CSMA 2026
17ème Colloque National en Calcul des Structures
18-22 Mai 2026, Presqu'île de Giens (Var)

Apprentissage des Propriétés d'Inertie de Solides Complexes via un Espace Latent Géométrique. (Learning Inertial Properties of Complex Solids via a Geometric Latent Space)

A. Schockaert[1,2,3], V. Magnier[1], J-F. Witz[1], G. Dufaye[3], H. Wannous[2]

[1] *Univ. Lille, CNRS, Centrale Lille, UMR 9013 - LaMcube - Laboratoire de Mécanique, Multiphysique, Multiéchelle, F-59000 Lille, France*
[2] *CERI SN, IMT Nord Europe, Villeneuve D'Ascq, 59650 (France)*
[3] *Downs, 59670 Sainte-Marie-Cappel (France)*

Résumé — Nous présentons une méthode d'estimation des propriétés inertielles d'objets à géométrie complexe (*patatoïde*) à partir d'images, fondée sur une représentation implicite de leur forme (SDF) et un espace latent géométrique appris. Notre approche repose sur l'optimisation d'un vecteur latent représentant la forme de l'objet, obtenu à partir d'un ensemble d'observations visuelles. Nous développons tout d'abord **SDFformer**, un modèle de reconstruction 3D basé sur l'apprentissage d'une représentation implicite de la surface des objets via leur Fonction de Distance Signée (SDF). Nous proposons ensuite un modèle basé sur l'apprentissage profond entraîné pour prédire les propriétés physiques de l'objet observé. Ce second modèle se base sur le premier et utilise le vecteur latent comme descripteur compact de la géométrie du solide observé. Les résultats montrent que cet espace latent capture efficacement les informations géométriques nécessaires à la description complète de la forme mais également des propriétés inertielles. Cette approche permet d'envisager une caractérisation temps réel d'objets observés, directement exploitable dans un jumeau numérique.
Mots clés — Représentations implicites, Modélisation 3D, jumeaux numériques.

1 Introduction et enjeux

L'étude de la mécanique des corps rigides constitue un pilier fondamental de nombreuses disciplines scientifiques, allant de la robotique à la simulation numérique, en passant par la réalité augmentée et les jumeaux numériques. Dans un environnement connu ou observable, la prédiction du mouvement d'un objet via les équations de Newton–Euler nécessite la connaissance préalable de ses propriétés inertielles. Pour des géométries complexes, c'est-à-dire non représentables par des primitives analytiques simples (*i.e.patatoïdes*), l'obtention de ces grandeurs (masse, centre de masse, tenseur d'inertie) demeure un défi majeur.

Les approches classiques reposent sur la discrétisation de la géométrie et l'intégration numérique des équations de volume. Typiquement, un modèle 3D du solide étudié est converti en une représentation discrète, telle qu'un maillage ou une grille de voxels. Les intégrales de volume définissant les propriétés inertielles sont alors calculées numériquement à partir de ces discrétisations. Ces méthodes, bien que robustes, présentent plusieurs limites : (i) un coût de calcul élevé pour des maillages fins, (ii) la nécessité d'une géométrie explicite et propre (fermée, sans auto-intersections), et (iii) l'impossibilité de réutiliser efficacement les calculs en cas de modification géométrique.

Dans ce travail, nous proposons une approche visant à lever ces verrous en estimant les propriétés inertielles à partir de simples observations visuelles. Notre méthode repose sur un modèle de reconstruction 3D qui apprend à encoder l'information géométrique d'un objet à partir d'un ensemble d'images, sous la forme d'un vecteur latent de faible dimension. Ce vecteur, expressif et compact, permet ensuite la prédiction directe des grandeurs inertielles via un réseau de neurones léger, agissant comme un modèle de substitution (*surrogate model*) des méthodes traditionnelles.

