Synthesizing Realistic Human Dance Motions Conditioned by Musical Data using Graph Convolutional Networks

João P. M. Ferreira*, Renato Martins*,†, Erickson R. Nascimento*
*Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Brazil
†VIBOT EMR CNRS 6000, ImViA, Université Bourgogne Franche-Comté, Le Creusot, France
E-mails: {joaoferreira, renato.martins, erickson}@dcc.ufmg.br

***Abstract*—Learning to move naturally from music, *i.e.*, to dance, is one of the most complex motions humans often perform effortlessly. Synthesizing human motion through learning techniques is becoming an increasingly popular approach to alleviating the requirement of new data capture to produce animations. Most approaches, addressing the problem of automatic dance motion synthesis with classical convolutional and recursive neural models, undergo training and variability issues due to the non-Euclidean geometry of the motion manifold structure. In this thesis[1], we design a novel method based on graph convolutional networks, that overcome the aforementioned issues, to tackle the problem of automatic dance generation from audio information. Our method uses an adversarial learning scheme conditioned on the input music audios to create natural motions preserving the key movements of different music styles. We also collected, annotated and made publicly available a novel multimodal dataset with paired audio, motion data and videos of people dancing three different music styles, as a common ground to evaluate dance generation approaches. The results suggest that the proposed GCN model outperforms the state-of-the-art dance generation method conditioned on music in different experiments. Moreover, our graph-convolutional approach is simpler, easier to be trained, and capable of generating more realistic motion styles regarding qualitative and different quantitative metrics. It also presents a visual movement perceptual quality comparable to real motion data. The dataset, source code, and qualitative results are available on the project's webpage: https://verlab.github.io/Learning2Dance_CAG_2020/.**

I. Introduction

Synthesizing motions through learning techniques is becoming an increasingly popular approach to alleviating the requirement of capturing new real motion data to produce animations. The motion synthesis has been applied to a myriad of applications such as graphic animation for entertainment, robotics, and multimodal graphic rendering engines with human crowds [1], to name a few. A crucial step to achieve plausible animation is to learn the motion distribution and then draw samples (*i.e.*, new motions) from it. For instance, a challenging human movement is dancing, where the animator does not aim to create avatars that mimic real poses but to produce a set of poses that match the music's choreography, while preserving the quality of being individual.

In this thesis, we address the problem of synthesizing dance movements from music using adversarial training and a convolutional graph network architecture (GCN). In dance moves, both the particularities of the dancer and the characteristics of the movement play an essential role in recognizing the dance style. Thus, a central challenge in our work is to synthesize a set of poses taking into account three main aspects: firstly, the motion must be plausible, *i.e.*, a blind evaluation should present similar results when compared to real motions; secondly, the synthesized motion must retain all the characteristics present in a typical performance of the music's choreography; third, each new set of poses should not be strictly equal to another set, in other words, when generating a movement for a new avatar, we must retain the quality of being individual.

Creating motions from sound relates to the paradigm of embodied music cognition. It couples perception and action, physical environmental conditions, and subjective user experiences (cultural heritage) [2]. Therefore, synthesizing realistic human motions regarding embodying motion aspects remains a challenging and active research field [3], [4]. In particular, advances in the deep learning techniques yielded an unprecedented combination of effective and abundant techniques able to predict and generate data, from high accuracy scores in image classification using convolutional neural networks (CNN) to photo-realistic image generation with the generative adversarial networks (GAN) [5]. In the same direction, synthesizing dance motions from audio information with recursive models and transformers is receiving a broad attention [6]–[8].

Most recently, networks operating on graphs have emerged as promising and effective approaches to deal with problems which structure is known *a priori*. A representative approach is the work of Kipf and Welling [9], where a convolutional architecture that operates directly on graph-structured data is used in a semi-supervised classification task. We argue that movements of a human skeleton, which has a graph-structured model, follow complex sequences of poses that are temporally related, and the set of defined and organized movements can be better modeled using a convolutional graph network trained using adversarial regime.

In this context, we propose an approach to synthesize motions with three main components. Our method starts encoding a sound signal to extract the music style using a CNN architecture. The music style and a spatial-temporal latent vector are used to condition a GCN architecture that is trained in an adversarial regime to predict 2D human body joint positions over time. Experiments with a user study and quantitative metrics shows that our approach outperforms the state-of-the-art method and provides plausible movements while maintaining the characteristics of different dance styles.

