35] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. 2012.

[36] R. J. Hocken and P. H. Pereira. Coordinate measuring machines and systems. 2012.

[37] D. Honzátko, E. Türetken, P. Fua, and L. Dunbar. Leveraging Spatial and Photometric Context for Calibrated Non-Lambertian Photometric Stereo. *Proceedings of the International Conference on 3D Vision (3DV)*, 2021.

[38] B. K. P. Horn. Shape from shading : A method for obtaining the shape of a smooth opaque object from one view. *MIT Artificial Intelligence Laboratory*, 1970.

[39] K. Hu, T. Wang, C. Shen, C. Weng, F. Zhou, M. Xia, and L. Weng. Overview of underwater 3d reconstruction technology based on optical images. *Journal of Marine Science and Engineering*, 2023.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[41] S. Ikehata. PS-transformer : Learning sparse photometric stereo network using self-attention mechanism. *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[42] S. Ikehata. Universal photometric stereo network using global lighting contexts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] S. Ikehata. Scalable, Detailed and Mask-free Universal Photometric Stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[44] S. Ikehata, D. Wipf, Y. Matsushita, and Kiyoharu Aizawa. Photometric stereo using sparse bayesian regression for general diffuse surfaces. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[45] Satoshi Ikehata. Cnn-ps : Cnn-based photometric stereo for general non-convex surfaces. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2018.

[46] Y. Ju, Y. Peng, M. Jian, F. Gao, and J. Dong. Learning conditional photometric stereo with high-resolution features. *Computational Visual Media (CVM)*, 2022.

[47] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam. Normattention-psn : A high-frequency region enhanced photometric stereo network with normalized attention. *International Journal of Computer Vision*, 2022.

[48] Ali Karami, Fabio Menna, and Fabio Remondino. Combining photogrammetry and photometric stereo to achieve precise and complete 3d reconstruction. *Sensors*, 2022.

[49] M. Kasser and Y. Egels. Photogrammétrie numérique. *Hermès-sciences*, 2001.

[50] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and Van G. Uncalibrated neural inverse rendering for photometric stereo of general surfaces. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[51] D. P. Kingma and J. Ba. Adam : a method for stochastic optimization. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[52] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai. An image is worth 16x16 words : Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[53] G. Konecny. The international society for photogrammetry and remote sensing-75 years old, or 75 years young. 1985.

[54] K. Kraus and P. Waldhäusl. Manuel de photogrammétrie, principes et procédés fondamentaux. *Hermès-sciences*, 1998.

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2012.

[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324*, 1998.

[57] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set Transformer : A Framework for Attention-based Permutation-Invariant Neural Networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

[58] J. Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. *Proceedings of the IEEE/CVF European Conference on Computer Vision (EECV)*, 2022.

[59] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita. Learning to minify photometric stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo : A robust solution and benchmark dataset for spatially varying isotropic materials. *Proceedings of the IEEE Transactions on Image Processing (TIP)*, 2020.

[61] Z. Li, Q. Zheng, B. Shi, G. Pan, and X. Jiang. Dani-net : Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[62] D. Lichy, S. Sengupta, and D. W. Jacobs. Fast light-weight near-field photometric stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[63] D. Lichy, J. Wu, S. Sengupta, and D. W. Jacobs. Shape and material capture at home. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[64] Y. Liu, Y. Ju, M. Jian, F. Gao, Y. Rao, Y. Hu, and J. Dong. A deep-shallow and global–local multi-feature fusion network for photometric stereo. *Image and Vision Computing*, 2022.

[65] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla. PX-net : Simple and efficient pixel-wise training of photometric stereo networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[66] Fotios Logothetis, Roberto Mecca, Ignas Budvytis, and Roberto Cipolla. A cnn based approach for the point-light photometric stereo problem. *Proceedings of the International Journal of Computer Vision (IJCV)*, 2022.

[67] W. Lorensen and H. Cline. Marching cubes : A high resolution 3D surface construction algorithm. *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques conference (SIGGRAPH)*, 1987.

[68] S. A.. Marhon, C. J. F. Cameron, and S. C. Kremer. Recurrent neural networks. *Handbook on Neural Information Processing*, 2013.

[69] R. Martin. Notions de photogrammétrie. *Editions Eyrolles*, 1985.

[70] W. Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003.

[71] J. McCarthy, J. Benjamin, T. Winton, and W. Van Duivenvoorde. The rise of 3d in maritime archaeology. *3D Recording and Interpretation for Maritime Archaeology*, 2019.

[72] R. Mecca, F. Logothetis, I. Budvytis, and R. Cipolla. Luces : A dataset for near-field point light source photometric stereo. *Proceedings of the British Machine Vision Conference (BMVC)*, 2021.

[73] J. Mélou, A. Laurent, C. Fritz, and J.-D. Durou. 3d digittization of heritage : Photometric stereo can help. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022.

[74] Nicolas Prouteau, Clément Joubert, Benjamin Bringier, and Majdi Khoudeir. Continuous material reflectance map for deep photometric stereo. *Journal of the Optical Society of America A*, 2023.

[75] Y. Quéau, J. Durou, and J. Aujol. Variational Methods for Normal Integration. *Journal of Mathematical Imaging and Vision*, 2018.

