This is the pre-printed version. The accepted version please check: https://ieeexplore.ieee.org/document/10096824

CONTRASTIVE SELF-SUPERVISED LEARNING FOR AUTOMATED MULTI-MODAL DANCE PERFORMANCE ASSESSMENT

Yun Zhong, Fan Zhang, Yiannis Demiris

Personal Robotics Laboratory, Department of Electrical and Electronic Engineering,
Imperial College London

ABSTRACT

A fundamental challenge of analyzing human motion is to effectively represent human movements both spatially and temporally. We propose a contrastive self-supervised strategy to tackle this challenge. Particularly, we focus on dancing, which involves a high level of physical and intellectual abilities. Firstly, we deploy Graph and Residual Neural Networks with Siamese architecture to represent the dance motion and music features respectively. Secondly, we apply the InfoNCE loss to contrastively embed the high-dimensional multimedia signals onto the latent space without label supervision. Finally, our proposed framework is evaluated on a multi-modal Dance-Music-Level dataset composed of various dance motions, music, genres and choreographies with dancers of different expertise levels. Experimental results demonstrate the robustness and improvements of our proposed method over 3 baselines and 6 ablation studies across tasks of dance genres, choreographies classification and dancer expertise level assessment.

***Index Terms*—** Human Motion analysis, Action Performance Assessment, Multi-modal Signal Processing

1. INTRODUCTION

It takes time and effort for people to become proficient in the art of dance. Analysing one's dance motion is essential for improving his/her expertise level. Personal trainers' feedback is usually costly due to time and location restrictions. In this work, we propose an automated multi-modal dance performance assessment model that is capable of continuously monitoring dancers' movement quality.

Automatic dance performance assessment lies in the intersection of human motion analysis, skill determination and action performance assessment (AQA). Many seminal works of skill determination and AQA [1, 2] have focused on Olympic sports or surgeries. Other human dance motion analysis studies [3, 4] have been successful in evaluating elementary dance movements but do not take the musical features into consideration. Generalizing these methods to real-world dance scenarios remains challenging.

In this paper, we present a method that addresses the abovementioned challenges by combining both motion and music signals, and developing a contrastive self-supervised network to identify individuals' expertise levels conditioned on diverse dance music, genres and choreographies. We adopt the InfoNCE contrastive loss in this paper as it is capable of learning feature representations in representation learning tasks due to its inherent information maximization objective [5]. The contributions of this paper are: (i) We propose a contrastive self-supervised strategy to assess human dance performance. Specifically, we deploy Spatial-Temporal Graph Convolutional Network (ST-GCN) [6] and Deep Residual Network (ResNet-18) [7] encoder to represent dance motion and music features. (ii) We evaluate our proposed strategy on the Dance-Music-Level dataset that contains 5 different dance genres, 20 dance choreographies and 3 expertise levels for each choreography. The corresponding music of each performed choreography is also included. (iii) We conduct quantitative and qualitative experiments to verify the effectiveness and generalizability of our model on multimedia dance-music scenarios.

To our knowledge, the paper is the first attempt to formulate a novel contrastive learning method, which is capable of not only determining human dance expertise levels, but also monitoring the dancers' level improvements over times. The method outperforms 3 popular classification algorithms [8, 5, 9] in terms of the dance expertise level classification accuracy.

2. LEARNING TO EVALUATE DANCE SKILLS

In this section, we first formulate the dance motion analysis problem. We present the self-supervised learning framework to tackle this problem. Augmentations for motion and music sequences is also explained. Moreover, we introduce our network that encodes motion and music features respectively. Fig. 1 shows our framework.

2.1. Problem Definition

Our goal is to determine the expertise level of the dance sequences performed by multiple individuals conditioned on different dance music, genres and choreographies. The expertise level determination involves evaluating the motion quality and the melody matching between motions and music. Hence, the model input is a set of sequences $\mathcal{S} = \{\mathcal{S}_{motion}, \mathcal{S}_{music}\}$ of dance motion and music.

