Étude de modèles de diffusion pour la modification d'images de danse sportive

Lucas Bondietti, Ismaël Viennot, Juliette Chevallier, Emmanuelle Claeys

ÉTUDE DE MODÈLES DE DIFFUSION POUR LA MODIFICATION D'IMAGES DE DANSE SPORTIVE

Lucas Bondietti [1], Ismaël Viennot [1], Juliette Chevallier [1,2] & Emmanuelle Claeys [3,4]

[1] *INSA Toulouse, Université de Toulouse, France*
[2] *Institut de Mathématiques de Toulouse, UMR 5219*
[3] *Université Paul Sabatier, Université de Toulouse, France*
[4] *Institut de Recherche en Informatique de Toulouse, UMR 5505*

Résumé. La génération d'images et de vidéos à travers des modèles génératifs et de diffusions est un domaine très actif de la recherche, notamment à travers la popularisation de modèles tels que Chat-GPT ou LAMA. L'augmentation des ressources de calcul a permis ces dernières années le développement de modèles capables de générer de nouvelles images et vidéos à partir de sources originales, ou encore de modifier ces dernières en suivant des contraintes fournies sous forme de prompts. Pourtant, ces modèles ont encore des difficultés à préserver l'anatomie humaine lorsqu'on les applique à des images présentant des corps humains effectuant des mouvements complexes, tels que ceux présents en gymnastique ou en danse. Dans cet article, nous proposons un retour d'expérience sur l'utilisation des dernières méthodes de référence pour de la modification d'images de danse artistique, ainsi qu'un nouveau cadre d'étude basé sur du *Poisson Blending* permettant un meilleur respect de la morphologie des danseurs.

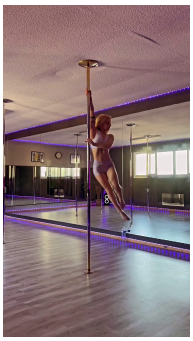
Mots-clés. Machine learning pour l'analyse sportive, modèles de diffusion, modèles génératifs, danse sportive.

Abstract. Generating images and videos using generative and diffusion models is a very active area of research, notably through popularizing models such as Chat-GPT or LAMA. In recent years, increased computational resources have enabled the development of models to generate new images and videos from original sources, or modify them according to constraints provided by prompts. However, these models still have difficulty in preserving human anatomy when applied to images displaying human bodies performing complex movements, such as those found in gymnastics or dance. In this article, we provide experiments relying on the latest reference methods for image modification in the context of artistic dance, as well as a new framework based on *Poisson Blending* allowing better respect of dancers' morphology.

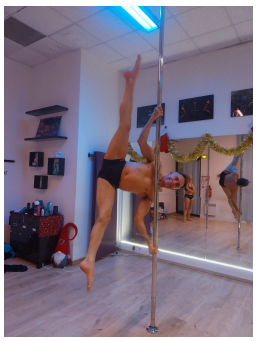
Keywords. Machine learning for sports analysis, diffusion models, generative models.

1 Introduction

Dans le contexte de la danse artistique et sportive, les athlètes sont très souvent amenés à faire la promotion de leur activité via la diffusion de courtes vidéos. Ces vidéos sont généralement capturées en intérieur, dans des studios de danse et sur un dispositif photographique de faible qualité, typiquement un téléphone portable. Aussi, leur qualité artistique reste limitée.



(a) Julie Cuirans



(b) Didier Benac

Figure 1: Deux athlètes professionnels photographiés dans leur studio respectif.

L'objectif de ce travail est de permettre la modification de ces vidéos, notamment de leur fond, à l'aide d'un *prompt* de commandes textuelles, et ceci afin d'améliorer l'esthétisme de ces productions. Nous nous concentrons ici sur la pole dance ; mais, les techniques développées s'appliquent directement à toutes disciplines présentant des caractéristiques similaires : gymnastique, danse plus généralement, *etc.*

On trouve dans la littérature beaucoup de méthodes open-source permettant de modifier une image ou bien une vidéo à partir d'un *prompt* [9]. Nos premiers essais ont montré que ces modèles sont performants sur des positions simples ou des animaux ; mais, lorsqu'ils sont utilisés sur des mouvements assimilés à de la gymnastique, ils conduisent à des artefacts et dénaturent complètement le corps humain : on parle de *body horror*. Ce résultat était attendu. En effet, les modèles génératifs ont encore beaucoup de difficultés à comprendre la biophysique basique, et cela reste une des limitations à leur usage [10].

