Supplementary Material

Anonymous CVPR submission

Paper ID 5880

SUMMARY

This appendix contains additional materials for the paper "*Music-Driven Group Choreography*". The appendix is organized as follows:

- Section 1 describes group dance evaluation metrics.
- Section 2 discusses the annotation process and manual correction when we build the dataset.
- Section 3 provides optimization details.
- Section 4 shows more examples from our dataset.
- Section 5 discusses the limitation and impact of our work.

1. Group Choreography Evaluation Metrics

Group Motion Realism. To calculate the realism between generated and ground-truth group motion, we need to find a single unified representation for all dancers' motion in the scene Based on the kinetic features of a single motion sequence [6], we propose to calculate Group Motion Realism (GMR), smaller is better. For each entity, we compute the velocity of each element j of the pose vector: $v_t^n = \frac{y_{t+1}^n - y_t^n}{\Delta t}$ where Δt is the time period between two consecutive frames. Note that the pose vector of each entity at each frame consist of the root orientation, root position and the joint angles. The group kinetic features of a sequence is approximated by taking logarithm of the total kinetic energy of all group entities as:

$$e_j = \log\left(1 + \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{n=1}^{N} m_j (v_{t,j}^n)^2\right), \qquad (1)$$

where m_j is the moment of inertia or mass of each joint. As in [6], we assume that m_j are constant with respect to time and entity. Then, we split the sequence into smaller chunks and calculate the features of these chunks. This process is identical for both the generated and ground-truth sequences. Finally, we utilize these sets of features (from generated and ground-truth group dance) to calculate the GMR using the original FID formulation introduced by [1].

Group Motion Correlation (GMC). We also evaluate the synchrony and the correlation between dancers within the generated group. The correlation of movements between individuals may reflect their interaction (in the choreography) [9]. For every pair of motions within a group, we first align the two motion sequences using Dynamic Time Warping algorithm [5] based on the Euclidean distance in the joint position space (obtained by SMPL joint regressor). We then calculate the mean cross-correlations between the time-aligned motion pairs using the kinetic features [6]. The generated group motion correlation degree is then calculated as the average of all motion pairs.

Trajectory Intersection Frequency (TIF). For the generated group sequences, the intersection rate is calculated over all F frames as:

$$\text{TIR} = \frac{\sum_F \sum_{i,j:i\neq j} \mathbb{1}[\text{intersect}(M(y^i), M(y^j))]}{F}, \qquad (2)$$

where M is the SMPL skinning function [4] which can output a 6890-vertices human mesh from the input pose parameters y. For TIF, smaller is better.

2. Manual Annotation and Correction Process

We rely on the optimization (Section 3 in our main paper) with humans in the loop at each step to annotate the 3D ground truth for GDANCE dataset. Even so, the final results may have two problematic cases: *(i)* Minor problem: The 3D output has minor wrong motions in some parts, but the overall motion of dancers is clear and reasonable, and *(ii)* Severe problem: The 3D outputs can not show the movement of dancers. For minor cases, we rely on standard 3D computer animation technique to fix the slightly wrong motion using industrial animation software. For the severe cases, we simply discard them as it is not feasible to recover and maintain the motion of the dancers. Figure 1 illustrates a minor case and Figure 2 shows a severe case.

3. Group Motion Fitting Details

For the Local Mesh Fitting (Equation 1 in the main paper), we follow a multi-stage strategy similar to [8]. We first estimate the root translation and orientation by minimizing only E_{J} over the torso and hip joints. Subsequently, we fix the translation and orientation and start optimizing the body pose parameters θ and shape parameters β with high regularization ($\lambda_\theta = 50$, $\lambda_\beta = 100$). We then optimize all parameters with $\lambda_\theta = 10$, $\lambda_\beta = 10$, $\lambda_{\mathrm{S}} = 200$, and $\lambda_{\mathrm{F}} = 100$. To speed up convergence and avoid bad solution, we also leverage a Human Mesh Recovery method [2] to initialize the pose parameters θ and shape parameters β.

