English version below

Modèles de diffusion sous des contraintes anatomiques réalistes

Mots-Clés: Modèles génératifs, Modèles de diffusion, Espace de Hilbert à noyau reproduisant (RKHS), Appariement difféomorphique avec grandes déformations (LDDMM).

Localisation: Université de Toulouse, Institut de Mathématiques de Toulouse (IMT).

Encadrement:

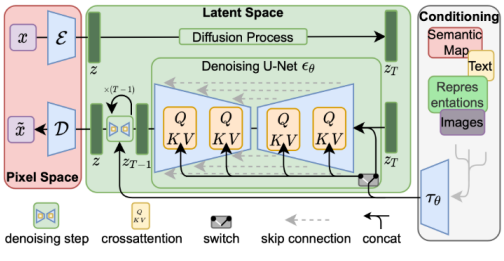
Emmanuelle Claeys (Université de Toulouse, IRIT)
email: emmanuelle.claeys@irit.fr
web: emmanuelle-claeys.com

Juliette Chevallier (INSA Toulouse, IMT)
email: juliette.chevallier@math.univ-toulouse.fr
web: juliette-chevallier.pages.math.cnrs.fr

Contexte: La génération d'images et de vidéos à travers des modèles génératifs, notamment de diffusion, constitue un domaine très actif de la recherche, en particulier depuis la popularisation des modèles linguistiques à grande échelle ou *large language models* (LLMs). L'augmentation des capacités de calcul a permis le développement de modèles capables de générer des contenus visuels, images ou vidéos, à partir de sources originales, ou encore de modifier ces dernières en suivant des contraintes fournies sous forme de prompts.

Lorsque la génération porte sur un avatar, on parle alors de *digital twins*. Ces jumeaux numériques génératifs permette de simuler, anticiper et optimiser des mouvements, postures et efforts physiques, tout en intégrant les contraintes physiologiques, géométriques et environnementales propres à l'être humain et à l'espace dans lequel il évolue (cf. figure ci-contre avec le modèle PRIMAL [13]).

Cependant, ces modèles rencontrent encore des difficultés à préserver la cohérence anatomique lorsqu'ils sont appliqués à des images représentant des corps humains en mouvement complexe (comme en danse, en gymnastique, ou dans d'autres contextes atypiques). Une étude préliminaire [1] a notamment mis en lumière que les modèles à l'état de l'art pour la génération d'images produisent des distorsions anatomiques sévères, ou *body horror*, dans ce contexte.



Une technique prometteuse pour l'édition d'images est le recours à des modèles de diffusion [4] et particulièrement aux modèles de diffusion latente ou *latent diffusion models* (LDM [6], cf. figure ci-contre).

La particularité de ces modèles repose sur la construction d'un espace latent spécifiquement conditionné par le problème considéré, typiquement via des prompts ou des caractéristiques intrinsèques aux images (carte de profondeur ou de contours calculée automatiquement, masques de segmentation, etc.). Le processus de diffusion, c'est-à-dire le mécanisme à même de modifier les images est alors réalisé dans ledit espace latent via le recours à un réseau de type U-Net [7].

Objectifs du stage: Ce stage poursuit un double objectif :

1. Générer des mouvements anatomiques complexes mais réalistes, y compris lorsqu'ils sont atypiques ;
2. Permettre la modification automatique de vidéos mettant en scène des corps en mouvement, sans altérer leur cohérence anatomique.

Dans [1], nous avons proposé une méthode capable de modifier des photos de danse sportive en se basant sur des algorithmes de segmentation et une technique de recollement particulière appelée Poisson Blending [5] permettant d'intégrer des contraintes de régularité. Ces premiers résultats, bien qu'encourageants, ne sont malheureusement pas directement transposables aux vidéos. Nous soupçonnons que cela est dû à l'absence d'une intégration suffisante des contraintes anatomiques dans le conditionnement des espaces latents. Dans ce but, nous souhaiterions explorer dans ce stage deux pistes : une première basée sur l'intégration de contraintes d'appariement inspirées de la théorie des espaces de formes [11], une seconde fondée sur l'apprentissage par renforcement.

La théorie des espaces de formes, largement utilisée en anatomie computationnelle, constitue une piste prometteuse. Elle permet de construire des métriques entre différentes formes anatomiques – ici des corps humains en mouvement – via des difféomorphismes quantifiant les déformations nécessaires à l'appariement d'une forme sur une autre.

