# Innovative Teaching Method for Dunhuang Dance Incorporating Pose Estimation Fusion

Zhenjie Liu[1,a], Shuaishuai Li[2,b], Yuezhou Zhang[2,c], Jiaxin Wang[2,d], Xihong Luo[3,e], Xiangzhen He[1,f*]

[1]Key Laboratory of Minzu Languages and Cultures Intelligent Information Processing, Gansu Province(Northwest Minzu University), Lanzhou, China
[2]Key Laboratory of Linguistic and Cultural Computing Ministry of Education(Northwest Minzu University), Lanzhou, China
[3]Department of Dance, Northwest Minzu University, Lanzhou, China

[a]holmexing@163.com, [b]1784545874@qq.com
[c]zyz_jinqu@qq.com, [d]610407262@qq.com
[e]619747959@qq.com, [f*]hexiangzhen@163.com

**Abstract.** To overcome the challenges of high costs, venue requirements, and lack of portability in current digital teaching for Dunhuang dance, this study uses Vicon optical motion capture to create a Dunhuang dance motion database and virtual dancers. Using lightweight Convolutional Neural Networks (CNN) for single-view pose estimation in Unity, it captures students' 3D body movements in real-time, driving their virtual avatars. The result is a virtual scene replicating the "one-on-one" teacher-student interaction. This approach enhances user engagement, offering a fresh learning experience and introducing a new teaching paradigm for Dunhuang dance.

**Keywords:** Dunhuang dance, motion capture, gesture estimation, dance learning.

## 1 Introduction

In recent years, the rapid development of the internet has spurred changes in education. The effective use of motion capture technology in teaching has seamlessly combined artificial intelligence with reality. Motion capture (MoCap) tracks body movements and joint positions in 3D space, making it a preferred choice for educational experiments due to its high fidelity, detailed motion capture, and cross-platform ease.

From the perspective of preserving intangible cultural heritage, researchers focused on Dunhuang dance [1,2], emphasizing immersive learning and precise digitization. They creatively bound Dunhuang dance motion data to 3D virtual characters, designing Dunhuang-style animations [3], contributing significantly to Dunhuang culture. An innovative 3D digital teaching system for Dunhuang dance was developed using a virtual game engine [4], achieving the digitization of Dunhuang artistic works.


M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_4

Recent methods using depth cameras and sensors to capture body poses gained attention. Researchers at Shanghai University, led by Zhu and Li [5], used Kinect depth photography to obtain users' motion and skeletal data, creating a dance teaching system. However, these methods have shortcomings: (1) Digital teaching still relies on "observation," lacking participation for dance students, leading to ineffective learning. (2) The use of high-definition depth cameras with motion sensors is hindered by expensive, non-portable equipment and sensitivity to environmental conditions[6,7], limiting Dunhuang dance promotion.

To meet the needs of enthusiasts and professional students, this paper combines pose estimation[8-11] and MoCap[12,13]. High-precision MoCap establishes a Dunhuang dance motion database, binding animations to 3D teacher models. Using an improved lightweight Convolutional Neural Network (CNN) model[14], student motion capture requires only a regular RGB camera. Real-time skeleton reconstruction in Unity drives student avatars, creating a simple "one-on-one" virtual teaching scene. This innovative fusion provides a new learning experience, offering a fresh approach to Dunhuang dance teaching.

## 2 Virtual Teacher Motion Driving Module

### 2.1 Dunhuang dance database

#### 2.1.1.Hardware System.

The study uses the Vicon passive optical motion capture system, commonly applied in research areas like digitization preservation and virtual reality[15-19]. It functions by capturing light reflections from markers attached to motion nodes through cameras. The computer interprets this light information, facilitating data collection and recording. The markers, also called Marker points, have a spherical shape and are coated with retroreflective material. Marker point size is selected based on the corresponding joints of the captured object in experiments.