For the Global Optimization (Equation 2 in the main paper), we also optimize the objective in two stages. We first optimize all parameters except for the ground term E_{gc} and ground parameters ($\mathbf{n}^*$, $\mathbf{f}$). Next, we fit the ground parameters as in Equation 6. Finally, we apply the full optimization for all dancers with the objective coefficients $\lambda_{\mathrm{pen}} = 1000$, $\lambda_{\mathrm{reg}} = 100$, $\lambda_{\mathrm{dep}} = 100$, $\lambda_{\mathrm{gc}} = 20$. The whole optimization process is implemented in PyTorch [7] using the L-BFGS [3] algorithm with a learning rate of 1.0 to optimize the whole video at each stage.

4. GDANCE Visualization

Figure 3 shows some example from our dataset with different number of dancers. Figure 4 illustrates some dance styles from our dataset.

Our GDANCE dataset has 16 music genres and 7 dance styles. Table 1 provides the detail description of music genres. Table 2 explains the dance styles.

5. Limitation and Broader Impacts

Limitation. Although our group dance dataset has provide group dance motions of all dancers. We currently do not consider the detailed finger movements and the facial expressions of the dancers. Accurately tracking the hands and facial motions in in-the-wild group dance videos is not a trivial task. Thus, we leave these problems for future work.

Broader Impacts. In this work, we have introduced a new largest dataset for group choreography generation. We also propose a baseline architecture that can successfully generate group dance motions from only the input audio and the initial positions. Our dataset is the currently largest music-driven group dance dataset with various dancing styles and music genres. The dataset can also be useful for other challenging applications such as dance recognition and interaction between dancers, dance information processing, dance style transfer, and driving group motion of virtual idols.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 2017. 1

[2] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 2

[3] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 1989. 2

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 2015. 1

[5] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, 2007. 1

[6] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics*, 2008. 1

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. 2019. 2

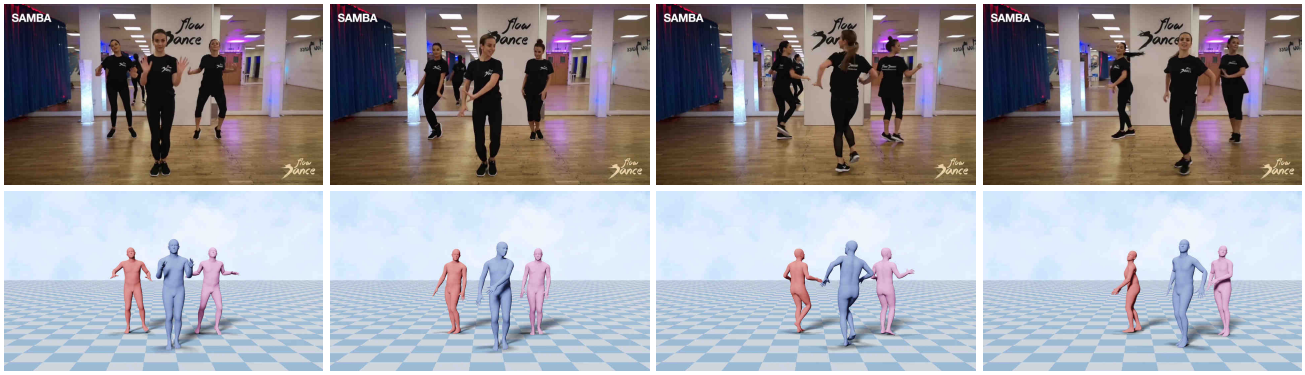
[8] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[9] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 2011. 1

