You Can Dance! Generating Music-Conditioned Dances on Real 3D Scans

Elona Dupont[a], Inder Pal Singh[b], Laura Lopez Fuentes[c], Sk Aziz Ali[d], Anis Kacem[e], Enjie Ghorbel[f] and Djamila Aouada[g]
SnT, University of Luxembourg, Luxembourg
{*elona.dupont, inder.singh*}*@uni.lu*

Keywords: Dance Generation, 3D Human Scan, 3D Animation, 3D Human Body Modeling.

Abstract: The generation of realistic body dances that are coherent with music has recently caught the attention of the Computer Vision community, due to its various real-world applications. In this work, we are the first to present a fully automated framework *'You Can Dance'* that generates a personalized music-conditioned 3D dance sequence, given a piece of music and a real 3D human scan. *'You Can Dance'* is composed of two modules: (1) The first module fits a parametric body model to an input 3D scan; (2) the second generates realistic dance poses that are coherent with music. These dance poses are injected into the body model to generate animated 3D dancing scans. Furthermore, the proposed framework is used to generate a synthetic dataset consisting of music-conditioned dancing 3D body scans. A human-based evaluation study is conducted to assess the quality and realism of the generated 3D dances. This study along with the qualitative results shows that the proposed framework can generate plausible music-conditioned 3D dances.

1 INTRODUCTION

With the recent advances in Artificial Intelligence (AI), the creation of artistic content has impressively progressed (Hernandez-Olivan and Beltran, 2021; Ramesh et al., 2022; Pu and Shan, 2022). In particular, the task of music-conditioned body dance generation has attracted a lot of interest (Tang et al., 2018b; Sun et al., 2020; Li et al., 2021; Siyao et al., 2022; Pu and Shan, 2022). Given a music piece, the aim of this task is to generate realistic human body dances that are coherent with music. Developing automatic systems for generating such dances plays an important role in numerous real-world applications, *e.g.*, assisting in the design of choreography and driving virtual characters performance (Siyao et al., 2022; Kärki, 2021). Previous works have focused on the generation of realistic music-conditioned dance motions in the form of 2D or 3D skeletons, and 3D template characters (Lee et al., 2019; Huang et al., 2020; Li et al., 2021; Siyao et al., 2022). However, the problem of dance generation without the prior knowledge of the 3D character to animate remains underexplored.

Earlier attempts have generated 2D dancing skeletons by exploiting internet videos (Lee et al., 2019; Huang et al., 2020). In particular, real video clips with music and professional dances have been collected and processed to extract 2D skeletons. This data along with the corresponding music pieces have been used for training different types of neural networks. In addition, Huang et al. (Huang et al., 2020) have demonstrated the practical interest of producing music-conditioned 2D dancing skeletons. The latter have been used for generating realistic RGB videos of dances. This was achieved by transferring motion patterns of the 2D dancing skeletons to a person in an input video/image using video-to-video translation techniques (Wang et al., 2018). Another branch of methods have focused on generating 3D dancing skeletons (Tang et al., 2018b; Sun et al., 2020; Li et al., 2021; Siyao et al., 2022; Pu and Shan, 2022) using dedicated datasets. To lift the ground-truth from 2D internet video clips to 3D, the authors in (Sun et al., 2020) have used a 3D pose estimation method and have employed the resulting 3D skeletons to train a Generative Adversarial Net-

[a] https://orcid.org/0000-0003-4045-5651
[b] https://orcid.org/0000-0002-4870-1426
[c] https://orcid.org/0000-0001-7850-4776
[d] https://orcid.org/0000-0002-6396-8436
[e] https://orcid.org/0000-0003-0640-9862
[f] https://orcid.org/0000-0002-6878-0141
[g] https://orcid.org/0000-0002-7576-2064