#### 2.1.2.Motion Capture.

Creating a Dunhuang Dance Motion Database through extensive capture and recording of Dunhuang Flying dance movements using motion capture devices. The main components include pre-capture preparations, the specific process and methods of motion capture data collection, optimization of motion capture data, and the binding of motion capture data with virtual dance character models. The process of capturing dance movements using motion capture devices is illustrated in Figure 1.

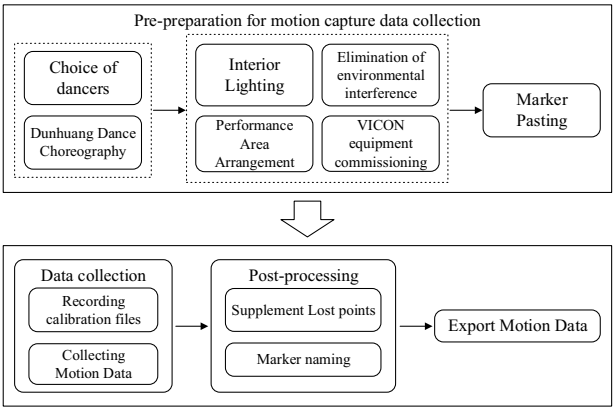


**Fig. 1.** Dunhuang Dance Acquisition process.

This study employs Vicon Shōgun for capturing dance movements, with two PC workstations working collaboratively to ensure no data loss and enhance data quality. Dance performers wear specialized motion capture suits and choose markers of appropriate specifications. Following ergonomic principles, 63 markers are applied to the entire body. Once everything is set up, performers execute pre-arranged dance movements within the capture space. The scene for collecting motion data is illustrated in Figure 2.

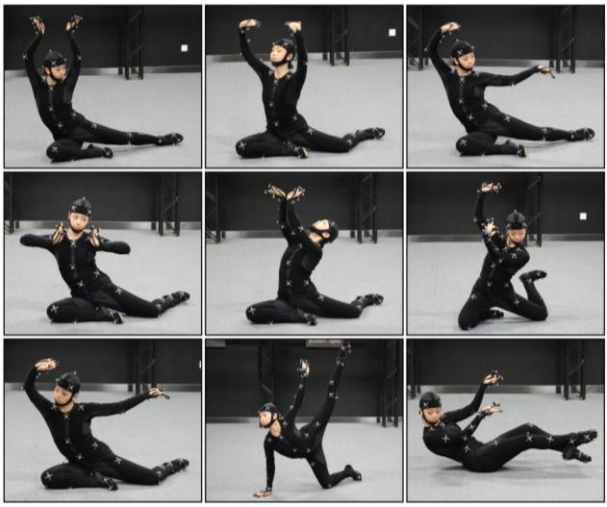

**Fig. 2.** Motion Capture Dance Data Acquisition.

### 2.1.3.Data Processing.

Motion capture (MoCap) data relies on markers, and occlusion can affect data quality, especially in dances with intricate finger movements like the Feitian dance. To address this, the Shōgun post system's built-in calibration and repair functions are used

for data optimization. The processed data is exported and cleaned to remove scattered noise, resulting in smoother original data. Finally, the point cloud skeleton data is converted into Unity-compatible .fbx format for driving the motion of a 3D virtual teacher in Unity.

## 2.2 MoCap Data Driving Virtual Teacher

Building upon the Dunhuang dance database, the core development of the virtual teacher dance-driving module is done in Unity. Skeletal data is presented in. FBX skeletal animations. Directly importing character models doesn't allow skeletal animation control. Hence, mapping the character model's skeleton to a Unity-created standard skeleton is essential. This precise binding to the Avatar enables seamless movement based on captured human skeletal data, facilitating the reuse of Dunhuang dance animation data.