[76] Y. Quéau, T. Wu, F. Lauze, J.-D. Durou, and Daniel Cremers. A Non-convex Variational Approach to Photometric Stereo under Inaccurate Lighting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[77] Yvain Quéau, Jean-Denis Durou, and Jean-François Aujol. Normal integration : A survey. *Journal of Mathematical Imaging and Vision*, 2018.

[78] T. Raj, F. H. Hashim, A. B. Huddin, M. F. Ibrahim, and A. Hussain. A survey on lidar scanning mechanisms. *Electronics*, 2020.

[79] J. Ren, X. Wang, Z. Jian, and M. Ren. Multiscale Convolutional Fusion Network for Non-Lambertian Photometric Stereo. *Proceedings of the IEEE Signal Processing Letters*, 2020.

[80] Jieji Ren, Feishi Wang, Jiahao Zhang, Qian Zheng, Mingjun Ren, and Boxin Shi. Diligent102 : A photometric stereo benchmark dataset with controlled shape and material variation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[81] H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 1951.

[82] O. Ronneberger, P. Fischer, and T. Brox. U-net : Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

[83] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo network. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[84] L. E. Santos Araújo Filho, G. P. Crestani, C. L. Nascimento Júnior, P. Daniel de Cerqueira Gava, J. R. Belchior de França Silva, T. M. Mancilha, W. Vieira, and G. J. Adabo. 3d reconstruction of a small dam using a profiling sonar and an uuv. *Proceedings of the IEEE International Systems Conference (SysCon)*, 2022.

[85] J. Schlarp, E. Csencsics, and G. Schitter. Optical scanning of a laser triangulation sensor for 3-d imaging. *IEEE Transactions on Instrumentation and Measurement*, 2020.

[86] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2013.

[87] B. Shi, Z. Wu, Z. Mo, D. Duan, S. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[88] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[89] T. Taniai and T. Maehara. Neural Inverse Rendering for General Reflectance Photometric Stereo. *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[90] S. Tozza, R. Mecca, M. Duocastella, and A. Del Bue. Direct Differential Photometric Stereo Shape Recovery of Diffuse and Specular Surfaces. *Journal of Mathematical Imaging and Vision*, 2016.

[91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[92] M.M.P.A. Vermeulen, P.C.J.N. Rosielle, and P.H.J. Schellekens. Design of a high-precision 3d-coordinate measuring machine. *CIRP Annals*, 1998.

[93] Oleg Voynov, Gleb Bobrovskikh, Pavel Karpyshev, Saveliy Galochkin, Andrei-Timotei Ardelean, Arseniy Bozhenko, Ekaterina Karmanova, Pavel Kopanev, Yaroslav Labutin-Rymsho, Ruslan Rakhimov, Aleksandr Safin, Valerii Serpiva, Alexey Artemov, Evgeny Burnaev, Dzmitry Tsetserukou, and Denis Zorin. Multi-sensor large-scale dataset for multi-view 3d reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[94] U. Wandinger. Introduction to lidar. *Lidar : Range-Resolved Optical Remote Sensing of the Atmosphere*.

[95] F. Wang, J. Ren, H. Guo, M. Ren, and B. Shi. DiLiGenT-Pi : Photometric Stereo for Planar Surfaces with Rich Details - Benchmark Dataset and Beyond. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[96] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. Neus : Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2021.

[97] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer : A versatile backbone for dense prediction without convolutions. *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021.

[98] X. Wei, K. Zhang, S. Bi, H. Tan, F. Luan, V. Deschaintre, K. Sunkavalli, H. Su, and Z. Xu. Meshlrm : Large reconstruction model for high-quality mesh. *arXiv preprint arXiv :2404.12385*, 2024.

[99] X. Wei, R. Zhang, J. Wu, J. Liu, M. Lu, Y. Guo, and S. Zhang. Nto3d : Neural target object 3d reconstruction with segment anything. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[100] R. J. Woodham. Photometric stereo : A reflectance map technique for determining surface orientation from image intensity. *Image Understanding Systems and Industrial Applications I*, 1979.

[101] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 1980.

[102] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2011.

[103] T.-P. Wu and C.-K. Tang. Photometric Stereo via Expectation Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[104] Z. Yao, K. Li, Y. Fu, H. Hu, and B. Shi. Gps-net : Graph-based photometric stereo network. *Proceedings of the Neural Information Processing Systems (NIPS)*, 2020.

[105] C. Yu and S. W. Lee. Deep Photometric Stereo Network with Multi-Scale Feature Aggregation. *Sensors*, 2020.

[106] Z. Zhang, Y. Chen, and V. Saligrama. Efficient Training of Very Deep Neural Networks for Supervised Hashing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[107] S. Zhao, F. Kang, J. Li, and C. Ma. Structural health monitoring and inspection of dams based on UAV photogrammetry with image 3D reconstruction. *Automation in Construction*, 2021.

[108] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L. Duan, and A. Kot. Spline-net : Sparse photometric stereo through lighting interpolation and normal estimation networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[109] Q. Zheng, A. Kumar, B. Shi, and G. Pan. Numerical Reflectance Compensation for Non-Lambertian Photometric Stereo. *IEEE Transactions on Image Processing*, 2019.

[110] Z. Zhuang Liu, H. Mao, C.-H. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.