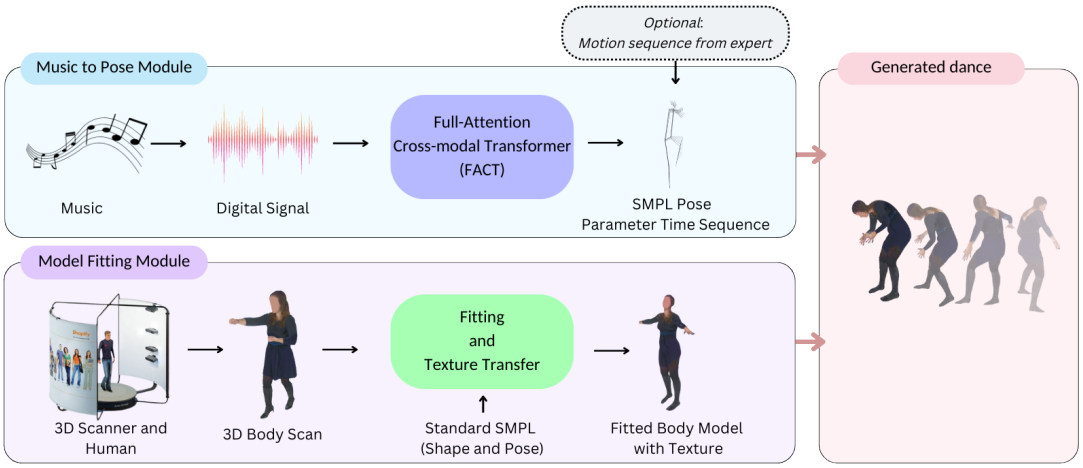

Figure 1: The *'You Can Dance'* framework is composed of two main modules. The *music to pose* module generates a sequence of pose parameters. This sequence represents a choreography obtained using the FACT model (Li et al., 2021) given an input music. The *model fitting* module fits the pose, shape, and texture to a template SMPL model (Loper et al., 2015). Combining these outputs, a music-conditioned animated 3D scan is obtained.

work (GAN). In (Tang et al., 2018a), an accurate motion capture sensor has been utilized to acquire 3D skeletal dances performed by professional actors. The collected data has been in turn employed to train a Long-Short Term Memory (LSTM) autoencoder. More recently, Li et al. (Li et al., 2021) have introduced the AIST++ dataset which extends the AIST dance benchmark (Tsuchida et al., 2019) by supplying it with richer motion annotations such as 2D and 3D skeletons, rigid body transformations, and SMPL motion parameters (Loper et al., 2015). These annotations were particularly useful for bridging the gap between the generation of dance motions and the actual animation of 3D characters dances. Multiple works have proposed transformer-based architectures for generating dances (Li et al., 2021; Huang et al., 2022; Pu and Shan, 2022; Siyao et al., 2022), leveraging the prediction of SMPL motion parameters to animate a 3D character template. However, to the best of our knowledge, none of them have considered the animation of personalized 3D characters, such as real 3D human scans acquired by dedicated sensors.

With the recent advances in 3D sensing and vision-based human body modelling, it became possible to obtain human body models with high accuracy that can represent real humans in virtual environments (Pavlakos et al., 2019; Loper et al., 2015; Saint et al., 2018; Saint et al., 2019). In this work, we propose to couple these 3D human modeling techniques with dance generation approaches to produce music-conditioned real 3D scans dancing in virtual environments. Given a textured 3D scan and a piece of music, the proposed system, *'You Can Dance'* , can automatically animate 3D scans with synthetically generated dance motions. This is achieved by first fitting an SMPL body model on a 3D scan to model its shape, texture, and initial pose. Then, the dance motions are generated using a recent method (Li et al., 2021) in the form of SMPL pose parameters. Finally, the dancing scan is obtained by replacing the initial SMPL pose with the predicted SMPL poses of the dance over time. Hence, the *'You Can Dance'* framework generates an animated 3D scan that realistically follows a uniquely generated choreography that is consistent with an input piece of music. Alternatively, the ground-truth dances from the AIST++ dataset (Li et al., 2021) can also be used to animate 3D scans as part of the *'You Can Dance'* framework. As a result, a novel dataset is synthetically generated by animating 3D real scans with different types of professional choreography. The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, we propose the *'You Can Dance'* framework as the first automatic system for music-conditioned dance animation of real 3D scans.
- we introduce a synthetic dataset consisting of animated 3D scans with dances performed by experts following music. The *'You Can Dance'* system has been used to animate 3D scans from the 3DBodyTex.v2 dataset (Saint et al., 2020b) based on dance motions from the AIST++ dataset (Li et al., 2021).
- we present a human-based study assessing the quality of the automatically generated 3D dances against the ground truth dances.

The rest of the paper is organized as follows. Section 2 formulates the problem of producing music-conditioned dances on real 3D human

scans. In Section 1, the proposed framework called *'You Can Dance'* is described. The experiments and discussions are sketched in Section 4. Finally, Section 5 concludes the paper and draws some perspectives.

