

Modèles génératifs image vers 3D avec contraintes géométriques et physiques

Contrat Doctoral PR[AI]RIE-PSAI

Mots clés : Vision 3D ; modèle génératif ; image de profondeur ; LiDAR ; fidélité géométrique.

Contexte de la thèse

La thèse proposée se déroulera dans le cadre du projet [PR\[AI\]RIE-PSAI](#) (Paris School of Artificial Intelligence (AI)). Ce projet est le grand lauréat du programme “IA Cluster”, porté par l’[Université PSL](#), et a pour objectif de faire progresser les connaissances en matière d’IA, à proposer un enseignement supérieur de niveau international et à produire des innovations de rupture.

La thèse aura lieu au [Centre de Robotique de Mines Paris - PSL](#). Le Centre de Robotique est un laboratoire de recherche spécialisé dans l’IA temps réel et les interactions Humain-Machine, appliquées aux véhicules automatisés, à la robotique mobile et collaborative, ainsi qu’à l’Industrie du Futur. Au sein de ce centre, l’axe Nuages de Points et Modélisation 3D (NPM3D) développe des techniques d’acquisition, traitement et rendu de nuages de points 3D, issues de la photogrammétrie ou du LiDAR, pour des applications variées (cartographie, robotique, patrimoine, archéologie).

Personnes impliquées dans l’encadrement de la thèse

- **Directeur de thèse : [Jean-Emmanuel DESCHAUD](#), chargé de recherche, HDR, Fellow IA PR[AI]RIE-PSAI, Centre de Robotique de Mines Paris - PSL.**
Jean-Emmanuel DESCHAUD est spécialisé dans le domaine de la Vision 3D à partir d’images, de capteurs RGB-D et LiDAR. Il s’intéresse aux méthodes de rendu et de reconstruction 3D à partir d’images, aux réseaux neuronaux pour nuages de points et plus généralement à la perception 3D pour la robotique et la conduite autonome.
- **Co-directeur de thèse : [Santiago VELASCO-FORERO](#), chargé de recherche, HDR, Fellow IA PR[AI]RIE-PSAI, Centre de Morphologie Mathématique de Mines Paris - PSL.**
Santiago VELASCO-FORERO est spécialisé dans la compréhension et la conception de méthodes pour résoudre des problèmes sur des images, des nuages de points 3D et des graphes. Il s’intéresse plus particulièrement aux méthodes basées sur les opérateurs non linéaires avec une interprétation géométrique et celles basées sur l’apprentissage profond.

Description du sujet de thèse

Au cours des dernières années, la vision 3D est devenue un domaine clé de la vision par ordinateur, jouant un rôle fondamental dans de nombreuses applications telles que la conduite autonome, la robotique, la réalité augmentée (RA) et l’imagerie médicale. Une compréhension approfondie des représentations 3D est en effet indispensable pour l’interprétation et l’interaction avec l’environnement physique, en particulier en robotique. Ainsi, la capacité à estimer une carte de profondeur ou un nuage de points à partir d’une image couleur constitue encore un défi majeur (survey dans [10]).

Des avancées récentes ont considérablement amélioré la génération de données 3D (cartes de profondeur et nuages de points) à partir d’images, comme en témoignent des modèles tels que Depth Anything [11] et Marigold [4]. Ces approches reposent sur des modèles pré-entraînés (DINOv2 [6] pour Depth Anything [11], Stable Diffusion [8] pour Marigold [4]), qui ont été entraînés sur de très vastes jeux de données d’images 2D. Toutefois, ces modèles ne possèdent pas de connaissance explicite du monde 3D et de ses contraintes géométriques et physiques intrinsèques.

L’objectif de cette thèse est de concevoir de nouveaux modèles génératifs 3D pour l’estimation de profondeur et la reconstruction de nuages de points à partir d’images couleur, en intégrant des contraintes géométriques et physiques. Cette approche vise à améliorer la fidélité des données générées dans les régions visibles et à renforcer la cohérence et la plausibilité des zones occultées, permettant ainsi une meilleure qualité des représentations 3D obtenues.

Etat de l’art

Les premières approches d’estimation de profondeur à partir d’images [3] étaient principalement restreintes au domaine sur lequel elles avaient été entraînées, comme les environnements intérieurs. Cependant, avec l’augmentation de la capacité expressive des modèles, de nouvelles méthodes ont émergé, permettant la prédiction de cartes de profondeur ou de représentations 3D in the wild, c’est-à-dire sur des images issues de domaines inconnus lors de l’entraînement.