Dans cet article, à partir de deux photos de danseurs (*cf.* Fig. 1), nous proposons un retour d'expérience ainsi qu'un nouveau cadre permettant l'édition du fond des photos, tout en préservant autant que possible le corps des deux athlètes. La section 2 propose un bref état de l'art sur les modèles de diffusion, la section 3 détaille notre nouveau cadre et évalue sa performance sur les deux images de référence susmentionnées.

2 État de l'art

Les modèles de diffusion [4, 17, 18] sont une catégorie de modèles génératifs basés sur des réseaux de neurones convolutionels (CNNs, [6]). Étant donnée une image originale, ces modèles fonctionnent en deux phases :

- ***Forward diffusion*** : L'image est progressivement bruitée par l'ajout progressif et markovien d'un bruit gaussien ;
- ***Reverse diffusion*** : À partir de l'image complètement bruitée, le modèle la débruite progressivement pour reconstruire une image réaliste, c'est-à-dire labellisée comme conforme par un classifieur.

Nous décrivons ci-après plus en détail le fonctionnement de ces réseaux.

Entraînement. Pour chacune des images composant le jeu d'entraînement, un bruit gaussien leur est ajouté, en plusieurs étapes $t \in [\![1, T]\!]$. On note x_0 l'image avant modifications, x_t l'image obtenue à la t-ième étape du processus. En particulier, x_T est complètement bruitée et ininterprétable. Plus précisément, le processus de *forward diffusion* est donné par l'équation

$$x_t \;=\; \sqrt{1-\beta_t}\, x_{t-1} \;+\; \sqrt{\beta_t}\, \varepsilon_t \,,$$

où β_t contrôle le niveau de bruit à chaque étape, et $\varepsilon_t \sim \mathcal{N}(0, I)$. Ainsi, par récurrence, on peut définir x_t à partir de x_0 :

$$x_t \;=\; \left(\prod_{s=1}^{t} \sqrt{1-\beta_s} \right) x_0 \;+\; \sum_{k=1}^{t} \left(\sqrt{\beta_k} \prod_{s=k+1}^{t} \sqrt{1-\beta_s} \right) \varepsilon_k \,. \tag{1}$$

Considérons un réseau de neurones paramétrisé par θ et visant à approcher ce processus. Historiquement [17], ces réseaux étaient entraînés par inférence variationelle via une perte ELBO, ou *evidence lower bound*, portant sur une approximation par le réseau de la loi de x_0. Des travaux plus récents [4] ont permis d'obtenir de meilleurs résultats empiriques en contrôlant l'erreur quadratique moyenne, ou *mean squared error* (MSE), entre le bruit réel associé à l'image x_t et celui prédit par le réseau noté $\varepsilon_\theta(x_t, t)$. Remarquons qu'en notant $\bar{\beta} := \left(\prod_{s=1}^{t} \sqrt{1-\beta_s}\right)$ et $\sigma_t := \sum_{k=1}^{t} \left(\sqrt{\beta_k} \prod_{s=k+1}^{t} \sqrt{1-\beta_s}\right)$, par propriété des vecteurs gaussiens, l'équation (1) se réécrit $x_t = \bar{\beta}\, x_0 + \sigma_t\, z$, où z désigne un bruit gaussien inconnu, et on remarque bien une symétrie entre la compréhension de x_0 et celle de z. Ceci conduit à la perte dite simple

$$\mathcal{L}\,(\theta) \;:=\; \sum_{t=1}^{T} \mathbb{E}_{x_0 \sim q,\, z \sim \mathcal{N}(0,I)} \left[\left\| \varepsilon_\theta(x_t, t) - z \right\|^2 \right] .$$

Inférence. Après entraînement, un modèle de diffusion est à même de générer de nouvelles observations réalistes, ou de modifier des images existantes, ceci via un usage judicieux du processus de *reverse diffusion* précédemment décris.

La génération pure s'effectue à partir d'un bruit blanc gaussien x_T, débruité étape par étape. Autrement dit, au cours de T étapes, le modèle prédit et supprime incrémentiellement le bruit $\varepsilon_\theta(x_t, t)$ de l'image x_t :

$$x_{t-1} \;=\; \frac{1}{\sqrt{1-\beta_t}} \big(x_t - \beta_t\, \varepsilon_\theta(x_t, t) \big) + \sqrt{\beta_t}\, \zeta \,, \tag{2}$$

où ζ est un bruit gaussien qui assure la diversité dans les modifications.