Notre contribution est double : (1) nous introduisons un modèle de reconstruction 3D basé sur une représentation implicite neuronale, **SDFformer**, capable de capturer fidèlement la géométrie d'objets complexes dans un espace latent continu ; (2) nous démontrons que cet espace latent contient suffisamment d'informations pour fournir un réseau de neurones capable de prédire avec précision les propriétés

inertielles des solides observés, ouvrant la voie à la génération de jumeaux numériques à partir de simples observations visuelles.

2 Travaux connexes

2.1 Calcul des propriétés inertielles

La détermination des propriétés inertielles d'un solide est un prérequis fondamental à toute analyse dynamique. Typiquement, pour un corps rigide, ces propriétés sont définies par des intégrales de volume. Si la géométrie étudiée est décrite par des fonctions analytiques simples (sphère, cube, etc.), ces intégrales admettent des solutions fermées. Cependant, pour des géométries complexes ou arbitraires, un calcul numérique est inévitable. Ce calcul demande typiquement un maillage du solide étudié. Bien qu'efficace, les méthodes de simulation numérique basées sur cette discrétisation souffre de deux verrous majeurs. Elles nécessitent une géométrie explicite du solide, typiquement un maillage de surface de haute qualité, qui doit être fermé et sans auto-intersection. De plus l'intégration numérique sur un maillage dense est une opération chronophage, incompatible avec des applications interactives ou des simulations en temps réel.

2.2 Représentation neuronales implicites

Des travaux récents [1, 2, 3] ont montré la puissance de l'utilisation d'une réseau de neurones pour approcher un signal. Ces représentations, nommées représentation neuronales implicites, ont démontré le potentiel des réseaux denses pour représenter tout type de signaux (son, images, forme 3D...). Ces réseaux sont typiquement entraînés à partir de représentations 3D connues comme la Signed Distance Function (SDF) [2] ou l'occupation [4] (occupancy). Récemment des travaux ont montré que ces méthodes pouvaient être améliorées pour représenter la texture des objets [3] mais également pour les rendre dynamiques [5, 6].

2.3 Reconstruction 3D à partir de vues 2D multiples

Les méthodes traditionnelles [7] reposent sur des pipelines en deux étapes comme la Structure-from-Motion (SfM) suivie de la Stéréo Multi-Vues (MVS). Bien que puissante, ces méthodes sont limitées par leur faible rapidité et par leur dépendance à des données de qualité. Afin d'obtenir un modèle 3D qualitatif, elles ont besoin de nombreuses vues de l'objet étudié ainsi que d'un large recouvrement entre ces dernières. Des approches par apprentissage profond ont ensuite émergées pour palier à ces limitations. Typiquement, ces méthodes mettent en place une architecture encodeur-decodeur et apprennent un *a priori* sur la structure des formes 3D pour inférer la géométrie directement à partir des images. Les premières méthodes [8, 9, 10] utilisaient des réseaux de neurones convolutifs (CNN) ou des réseaux de neurones récurrent (RNNs), mais celles-ci montraient des performances limitées pour mettre en correspondance des vues éloignées. Plus récemment, les architectures basées sur les transformers [11, 12, 13, 14] se sont imposées grâce à leur capacité à capturer des dépendances à longue portée, permettant une extraction des informations de données plus robuste à travers un ensemble non-ordonné de vues. Yang *et al.* [14], propose une stratégie basée sur l'attention de groupe permettant d'extraire avec précision des caractéristiques à partir d'une série d'images. Cependant, leur méthode reposent sur une représentation en grille de voxels qui est limité par la discrétisation qu'elle réalise de l'espace 3D. Notre approche s'appuie sur ces méthodes de reconstruction 3D, cependant nous avons fait le choix de représenter les objets étudiés par leur SDF. Cela permet une représentation continue qui n'est pas limité en résolution.

3 Modèle de reconstruction 3D et apprentissage de l'espace latent

Notre modèle de reconstruction se décompose en deux sous-modules, illustrés en Figure 1. Dans cette section, nous détaillons d'abord la modélisation implicite de la géométrie des objets via la SDF 3.1, puis l'encodeur multi-vues permettant de prédire le vecteur latent associé à chaque objet 3.2.

