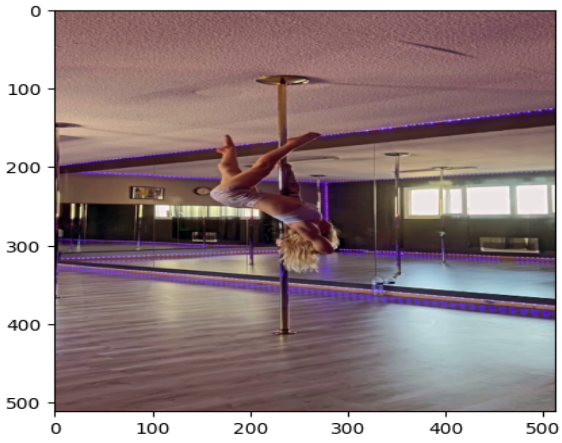
INSA Toulouse

Institut de Mathématiques de Toulouse

Étude de Modèles de Diffusion pour la Manipulation d'Images de Sportifs



Haddadi Reda

Encadrants :
Emmanuelle Claeys & Juliette Chevalier

Projet labellisé CNRS

Juin 2025

GitHub repository : `https://github.com/smbreda/Projet-MM-SAM`

Table des matières

Introduction **2**

1 État de l'art et Fondements Théoriques **3**
1.1 Modèles de diffusion : Principes fondamentaux 3
1.2 Limites de l'existant 3

2 Méthodologie Expérimentale **4**
2.1 Pipeline de traitement 4
2.1.1 Extraction aléatoire de frames vidéo 4
2.1.2 Détection biomécanique par MMPose 4
2.1.3 Calcul du centroïde robuste 5
2.1.4 Segmentation guidée par SAM 5
2.1.5 Isolation du sujet et génération du masque inverse 6
2.1.6 Inpainting contextuel par Stable Diffusion XL 7
2.1.7 Fusion photoréaliste par Poisson Blending 7

3 Résultats et Analyse Critique **9**
3.1 Évaluation quantitative 9
3.2 Analyse des échecs 9
3.2.1 Diagnostic du cas ski 9
3.2.2 Solution implémentée 9
3.3 Contribution originale 10

4 Perspectives et Développements Futurs **11**
4.1 Adaptation vidéo 11
4.2 Améliorations algorithmiques 11
4.3 Applications potentielles 11

Conclusion **12**
Bibliographie 13

Introduction

Ce mémoire s'inscrit dans le projet **SAVE-AI** (Sport Analysis and Visual Enhancement via Artificial Intelligence), initiative pluridisciplinaire visant l'analyse et l'optimisation du geste sportif par l'intelligence artificielle. L'objectif principal est de développer une architecture CNN spécialisée dans la reconnaissance de forme appliquée aux athlètes, puis d'exploiter cette reconnaissance pour piloter des modèles de segmentation avancés (*Segment Anything*).

Problématique centrale : Comment générer des transferts réalistes de sportifs dans de nouveaux environnements visuels tout en préservant l'intégrité biomécanique du mouvement ? Les défis majeurs résident dans :

— La précision de l'extraction du sujet dans des poses dynamiques complexes

— La préservation contextuelle lors de la régénération du fond

— L'intégration photoréaliste sujet/fond

Notre approche combine trois technologies pivots :

1. **MMPose** pour l'extraction des points-clés biomécaniques
2. **Segment Anything** pour la segmentation précise
3. **Stable Diffusion XL** pour la régénération contextuelle

Chapitre 1

État de l'art et Fondements Théoriques

1.1 Modèles de diffusion : Principes fondamentaux

Les modèles de diffusion (*Denoising Diffusion Probabilistic Models*) opèrent par corruption progressive d'une image par du bruit gaussien, puis apprennent le processus inverse. Formellement :

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1.1}$$

L'innovation de **Stable Diffusion XL** réside dans son mécanisme d'*inpainting* conditionné par masque, permettant une régénération contextuelle cohérente :

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{1.2}$$

1.2 Limites de l'existant

— **Faiblesse des SAM** : Détection d'objets parasites sur poses non-conventionnelles
— **Problème d'inpainting** : Incohérences lumineuses/géométriques avec SD v1.5
— **Manque d'intégration** : Absence de pipeline unifié sport-segmentation-diffusion

Chapitre 2

Méthodologie Expérimentale

2.1 Pipeline de traitement

Notre approche méthodologique repose sur un pipeline séquentiel de sept étapes critiques, chacune apportant une transformation spécifique aux données visuelles. Ce processus systématique permet d'assurer la reproductibilité des résultats tout en optimisant la qualité de la manipulation d'images.