Dans notre contexte, nous souhaitons intégrer ces métriques dites LDDMM, pour *large deformation diffeomorphic metric mapping*, dans la fonction de perte du réseau diffusif impliqué dans la modification des images, sous forme de contraintes explicites. En particulier, des travaux récents en apprentissage profond [8] ont montré l'efficacité de ces métriques pour contraindre l'apprentissage de réseaux convolutifs en s'appuyant sur une structure de graphe. Nous souhaiterions tirer profit de cette idée en l'adaptant à notre contexte, les différentes articulations d'un corps en mouvement pouvant être vues comme les noeuds d'un graphe. L'extraction de tels graphes peut notamment être réalisée via Neural Localizer Fields[9], PRIMAL [13], DensePose [3] ou encore OpenPose [2], en utilisant les infrastructures du LAAS.

Parallèlement, nous envisageons de recourir à de l'apprentissage par renforcement afin de corriger en temps réel les incohérences anatomiques. En effet, les réseaux de neurones profonds sont aujourd'hui à l'état de l'art pour la génération d'images anatomiques complexes [10, 12]. Les métriques LDDMM pourraient notamment servir à quantifier un "écart anatomique", utilisé alors comme signal de récompense ou de pénalité au cours de l'apprentissage. Une thèse sera proposée à l'issu du stage sous réserve de résultats.

Pré-requis: La personne recrutée devra avoir suivi des cours d'apprentissage profond ainsi que des cours de mathématiques d'un niveau au moins master afin de pouvoir appréhender le contexte des espaces de formes. Un bagage en géométrie riemannienne et/ou en apprentissage par renforcement sera apprécié, mais ces notions pourront entièrement être introduites en début de stage.

Environnement de travail: Le stage se déroulera à l'Institut de Mathématiques de Toulouse (IMT) et à l'Institut de Recherche en Informatique de Toulouse (IRIT) ; la personne recrutée pourra prendre part aux activités des équipes Statistique et Optimisation (SO) et MISFIT, et sera associée à un groupe de travail autour de l'usage de métriques LDDMM en apprentissage profond. La participation au séminaire hebdomadaire de l'équipe SO, ainsi qu'aux autres séminaires du site, tel que le séminaire SPOT, sera vivement encouragée.

Candidature: Envoyer un CV et un relevé de notes des deux dernières années, ainsi qu'une lettre de candidature, par email à Emmanuelle Claeys et Juliette Chevallier.

English version

Diffusion Models Under Realistic Anatomical Constraints

Keywords: Generative models, Diffusion models, Reproducing Kernel Hilbert Space (RKHS), Large Deformation Diffeomorphic Metric Mapping (LDDMM).

Location: University of Toulouse, Institut de Mathématiques de Toulouse (IMT), France.

Supervision:

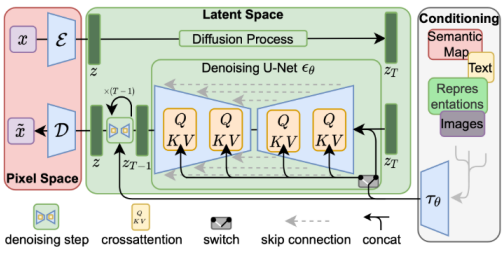
Emmanuelle Claeys (University of Toulouse, IRIT)
email: emmanuelle.claeys@irit.fr
web: emmanuelle-claeys.com

Juliette Chevallier (INSA Toulouse, IMT)
email: juliette.chevallier@math.univ-toulouse.fr
web: juliette-chevallier.pages.math.cnrs.fr

Context: The generation of images and videos using generative models—especially diffusion-based models—is currently a very active research area, particularly following the rise of large language models (LLMs). The increasing computational power available in recent years has enabled the development of models capable of generating visual content, such as images or videos, from original source material, or of modifying such material according to constraints provided through prompts.

When generation focuses on an avatar, we refer to *digital twins*. These generative digital twins make it possible to simulate, anticipate and optimise movements, postures and physical efforts, while integrating the physiological, geometric and environmental constraints specific to human beings and the space in which they evolve (see figure opposite with the PRIMAL model [13]).