2 PROBLEM FORMULATION

Let $\mathbf{m} = \{m_t\}_{t \in K} \in \mathcal{M}$ be a given digital signal representation of a music piece, with t referring to the time-stamp, K to the set of sampled time-stamps, n to the cardinality of K, and $\mathcal{M}$ to the space of digital signal representation of all music pieces. Let us denote by $\mathbf{X} \in \mathcal{X}(\mathcal{S}, \mathcal{P}, \mathcal{U})$ a textured human body parametrized by a shape $\mathbf{s} \in \mathcal{S}$, a pose $\mathbf{p} \in \mathcal{P}$, and a texture $\mathbf{u} \in \mathcal{U}$, with $\mathcal{S}$ being the space of human shapes, $\mathcal{P}$ the space of human poses, $\mathcal{U}$ the space of human body textures, and $\mathcal{X}(\mathcal{S}, \mathcal{P}, \mathcal{U})$ the human body space. Let $\mathcal{X}(\mathbf{s}, \mathcal{P}, \mathbf{u}) \subset \mathcal{X}(\mathcal{S}, \mathcal{P}, \mathcal{U})$ be the human body space with fixed shape and texture. Let $\hat{K} = \{\hat{k}\}_{1 \leq \hat{k} \leq n'}$ be an under-sampling of K at a given rate and n' the cardinality of $\hat{K}$. In this paper, our goal is to estimate a function $f : \mathcal{X}(\mathbf{s}, \mathcal{P}, \mathbf{u}) \times \mathcal{M} \times \hat{K} \mapsto \mathcal{X}(\mathbf{s}, \mathcal{P}, \mathbf{u})$ at each time-stamp $t \in \hat{K}$ that computes the animated textured body shape $\mathbf{X}_t$ coherently with the music $\mathbf{m}$ as follows,

$$\mathbf{X}_t = f(\mathbf{X}, \mathbf{m}, t) \; . \tag{1}$$

3 PROPOSED FRAMEWORK

3.1 Overview of the Framework

The *'You Can Dance'* framework aims at automatically animating a real 3D scan $\mathbf{X}_s$ with a generated dance that is coherent with a given piece of music $\mathbf{m}$. As depicted in Figure 1, the framework is composed of two main components, namely, a *model fitting* module followed by a *music to pose* module. The function f defined in Equation (1) implies updating the pose parameters of the input body while fixing the shape and texture. However, a 3D scan $\mathbf{X}_s = (\mathbf{V}_s, \mathbf{F}_s, \mathbf{u}_s)$ is often represented by a set of n_v vertices $\mathbf{V}_s \in \mathbb{R}^{n_v \times 3}$, a set of n_f faces $\mathbf{F}_s \in \mathbb{N}^{n_f \times 3}$, a texture parametrization $\mathbf{u}_s \in \mathbb{R}^{n_v \times 2}$, and a texture-atlas $\mathbf{A}_s$ in the form of an RGB image. Consequently, $\mathbf{X}_s$ does not encode any pose or shape parametrization. The proposed *model fitting* module parametrizes the input scan $\mathbf{X}_s$ with shape and pose parameters through a function h_1 that is defined in Section 3.2. The *music to pose* module takes as input a piece of music $\mathbf{m}$ and generates a pose at each time-stamp t through a function h_2 which is defined in Section 3.3. Finally, the output poses generated by the *music to pose* module are used to update the pose parameters of the parametrized scan through a function g that is described in Section 3.4. Overall, the function f defined in Equation (1) is decomposed as follows,

$$\begin{aligned} \mathbf{X}_t &= f(\mathbf{X}_s, \mathbf{m}, t) \\ &= g(h_1(\mathbf{X}_s), h_2(\mathbf{m}, t)) \; . \end{aligned} \tag{2}$$