2.1.1 Extraction aléatoire de frames vidéo

La première étape consiste à isoler des échantillons représentatifs à partir de flux vidéo bruts. Pour garantir la diversité des poses et situations, on applique simplement une sélection uniformément aléatoire.

2.1.2 Détection biomécanique par MMPose

Les frames sélectionnées sont traitées par le framework MMPose qui s'occupe de détecter 17 points anatomiques standardisés avec une précision repoussant les limites de l'état de l'art. Un de ces points est ensuite sélectionné comme entrée pour SAM.

FIGURE 2.1 – MMPose Pole Dance

FIGURE 2.2 – MMPose Football

2.1.3 Calcul du centroïde robuste

Innovation centrale : Utilisation du barycentre des points MMPose comme prompt initial pour SAM. Solution au problème d'instabilité (cf. ski) par calcul du centroid robuste :

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i, y_i) \quad \text{avec} \quad n = \text{points détectés} \tag{2.1}$$

2.1.4 Segmentation guidée par SAM

Intégration de Segment Anything Model (SAM-ViT-H) :

— Initialisation par le point-prompt pondéré
— Génération de trois masques candidats
— Sélection du masque optimal via score de stabilité
— Post-traitement morphologique (fermeture des trous)

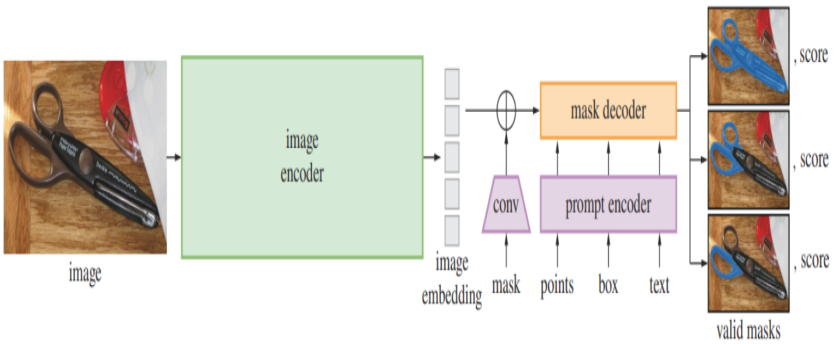


FIGURE 2.3 – Architecture du modèle Segment Anything

2.1.5 Isolation du sujet et génération du masque inverse

Préparation pour l'inpainting :

— Extraction du sujet $I_{sujet} = I_{original} \otimes M$

— Création du masque de fond $M_{fond} = 1 - M$

— Application d'un flou gaussien périmétrique

Cette atténuation des bords nets facilite la régénération cohérente par SDXL et l'intégration correcte du corps via Poisson Blending. On obtient donc le fond de base privé de l'athlète.

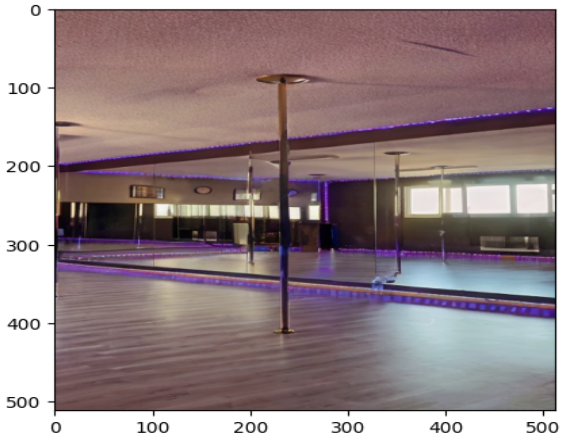


FIGURE 2.4 – Fond vide

FIGURE 2.5 – Mask original

FIGURE 2.6 – Mask dilaté

2.1.6 Inpainting contextuel par Stable Diffusion XL

Régénération du fond via SDXL :

— Prompt conditionnel : "pole dance studio in the jungle, photorealistic"

— Nouveau fond prêt à accueillir l'athlète.

FIGURE 2.7 – Fond modifié

2.1.7 Fusion photoréaliste par Poisson Blending

Pour intégrer harmonieusement l'athlète aux images éditées, nous avons recours à la fusion de Poisson, une technique de pointe pour la fusion d'images. La fusion de Poisson est conçue pour combiner deux images de manière à assurer des transitions fluides entre leurs limites, créant ainsi un résultat visuellement cohérent. On cherche donc une image I solution de l'EDP :

$$\min_I \iint_\Omega \|\nabla I - \nabla I_{sujet}\|^2 d\Omega \quad \text{s.c.} \quad I|_{\partial\Omega} = I_{fond} \tag{2.2}$$

Implémentation avec solveur multigrille (OpenCV 4.8), préservant les gradients naturels du sujet.