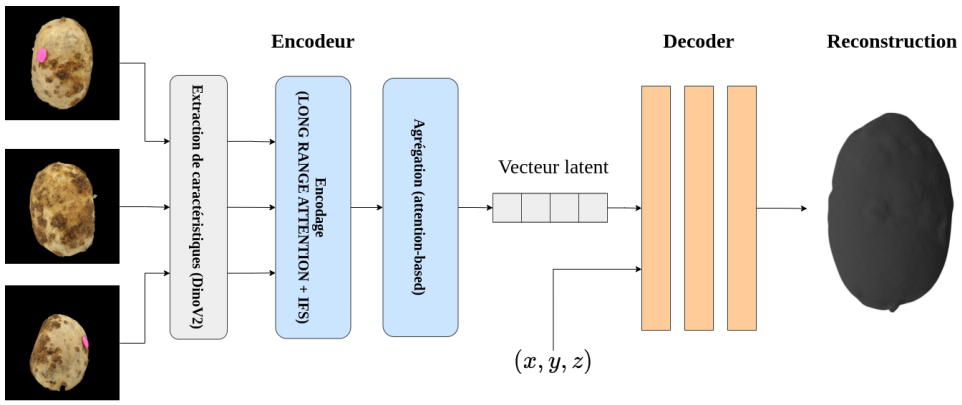

FIGURE 1 – Architecture de notre modèle de reconstruction

3.1 Principes de la SDF (Signed Distance Function)

Nous représentons la géométrie d'un objet par sa *Signed Distance Function* (SDF), une fonction continue qui associe à chaque point de l'espace sa distance à la surface de l'objet :

$$\mathrm{SDF}(\mathbf{x}) = \begin{cases} +d(\mathbf{x}, \partial\Omega), & \text{si } \mathbf{x} \notin \Omega, \\ -d(\mathbf{x}, \partial\Omega), & \text{si } \mathbf{x} \in \Omega, \end{cases} \tag{1}$$

où $d(\mathbf{x}, \partial\Omega)$ désigne la distance euclidienne minimale entre le point $\mathbf{x}$ et la surface $\partial\Omega$. Le niveau zéro de cette fonction, $\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 \,|\, \mathrm{SDF}(\mathbf{x}) = 0\}$ correspond donc à la surface de l'objet.

Les récentes méthodes des représentation neuronales implicites [2, 1] montrent qu'un signal comme la SDF des objets par un réseaux de neurones. Inspiré par *Park et al.* [2], nous proposons un réseau de neurones entièrement connecté (MLP) F_θ qui prend en entrée un point spatial $\mathbf{x} \in \mathbb{R}^3$ et un code latent $\mathbf{z} \in \mathbb{R}^d$ contenant les caractéristiques de l'objet, et prédit la valeur de la SDF correspondante :

$$\hat{SDF}(x) = \mathrm{F}_\theta(\mathbf{x}, \mathbf{z}), \tag{2}$$

où θ regroupe l'ensemble des paramètres du réseau. L'entraînement conjoint de θ et $\mathbf{z}$ s'effectue en minimisant la perte :

$$\mathcal{L}(\theta, \mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} \left\| F_\theta(\mathbf{x}_i, \mathbf{z}) - s_i \right\|_1, \tag{3}$$

où s_i est la valeur de référence de la SDF pour le point $\mathbf{x}_i$. À l'issue de l'entraînement, chaque objet est décrit par un vecteur latent $\mathbf{z}$ capturant sa géométrie globale dans un espace continu et différentiable. Afin d'appliquer cette modélisation à la reconstruction, nous proposons un encodeur d'images multi-vues capable de projeter directement les observations visuelles d'un objet dans cet espace latent, garantissant ainsi la cohérence entre les vues et la représentation implicite apprise.