Par opposion, pour la modification d'image, l'état initial bruité x_T est conditionné par l'image originale ayant été labellisée automatiquement. Les modifications apportées à l'image peuvent également être conditionnées par une base d'entrées externes, telles que des descriptions textuelles ou *prompts*. Cela permet au modèle de conserver des éléments de l'image originale, tout en appliquant les modifications souhaitées. Concrètement, en notant c l'entrée de conditionnement, le processus de *reverse diffusion* décrit Eq. (2) est inchangé, sauf à tenir compte de c dans la définition du bruit prédit $\varepsilon_\theta(x_t, t, c)$.

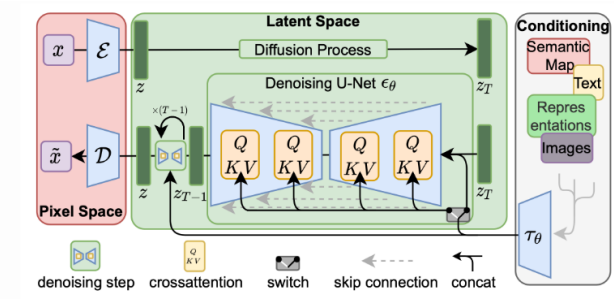


Figure 2: *Latent diffusion models* ou LDM – Image tirée de [15].

À droite, les contraintes de conditionnement sont intégrées à l'image originale, conduisant à une représentation latente combinée z_t hautement sémantique. La diffusion porte alors sur ces représentations latentes.

Dans notre contexte, nous souhaitons éditer le fond d'une photo de danse, tout en préservant le corps des athlètes, et en contrôlant la manière dont est modifiée le reste de l'image. Pour ce faire, nous allons recourir à deux types de conditionnement :

- **Descriptions textuelles** ou *prompts* : Afin d'être interprétées, les descriptions textuelles sont préalablement transformées en vecteurs numériques multidimensionnels par *embedding*. Ceci conduit à une représentation combinée de l'image, qui peut ensuite être utilisée pour préserver des contraintes dans chaque étape du modèle de diffusion.
- **Caractéristiques explicites** ou *feature conditioning* : Afin de fournir des contraintes localisées, on peut recourir à l'usage de masques de segmentation, de cartes de profondeur ou de bords. Le modèle est ainsi contraint de respecter des informations structurelles ou spatiales encapsulées dans ces cartes de caractéristiques ou *features maps*, garantissant l'alignement de l'image finale sur les conditions spécifiées.

Ces *features* de conditionnement c sont combinées à la représentation latente de l'image bruitée x_t par le biais de mécanismes d'attention [7, 19], modifiant l'estimateur de bruit ε_θ comme suit :

$$\varepsilon_\theta(x_t, t, c) \;=\; \texttt{CrossAttention}\left(\varepsilon_\theta(x_t, t)\,,\, c\right).$$

Nous appelons ces modèles mêlant mécanismes d'attention et processus de diffusion modèles de diffusion latente ou *latent diffusion models* (LDM, [15]). Le schéma reproduit Fig. 2 et issus de l'article séminale [15] décrit leur architecture. L'idée de ces réseaux est de forcer le respect des contraintes spatiales via l'usage d'un réseau de type U-Net [16]. En particulier, l'attention vise ici à s'assurer de la présence de certaines "clés" (*keys*) K dans la représentation latente successivement modifiée par l'U-Net. Les clés à maintenir sont définies par un ensemble de "requêtes" (*queries*) Q et sont chacunes associées à une valeur V. Concrètement, $\texttt{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \cdot V$, pour une représentation latente de dimension d et des matrices Q, K, V à estimer.

3 Contribution

3.1 Retour d'expérience

Nous présentons dans cette section les résultats que nous avons obtenus sur trois modèles de diffusion latente ou LDMs, plébiscités pour leurs performances :

- *Stable Diffusion XL* (SDXL, [12]) incorpore des couches supplémentaires au LDM et améliore l'U-Net sous-jacent. Cette architecture, plus profonde, permet un traitement plus détaillé et multi-échelle des informations par SDXL, conduisant à la génération d'images de meilleure qualité, avec un niveau de détails plus fins et une meilleure gestion des textures complexes par rapport à un LDM standard.
- *Image-to-Image* [20] permet l'ajout de paramètres optionnels tels que le nombre d'étapes d'inférence ou la force de la transformation. Plus précisément, le paramètre "force" contrôle la déviation du modèle par rapport à l'image d'entrée : une valeur faible maintient la sortie proche de l'original, tandis qu'une valeur plus élevée autorise des changements plus importants.
- *Depth-to-Image* [14] estime la distance entre l'appareil photo et les objets présents dans l'image en analysant les pixels, conduisant à la création d'une carte de profondeurs faisant office de conditionnement pour le mécanisme de diffusion. Cela permet la conservation de la structure spatiale des images lors de leur modification.