*For the virtual teacher avatar, an Animator Controller component is bound, adding processed Dunhuang dance. fbx animations. Animation switching is controlled through triggers. This creates a professionally modeled virtual teacher avatar capable of gracefully dancing Dunhuang dance. It allows the demonstration and teaching of Dunhuang dance movements in a virtual environment. Figure 3 illustrates the virtual teacher avatar created by binding motion capture data of partial Feitian basic poses to a 3D model.*

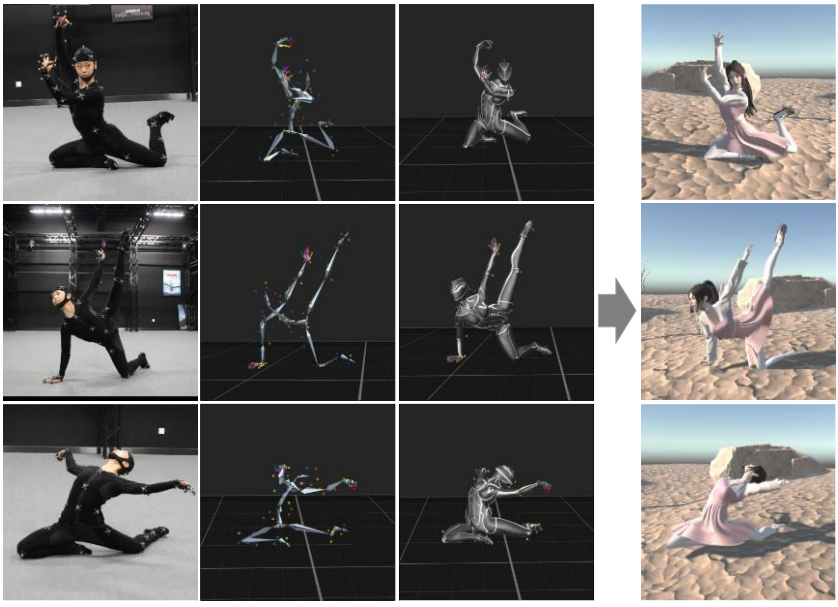


**Fig. 3.** Motion Capture Data-Driven Virtual Teachers.

# 3 Real-time Student Motion Data Driving Module

## 3.1 Pose Estimation Neural Model

To overcome the challenges of expensive large-scale equipment and limited portability of posture sensors for widespread use in synchronous teaching, a single-view real-time human pose estimation technique based on the VNect strategy is employed. This method can simultaneously regress 2D and 3D joint positions in real-time without the need for high-quality cameras. In contrast to Kinect, it achieves quality pose estimation results with low-cost RGB cameras, webcams, or even mobile phones.

The CNN under the VNect strategy primarily utilizes a Residual Neural Network (Resnet50). In this study, to reduce computational costs, a lighter Resnet34 network is used for effective human pose estimation. Figure 4 illustrates the pose estimation results for the fundamental postures of the Dunhuang Feitian dance. It is evident that the movements of the dance performer have been comprehensively captured.

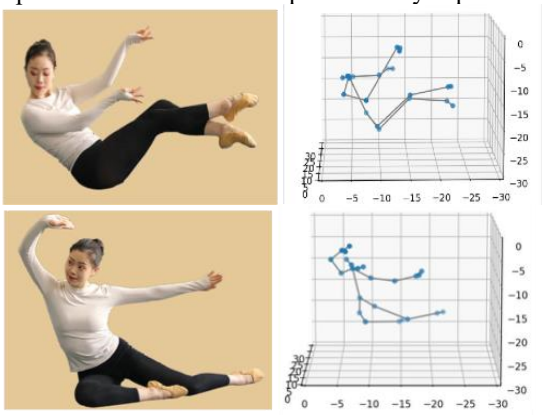

**Fig. 4.** Estimated effect of flying stylized gestures.

## 3.2 Human Pose Prediction

In Unity, the Barracuda package reads the neural network model and uses T-pose images as input for initial pose. The system captures real-time human motion images through a monocular camera with the Video-capture script. Defining 24 body joints, each frame of the two-dimensional image is input into the neural network, producing 3D heatmaps and offset maps for joints. The heatmap divides the original image into a (a, a) two-dimensional grid (where 'a' is 28), influencing the confidence accuracy for joints. By finding the most likely feature map and grid in the heatmap, a rough joint position is determined. The corresponding 3 feature maps in the offset represent precise joint offsets.