3.2 Encodage des observations visuelles

Inspiré des travaux récents sur la reconstruction 3D à partir d'images [12, 14], l'encodeur que nous proposons est conçu pour traiter un ensemble non ordonné de vues d'un même objet. Il suit une architecture en deux étapes :

— une extraction indépendante des caractéristiques locales pour chaque vue ;
— une fusion hiérarchique des informations entre vues afin de capturer la cohérence spatiale globale de l'objet.

Dans la première étape, chaque image est traitée par un réseau pré-entraîné, ici *DINOv2* [15], qui extrait des descripteurs visuels robustes aux variations d'éclairage, de texture et de point de vue. Ces descripteurs représentent les structures locales significatives observées dans chaque image. Dans la seconde étape, ces descripteurs sont enrichies par un mécanisme d'attention à longue portée (*Long Range Attention*, LRA) [14], qui établit des corrélations entre les différentes vues. Ce mécanisme permet à

chaque image de bénéficier d'informations contextuelles issues des autres points de vue. Enfin, nous intégrons un module d'*Inter-View Feature Signatures* (IFS) qui aide à distinguer les contributions spécifiques de chaque vue tout en évitant les redondances d'information. Cette combinaison d'attention intra-vue et inter-vue permet à l'encodeur d'extraire une représentation globale et cohérente de la géométrie de l'objet, même lorsque les vues sont partiellement recouvrantes ou incomplètes. Les caractéristiques extraites par l'encodeur multi-vues sont finalement fusionnées par un *Attention-based Aggregation Module* (AAM). Ce module apprend à pondérer les informations issues de chaque vue en fonction de leur pertinence, afin de produire un vecteur latent unique $\mathbf{z}$ représentant la géométrie globale de l'objet. Cette opération d'agrégation agit comme une réduction dimensionnelle non-linéaire, condensant les descripteurs visuels en une représentation compacte et cohérente, directement utilisée pour conditionner le décodeur SDF lors de la reconstruction implicite.

Cette architecture couplée (**encodeur + AAM + décodeur SDF**) permet d'obtenir une reconstruction implicite continue et différentiable de la géométrie d'un objet à partir d'un simple ensemble d'images non ordonnées, tout en assurant la compatibilité avec l'espace latent appris durant la phase de pré-entraînement du décodeur.

4 Prédiction des propriétés inertielles par modèle de substitution

Notre objectif final est d'étudier la mécanique des corps rigides à partir d'observations de ce corps en mouvement. Pour cela, nous tirons à profit l'espace latent appris par notre modèle de reconstruction SDFformer 2.3. Pour cela, nous prenons comme hypothèse que les vecteurs latents z encodés par notre modèle sont des descripteurs complets de la géométrie des objets étudiés. Dans ce cas, s'il existe une existe une application H reliant la géométrie du corps rigide à ces différentes propriétés inertielles (masse, centre de masse, tenseur d'inertie) alors il doit exister une application H' reliant z à ces mêmes propriétés. Cependant, cette fonction H' ne possède pas de forme analytique connue et est probablement complexe à définir puisqu'elle encapsule la double opération de z vers la géométrie puis de la géométrie vers les propriétés inertielles. Nous proposons un modèle de substitution (surrogate model) H_θ capable d'apprendre l'application H'. Ce modèle est un *MLP* prenant en entrée un vecteur latent décrivant un solide et prédit en sortie un ensemble de propriétés inertielles. Dans ce travail, nous choisissons de prédire la masse m, le centre de masse C et le tenseur d'inertie $\bar{\bar{I}}$ mais nous supposons que notre méthode peut être adaptée pour prédire toutes les caractéristiques physiques induites de la forme et du mouvement de l'objet. Nous proposons donc le modèle suivant :

$$\hat{m}_i, \hat{C}_i, \hat{I}_i = H_\theta(z_i) \tag{4}$$

où $\hat{m}_i, \hat{C}_i, \hat{I}_i$ sont les propriétés inertielles prédites à partir de z_i le vecteur latent encodant la géométrie d'un solide observé.