3.2 Model Fitting Module

We employ an integrated rigid alignment and non-rigid surface fitting method (Saint et al., 2019) for automatic body model to body scan fitting task. More formally, the role of this module is to fit a parametric body model $\mathbf{X}_f$, with pose and shape parameters, to the input body scan $\mathbf{X}_s$ through a function $h_1 : \mathbb{R}^{n_v \times 3} \times \mathbb{N}^{n_f \times 3} \times \mathbb{R}^{n_v \times 2} \mapsto \mathcal{X}(\mathcal{S}, \mathcal{P}, \mathbf{u})$ such that,

$$\mathbf{X}_f = h_1(\mathbf{X}_s) \; . \tag{3}$$

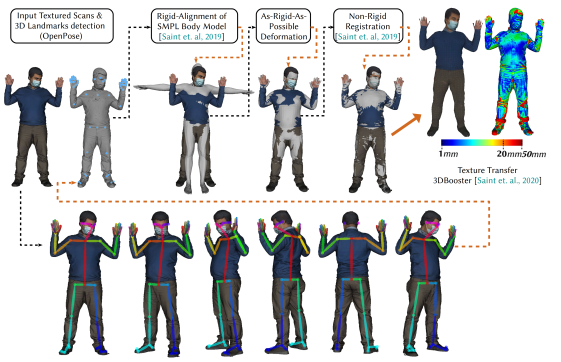

Figure 2: Body model to 3D scan fitting and texture transfer pipeline.

In practice, this function h_1 is defined using an SMPL body model fitting (Loper et al., 2015). As performed in (Saint et al., 2019), this is achieved by quantifying the SMPL pose parameters, which allows the estimation of the shape parameters. Note that SMPL has a fixed texture parametrization $\mathbf{u}$, which is different from that of the input scan. A texture transfer from the scan to the fitted SMPL body model is therefore necessary. Figure 2 shows the different stages of our fitting and texture transfer process. In order to estimate the pose parameters $\mathbf{p}$, 3D keypoints are predicted on the input 3D scan $\mathbf{X}_s$. These are obtained by firstly predicting 2D body keypoints from different 2D views of the scan using OpenPose[1] (bot-

[1] https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/pages.html

tom part of Figure 2). The 2D keypoints are then projected back to the 3D scan. Furthermore, a rigid alignment is used to align the pose of the body model $\mathbf{X}_f$ with that of the input scan $\mathbf{X}_s$. Next, an As-Rigid-As-Possible (ARAP) deformation is applied on the segments of the SMPL body that correspond to the keypoints detected on scan. The ARAP deformation smoothly deforms the vertices around joints. The shape parameters $\mathbf{s}$ are estimated such that the shape of the model matches the scan. This shape fitting is further refined using a non-rigid registration technique (Saint et al., 2019). The final step consists of transferring the texture of the scan $\mathbf{X}_s$ to the fitted SMPL model $\mathbf{X}_f$ using a ray-casting algorithm that propagates the texture information along the surface normal directions as described in (Saint et al., 2020a). Note that this step will preserve the texture parametrization $\mathbf{u}$ of the SMPL body model.

3.3 Music to Pose Module

The role of the *music to pose* module is to generate a temporal sequence of SMPL pose parameters $\{\mathbf{p}_t\}_{t\in\hat{K}}$, with $\mathbf{p}_t \in \mathcal{P}$, given an input piece of music $\mathbf{m}$. This can be modelled as a function $h_2 : \mathcal{M} \times \hat{K} \mapsto \mathcal{P}$ that computes,

$$\mathbf{p}_t = h_2(\mathbf{m}, t) \ . \tag{4}$$

Note that since the audio sampling rate is different from the video frame rate, the number of audio samples $card(K) = n$ in $\mathbf{m}$ differs from the number of corresponding poses $card(\hat{K}) = \hat{n}$. The output pose parameter sequence, $\{\mathbf{p}_t\}_{t\in\hat{K}}$, represents a succession of realistic human movements making a coherent choreography with respect to the style and tempo of the input music. As a proof of concept, the Full-Attention Cross-modal Transformer (FACT) (Li et al., 2021) model is chosen, and in practice any model that predicts 3D body keypoints or SMPL parameters could be used. Moreover, a motion sequence captured from a dancer could also be directly used to animate the 3D body scan.

In the FACT model (Li et al., 2021), audio features, such as envelope, MFCC, chroma, one-hot peaks, and one-hot beats, are first extracted from the input audio signal using a predefined temporal window size. These features are passed through an audio transformer to obtain embeddings for each window. Similarly, a motion transformer is used to compute the corresponding per-frame motion embeddings from the rotation matrices of 24 joints and a global translation matrix. The resulting audio and motion embeddings are then concatenated before being passed into a cross-modal transformer that outputs the predicted pose parameters for the following 20 frames. Given a two-second seed motion, the FACT model can output a realistic 3D dance motion for over 1200 frames with a resolution of 60 frames per second.