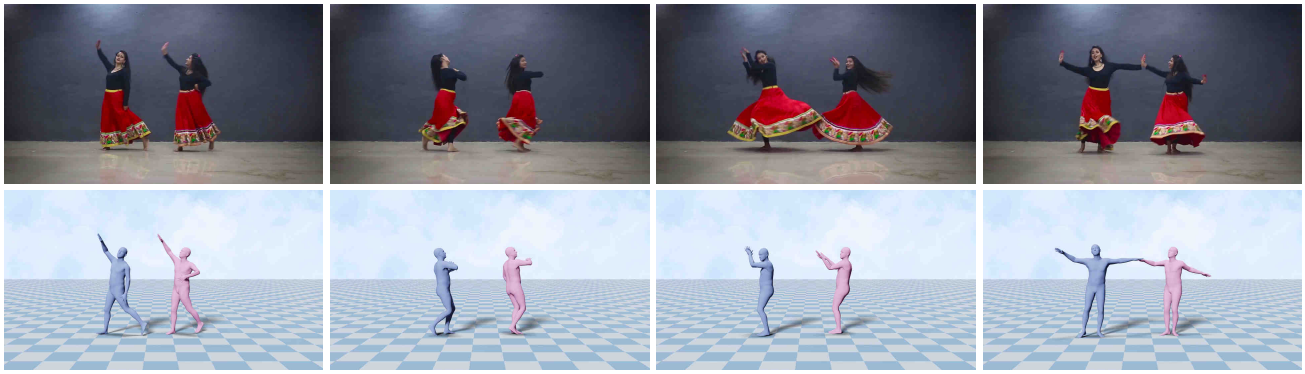
Figure 1. Manual annotation and correction process. From top to bottom: *Input Images*, *Detected Joints*, *Refined Joints*, *Fitted Mesh*, *Refined Mesh*. Red boxes denote the wrong joints where we manually correct them.

Figure 2. Failure cases of our semi-labeling method due to extreme occlusions and extreme motions.

Figure 3. Some examples from our GDANCE dataset.



(a) Samba



(b) Bollywood

(c) Rumba

Figure 4. Some dance styles from our dataset.

Music Genre	Description
Rock	has syncopated rhythms with a repetitive snare drum back beat on beats two and four.
Electronic	employs electronic musical instruments, digital instruments, or circuitry-based music technology.
Folk	has strong story telling elements.
Ballad	has slow tempo that deals with themes of love and loss.
Reggae	has a characteristic rhythm with amplified bass guitar riffs, and moderate tempo with an accent on the offbeat.
Bachata	has a heavy guitar emphasis and heartrending love stories as its basis.
Latin	includes other various styles of music from Latin America, Spain, Portugal, and the United States, except the two most popular, i.e., Bachata and Reggae.
Pop	marked by a consistent and noticeable rhythmic element, a mainstream style, and a simple traditional structure.
Rap	contains stylized rhyming speech that is chanted.
R&B	is also called rhythm and blues, includes soulful singing over a strong backbeat, commonalities in rhythm, bands divided into a rhythm and horn section, repetition of rhythms, verses and notes, and often complex blending of instruments.
Indian	is characterized by microtones (or shruti), notes (or swara), ornamentations (or alankar), among others.
Casual	or Classical, contains melodies with clear-cut phrases, and clearly marked cadences, emphasises on beauty, elegance and balance.
Blues	is known for being microtonal, using pitches between the semitones defined by a piano keyboard.
Funk	driven by hard syncopated bass lines and drumbeats and accented by any number of instruments involved in rhythmic counterplay.
Metal	characterized by loud distorted guitars, emphatic rhythms, dense bass-and-drum sound, and vigorous vocals.
Disco	is typified by four-on-the-floor beats, syncopated bass-lines, string sections, horns, electric piano, synthesizers, and electric rhythm guitars.

Table 1. Detailed description of music genres in GDANCE dataset.

Dance Style	Description
Zumba	features high- and low-intensity intervals, involves high-impact moves like bouncing and jumping
Aerobic	is a full-body movement, uses large muscle groups, is rhythmic in nature, and can be maintained continuously for several minutes
Commercial	include floor work and unpredictability as dancers are not moving to repetitive rhythms and will freely dance in contrasting directions and styles.
Bollywood	incorporates head, neck and body movements, mudras and footwork. There are about 108 mudras.
Irish	dancers must land from difficult moves without letting their knees bend or heels touch the ground, causing large forces to be absorbed by the body.
Rumba	contains subtle side-to-side hip movements with the torso erect, and is danced with a basic pattern of two quick side steps and a slow forward step.
Samba	characterized by simple forward and backward steps and tilting, rocking body movements.

Table 2. Detailed description of dance styles in GDANCE dataset.