This two-stage process, using heatmap for rough positioning and offset for accurate adjustment, achieves real-time prediction of the learner's 2D and 3D joint positions. The workflow for driving the 3D model with CNN-based pose estimation is shown in Figure 5.

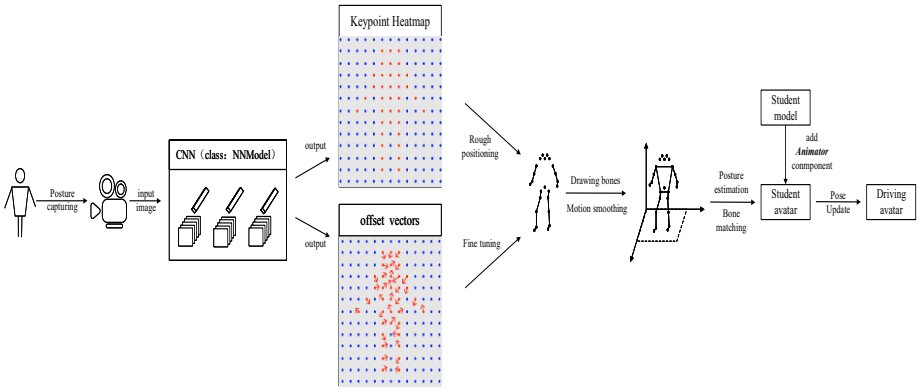


**Fig. 5.** Monocular pose estimation technology driving 3D avatar.

## 3.3 Real-time Driving of Virtual Student Motion

### 3.3.1.Definition and Model Matching of Joints.

Using the calculated 3D coordinates and bones to bind and drive the avatar. After predicting the 3D spatial positions of the model's pose, the joints of the avatar are predefined and named. An Animator component is added to the user's avatar model, and its built-in methods are used to obtain the current body skeleton's displacement and size changes. The avatar's bones are then bound to the defined joint numbers in a one-to-one correspondence, establishing a hierarchical relationship between the joints. Figure 6 illustrates the positions of the 24 joints, and Table 1 provides the joint names corresponding to each number.

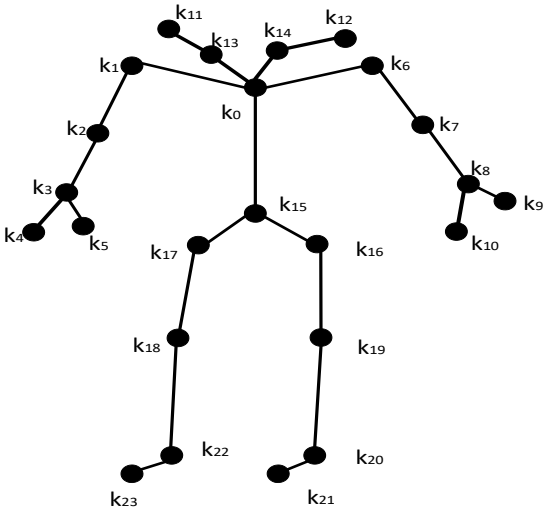


**Fig. 6.** 24 human joints.

**Table 1.** Corresponding number for each joint

| Joint Name Definition | Joint Corresponding Number |
|---|---|
| Right Arm | Shoulder $k_1$, Elbow $k_2$, Wrist $k_3$, Middle Finger $k_4$, Thumb $k_5$ |
| Left Arm | Shoulder $k_6$, Elbow $k_7$, Wrist $k_8$, Middle Finger $k_{10}$, Thumb $k_9$ |
| Right Leg | Knee $k_{18}$, Ankle $k_{22}$, Sole $k_{23}$ |
| Left Leg | Knee $k_{19}$, Ankle $k_{20}$, Sole $k_{21}$ |
| Left Hip | $K_{16}$ |
| Right Hip | $K_{17}$ |
| Root | $K_{15}$ |
| Neck | $K_0$ |
| Ear | Left ear $k_{12}$, Right ear $k_{11}$ |
| Eyes | Left eye $k_{14}$, Right eye $k_{13}$ |