L'entraînement de notre modèle est réalisé en minimisant une norme ℓ_1 entre les grandeurs de référence et les prédictions du modèle H_θ :

$$\mathcal{L}_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\|\hat{m}_i - m_i\|_1 + \|\hat{C}_i - C_i\|_1 + \|\hat{I}_i - I_i\|_1\Big). \tag{5}$$

Cet apprentissage repose sur un ensemble de données que nous qualifions de références, bien que des imprécisions peuvent exister. Les grandeurs physiques sont obtenues par intégration numérique sur la géométrie connue, en supposant une densité homogène. Cette approche permet de constituer un grand jeu de données et offre un contrôle direct sur les conditions physiques de référence. Cependant, elles introduisent donc des imprécisions systématiques, liées à la discrétisation volumique et aux hypothèses physiques adoptées. Le modèle doit ainsi être capable de généraliser malgré ces biais numériques.

Enfin, pour limiter la sur-adaptation à ces données imparfaites et favoriser la cohérence physique des prédictions, nous introduisons un terme de régularisation $\mathcal{L}_{\text{reg}}$ ajouté à la fonction de perte totale. Cette régularisation permet de contrôler la complexité des paramètres du modèle (éviter des poids trop grands ou instables) et encourage la cohérence du tenseur d'inertie prédit, en pénalisant les écarts à la symétrie. La régularisation est définie comme :

$$\mathcal{L}_{\text{reg}} = \|\theta\|_2^2 + \beta\,\|\hat{I}_i - \hat{I}_i^\top\|_F^2, \tag{6}$$

où $\|\theta\|_2^2$ est une régularisation L_2 classique, et le second terme contrôle la symétrie du tenseur d'inertie prédit, avec β un coefficient de pondération. La fonction de perte totale utilisée pour l'entraînement de H_θ combine la fidélité aux données et la régularisation :

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{data}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{7}$$

où λ_{reg} est un facteur de pondération réglé au préalable.

5 Résultats et interprétations

5.1 Dataset

Pour l'entraînement et la validation de nos méthodes, nous utilisons *3dPotatoTwin* [16], un jeu de données contenant des informations sur 339 tubercules de pommes de terre *(autant prendre de vraies patatoïdes...)*. Pour chaque tubercule, nous disposons :

— d'images RGB multi-vues de la tubercule en rotation ;
— d'un modèle 3D fidèlement reconstruit à partir d'une méthode SfM [7] ;
— Plusieurs grandeurs caractéristiques telles que le volume.

Afin d'évaluer la capacité de généralisation de nos modèles, nous avons séparé le jeu de données en trois sous-ensembles ; 70% des tubercules sont utilisés pour l'entraînement, 15% pour la validation et 15% pour le test. Tous les résultats présentés dans cette section sont obtenus sur l'ensemble de test, c'est-à-dire sur des objets jamais vus pendant l'entraînement.