3.4 Dancing 3D Scans

In the last step of the *'You Can Dance'* framework, the pose parameter sequence $\{\mathbf{p}_t\}_{t\in\hat{K}}$ from the *music to pose* module are applied to the fitted model $\mathbf{X}_f$ from the *model fitting* module. This process can be modelled by the function $g : \mathcal{P} \times \mathcal{X}(\mathbf{s},\mathbf{p},\mathbf{u}) \mapsto \mathcal{X}(\mathbf{s},\mathcal{P},\mathbf{u})$, which computes,

$$\mathbf{X}_{f_t} = g(\mathbf{p}_t, \mathbf{X}_f, t) \ , \tag{5}$$

at each time-stamp t. This function replaces the pose parameters of the fitted SMPL body model $\mathbf{X}_f$ with the ones generated by the *music to pose* module. As a result, a realistic animation of a 3D scan representing a coherent choreography is obtained.

4 EXPERIMENTS

In this section, we describe the used datasets, the implementation details as well as the obtained results.

4.1 Datasets

Two datasets have been used in our experiments. The first one called AIST++ is used to train the music to pose module. The second dataset termed 3DBodytex is formed by real 3D human scans and has been employed to generate the proposed synthetic 3D scan-based dance dataset.

AIST++ Dance Motion: It is a large-scale 3D dance motion dataset (Li et al., 2021) that contains a wide variety of 3D dances paired with music pieces. More specifically, it is formed by a total of 1408 sequences containing 30 subjects and 10 dance genres with 85% of basic and 15% of advanced choreography. The 10 dance genres are Old School (Break, Pop, Lock and Waack) and New School (Middle Hip-hop, LA-style Hip-hop, House, Krump, Street Jazz and Ballet Jazz).

3DBodyTex: To generate the proposed synthetic 3D dance motion dataset, the 3D scans proposed in 3DBodyTex.v2 dataset (Saint et al., 2020b) are used. This dataset contains nearly 3000 static 3D scans of

Figure 3: Samples of generated dances. Each row represents 15 frames corresponding to 1.5 s of the generated dance. The dances are generated with a frame rate of 60 fps.

around 500 different human subjects with a high-resolution texture. Each participant has been captured in at least 3 poses. The scans are captured in both close-fitting clothing and arbitrary casual clothing. Additionally, the texture is evenly illuminated and reflects the natural appearance of the body.

4.2 Implementation Details

The *'You Can Dance'* system is designed in a modular and ease-of-use fashion.

A pre-trained version of the FACT model (Li et al., 2021) on the AIST++ dataset has been used for predicting pose parameters from a given piece of music.

The implementation of the model fitting and texture transfer method is based on several cross-platform functional components – *e.g.,* ***(i)*** OpenPose docker NVIDIA container [2], ***(ii)*** a Python SMPL model-to-scan fitting module, and lastly, ***(iii)*** a C++ ray-casting based proprietary algorithm for texture-transfer. As a result, the *'You Can Dance'* framework supports future updates as each component can easily be upgraded.

4.3 Qualitative Results

In this section, a qualitative analysis of the generated 3D scan dances is presented. Figure 3 shows three examples of generated dances using different input music pieces applied to different input 3D scans.

The top row of Figure 3 depicts an example of a realistic sequence of dance moves. In this example, the body fitting appears to retain the characteristic of the original 3D body scan. However, the edges of the clothes (bottom of the white T-shirt) do not appear as smooth and sharp as expected from real clothing. This flaw is a consequence of the texture transfer technique, which fails to capture fine details.

The example in the middle row shows ballet-like dance moves. Indeed, the input music consists of a slow piano melody, and the dance moves are synchronized with the tempo of the music. As a result, it can be concluded that the music-to-pose module is able to generate dance moves that are coherent with the music style. This example, however, highlights one of the main limitations of the model fitting module. The original model wears a dress and shoes with heels as seen in Figure 1. Since the model fitting module aims at fitting a 3D scan on one of the male, female, or on a neutral SMPL template, any details such as large clothing or long floating hair are lost at the fitting stage. This limitation also reveals the inability of the framework to adapt to gender non-conforming individuals. Hence, we acknowledge the limitation of our work in representing gender non-conforming persons, which can cause real-world harm (Dodik et al., 2022).