[1]This work relates to an M.Sc. dissertation.

Contributions. The contribution of this thesis can be summarized as follows:

i) A new conditional GCN architecture to synthesize human dance motions based on auditory data. In our method, we push further the adversarial learning by providing multimodal data with temporal dependence;
ii) A novel multimodal dataset with paired audio, motion data and videos of people dancing different music styles.

II. Related Work

Sound & Motion. Recently, we have witnessed an overwhelming growth of new approaches to deal with the tasks of transferring motion style and building animations of people from sounds. For example, Cudeiro *et al*. [10] presented an encoder-decoder network that uses audio features extracted from DeepSpeech [11]. The network generates realistic 3D facial animations conditioned on subject labels to learn different individual speaking styles. Ginosar *et al*. [3] enable translation from speech to gesture, generating arms and hand movements by mapping audio to pose. They used an adversarial training, where a U-Net architecture transforms the encoded audio input into a temporal sequence of 2D poses. However, their method is subject-specific and does not generalize to other speakers.

A close related work to ours is the approach proposed by Lee *et al*. [12]. The authors use a complex architecture to synthesize dance movements (expressed as a sequence of 2D poses) given a piece of input music. Their architecture is based on an elaborated decomposition-to-composition framework trained with an adversarial learning scheme. Our graph-convolutional based approach, on its turn, is simpler, easier to be trained, and generates more realistic motion styles regarding qualitative and different quantitative metrics.

Generative Graph Convolutional Networks. Since the seminal work of Goodfellow *et al*. [5], generative adversarial networks (GAN) have been successfully applied to a myriad of hard problems. Mirza and Osindero [13] proposed Conditional GANs (cGAN), which provides some guidance into the data generation. Graph Convolutional Networks (GCN) recently emerged as a powerful tool for learning from data by leveraging geometric properties that are embedded beyond n-dimensional Euclidean vector spaces, such as graphs and simplicial complex. In our context, conversely to classical CNNs, GCNs can model the motion manifold space structure [4], [14]. Yan *et al*. [14] applied GCNs to model human movements and classify actions. After extracting 2D human poses for each frame from the input video, the skeletons are processed by a Spatial-Temporal Graph Convolutional Network (ST-GCN). Yan *et al*. proceeded in exploiting the representation power of GCNs and presented the Convolutional Sequence Generation Network (CSGN) [4]. By sampling correlated latent vectors from a Gaussian process and using temporal convolutions, the CSGN architecture was capable of generating temporal coherent long human body action sequences as skeleton graphs. Our method takes one step further than [4], [14]. It generates human skeletal-based graph motion sequences conditioned on acoustic data, *i.e*., music. By conditioning the movement distributions, our method learns not only to create plausible human motions, but it also learns the music style signature movements from different domains.

Estimating and Forecasting Human Pose. Motion synthesis and motion analysis problems have been benefited from the improvements in the accuracy of human pose estimation methods. Human pose estimation from images, for its turn, greatly benefited from the recent emergence of large datasets [15]–[17] with annotated positions of joints, and dense correspondences from 2D images to 3D human shapes [17]–[21]. This large amount of annotated data has made possible important milestones towards predicting and modeling human motions [22]–[25]. The recent trend in time-series prediction with recurrent neural networks (RNN) became popular in several frameworks for human motion prediction [23], [24]. Nevertheless, the pose error accumulation in the predictions allows mostly predicting over a limited range of future frames [22]. Gui *et al*. [22] proposed to overcome this issue by applying adversarial training using two global recurrent discriminators that simultaneously validate the sequence-level plausibility of the prediction and its coherence with the input sequence. Wang *et al*. [25] proposed a network architecture to model the spatial and temporal variability of motions through a spatial component for feature extraction. Yet, these RNN models are known to be difficult to train and computationally cumbersome [26]. As also noted by [12], motions generated by RNNs tend to collapse.