5.2 Qualité de la reconstruction

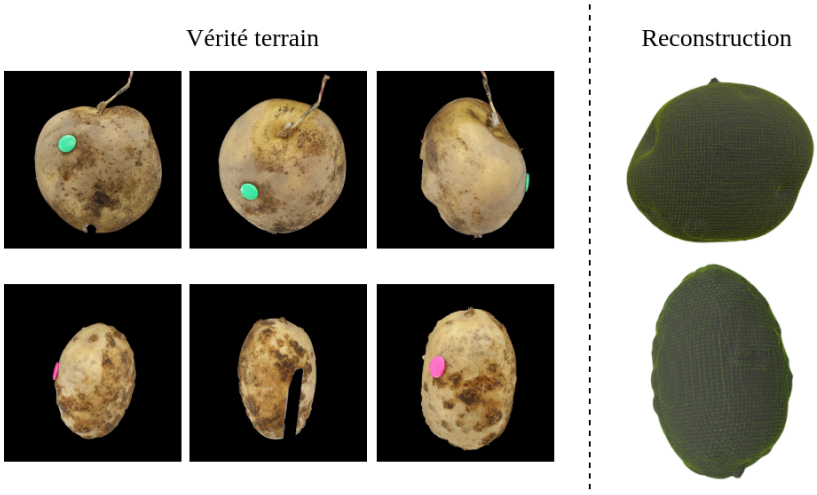

FIGURE 2 – Résultats quantitatifs de notre modèle de reconstruction

La première étape de notre validation consiste à évaluer la capacité de notre modèle SDFformer à reconstruire fidèlement la géométrie d'un objet à partir d'images. La qualité de cette reconstruction conditionne directement la pertinence des vecteurs latents z produits par l'encodeur.
Qualitativement, comme illustré dans la figure 2, notre modèle reproduit avec précision les formes globales ainsi que les irrégularités locales caractéristiques des tubercules.
Quantitativement, nous utilisons la *Chamfer Distance* (CD) comme métrique d'évaluation, calculée entre les nuages de points de référence $\mathcal{P}$ et celui extrait de la SDF prédite $\hat{\mathcal{P}}$:

$$\text{CD}(\mathcal{P}, \hat{\mathcal{P}}) = \frac{1}{|\mathcal{P}|} \sum_{x \in \mathcal{P}} \min_{\hat{x} \in \hat{\mathcal{P}}} \|x - \hat{x}\|_2^2 + \frac{1}{|\hat{\mathcal{P}}|} \sum_{\hat{x} \in \hat{\mathcal{P}}} \min_{x \in \mathcal{P}} \|x - \hat{x}\|_2^2. \tag{8}$$

Cette métrique mesure la proximité géométrique entre deux surfaces, une valeur faible indiquant une reconstruction fidèle. Une erreur inférieure à 1×10^{-3} sur des objets normalisés est généralement considérée comme excellente dans les tâches de reconstruction implicite.

TABLE 1 – Qualité de reconstruction.

Métrique	Moyenne	Écart-type
Chamfer Distance	0.007	0.008

Les valeurs rapportées dans le Tableau 1 confirment la très bonne précision géométrique atteinte par SDFformer. La faible variance entre échantillons traduit une stabilité du modèle sur l'ensemble des morphologies.

5.3 Précision de la prédiction des propriétés inertielles

Le cœur de notre contribution réside dans la capacité à prédire du modèle de substitution H_θ à prédire précisément les propriétés inertielles directement à partir du vecteur latent z. La figure 3 illustre la convergence de la perte totale au cours des itérations d'entraînement. On observe une décroissance rapide de l'erreur sur les trois grandeurs, puis une stabilisation après environ 100 époques, signe d'une convergence stable.Ces résultats montrent que le modèle H_θ parvient à inférer des propriétés inertielles cohérentes à partir du vecteur latent z uniquement.

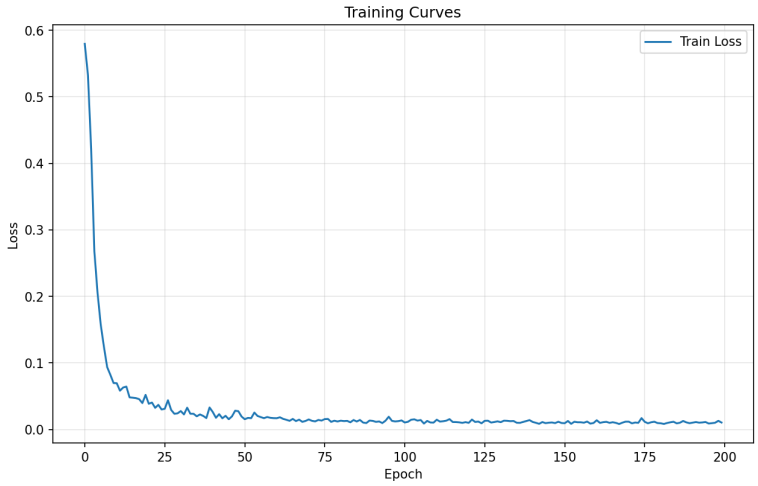

FIGURE 3 – Courbe d'apprentissage

5.4 Gain en temps de calcul

Enfin, nous comparons le coût d'inférence de notre approche à celui d'une chaîne de calcul classique comprenant :

— la reconstruction 3D par SfM,
— le maillage volumique de la géométrie,
— et l'intégration numérique pour le calcul des propriétés inertielles.