The whole learning process is explained as follows. We first extract the features $\mathcal{F} = \{\mathcal{F}_{motion}, \mathcal{F}_{music}\}$ from recorded motion and music. With these collective features, we train a model $\mathcal{L} = \mathcal{M}(\{\mathcal{F}_{motion}, \mathcal{F}_{music}\})$ that assesses the dance performance, where $\mathcal{L}$ stands for the expertise levels of the trained sequences. The ground truth of $\mathcal{L}$ is annotated by two expert dancers based on the assessment criteria of body competency, movement fluidity and musical interpretation from the Royal Academy of Dance [10], which can be divided into three categories: beginner, intermediate and expert. We represent the underlying semantic features of the input se-

{y.zhong20,f.zhang16,y.demiris}@imperial.ac.uk.
This research is supported by a Royal Academy of Engineering Chair in Emerging Technologies to YD.

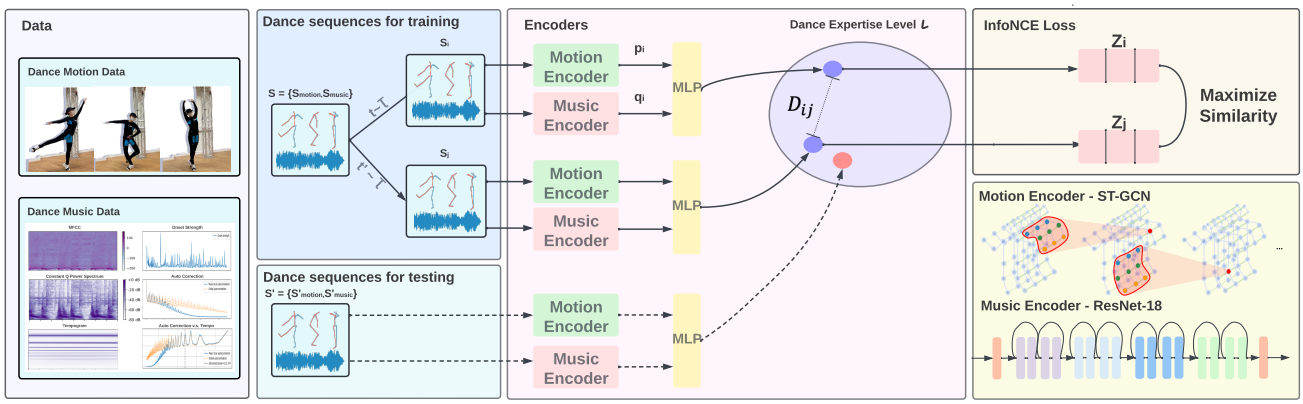

Fig. 1: Description of our contrastive self-supervised learning framework. The dance sequences are mapped onto the lower-dimensional latent space by using ST-GCN and ResNet encoders. The InfoNCE loss is applied to maximize the similar representation in the latent space.

quences in a 2D latent space, which is capable of efficiently separating different dance genres, choreographies and expertise levels. The Euclidean distance (D_{ij}) between two points in the latent space represents the data similarity. If the distance is smaller, it represents that two dancing sequences have more chance to belong to the same class of expertise level. In the pre-train phase, we use InfoNCE loss to optimize such distances. In the downstream phase, we map the dancing sequence onto the latent space, followed by using logistic regression to learn its expertise level ($\mathcal{L}$).

2.2. Spatial-Temporal Data Augmentation

Data augmentation has been proven to be critical in self-supervised image classification tasks [5]. However, the augmentation methods in most predictive tasks are only applied to the spatial domain of images. In this study, we enhance the augmentation methods, where the augmentations are also applied to the temporal aspect of the sequential data. Additionally, we extend RGB image augmentation techniques to 3D-skeleton data.

Temporal Augmentation: Timing is important to dance. Thus we adopt temporal augmentation on the sequential dancing data, aiming to improve temporal robustness. The two augmentation techniques we propose are: (i) adjusting playback speed and (ii) temporal crop. Temporal augmentation is applied in both motion and music data to achieve temporal alignment. Compared to other time-series augmentation methods, these 2 methods are exclusively designed for our dance performance assessment task to guarantee the expertise level of the dance sequences remaining same after augmentations.

Spatial Augmentation: We introduce (i) pose translation and (ii) joint jittering that centered on skeleton data. The pose translation allows the data to be robust to variant viewpoints or camera positions. We also propose joint jittering, where a certain joint position in motion sequences are randomly disrupted. This allows the proposed method to be noise-invariant to the input skeleton data which might be occasionally partially occluded or disrupted.