Transferring Style and Human Motion. Synthesizing motion with specific movement style has been studied in a large body of prior works [27]–[29]. Most methods formulate the problem as transferring a specific motion style to an input motion [30], [31], or transferring the motion from one character to another, commonly referred as motion retargeting [32]–[34]. Another active research direction is transferring motion from video-to-video [27]–[29]. However, the generation of stylistic motion from audio is less explored, and it is still a challenging research field. Wang *et al*. [35] discussed how adversarial learning could be used to generate human motion by using a sequence of autoencoders. The authors focused on three tasks: motion synthesis, conditional motion synthesis, and motion style transfer. As our work, their framework enables conditional movement generation according to a style label parameterization, but there is no multimodality associated with it. Jang *et al*. [36] presented a method inspired by sequence-to-sequence models to generate a motion manifold. As a significant drawback, the performance of their method decreases when creating movements longer than 10s, which makes the method inappropriate to generate long sequences. Our approach, on the other hand, can create long movement sequences conditioned on different music styles, by taking advantage of the adversarial GCN's power to generate new long, yet recognizable, motion sequences.

III. Methodology

Our method has been designed to synthesize a sequence of 2D human poses resembling a human dancing according to a

music style. Specifically, we aim to estimate a motion $\mathcal{M}$ that provides the best fit for a given input music audio. $\mathcal{M}$ is a sequence of N human body poses defined as:

$$\mathcal{M} = [\mathbf{P}_0, \mathbf{P}_1, \cdots, \mathbf{P}_N] \in \mathbb{R}^{N \times 25 \times 2}, \quad (1)$$

where $\mathbf{P}_t = [\mathbf{J}_0, \mathbf{J}_1, \cdots, \mathbf{J}_{24}]$ is a graph representing the body pose in the frame t and $\mathbf{J}_i \in \mathbb{R}^2$ the 2D image coordinates of i-th node of this graph. The graph topology (the edges and nodes' connections) follows the standard defined by OpenPose [18].

Our approach consists of three main components, outlined in Figure 1. We start training a 1D-CNN classifier to define the input music style. Then, the result of the classification is combined with a spatial-temporal correlated latent vector generated by a Gaussian process (GP). At last, we perform the human motion generation from the latent vector. In the training phase of the generator, we use the latent vector to feed a graph convolutional network that is trained in an adversarial regime on the dance style defined by an oracle algorithm. In the test phase, we replace the oracle by the 1D-CNN classifier. Thus, our approach has two training stages: *i)* The training of the audio classifier to be used in the test phase and *ii)* The GCN training with an adversarial regime that uses the music style to condition the motion generation.

Sound Processing and Style Feature Extraction. Our motion generation is conditioned by a latent vector that encodes information from the music style. In this context, we used the SoundNet [37] architecture as the backbone to a one-dimensional CNN, because of its displayed capabilities to learn representations from audio to visual tasks. The 1D-CNN receives a sound in waveform and outputs the most likely music style considering three classes. The classifier is trained in a dataset composed of 107 music files and divided into three music-dance styles: *Ballet*, *Salsa*, and *Michael Jackson (MJ)*. To find the best hyperparameters, we run a 10-fold cross-validation and kept the best model.

Latent Space Encoding for Motion Generation. In order to create movements that follow the music style, while keeping particularities of the motion and being temporally coherent, we build a latent vector that combines the extracted music style with a spatiotemporal correlated signal from a Gaussian process. The information used to condition the motion generation, and to create our latent space, is a trainable dense feature vector representation of each music style.

Then, we combine a temporal coherent random noise with the music style representation in order to generate coherent motions over time. Therefore, the final latent vector is the result of concatenating the dense trainable representation of the audio class with the coherent temporal signal in the dimension of the features. This concatenation plays a key role in the capability of our method to generate synthetic motions with more than one dancing style when the audio is a mix of different music styles. In other words, unlike a vanilla conditional generative model, which conditioning is limited to one class, we can condition over several classes over time. The coherent temporal signals are sampled from Radial Basis Function kernel (RBF) [38] to enforce temporal relationship among the N frames.

The Gaussian process generates our random noise z and the dense representation of the dance style is the variable used to condition our model y. The combination of both data is used as input for the generator.

Conditional Adversarial GCN for Motion Synthesis. To generate realistic movements, we use a graph convolutional neural network (GCN) trained with an adversarial strategy. Figure 1 illustrates the training scheme.

a) ***Generator****:* The architecture of our generator G is mainly composed of three types of layers: temporal and spatial upsampling operations, and graph convolutions. When using GCNs, one challenge that appears in an adversarial training is the requirement of upsampling the latent vector in the spatial and temporal dimensions to fit the motion space $\mathcal{M}$ (Equation 1).