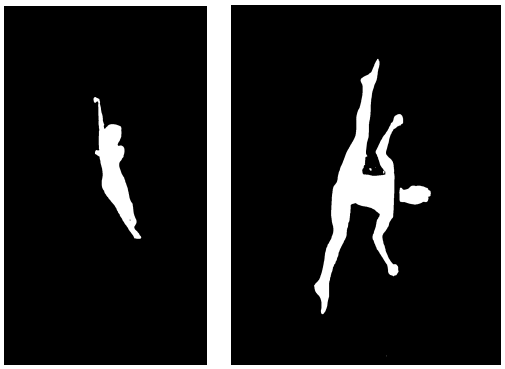
Les meilleurs résultats obtenus par ces modèles sont fournis en annexe, Figs. 4, 5, et 6. À noter qu'aucun des modèles précédemment listés n'a produit de résultats concluants. Nous intuitons que cela provient de leur entraînement : ces modèles ne sont pas entraînés sur un grand ensemble d'images de danseurs ou de gymnastes, ce qui limite leur compréhension de l'anatomie des athlètes en mouvement. Dans des cas plus simples, tels que l'édition d'images d'animaux ou de paysages, les résultats obtenus étaient alors tout à fait satisfaisants.

Dans la section suivante, nous proposons un nouveau *framework* couplant l'utilisation d'un masque pour préserver le corps des athlètes, la segmentation de ce dernier, la complétion de l'image évidée par *inpainting*, et enfin la fusion du masque et de l'image complétée, tout en préservant la cohérence de l'image finale.

3.2 Nouveau cadre de travail

La première étape consiste en la localisation et la segmentation automatique du danseur de l'image à l'aide de *Segment Anything* (SAM, [5]). Étant donnée une image en couleurs de dimension $h \times w$, où h et w représentent respectivement la hauteur et la largeur, l'image est divisée en n patchs [8] de dimensions $p \times p$ tels que $n = \frac{hw}{p^2}$. Les patchs sont redimensionnés et projetés linéairement par une matrice de poids W_p, de sorte à définir le vecteur de caractéristiques x_i associé au i-ième patch P_i par :

$$x_i \;=\; W_p \cdot \texttt{Flatten}(P_i) + E_{pos,i}\,,$$



(a) Masques de ségmentation obtenu par SAM

(b) Poisson Blending

Figure 3: Segmentation et remplissage des deux images d'origine par *Poisson Blending*.

où $E_{pos,i}$ désigne le codage positionnel et correspond à l'information spatiale. Ces vecteurs de caractéristiques sont alors traités comme des séquences d'entrée par un transformeur afin de produire des représentations des caractéristiques de l'image tenant compte du contexte. Les interactions entre les patchs sont modélisées à l'aide de la fonction `CrossAttention` décrite section 2. Le résultat de cette segmentation pour nos images de références est donnée Fig. 3a.

Une fois la segmentation réalisée, l'image évidée de l'athlète est complétée par *inpainting* [1, 13] ; puis, l'image ainsi complétée est modifiée via *Stable Diffusion XL* (*cf.* Section 3.1) conditionné par le masque de segmentation. Ce conditionnement permet de s'assurer de la cohérence et de la conservation de la structure globale de l'image, les parties non masquées agissant alors comme des contraintes. À ce stade, l'image modifiée consiste en une représentation alternée, mais cohérente, du fond de l'image (ici, un studio de danse). En particulier, l'athlète n'est plus présent sur l'image.