2.3. Motion and Music Encoder

ST-GCN Motion Encoder: We adopt ST-GCN [6] to extract the motion features, which takes advantage of the Graph Neural Network that is capable of automatically learning the spatial and temporal patterns from human motion. The input of the spatial-temporal graphs is the sequence of motion $\boldsymbol{m}_i^k$, where i is the index of the skeletal joints and k is the frame number. Each ST-GCN unit consists of a graph convolutional network (GCN) layer and a temporal convolution network (TCN) layer. For GCN, graph nodes are constructed according to the connectivity of the human body's skeletal structure, and each node consists of 3D coordinate vectors. As for TCN, each node of the graphs is connected to the same node in the consecutive frame. We thus have a graph of the skeleton sequences that are connected by convolution operations frame by frame. The convolution is performed spatially first then temporally next, which encodes the complex high-dimensional motions as the reduced-dimensional feature vector. We denote the ST-GCN encoder as $\phi(\cdot)$:

$$\boldsymbol{p}_i = \textit{ST-GCN}(\boldsymbol{m}_i^k) = \phi(\boldsymbol{m}_i^k)$$

ResNet-18 Music Encoder: To represent the acoustic features from music, we first perform acoustic feature extraction to facilitate the subsequent residual operation. The acoustic feature is expected to be more informative and non-redundant than the raw music, where the information is retrieved from given music by using the audio signal processing library - *librosa* [11]. Specifically, there are 5 categories of features $\mathcal{F}(\boldsymbol{a}_i) = \{\boldsymbol{a}_0, \boldsymbol{a}_1, ..., \boldsymbol{a}_4\}$ that are extracted in this study: the Mel-frequency cepstral coefficients (MFCC), MFCC-delta, constant-Q chromagram, tempogram and onset strength. Given the constructed acoustic features, we adopt ResNet-18 $\psi(\cdot)$ to abstract acoustic features as in [12, 13]. The ablation study conducted in Section 4.3 proves the best results of feature representation learning by using ST-GCN and ResNet-18 encoders for motion and music.

$$\boldsymbol{q}_i = \textit{ResNet}(\boldsymbol{a}_i) = \psi(\boldsymbol{a}_i)$$

The feature embedding $\boldsymbol{z}_i$ of the motion and music is obtained by concatenating outputs of 2 encoders followed by an MLP layer.

2.4. InfoNCE Contratstive Loss

After obtaining the feature embedding $\boldsymbol{z}_i$, the InfoNCE contrastive loss [20] is applied because it can efficiently learn latent representations and achieve generalizability in various domains [5, 21]. The loss compares the distance to positive examples with the distance to negative examples for each positive pair of the network

Table 1: Comparison of state-of-art human motion skill and dance dataset

Dataset	3D-Skeleton	Music	Genres	Choreography	**Expertise Level**	Subjects	Sequences	**Repeated Motion**	Fps	Seconds
BEST[14]	N	N	0	0	3	500	500	100	10	25000
JIGSWAS[15]	N	N	0	0	3	7	103	30	30	515
DanceNet[16]	Y	Y	2	2	0	2	2	1	60	3472
Dance with Melody[17]	Y	Y	0	40	0	-	61	1	25	5640
GrooveNet[18]	Y	Y	4	4	0	1	2	1	100	1380
AIST+[19]	Y	Y	10	101	0	10	1408	2	60	18694
Dance-Music-Level	**Y**	**Y**	**5**	**20**	**3**	**2**	**6000**	**100**	**100**	**60000**

output. There are N examples that are randomly chosen within the minibatch, and the training is formulated based on the augmented pairs that are obtained from each example, which resulting $2N$ data points. According to [22], apart from the positive example, we treat the other $2(N-1)$ examples within a minibatch as negative examples. If (i, j) is a positive pair, the loss function is defined as:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{r=1}^{2N} 1_{[r \neq i]} \exp(\text{sim}(z_i, z_r)/\tau)} \quad (1)$$

where the indicator function $1_{[r \neq i]} \in \{0, 1\}$ is estimated to be 1 if $r = i$. The final contrastive loss is computed across all positive pairs within minibatch for both (i, j) and (j, i). Specifically, we can conclude from Equation (1) that, the loss value is low if positive samples are encoded to similar (closer) representations in latent space and negative samples are encoded to dissimilar (further) representations.