The temporal upsampling layer consists of transposed 2D convolutions that double the time dimension, ignoring the input shape of each layer. Inspired by Yan *et al*. [4], we also included in our architecture a spatial upsampling layer. This layer operates using an aggregation function defined by an adjacency matrix A^ω that maps a graph $S(V, E)$ with V vertices and E edges to a bigger graph $S'(V', E')$. The network can learn the best values of A^ω that leads to a good upsampling of the graph by assigning different importance of each neighbor to the new set of vertices.

In the first layer of the generator, we have one node containing a total of N features; these features represent our latent space (half from the Gaussian Process and a half from the audio representation). The features of the subsequent layers are computed by the operations of upsampling and aggregation. The last layer outputs a sequence of graphs, each one with 25 nodes containing the (x, y) coordinates of each skeleton joint.

b) ***Discriminator****:* The discriminator D has the same architecture used by the generator but using downsampling layers instead of upsampling layers. Thus, all transposed 2D convolutions are converted to standard 2D convolutions, and the spatial downsampling layers follow the same procedure of upsampling operations but using an aggregation matrix B^ϕ with trainable weights ϕ, different from the weights learned by the generator.

In the discriminator network, the feature vectors are assigned to each node as follows: the first layer contains a sequence of graphs, each one with 25 nodes, where their feature vectors are composed of the (x, y) coordinates on a normalized space and the class of the input motion. In the subsequent layers, the features of each node are computed by the operations of downsampling and aggregation. The last layer contains only one node that outputs the classification of the input data being fake or real. Figure 1-(b) illustrates the discriminator architecture.

c) ***Adversarial training****:* Given the motion generator and discriminator, our conditional adversarial network aims at

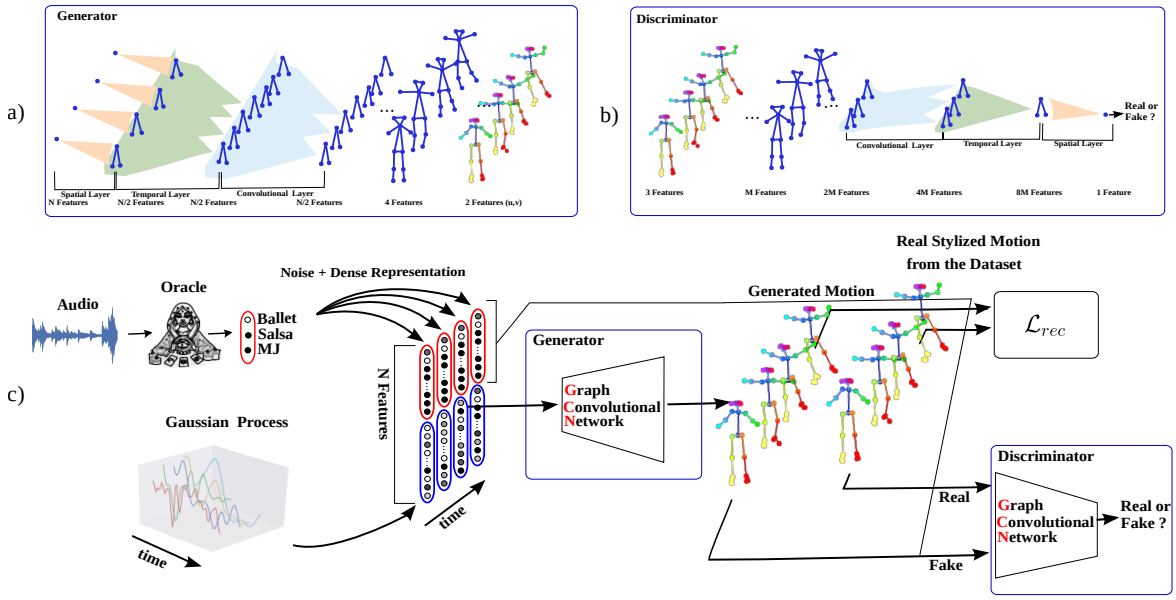

Fig. 1. **Proposed methodology**: (a) GCN Motion Generator G; (b) GCN Motion Discriminator D; and (c) an overview of the adversarial training regime.