While the example on the bottom row of Figure 3 appears to be realistic, it can be observed that the different body-parts barely move over time. This near-static body motion cannot be considered to be a dance choreography. This freezing motion is a common problem in predictive models (Aksan et al., 2021). Nevertheless, it was found that this phenomenon only occurs in few instances.

From the different examples of Figure 3, it can be concluded that the combination of the music to pose and the model fitting modules create 3D body dances that are realistic. However, it was observed that on very rare occasions the model fitting module led to body poses that are unnatural. Figure 4 shows two failure cases of fitted bodys obtained from a real

[2] https://hub.docker.com/r/cwaffles/openpose

(a) (b)

Figure 4: Failure cases of the SMPL based animation using dances from experts.

dance choreography. In particular, the chest is positioned unrealistically forward in Figure 4a and the rotation of the left elbow appears to be unnatural in Figure 4b.

In summary, it can be concluded that the generated dance choreographies appear to be mostly realistic and coherent with respect to the input music.

4.4 Human Perception Study

Evaluating the quality of the generated dances objectively is not straightforward. In fact, dance being an art form, its appreciation may depend widely on the audience. Therefore, we propose to conduct a human perception study involving 25 participants between 22 and 50 years old. In this context, the participants are asked to give their feedback on the generated 3D body models. More specifically, the participants are exposed to:

- **AI-generated Videos:** they correspond to two videos that have been generated by animating 3D scans with synthetically generated dances (Video 1 and Video 2),
- **Real Dances:** they represent two videos (that can be seen as the ground-truth of Video 1 and Video 2), where the 3D scans are animated with real dances.

Note that we do not disclose the video labels (real or AI-generated) to the participants. Hence, we record the opinion of the participants with respect to four main criteria:

- **The Overall Smoothness (Smoothness):** the participants are asked to rate the smoothness of each video on a scale from 1 to 5 (with 5 being the highest score);
- **The Synchronization Between Dance Steps and Audio Beats (Synchronization):** the participants are asked to rate the synchronization between dance steps and audio beats for each video on a scale from 1 to 5 (with 5 being the highest score).

Table 1: Results from the human study survey.

Name	Type	Average Scores			
		Smoothness	Synchronization	Realism	Reality confidence
Video 1	AI-based	2.7	2.7	2.3	34.6%
	Real	3.0	2.5	2.3	50%
Video 2	AI-based	2.9	2.9	2.7	46.2%
	Real	3.0	2.3	2.5	42.3%

- **The Overall Realism within the Video (Realism):** the participants are asked to rate the overall realism within each video on a scale from 1 to 5 (with 5 being the highest score).
- **The Reality Confidence:** the participants are asked whether they find the video real or not. The percentage of positive answers is then computed.

Table 1 summarizes the obtained results. It reports the average scores of smoothness synchronization and realism across all the participants, as well as the percentage of reality confidence. In general, the participants find that the AI-generated videos are slightly more synchronized than the real ones. In addition, it can be noted that real videos do not seem more realistic than AI-generated ones. This is also confirmed by the nuanced results obtained for reality confidence that differs for Video 1 and Video 2. Nevertheless, the results also suggest that the real videos show smoother dances as compared to AI-generated ones.

4.5 Limitations and Future Directions

As mentioned in Section 4.3 and pointed out by the human perception study in Section 4.4, the proposed framework have some limitations. First, the proposed fitting of static 3D scans using SMPL was not able to fit 3D scans with large clothing or long hair, as illustrated in the second row of Figure 3. This is due to the fact that SMPL body model is supposed to only model body shapes (*i.e.,* without clothing or hair). A separate modeling of the clothing and the hair could improve the quality of the generated dances. Second, the animation using SMPL has some limitations even with dances performed by experts. As shown in Figure 4, the fitting of SMPL was unable to realistically fit some poses. A solution for this problem could consist of predicting blend weights for a specific 3D scan given a pose, as done in (Tiwari et al., 2021). Finally, we argue that the FACT model (Li et al., 2021) that was used to generate music-conditioned dance motions has its own limitations, such as unrealistic frozen motions. This module can be improved by including more recent models such as (Siyao et al., 2022) that have shown a better motion quality than (Li et al., 2021).