However, these models still struggle to preserve anatomical coherence when applied to images depicting human bodies in complex motion – such as in dance, gymnastics, or other unusual contexts. A preliminary study [1] has highlighted that state-of-the-art image generation models often produce severe anatomical distortions, or *body horror*, in such settings.



A promising technique for image editing relies on diffusion models [4], and in particular on latent diffusion models (LDMs [6], see the figure on the right). The specificity of these models lies in the construction of a latent space conditioned on the problem at hand, typically through prompts or intrinsic image features (automatically computed depth or edge maps, segmentation masks, etc.). The diffusion process—i.e., the mechanism responsible for modifying the images—is then performed in this latent space using a U-Net architecture [7].

Internship objectives: This internship has two main objectives:

1. Generate anatomically complex yet realistic human motions, including atypical ones;
2. Enable automatic editing of videos involving moving human bodies while preserving their anatomical coherence.

In [1], we proposed a method for editing images of competitive dance using segmentation algorithms and a specific compositing technique known as Poisson Blending [5], allowing regularity constraints to be incorporated. Although promising, these results cannot be directly transferred to video. We suspect that this limitation stems from insufficient integration of anatomical constraints in the conditioning of the latent spaces. To address this issue, we aim in this internship to explore two directions: one based on incorporating matching constraints inspired by shape space theory [11], and another based on reinforcement learning.

Shape space theory, widely used in computational anatomy, offers a promising avenue. It allows the construction of metrics between different anatomical shapes—here, human bodies in motion—via diffeomorphisms quantifying the deformations required to map one shape onto another.

In our setting, we aim to integrate these metrics, known as LDDMM (*Large Deformation Diffeomorphic Metric Mapping*), into the loss function of the diffusion network involved in image modification, in the form of explicit constraints. In particular, recent deep learning work [8] has demonstrated the effectiveness of such metrics for constraining the training of convolutional networks through graph structures. We intend to leverage this idea in our context, viewing the various joints of a moving human body as nodes of a graph. Such graphs can be extracted using Neural Localizer Fields[9], PRIMAL [13], DensePose [3] or OpenPose [2], using the infrastructure provided by LAAS-CNRS.

In parallel, we plan to use reinforcement learning to dynamically correct anatomically inconsistent generations. Indeed, deep neural networks currently represent the state of the art for generating complex anatomical images [10, 12]. LDDMM metrics could be used to quantify an "anatomical discrepancy," which would then serve as a reward or penalty signal during training. A thesis will be proposed at the end of the internship, subject to results.

Requirements: The selected candidate should have completed coursework in deep learning, as well as mathematics at least at the master's level, in order to understand the context of shape spaces. Prior knowledge of Riemannian geometry and/or reinforcement learning would be appreciated, but these notions can be introduced at the beginning of the internship if needed.

Working environment: The internship will take place at the Institut de Mathématiques de Toulouse (IMT) and at the Institut de Recherche en Informatique de Toulouse (IRIT). The selected candidate will be able to take part in the activities of the Statistique et Optimisation (SO) and MISFIT teams, and will also be associated with a working group on the use of LDDMM metrics in a deep learning context.

Participation in the SO team's weekly seminar will be strongly encouraged, as well as attendance at other seminars on site, such as the SPOT seminar.

Application: Please send a CV and transcripts from the last two academic years, as well as a cover letter, by email to Emmanuelle Claeys and Juliette Chevallier.

References

[1] Lucas Bondietti, Ismaël Viennot, Juliette Chevallier, and Emmanuelle Claeys. Étude de modèles de diffusion pour la modification d'images de danse sportive. In *56èmes Journées de Statistiques de la Société Française de Statistique*, 2025.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[3] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild, 2018.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[5] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. 2023.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[8] Camilo Ruiz, Hongyu Ren, Kexin Huang, and Jure Leskovec. High dimensional, tabular deep learning with an auxiliary knowledge graph. *Advances in Neural Information Processing Systems*, 36:26348–26371, 2023.

[9] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation, 2024.

[10] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. *arXiv preprint arXiv:2410.03441*, 2024.

[11] Alain Trouvé. Diffeomorphisms groups and pattern matching in image analysis. *International journal of computer vision*, 28(3):213–221, 1998.

[12] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025.

[13] Yan Zhang, Yao Feng, Alpár Cseke, Nitin Saini, Nathan Bajandas, Nicolas Heron, and Michael J. Black. Primal: Physically reactive and interactive motor model for avatar learning, 2025.