minimizing the binary cross-entropy loss:

$$\mathcal{L}_{cGAN}(G,D) = \min_G \max_D \left(\mathbb{E}_{x\sim p_{data}}(x)[\log D(x|y)] + \mathbb{E}_{z\sim p_z}(z)[\log(1 - D(G(z|y)))]\right), \quad (2)$$

where the generator aims to maximize the error of the discriminator, while the discriminator aims to minimize the classification fake-real error shown in Equation 2. In particular, in our problem, p_{data} represents the distribution of real motion samples, $x = \mathcal{M}_\tau$ is a real sample from p_{data}, and $\tau \in [0 - \mathcal{D}_{size}]$ and $\mathcal{D}_{size}$ is the number of real samples in the dataset. Figure 1-(c) shows a concise overview of the steps in our adversarial training.

The latent vector, which is used by the generator to synthesize the fake samples x', is represented by the variable z, the coherent temporal signal. The dense representation of the dance style is determined by y, and p_z, which is a distribution of all possible temporal coherent latent vectors generated by the Gaussian process.

To improve the generated motion results, we use a motion reconstruction loss term ($\mathcal{L}_{rec}$) applying L_1 distance in all skeletons over the N motion frames. Thus, our final loss is a weighted sum of the motion reconstruction and cGAN discriminator losses given by

$$\mathcal{L} = \mathcal{L}_{cGAN} + \lambda \mathcal{L}_{rec}, \quad (3)$$

where λ weights the reconstruction term. The λ value was chosen empirically, and was fixed throughout the training stage. The initial guess regarding the magnitude of λ followed the values chosen by Wang *et al*. [27]. We then apply a cubic-spline interpolation in the final motion to remove eventual high frequency artifacts from the generated motion frames $\mathcal{M}$.

IV. Audio-Visual Dance Dataset

We build a new dataset composed of paired videos of people dancing different music styles[2]. Our dataset differs from existing ones mainly in the quality of the annotated movements. For instance, the work of [12] also present a large dataset for automatic dance generation, however they collected the data without careful selection of representative dances to the desired styles. On the other hand, we carefully selected characteristic movements to compose our dataset. The dataset is used to train and evaluate the methodologies for motion generation from audio. We split the samples into training and evaluation sets that contain multimodal data for three music/dance styles: Ballet, Michael Jackson, and Salsa. These two sets are composed of two data types: visual data from careful-selected parts of publicly available videos of dancers performing representative movements of the music style and audio data from the styles we are training. Figure 2 shows some data samples of our dataset.

In order to collect meaningful audio information, several playlists from YouTube were chosen with the name of the style/singer as a search query. For the visual data, we started by collecting videos that matched the music style and had representative moves. Each video was manually cropped in parts of interest, by selecting representative moves for each dance style present in our dataset. We annotate the 25 2D human joint poses for each video by estimating the pose with OpenPose [18]. Each motion sample is defined as a set of 2D human poses of 64 consecutive frames. To improve the quality of the estimated poses in the dataset, we handled the miss-detection of joints by exploiting the body dynamics in the video. We also performed motion data augmentation to increase the variability

[2]The dataset and source code are publicly available at https://verlab.github.io/Learning2Dance_CAG_2020/.