TABLE 2 – Comparaison du temps moyen de calcul par échantillon.

Méthode	**Temps moyen**	**Gain**
Pipeline classique (SfM + maillage + intégration)	$\sim$ 240 s	—
Notre méthode (encodage + H_θ)	0.045 s	$\times$5300

Les résultats (Tableau 2) montrent que notre approche offre un gain de plusieurs ordres de grandeur en temps de calcul, rendant possible une estimation quasi-instantanée des propriétés inertielles à partir d'observations visuelles.

6 Discussions

6.1 Limitations

Bien que les résultats obtenus démontrent la pertinence de notre approche pour la reconstruction et la prédiction de propriétés inertielles, plusieurs limitations méritent d'être discutées.

Tout d'abord, l'architecture actuelle du modèle est centrée sur la reconstruction et la caractérisation d'un unique solide à la fois. Dans un cadre applicatif où plusieurs corps rigides interagiraient (par exemple lors de collisions ou de contacts multiples), le temps de calcul pourrait devenir un facteur limitant.

Ensuite, notre modèle repose sur l'hypothèse que le vecteur latent z encode l'intégralité des informations géométriques nécessaires à la prédiction des grandeurs inertielles. Si cela s'avère suffisant pour des solides isolés, l'étude du mouvement réel d'un corps rigide exige la prise en compte d'autres grandeurs physiques, telles que les forces de contact, les frottements ou les contraintes de support. Ces phénomènes ne sont pas encore intégrés dans notre modélisation.

De plus, notre méthode fait le choix d'une représentation purement globale de la géométrie via un code latent z fixe. Une approche dynamique, exploitant un historique temporel des observations (par exemple à travers un modèle séquentiel ou récurrent), pourrait améliorer la robustesse et la cohérence des prédictions, notamment dans un contexte de mouvement.

Enfin, l'entraînement a été effectué sur un jeu de données limité, à savoir des tubercules de pommes de terre présentant des géométries similaires. Si cette base constitue une excellente preuve de concept, la généralisation à d'autres formes, topologies ou matériaux doit encore être validée expérimentalement. L'évaluation sur des jeux de données plus variés serait nécessaire pour confirmer la portée de notre approche.

6.2 Travaux futurs

Nos perspectives de recherche se déclinent selon plusieurs axes. Un premier axe consiste à étendre la méthode à l'étude du mouvement des corps rigides, en intégrant des modèles de contact et des interactions dynamiques. L'objectif est de relier directement les descripteurs latents z à la prédiction du mouvement sous l'action de forces extérieures, ouvrant la voie à une modélisation physique complète à partir d'observations visuelles uniquement.

Un second axe vise à intégrer notre approche dans un cadre de jumeau numérique. En associant les reconstructions géométriques, les propriétés physiques apprises et les observations expérimentales, un tel système permettrait d'effectuer des simulations en temps réel ou semi-temps réel, utiles dans de nombreux domaines d'applications.

Enfin, des développements méthodologiques sont envisagés, notamment l'enrichissement du modèle de substitution H_θ par des contraintes physiques explicites (méthodes PINNS [17] ou par des stratégies d'apprentissage auto-supervisées. Cela permettrait de réduire la dépendance à des données étiquetées coûteuses et d'améliorer la robustesse face aux incertitudes des mesures expérimentales. Ce travail ouvre aussi la voie à une intégration des représentations implicites au sein de pipelines hybrides mêlant modèles physiques et espaces latents appris.