3. DANCE-MUSIC-LEVEL DATASET

We construct and annotate the Dance-Music-Level dataset which consists of 5 different dance genres: (i) Ballet, (ii) K-pop, (iii) Jazz, (iv) Hip-hop and (v) Urban. For Ballet and K-pop, there are 5 different choreographies within each genre, while 2 choreographies for Hip-Hop and 4 choreographies for Jazz and Urban, resulting in 20 diverse choreographies in total. Music for these 20 choreographies is also included. For different choreographies that belong to the same dance genre, the provided music and designed dance steps for each choreography are different. For each choreography, dance motions from individuals in 3 expertise levels are included, while each expertise level contains 100 dance sequences with the same music. Each sequence contains one individual's performance. To sum up, there are 6000 sequences provided in this dataset ($6000 = 20 \times 3 \times 100$).

Two subjects (1 female and 1 male) are invited to develop this dataset. Table 1 shows the dataset specification. The first 2 rows are the motion skill levels datasets that are widely used in AQA studies [1, 2]. However, there is no other public dataset that describes the dance motion level. While the rest rows are the dance-music dataset without different skills level included. There are 2 types of data in our dataset: motion and music. As for the motion, the data is collected by using OptiTrack motion capture devices, which record the coordinates of 21 skeletal joints in 3D space at 100fps. To this end, for a 10 seconds motion sequence, 63,000 feature dimension is obtained, which can be represented as $(10, 100, 21, 3)$. Musical signals are recorded and temporally aligned with the dance motions by utilizing the robot operating system (ROS)[23]. By using *librosa*, a 10 seconds music can be represented as 220500 feature dimensions.

4. EXPERIMENTS

We conduct 3 experiments to demonstrate: (i) Pre-train Feature Representation Evaluation: the validity of our proposed strategy on motion-music feature representation for different dance genres, choreographies and expertise levels. (ii) Downstream Tasks and Expertise Level Modeling: the effectiveness of our approach on expertise level classification and level improvement modeling, and (iii) Quantitative Analysis: the results of our method in comparison with different baselines. The Adam optimizer is used to train all models, with a batch size of 8 and a learning rate of 3×10^{-4} for 100 epochs. The model is trained in PyTorch using RTX 3080 GPUs.

4.1. Pre-train Feature Representation Evaluation

We first train a model using all sequences (6000) in our proposed Dance-Music-Level dataset. To visualize the training feature embedding, the features are mapped onto 2D space by using t-SNE. For clarity, we show the results from 5 dance genres, (each dance genre includes 1 choreography, each choreography contains 3 different expertise levels data, and each expertise level contains 100 dance sequences, resulting 1500 sequences $1500 = 5 \times 1 \times 3 \times 100$), in Fig. 2(a). Two observations can be summarized from Fig. 2(a): (i) The proposed method is able to separate different dance genres within the 2D latent space. The 5 selected dance genres are denoted in 5 colors (red, purple, blue, yellow and green) in the figure. (ii) The method is also capable of separating different expertise levels of each choreography. For example, the beginner, intermediate and expert data for Ballet choreography are classified into 3 clusters (purple), and far away from other choreographies data points.

4.2. Downstream Task and Expertise Level Modeling

In this experiment, we split the Dance-Music-Level dataset into two subsets: training (900) and testing (1200) sequences. We trained the model with the training dataset of choreographies and tested the latent representation with unseen choreographies. We select 3 different dance genres for training: Ballet, K-pop and Urban. Each genre contains 1 choreography, and each choreography contains 3 expertise levels with 100 dance sequences included. We then perform a downstream task to test the model using the testing set, which includes Hip-Hop, K-pop and Ballet dance. For Ballet and Hip-Hop, there is 1 choreography for each, while for K-Pop, two choreographies are included. Fig. 2(b) shows the feature representation of the downstream testing result. Three observations can be made from this result: (i) Our method shows the ability of clustering different unseen dance genres (denoted as purple, yellow and green colors) in downstream prediction tasks. (ii) The method is capable of clustering different choreographies (denoted as yellow color) of the same dance genres in the downstream task. (iii) The proposed strategy also shows robustness and generalizability in separating different expertise levels for each dance choreography. As we can see from the Fig. 2(b), for all dance choreographies, including Ballet (purple), two K-Pop choreographies (yellow), and Hip-Hop (green), the three expertise levels for each choreography are well separated.