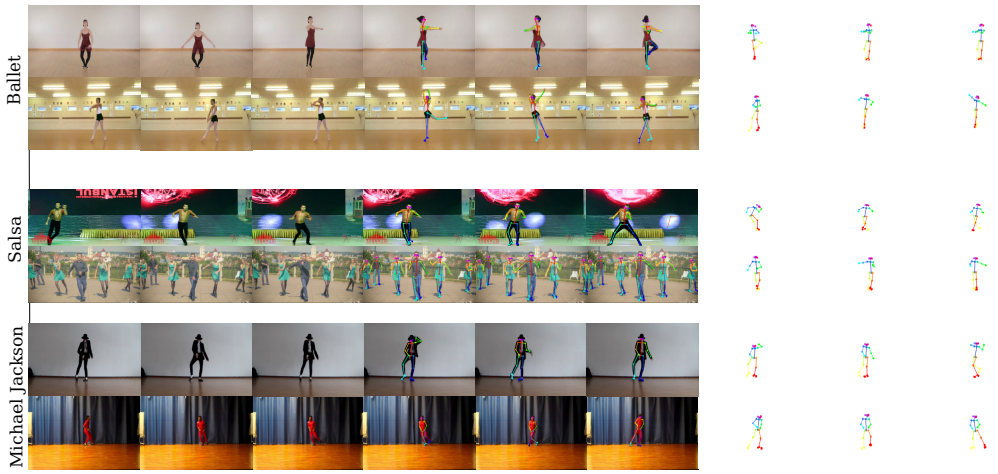

Fig. 2. **Audio-motion dataset.** Video samples of the multimodal dataset with carefully annotated audio and 2D human motions of different dance styles.

and number of motion samples. Experimental evaluations with the same architecture and hyperparameters, with and without data augmentation, showed the performance evaluated with the Fréchet Inception Distance (FID) metric improves when using data augmentation. Moreover, we observed improvements in the motions' variability, and body movements were easier to notice, when using data augmentation to train our method.

V. Experiments & Results

To assess the performance our method, we conduct several experiments evaluating different aspects of motion synthesis from audio information. We also compared our approach to the state-of-the-art technique proposed by Lee *et al.* [12], hereinafter referred to as D2M. We choose to compare our method to D2M since other methods have major drawbacks that make a comparison with our approach unsuitable, such as different skeleton structures [3]. The experiments are as follows: *i)* We performed a perceptual user study using a blind evaluation with users trying to identify the dance style of the dance moves. For a generated dance video, we ask the user to choose what style (Ballet, Michael Jackson (MJ), or Salsa) the avatar on the video is dancing; *ii)* Aside from the user study, we also evaluated our approach on commonly adopted quantitative metrics in the evaluation of generative models, such as Fréchet Inception Distance (FID), GAN-train, and GAN-test [39].

A. Implementation and Training Details

Audio and Poses Preprocessing. Our one-dimensional audio CNN was trained for 500 epochs, with batch size equal to 8, Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and learning rate of 0.01. Similar to [40], we preprocessed the input music audio using a $\mu - law$ non-linear transformation to reduce noise from audio inputs that were not appropriately recorded. We performed 10-fold cross-validation to choose the best hyperparameters.

In order to handle different shapes of the actors and to reduce the effect of translations in the 2D poses of the joints, we normalized the motion data used during the adversarial GCN training. We managed changes beyond body shape and translations, such as the situations of actors lying on the floor or bending forward, by selecting the diagonal distance of the bounding box encapsulating all 2D body joints $\mathbf{P}_t$ of the frame as scaling factor.

Training. We trained the GCN adversarial model for 500 epochs. We select 64 frames as the size of our samples to follow a similar setup presented in [3]. However, it is worth noting that our method can synthesize long motion sequences. We use a batch size of 8 motion sets of N frames each. We optimized the cGAN with Adam optimizer for the generator with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ with learning rate of 0.002. The discriminator was optimized with Stochastic Gradient Descent (SGD) with a learning rate of 2×10^{-4}. We used $\lambda = 100$ in Equation 3. Dropout layers were used on both generator and discriminator to prevent overfitting. The training process takes around 8 hours with a GTX 1080 GPU.

Avatar Animations. As an application of our formulation, we animate three virtual avatars using the generated motions to different music styles. The image-to-image translation technique vid2vid [27] was selected to synthesize videos. We trained vid2vid to generate new images for these avatars, following the multi-resolution protocol described in [27]. For inference, we feed vid2vid with the output of our GCN. We highlight that any motion style transfer method can be used with few adaptations, as for instance, the works of [28], [29].

B. User Study

We conducted a perceptual study with 60 users. The perceptual study was composed of 45 randomly sorted tests. For each test, the user watches a video (with no sound) synthesized by vid2vid using a generated set of poses. Then we asked them to associate the motion performed on the synthesized video as belonging to one of the audio classes: Ballet, Michael Jackson, or Salsa. In each question, the users were supposed to listen to one audio of each class to help them to classify the video. The set of questions was composed of 15 videos

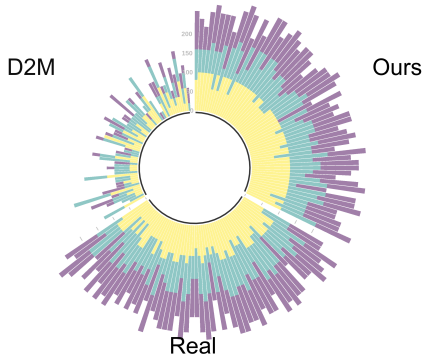

Fig. 3. **User study results.** Each stacked bar represents one user evaluation and the colors of each stacked bar indicates the dance styles (Ballet = yellow, Michael Jackson (MJ) = blue, and Salsa = purple). We show the results for all **60** users that fully answered our study.