Une première catégorie de méthodes repose sur des modèles discriminatifs, tels que DINOv2 [6]. Parmi celles-ci, Depth Anything [11] s’appuie sur un passage à l’échelle des données d’entraînement, tandis que Depth Anything V2 [12] exploite une stratégie de pseudo-étiquetage en transférant des labels de données synthétiques vers des données réelles afin d’améliorer leur qualité d’annotation.

Une seconde famille de méthodes s’appuie sur des modèles génératifs, notamment ceux issus de la diffusion, comme Stable Diffusion [8]. Un exemple récent est Marigold [4], qui se distingue par une meilleure capacité à modéliser les détails fins par rapport aux approches discriminatives. Marigold est une méthode basée sur la diffusion permettant une estimation monoculaire de la profondeur sur une grande diversité de jeux de données réels. Contrairement aux approches classiques, elle ne requiert aucune image de profondeur réelle et nécessite seulement d’affiner le U-Net de débruitage sur des données synthétiques, préservant ainsi les riches a priori visuels du modèle pré-entraîné.

La génération de nuages de points à partir d’images est restée longtemps confinée à la reconstruction de petits objets [5, 13]. Un progrès notable a été apporté par GECCO [9], l’une des premières méthodes capables de générer des nuages de points représentant des scènes 3D complètes, tout en assurant une cohérence sémantique avec l’image d’entrée. Contrairement aux méthodes d’estimation de profondeur, GECCO est en mesure de générer des hypothèses plausibles pour les zones occultées. Toutefois, cette approche demeure limitée par une capacité de génération restreinte à 2048 points.

À ce jour, l’intégration de contraintes géométriques reste marginale et se limite généralement à l’exploitation des normales de surface afin d’assurer une meilleure cohérence 3D, comme le propose IronDepth [1]. Cependant, cette approche se restreint à une hypothèse forte de planarité locale des surfaces pour garantir une reconstruction fidèle de la structure sous-jacente des scènes.

LiDAR-Diffusion [7] introduit certains concepts pour améliorer les modèles de diffusion pour la génération de données LiDAR. Ils intègrent une supervision point-à-point des coordonnées pour compenser la perte d’information due à la conversion des nuages de points en images de profondeur, afin d’accroître le réalisme des structures et des objets générés. Bien que cette méthode permette de produire des nuages de points plus proches des données capturées par des capteurs réels, elle ne contraint pas suffisamment la génération à partir d’une image donnée et demeure encore éloignée des résultats obtenus par des systèmes LiDAR physiques.

Programme de la thèse

Les représentations de la 3D sont nombreuses, comme les images de profondeur, les nuages de points, les maillages, les SDF et les champs de radiance. Les images de profondeur sont issues de caméras stéréo, de capteurs à lumière structurée ou de capteurs ToF, et sont souvent combinées avec des images RGB. Les nuages de points, produits par des dispositifs comme le LiDAR, capturent une représentation discrète et éparse des surfaces. Ces deux types de données (images de profondeur et nuages de points) sont largement

représentés en 3D, c'est pourquoi cette thèse se focalisera sur la génération de 3D sous forme d'images de profondeur et de nuages de points.

La fidélité géométrique constitue un enjeu central de cette thèse et nécessitera une analyse approfondie afin d'en préciser les implications et les critères d'évaluation. Il est aussi important d'arriver jusqu'à une génération de données 3D à l'échelle permettant une exploitation directe des résultats en vision robotique.

Voici une proposition de plan de travail pour la thèse :