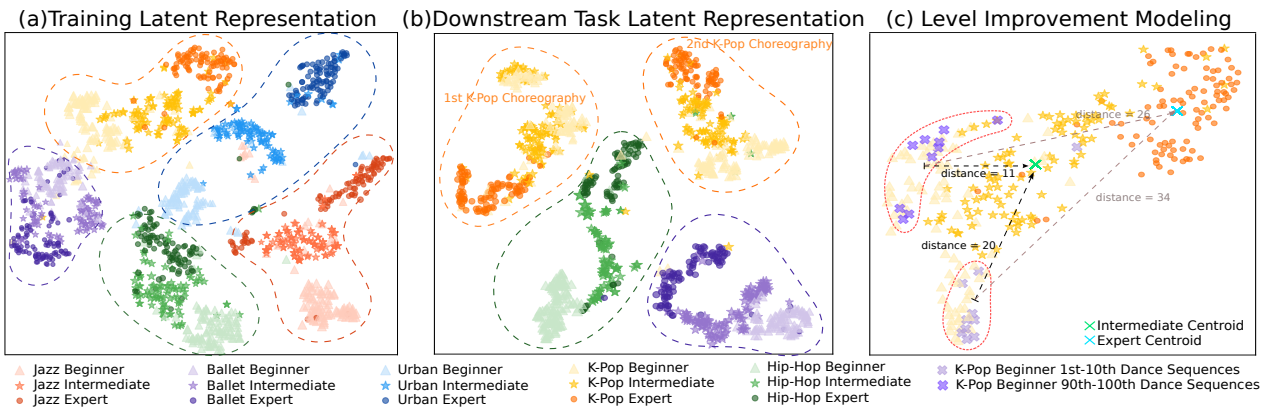

Fig. 2: Qualitative analysis. (a) Motion and music feature representation of the pre-train tasks. Five different dance genres and three expertise levels for each genre can be well separated simultaneously. (b) Latent representation of the downstream tasks. Four testing choreographies that are unseen in the pre-train phase are classified. Three distinct expertise levels for each choreography are also separated. (c) Level improvement modeling. The median Euclidean distance of a beginner's 90th-100th dance sequences to the intermediate and expert cluster center is closer than the distance between his/her 1st-10th attempts.

Table 2: Quantitative analysis

Method	Motion Encoder	Music Encoder	Top1	Top5
	ResNet-18	LSTM	28.43	72.79
	ResNet-50	LSTM	34.56	70.22
	ResNet-18	ResNet-18	58.51	87.62
	ResNet-50	ResNet-50	64.34	91.12
	ST-GCN	LSTM	67.83	93.38
	ST-GCN	ResNet-50	73.65	95.34
Ours	**ST-GCN**	**ResNet-18**	**74.26**	**95.59**
MoCo [8]	ResNet-50	ResNet-50	14.71	22.16
SimCLR [5]	ResNet-50	ResNet-50	64.34	91.12
Cross Entropy [9]	ResNet-50	ResNet-50	38.46	51.33
Ours	**ST-GCN**	**ResNet-18**	**74.26**	**95.59**

We then analyze the improvement made in dance capability for individual from a particular expertise level during downstream tasks. During the data collecting process, as the individual keeps repeating the same choreography 100 times, his/her performance is expected to experience an improvement. In 2(c), we first obtain the latent representation of an example of testing K-Pop choreography. Similar to the above experimental results, the latent space shows the clustered 3 expertise levels of data. We then visualize the 1st-10th beginner dance sequences (lilac) and last 90th-100th beginner dance sequences (dark purple), and measure their median Euclidean distance to the center of intermediate and expert clusters. Specifically, the distances of the 1st-10th sequences to the intermediate and expert clusters are 11 and 26 respectively. While the distances of the last 90th-100th sequences to the intermediate and expert clusters are 20 and 34 respectively, which is closer than the first 10 sequences. This indicates a level improvement of the beginner's dance capabilities when repeatedly performing the same choreography.