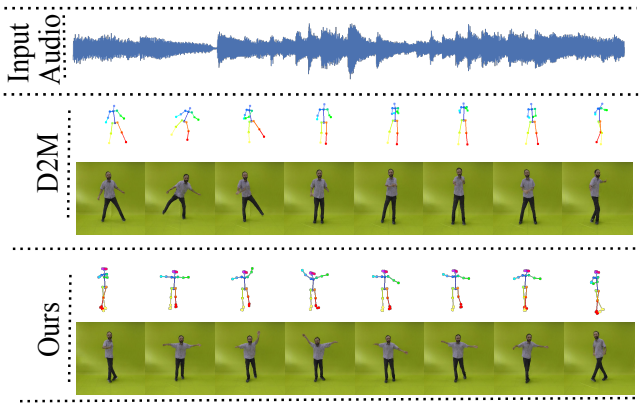

Fig. 4. **Qualitative results.** Results of our approach in comparison to D2M [12] for *Ballet*

of movements generated by our approach, 15 videos generated by D2M [12], and 15 videos of real movements extracted from our training dataset. We applied the same transformations to all data and every video had an avatar performing the motion with a skeleton with approximately the same dimensions. We split equally the 15 videos shown between the three dance styles. Figure 3 presents the user study results, and we can draw the following observations: first, our method achieved similar motion perceptual performance to the one obtained from real data. Second, our method outperformed D2M with a large margin. Thus, we argue that the proposed model is capable of generating realistic movements taking into account two of the following aspects: *i)* Our performance is similar to the results from real motion data in a blind study; *ii)* Users show higher accuracy in categorizing our generated motion.

C. Quantitative and Qualitative Evaluation

For a more detailed performance assessment regarding the similarity between the learned distributions and the real ones, we adopted the commonly used Fréchet Inception Distance (FID). We computed the FID values using motion features extracted from the action recognition ST-GCN model presented in [14], similar to the metric used in [4], [12]. We also computed the GAN-Train and GAN-Test metrics, two well-known GAN evaluation metrics [39]. The detailed quantitative results can be seen in the Sections 5.2 and 5.3 in the Masters' thesis or in [41]. Figure 4 shows some qualitative results. We can notice that the sequences generated by D2M presented some characteristics clearly inherent to the dance style, but they are not present along the whole sequence. Conversely, our method generates poses commonly associated with ballet movements such as rotating the torso with stretched arms.

VI. Conclusions

We proposed a new method for synthesizing human motion from music. Unlike previous methods, we explore graph convolutional networks trained in an adversarial regime to address the problem. We achieved better qualitative and quantitative performance as compared to a state-of-the-art dance generation technique. Our method outperformed Dancing to Music method proposed by [12], in terms of FID, GAN-Train, and GAN-Test metrics. We also conducted a user study, which showed that our method received similar scores to real dance movements, which was not observed in the competitor. Moreover, we presented a new dataset with audio and visual data, carefully collected to train and evaluate algorithms designed to synthesize human motion in dance scenarios. We observe several potential applications such as in the animation of human avatars and human crowds used in the gaming industry. Finally, the work presented in this thesis indicates that when working with learning techniques for motion synthesis, the model awareness of the geometric motion data structure results in simpler models, while leading to more realistic motions.

Limitations & Future Work: We plan to investigate the following extensions of our formulation: *i)* Since we only condition the motion to the style, the variability of the movements in one dance style is related only with the random noise from the Gaussian process. Simultaneously using the audio information to create this variability should be a more suitable approach; *ii)* We intend to extend our dataset in terms of dance styles. This extension will allow us to stress our approach and create samples to be used in realistic animations following the auditory data in broader contexts.

Acknowledgments: We would like to thank the PPGCC-UFMG, CAPES, FAPEMIg, and CNPq for funding different parts of this work.