- 1. Estimation d'images de profondeur par modèle génératif avec contraintes géométriques**
L'ajout de contraintes géométriques dans un modèle génératif d'estimation de profondeur vise à améliorer la cohérence structurelle des scènes reconstruites en imposant des régularisations tout au long du processus de diffusion. Une première approche consistera à dé-projeter les cartes de profondeur générées à chaque itération afin de reconstruire une représentation 3D explicite, permettant ainsi d'intégrer des contraintes (la préservation des bords qui garantit les discontinuités de profondeur, une régularisation locale des surfaces...) adaptées aux propriétés géométriques des objets et des surfaces. L'incorporation de ces contraintes structurelles permettra d'aligner les prédictions du modèle avec les principes de la perception 3D, améliorant ainsi la fidélité et la plausibilité des reconstructions.
- 2. Ajout de contraintes physiques aux modèles génératifs de profondeur et de nuages de points**
L'intégration de contraintes physiques dans les modèles génératifs permettrait d'améliorer la qualité et la cohérence des reconstructions 3D à partir d'images en imposant des régularisations fondées sur les lois fondamentales de la physique. En s'inspirant de Physics-Informed Diffusion Models [2], l'incorporation de principes tels que les lois de conservation, la gestion des collisions et les propriétés des matériaux permettrait d'ancrer les générations dans une réalité physique plus rigoureuse. Par exemple, la prise en compte des interactions physiques entre les objets, via des contraintes de non-interpénétration ou de rigidité, permettrait d'éviter des artefacts de génération incohérents. L'intégration des propriétés des matériaux, notamment la réflexion, l'absorption et la diffusion de la lumière, contribuerait à produire des surfaces plus réalistes et compatibles avec l'apparence observée dans l'image d'entrée.

En structurant ainsi l'apprentissage du modèle par des contraintes géométriques et des régularisations physiques, on réduirait les incohérences locales et globales tout en renforçant la fidélité des cartes de profondeur et des nuages de points, les rendant plus exploitables pour des applications telles que la simulation physique, la vision robotique ou la réalité virtuelle.

Diffusion et valorisation des résultats

Les contributions scientifiques issues de cette recherche seront publiées dans des conférences et revues internationales afin d'assurer leur diffusion au sein de la communauté académique. Par ailleurs, les codes et modèles développés seront publiés et mis à disposition en open-source.

Profil du candidat recherché pour la thèse

- Diplôme niveau Master 2 (BAC+5)
- Rigueur scientifique et autonomie
- Connaissances en Vision 3D, Machine Learning.
- Bon niveau de programmation en Python, CUDA.

Pour toute candidature, envoyez votre CV, vos notes des dernières années de formation (Master, école d'ingénieur), une copie des derniers diplômes et une lettre de motivation d'une page décrivant les ambitions pour le sujet de thèse et la pertinence de la candidature par rapport à la description du sujet.

Dossier de candidature à envoyer aux adresses emails suivantes avant le 25 Mai 2025 :

jean-emmanuel.deschaud@minesparis.psl.eu ; santiago.velasco@minesparis.psl.eu

L'ensemble des partenaires de PR[AI]RIE-PSAI s'engagent à soutenir et promouvoir l'égalité, la diversité et l'inclusion au sein de ses communautés. Nous encourageons les candidatures issues de profils variés, que nous veillerons à sélectionner via un processus de recrutement ouvert et transparent.

Références

- [1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth : Iterative refinement of single-view depth using surface normal and its uncertainty. In *British Machine Vision Conference (BMVC)*, 2022. URL : <https://arxiv.org/pdf/2210.03676>.
- [2] Jan-Hendrik Bastek, WaiChing Sun, and Dennis Kochmann. Physics-informed diffusion models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL : <https://arxiv.org/pdf/2403.14404>.
- [3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014. URL : <https://arxiv.org/pdf/1406.2283>.
- [4] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL : <https://arxiv.org/pdf/2312.02145>.
- [5] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3D point cloud generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL : <https://arxiv.org/pdf/2103.01458>.
- [6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2 : Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL : <https://arxiv.org/pdf/2304.07193>.
- [7] Haoxi Ran, Vitor Guizilini, and Yue Wang. Towards realistic scene generation with lidar diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL : <https://arxiv.org/pdf/2404.00815>.
- [8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL : <https://arxiv.org/pdf/2112.10752>.
- [9] Michał J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. Gecco : Geometrically-conditioned point diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL : <https://arxiv.org/pdf/2303.05916>.
- [10] Zhen Wang, Dongyuan Li, and Renhe Jiang. Diffusion models in 3D vision : A survey, 2024. URL : <https://arxiv.org/pdf/2410.04738>, [arXiv:2410.04738](https://arxiv.org/abs/2410.04738).
- [11] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything : Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL : <https://arxiv.org/pdf/2401.10891>.
- [12] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. URL : <https://arxiv.org/pdf/2406.09414>.
- [13] Linqi Zhou, Yilun Du, and Jiajun Wu. 3D shape generation and completion through point-voxel diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL : <https://arxiv.org/pdf/2104.03670>.