Afin de réintégrer les danseurs de manière cohérente dans les images éditées, nous procédons par *Poisson Blending* [11]. L'idée du *Poisson Blending* est de copier des zones d'une image source S (ici, l'athlète) dans une image cible T (ici, le studio modifié par SDXL) tout en imposant des contraintes de recollement et de régularité. Plus précisément, le résultat du *Poisson Blending* est l'image x obtenue via l'équation suivante :

$$\min_x \int_\Omega ||\nabla x - \nabla S||^2 \quad \text{tel que} \quad x_{\mathcal{D}\setminus\Omega} = T\,,$$

où Ω est un masque binaire représentant la zone de S à copier dans T et $\mathcal{D}$ est le domaine des images. Intuitivement, la régularité du recollement provient du fait que l'on copie les gradients spatiaux ∇S de l'image source dans T, et non les valeurs des pixels composants S. Cette équation est résolue à l'aide de solveurs itératifs classiques, par exemple par méthode du gradient conjugué [3]. Les résultats obtenus par cette méthode sont présentés Fig. 3b.

4 Conclusion

Nos expériences ont mis en lumière la difficulté que représente la modification d'une image, tout en conservant l'anatomie de corps en mouvements techniques, tels que ceux rencontrés en danse. En particulier, même les algorithmes à l'état de l'art pour la modification d'images ne parviennent pas à réaliser de telles modifications. Néanmoins, en combinant plusieurs méthodes différentes, nous avons pu obtenir des résultats satisfaisants.

Les perspectives de ce travail sont multiples. On pourra par exemple s'intéresser au passage de la photo à la vidéo, ce qui soulèvera de nouveaux verrous méthodologiques : En effet, les modifications d'un plan à l'autre de la vidéo devront alors être également cohérentes entre elles. Une autre piste de travail envisagée consiste en l'intégration des contraintes biophysiques directement dans le modèle de diffusion, probablement via du conditionnement.

References

[1] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting

[2] Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)

[3] Hestenes, M.R., Stiefel, E., et al.: Methods of conjugate gradients for solving linear systems, vol. 49. NBS Washington, DC (1952)

[4] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[5] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)

[6] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)

[7] Lin, H., Cheng, X., Wu, X., Shen, D.: Cat: Cross attention in vision transformer. In: IEEE international conference on multimedia and expo (ICME). pp. 1–6 (2022)

[8] Nguyen, D.K., Assran, M., Jain, U., Oswald, M.R., Snoek, C.G., Chen, X.: An image is worth more than 16x16 patches: Exploring transformers on individual pixels. Preprint arXiv:2406.09415 (2024)

[9] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. Preprint arXiv:2112.10741 (2021)

[10] O'Meara, J., Murphy, C.: Aberrant ai creations: co-creating surrealist body horror using the dall-e mini text-to-image generator. Convergence **29**(4), 1070–1096 (2023)

[11] Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 577–582 (2023)

[12] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. Preprint arXiv:2307.01952 (2023)

[13] Qin, Z., Zeng, Q., Zong, Y., Xu, F.: Image inpainting based on deep learning: A review. Displays **69**, 102028 (2021)

[14] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)

[15] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

[17] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)

[18] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. Preprint arXiv:2011.13456 (2020)

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

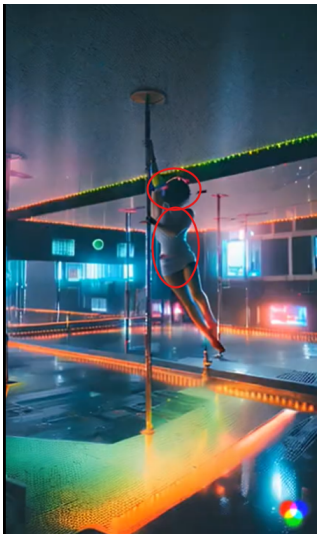
[20] Yu, X., Tian, J., Hu, Z.: An analysis for image-to-image translation and style transfer. Preprint arXiv:2408.06000 (2024)

Annexe

Figure 4: Images modifiées à l'aide d'*Image-to-Image* [20].

L'utilisation d'*Image-to-Image* sur des corps d'athlètes en mouvement conduit à des *body horror*.

(a) Images modifiées par *Depth-to-Image*



(b) Image modifiée par *GEN2*

Figure 5: Images modifiées par *Depth-to-Image* [14] et *GEN-2* [2].

Ces deux réseaux, bien qu'à l'état de l'art pour la modification d'images, altèrent la morphologie des athlètes: *Cf.* le visage et le buste Fig. 5a-Gauche, la musculature Fig. 5a-Droite, les cercles rouges Fig. 5b.

(a) `A woman doing pole dance in the jungle`

(b) `A man doing pole dance in the jungle`

Figure 6: Images générées par SDXL [12], à l'aide d'un texte simple.

La légende de chaque image reporte le *prompt* utilisé pour la génération. Ces exemples tendent à montrer que SDXL est capable de traiter ce type de données.