7 Conclusions

Nous avons présenté une approche complète de reconstruction et de caractérisation physique de corps rigides à partir d'observations visuelles uniquement. En combinant une représentation géométrique implicite basée sur la fonction de distance signée (SDF) et un modèle de substitution opérant dans l'espace latent, notre méthode permet d'estimer efficacement les propriétés inertielles d'un objet sans étape de maillage ni intégration volumique explicite. Les résultats obtenus sur le jeu de données *3dPotatoTwin* démontrent la capacité de notre modèle à capturer la complexité géométrique et à inférer des grandeurs physiques cohérentes. Contrairement aux approches existantes, nous exploitons l'espace latent géométrique non pour reconstruire la forme, mais pour relier celle-ci à des propriétés mécaniques mesurables. À terme, cette approche pourrait servir de fondement à une modélisation complète du mouvement de corps rigides observés, ouvrant des perspectives intéressantes en simulation, robotique et ingénierie numérique.

Références

[1] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473. Curran Associates, Inc., 2020.

[2] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF : Learning Continuous Signed Distance Functions for Shape Representation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, Long Beach, CA, USA, June 2019. IEEE.

[3] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF : Representing Scenes as Neural Radiance Fields for View Synthesis, August 2020. arXiv :2003.08934 [cs].

[4] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks : Learning 3D Reconstruction in Function Space, April 2019. arXiv :1812.03828 [cs].

[5] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF : Neural Radiance Fields for Dynamic Scenes, November 2020. arXiv :2011.13961 [cs].

[6] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy Flow : 4D Reconstruction by Learning Particle Dynamics. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5378–5388, October 2019. ISSN : 2380-7504.

[7] Johannes L. Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. pages 4104–4113, 2016.

[8] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2 : A Unified Approach for Single and Multi-view 3D Object Reconstruction, April 2016. arXiv :1604.00449 [cs].

[9] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2Vox++ : Multi-scale Context-aware 3D Object Reconstruction from Single and Multiple Images. *International Journal of Computer Vision*, 128(12) :2919–2935, December 2020.

[10] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN : Deep Implicit Surface Network for High-quality Single-view 3D Reconstruction. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[11] Zai Shi, Zhao Meng, Yiran Xing, Yunpu Ma, and Roger Wattenhofer. 3D-RETR : End-to-End Single and Multi-View 3D Reconstruction with Transformers, November 2021. arXiv :2110.08861 [cs].

[12] Zhenwei Zhu, Liying Yang, Ning Li, Chaohao Jiang, and Yanyan Liang. UMIFormer : Mining the Correlations between Similar Tokens for Multi-View 3D Reconstruction, August 2023. arXiv :2302.13987 [cs].

[13] Zhenwei Zhu, Liying Yang, Xuxin Lin, Lin Yang, and Yanyan Liang. GARNet : Global-aware multi-view 3D reconstruction network and the cost-performance tradeoff. *Pattern Recognition*, 142 :109674, October 2023.

[14] Liying Yang, Zhenwei Zhu, Xuxin Lin Jian Nong, and Yanyan Liang. Long-Range Grouping Transformer for Multi-View 3D Reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18211–18221, Paris, France, October 2023. IEEE.

[15] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO : DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, July 2022. arXiv :2203.03605 [cs].

[16] Haozhou Wang, Pieter M. Blok, James Burridge, Ting Jiang, Minato Miyauchi, Kyosuke Miyamoto, Kunihiro Tanaka, and Wei Guo. 3DPotatoTwin : a Paired Potato Tuber Dataset for 3D Multi-Sensory Fusion. *Plant Phenomics*, page 100123, October 2025.

[17] Chayan Banerjee, Kien Nguyen, Clinton Fookes, and George Karniadakis. Physics-Informed Computer Vision : A Review and Perspectives, May 2024. arXiv :2305.18035 [eess] version : 3.