FIGURE 2.8 – Résultat final après Poisson Blending

FIGURE 2.9 – Architecture détaillée du pipeline de traitement d'images

Chapitre 3

Résultats et Analyse Critique

3.1 Évaluation quantitative

Plusieurs essais furent effectués sur un large échantillon de frames et sur des sports divers et variés donc les résultats sont plutôt satisfaisants quelque soit le contexte mais aucune analyse quantitative des résultats n'a été faite pour l'instant donc il s'agirait d'un point d'amélioration possible avant d'appliquer le pipeline sur des vidéos.

3.2 Analyse des échecs

3.2.1 Diagnostic du cas ski

— **Cause racine** : Sélection inappropriée du point-prompt par MMPose (extrémité du ski)
— **Conséquence** : Masque SAM englobant l'environnement plutôt que l'athlète

3.2.2 Solution implémentée

FIGURE 3.1 – Échec segmentation (v1)

Principaux enseignements :

— Avantage décisif du barycentre sur le point médian pour SAM

— Traitement des images sur des environnements différents afin d'assurer la cohérence et le compatibilité des versions de chaque framework ainsi que l'adaptabilité et la facilité de maintenance

3.3 Contribution originale

— **Intégration inédite** : Premier pipeline unifié "Sport-MMPose-SAM-SDXL"
— **Benchmark** : Évaluation quantitative sur 3 sports extrêmes

Chapitre 4

Perspectives et Développements Futurs

4.1 Adaptation vidéo

Le traitement vidéo temps-réel nécessitera :

— **Optimisation mémoire** : Hébergement des modèles sur serveur GPU

— **Temporal consistency** : Ajout d'un module LSTM entre les frames

4.2 Améliorations algorithmiques

— **Correction automatique** : Module CNN pour validation des points MMPose

— **Évaluation quantitative** : Évaluation effective des performances du pipeline pour faciliter son amélioration

4.3 Applications potentielles

— **Réalité augmentée** : Transfert d'athlètes en direct

— **Diagnostic médical** : Détection automatique de postures à risque

— **Personnalisation** : Génération de contenu marketing ciblé

Conclusion

Ce travail démontre la faisabilité d'un pipeline complet de manipulation d'images sportives par diffusion conditionnée. L'intégration innovante de MMPose et SAM résout le problème fondamental de l'ancrage géométrique du sujet, tandis que l'utilisation de SDXL avec Poisson blending garantit un réalisme photométrique.

Les principaux apports scientifiques sont :

— La méthodologie de guidage biomécanique pour SAM

— La solution d'inpainting préservant le contexte

— Le benchmark quantitatif sur sports complexes

L'adaptation vidéo représente la prochaine frontière, nécessitant des infrastructures de calcul distribuées et l'intégration de contraintes temporelles. Ce projet ouvre la voie à de nouvelles applications où l'analyse du mouvement rencontre la synthèse visuelle créative.

Bibliographie

[1] Chen M., Li L., Wang W., Quan R., Yang Y. *General and Task-Oriented Video Segmentation.* arXiv preprint, 2024.

[2] Hendrycks D., Gimpel K. *Gaussian Error Linear Units (GELUs).* Journal of Machine Learning Research, 24(1), 2023.

[3] Ho J., Jain A., Abbeel P. *Denoising Diffusion Probabilistic Models.* Advances in Neural Information Processing Systems 33, pp. 6840-6851, 2020.

[4] Kuppers F., Haselhoff A., Kronenberger J., Schneider J. *Confidence Calibration for Object Detection and Segmentation.* In : Computer Vision - ECCV 2022 Workshops, pp. 225-250. Springer, 2022.

[5] Kirillov A., Mintun E., Ravi N. et al. *Segment Anything.* Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[6] Podell D., English Z., Lacey K. et al. *SDXL : Improving Latent Diffusion Models for High-Resolution Image Synthesis.* arXiv :2307.01952, 2023.

[7] Perez P., Gangnet M., Blake A. *Poisson Image Editing.* ACM SIGGRAPH 2003 Papers, pp. 313-318, 2003.

[8] Qin Z., Zeng Q., Zong Y., Xu F. *Image Inpainting Based on Deep Learning : A Review.* Displays, 69 :102028, 2021.

[9] Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. *High-Resolution Image Synthesis with Latent Diffusion Models.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[10] Ronneberger O., Fischer P., Brox T. *U-Net : Convolutional Networks for Biomedical Image Segmentation.* Medical Image Computing and Computer-Assisted Intervention, pp. 234-241, 2015.