VII. Publications

The results of this thesis were published in the international journal *Computers & Graphics* [41]. This journal has been recently ranked 4 in the top Computer Graphics publications[3]. We also would like to highlight that the student also contributed as co-author to two related publications on human motion synthesis: one in the international conference WACV'20 [29] and one in the *International Journal of Computer Vision* (IJCV'21) [42].

[3]https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computergraphics

REFERENCES

[1] K. Ikeuchi, Z. Ma, Z. Yan, S. Kudoh, and M. Nakamura, "Describing upper-body motions based on labanotation for learning-from-observation robots," *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1415–1429, 2018.

[2] M. Leman, "The role of embodiment in the perception of music," *Empirical Musicology Review*, vol. 9, no. 3-4, pp. 236–246, 2014.

[3] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *CVPR*, 2019.

[4] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin, "Convolutional sequence generation for skeleton-based action synthesis," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[5] 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*, 2014, pp. 2672–2680.

[6] X. Ren, H. Li, Z. Huang, and Q. Chen, "Music-oriented dance video synthesis with pose perceptual loss," *arXiv preprint arXiv:1912.06606*, 2019.

[7] J. Li, Y. Yin, H. Chu, Y. Zhou, T. Wang, S. Fidler, and H. Li, "Learning to generate diverse dance motions with transformer," *arXiv preprint arXiv:2008.08171*, 2020.

[8] R. Huang, H. Hu, W. Wu, K. Sawada, and M. Zhang, "Dance revolution: Long sequence dance generation with music via curriculum learning," in *ICLR 2021*, 2021.

[9] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[10] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 101–10 111.

[11] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014.

[12] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz, "Dancing to music," in *Advances in Neural Information Processing Systems*, 2019.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[14] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, 2018.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[16] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.

[17] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[18] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[19] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2252–2261.

[20] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5607–5616.

[21] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. Moura, "Adversarial geometry-aware human motion prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 786–803.

[23] P. Ghosh, J. Song, E. Aksan, and O. Hilliges, "Learning human motion models for long-term predictions," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 458–466.

[24] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *ICCV*, 2015, pp. 4346–4354.

[25] H. Wang, E. S. L. Ho, H. P. H. Shum, and Z. Zhu, "Spatio-temporal manifold learning for human motions via long-horizon modeling," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[26] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, p. III–1310–III–1318.

[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Conference on Neural Information Processing Systems*, 2018.

[28] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5932–5941.

[29] T. L. Gomes, R. Martins, J. Ferreira, and E. R. Nascimento, "Do as I do: transferring human motion and appearance between monocular videos with spatial and temporal constraints," in *IEEE Conference on Applications of Computer Vision (WACV)*, 2020.

[30] S. Xia, C. Wang, J. Chai, and J. Hodgins, "Realtime style transfer for unlabeled heterogeneous human motion," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, pp. 1–10, 2015.

[31] H. J. Smith, C. Cao, M. Neff, and Y. Wang, "Efficient neural networks for real-time motion style transfer," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 2, no. 2, pp. 1–17, 2019.

[32] M. Gleicher, "Retargetting motion to new characters," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 33–42.

[33] K.-J. Choi and H.-S. Ko, "On-line motion retargeting," *Journal of Visualization and Computer Animation*, vol. 11, pp. 223–235, 12 2000.

[34] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] Q. Wang, T. Artières, M. Chen, and L. Denoyer, "Adversarial learning for modeling human motion," *The Visual Computer*, vol. 36, no. 1, pp. 141–160, 2020.

[36] D.-K. Jang and S.-H. Lee, "Constructing human motion manifold with sequential networks," *Computer Graphics Forum*, 2020.

[37] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016.

[38] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.

[39] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my GAN?" in *ECCV*, 2018, pp. 213–229.

[40] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[41] J. P. Ferreira, T. M. Coutinho, T. L. Gomes, J. F. Neto, R. Azevedo, R. Martins, and E. R. Nascimento, "Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio," *Computers & Graphics*, 2020.

[42] T. L. Gomes, R. Martins, J. Ferreira, and E. R. Nascimento, "A shape-aware retargeting approach to transfer human motion and appearance in monocular videos," 2021, international Journal of Computer Vision - IJCV.