4.3. Quantitative Analysis

We compare our proposed contrastive self-supervised method against: (i) **Ablation Study:** Using InfoNCE loss with different encoder architectures. and (ii) **Baselines:** using the proposed encoders with different representation learning methods and loss functions. For both experiments, we perform downstream logistic regression to obtain the expertise level classification accuracy. We split the Dance-Level-Music dataset into 3 subsets: (i) training sequences (2700), (ii) downstream LR training sequences (900), and (iii) downstream LR testing sequences (600). Table 2 shows the detailed experimental results. For the ablation study, we compare

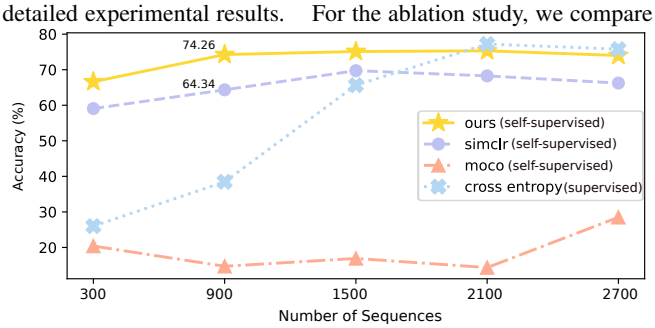

Fig. 3: Comparison of four learning mechanisms with varies number of training sequences used in logistic regression. Our method shows better accuracy when fewer training sequences are given. The accuracy of our approach exceeds by 15% compared to simCLR when trained with training sequences.

our proposed ST-GCN and ResNet-18 encoders with other architectures to show the superiority of the designed encoder combination in representing motion and music features. While for the baselines, we compare our contrastive self-supervised framework with state-of-the-art representation learning methods, where our proposed method achieves the highest classification accuracy. Additionally, we modify the number of logistic regression training sequences aiming to show the advantages of our proposed self-supervised method as a pre-train process. As shown in Fig. 3, our method shows capabilities of achieving the highest accuracy with fewer training sequences, which is potential to reduce the effort in labeling data.

5. CONCLUSION

We formulate a novel multimedia dance motion analysis problem and present a contrastive framework to solve it. We evaluate our proposed method by constructing a multi-modal Dance-Music-Level dataset including diverse expertise levels of dance motions. The framework is capable of not only evaluating human dance expertise levels based on different music, genres, choreographies qualitatively and quantitatively, but also monitoring the dancers' level improvements. The method has the potential to be extended to various applications such as music-aided dance teaching and motion generation.

6. REFERENCES

[1] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[2] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, pp. 443–455, 2018.

[3] D. S. Alexiadis and P. Daras, "Quaternionic signal processing techniques for automatic evaluation of dance performances from mocap data," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1391–1406, 2014.

[4] I. Kico and F. Liarokapis, "Comparison of trajectories and quaternions of folk dance movements using dynamic time warping," in *2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, 2019, pp. 1–4.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] A. Examinations Department of Royal Academy of Dance, "Qcf examinations information, rules and regulations - royal academy of dance," 2020. [Online]. Available: https://media.royalacademyofdance.org/media/2019/12/17135212/20191216-Specifications-SCOTLAND.pdf

[11] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[12] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[13] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6548–6552.

[14] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7862–7871.

[15] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, no. 3, 2014.

[16] W. Zhuang, C. Wang, J. Chai, Y. Wang, M. Shao, and S. Xia, "Music2dance: Dancenet for music-driven dance generation," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1–21, 2022.

[17] T. Tang, J. Jia, and H. Mao, "Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis," *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[18] O. Alemi, J. Françoise, and P. Pasquier, "Groovenet: Real-time music-driven dance movement generation using artificial neural networks," *networks*, vol. 8, no. 17, p. 26, 2017.

[19] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 401–13 412.

[20] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," *Advances in neural information processing systems*, vol. 29, 2016.

[21] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[22] T. Chen, Y. Sun, Y. Shi, and L. Hong, "On sampling strategies for neural network-based collaborative filtering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 767–776.

[23] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system." [Online]. Available: https://www.ros.org