5 CONCLUSIONS

In this paper, *'You Can Dance'* , an automatic framework for generating music-conditioned dances on real 3D scans, is proposed. The system is composed of two main modules. The first one generates dance motions that are coherent with a given music. The second translates the generated motion dances to real 3D scans using SMPL body model fitting. The fitting module has also been used to generate dancing 3D scans by translating the dance motions of AIST++ dataset to the real 3D scans of the 3DBodyTex.v2 dataset. A human-based evaluation study was conducted to assess the quality of the animated 3D scans when using dance motions performed by experts and generated by the music-conditioned dance module. The proposed framework achieved plausible qualitative results, but had some limitations that are mainly due to inaccurate 3D scan fitting and unrealistic generated dance motions. Nevertheless, the authors believe that this first attempt towards music-conditioned dance animation of 3D scans might open the doors for the community to investigate it further.

ACKNOWLEDGEMENTS

This work was funded by the National Research Fund (FNR), Luxembourg in the context of the PSP-F2018/12856451/Smart_Schoul_2025 project and by the Esch22 project entitled Sound of data. The authors are grateful to the contributors of the open source libraries used in this work.

REFERENCES

Aksan, E., Kaufmann, M., Cao, P., and Hilliges, O. (2021). A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE.

Dodik, A., Sellán, S., Kim, T., and Phillips, A. (2022). Sex and gender in the computer graphics literature. *ACM SIGGRAPH Talks*.

Hernandez-Olivan, C. and Beltran, J. R. (2021). Music composition with deep learning: A review. *arXiv preprint arXiv:2108.12290*.

Huang, R., Hu, H., Wu, W., Sawada, K., Zhang, M., and Jiang, D. (2020). Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*.

Huang, Y., Zhang, J., Liu, S., Bao, Q., Zeng, D., Chen, Z., and Liu, W. (2022). Genre-conditioned long-term 3d dance generation driven by music. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4858–4862. IEEE.

Kärki, K. (2021). Vocaloid liveness? hatsune miku and the live production of japanese virtual idol concerts. In *Researching Live Music*, pages 127–140. Focal Press.

Lee, H.-Y., Yang, X., Liu, M.-Y., Wang, T.-C., Lu, Y.-D., Yang, M.-H., and Kautz, J. (2019). Dancing to music. *Advances in neural information processing systems*, 32.

Li, R., Yang, S., Ross, D. A., and Kanazawa, A. (2021). Learn to dance with aist++: Music conditioned 3d dance generation.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Pu, J. and Shan, Y. (2022). Music-driven dance regeneration with controllable key pose constraints. *arXiv preprint arXiv:2207.03682*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Saint, A., Ahmed, E., Cherenkova, K., Gusev, G., Aouada, D., Ottersten, B., et al. (2018). 3dbodytex: Textured 3d body dataset. In *2018 International Conference on 3D Vision (3DV)*, pages 495–504. IEEE.

Saint, A., Kacem, A., Cherenkova, K., and Aouada, D. (2020a). 3dbooster: 3d body shape and texture recovery. In *European Conference on Computer Vision*, pages 726–740. Springer.

Saint, A., Kacem, A., Cherenkova, K., Papadopoulos, K., Chibane, J., Pons-Moll, G., Gusev, G., Fofi, D., Aouada, D., and Ottersten, B. (2020b). Sharp 2020: The 1st shape recovery from partial textured 3d scans challenge results. In *European Conference on Computer Vision*, pages 741–755. Springer.

Saint, A., Rahman Shabayek, A. E., Cherenkova, K., Gusev, G., Aouada, D., and Ottersten, B. (2019). Bodyfitr: Robust automatic 3d human body fitting. In *2019 IEEE International Conference on Image Processing (ICIP)*.

Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C. C., and Liu, Z. (2022). Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11050–11059.

Sun, G., Wong, Y., Cheng, Z., Kankanhalli, M. S., Geng, W., and Li, X. (2020). Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509.

Tang, T., Jia, J., and Mao, H. (2018a). Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *2018 ACM Multimedia*

Conference on Multimedia Conference, pages 1598–1606. ACM.

Tang, T., Mao, H., and Jia, J. (2018b). Anidance: Real-time dance motion synthesize to the song. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1237–1239. ACM.

Tiwari, G., Sarafianos, N., Tung, T., and Pons-Moll, G. (2021). Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718.

Tsuchida, S., Fukayama, S., Hamasaki, M., and Goto, M. (2019). Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018). Video-to-video synthesis. *Advances in Neural Information Processing Systems*, 31.