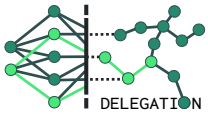
IRIMAS UNIVERSITÉ HAUTE-ALSACE
Institut de Recherche en Informatique, Mathématiques, Automatique et Signal

# Master Internship Position

## Deep Learning architectures for generating skeleton-based human motion




## 1 Information about the internship

- **Supervisors:**
  - Dr. Maxime Devanne ( https://maxime-devanne.com)
  - Dr. Jonathan Weber (https://www.jonathan-weber.eu)
  - Prof. Germain Forestier (https://germain-forestier.info)
- **Location:** UHA/IRIMAS EA 7499, Mulhouse, France
- **Duration:** 6 months (starting from February or March 2022)
- **Salary:** 600.60€ monthly
- **Keywords:** Deep Learning, generative models, human motion, time series

## 2 Context

Human motion analysis is crucial for studying people and understanding how they behave, communicate and interact with real world environments. Due to the complex nature of body movements as well as the high cost of motion capture systems, acquisition of human motion is not straightforward and thus constraints data production. Hopefully, recent approaches estimating human poses from videos offer new opportunities to analyze skeleton-based human motion [1, 2]. While skeleton-based human motion analysis has been extensively studied for behavior understanding like action recognition, some efforts are yet to be done for the task of human motion generation. Particularly, the automatic generation of motion sequences is beneficial for rapidly increasing the amount of data and improving Deep Learning-based analysis algorithms.

Since several years, new image generation paradigms have been possible thanks to the appearance of Generative Adversarial Networks (GAN) [3] which

have proved to be extremely efficient for many image generation tasks and human posture estimation [4]. Although these networks are very efficient, their explainability and control still remain challenging tasks. Differently, other generative models have also emerged by considering the data distribution during training like Variational AutoEncoder (VAE) [5] and Diffusion models [6]. First work addressing deep generative models for human motion have considered motion capture (mocap) data allowing to accurately extract body parts positions along the time. Hence, aforementioned generative architectures have been successively employed for generating mocap-based human motion sequences [7, 8].

Differently, we consider noisy skeleton data estimated from videos as it is easily applicable in real-world scenarios for the general public.

# 3 Goals

The goal of this internship is to provide guidelines in building deep generative models for skeleton-based human motion sequences. Inspiring from recent effective Deep Learning-based approaches [9, 10, 11, 12], the aim is to generate full skeleton-based motion sequences without access to successive poses as prior information as it can be done in prediction tasks. It is therefore crucial to investigate how deep generative models can handle such noisy and possibly incomplete data in order to generate novel motion sequences as natural and variable as possible.

In particular, the candidate will work on the following tasks:

- **Deep Learning architectures for skeleton-based human motion:** investigation and assessment of the influence of different deep network architectures for capturing complex human motion features. Particularly, the goal of this task is to theoretically and empirically analyze the performance of existing architectures like CNN, RNN GCN and Transformers for modeling skeleton-based human motion.
- **Deep generative models adapted to skeleton data:** based on studies from the previous task, the goal is to build generative models upon the previously identified meaningful spaces where skeleton sequences are represented. Therefore, the candidate will investigate different generative models, like GAN, VAE and Diffusion models, in order to propose and develop a complete Deep Learning model for generating skeleton-based human motions.
- **Evaluation of deep generative models:** in order to validate the proposed model, experimental evaluation is crucial. In comparison to motion recognition where classification accuracy is a natural way to assess an approach, evaluating the task of motion generation is not as straightforward. Dedicated metrics evaluating both naturalness and diversity of generated sequences as well as the impact of new generated sequences in a classification task will be considered.

## 4 Profile of applicant

The candidate must fit the following requirements:

- Registered in Master 2 or last year of Engineering School (or equivalent) in **Computer Science**
- Advanced skills in **Python programming** are mandatory
- Good skills in **Machine Learning & Deep Learning** using related libraries (scikit-learn, Tensorflow, Pytorch, etc.) are required
- Knowledge and/or a first experience in **human motion analysis** will be appreciated

## 5 Research environment

The proposed internship will be carried out within the MSD (Modeling and Data Science) team from the IRIMAS Institute. It will be part of the ANR DELEGATION project [1].

## 6 Application

For further information or for applying, candidates should send a **CV, academic records, personal projects (e.g. github repo) and a motivation letter** to maxime.devanne@uha.fr.

[1] https://maxime-devanne.com/delegation/

# References


[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[2] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[4] C.-J. Chou, J.-T. Chien, and H.-T. Chen, "Self adversarial training for human pose estimation," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 17–30, IEEE, 2018.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv preprint arXiv:2208.15001*, 2022.

[7] X. Lin and M. R. Amer, "Human motion modeling using dvgans," *preprint arXiv:1804.10652*, 2018.

[8] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, T. Komura, J. Saito, I. Kusajima, X. Zhao, M.-G. Choi, R. Hu, *et al.*, "A recurrent variational autoencoder for human motion synthesis.," in *BMVC*, 2017.

[9] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2900, 2017.

[10] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5226–5234, 2018.

[11] Y. Cai, L. Huang, Y. Wang, T.-J. Cham, J. Cai, J. Yuan, J. Liu, X. Yang, Y. Zhu, X. Shen, *et al.*, "Learning progressive joint propagation for human motion prediction," in *European Conference on Computer Vision*, pp. 226–242, Springer, 2020.

[12] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995, 2021.