### 3.3.2.Skeleton Alignment.

The position of each joint in the body can move independently or follow the motion of its parent joint. Real-time pose estimation updates the current joint position. To accurately calculate the real-time joint rotation of the current skeleton, we use gaze rotation (LookRotation) and intermediate alignment matrices.

Before driving the avatar's movement, we calculate the body orientation using the coordinates of the root joint and left-right hip joints in the initial pose. Each joint and its child joints are initialized for alignment, except for the head and palms, which require separate calculations due to their uniqueness.

$$j_i^{fy} = \mathrm{tr}(r,\ tb_l,\ tb_r) \tag{1}$$

$$j_i^{fy} = \mathrm{v}\left(j_i,\ j_{i_c},\ j_i^{fy}\right) \tag{2}$$

*Formulas (1) and (2) calculate LookRotation for the body. Here, i is the index of the body joint.* $j_i^{fy}$ *represents the forward y-vector for joint j, equivalent to the normal direction of the plane formed by the root joint r, left hip joint* $tb_l$ *,and right hip joint* $tb_r$.$j_i^{fz}$*represents the z-vector perpendicular to y for joint j, resulting from the combination of the current joint vector j, the vector of the child joint of j,and the vector* $j_i^{fy}$.

$$j_H^{fz} = v(j_H, j_n) \tag{3}$$

*Formula (3) similarly directly calculates the vector from the head to the nose as the z-vector for the LookRotation of the head.*

$$j_h^{f_y} = tr(w,t,mf) \tag{4}$$

$$j_{h_l}^{f_z} = v(t,mf) \tag{5}$$

$$j_{h_r}^{f_z} = v(mf, t) \tag{6}$$

*Formulas (4), (5), and (6) calculate the LookRotation vector for the hands. In these formulas, the coordinates of hand joints (wrist joint, thumb, and middle finger) collectively determine the normal vector of the plane representing the y-direction. The z-direction for the left wrist is formed by the vector from the thumb to the middle finger, and for the right wrist, it is formed by the vector from the middle finger to the thumb.*

$$J^f = J^{Init} * Q \tag{7}$$

$$A * A^{-1} = 1 \tag{8}$$

$$Q^{-1} = J^{Init} * (J^f)^{-1} \tag{9}$$

*Combining the LookRotation matrix of the character model with the model's initial rotation (InitRotation) using formulas (7) and (8) yields formula (9). Here, $J^f$ is the LookRotation matrix of the model, and $J^{Init}$ is the initial rotation (InitRotation) matrix. This step calculates the intermediate alignment matrix Q, which unifies the 3D model and predicted skeleton in a common world coordinate system for easier mapping and updating of joint positions.*

### 3.3.3.Real-time prediction of joint movements.

Calculate the LookRotation matrix for the user's skeleton obtained from pose estimation, denoted as $J^f$ . The current joint rotation for the user is represented by $J^{now}$ , and driving the avatar's movements using the skeleton involves aligning the two coordinate systems using the previously calculated intermediate matrix Q.

$$Q^{-1} = J^{now} * (J^f)^{-1} \tag{10}$$

$$J^{now} = Q^{-1} * J^f \tag{11}$$

*By applying formulas (10) and (11), the real-time rotation matrix $J^{now}$ for the user is obtained. Utilizing the PoseUpdate method, the current skeletal rotation of the user is aligned with LookRotation and bound to the 3D model skeleton, updating joint information, thereby achieving real-time driving of the student avatar.*

To test the use of a monocular camera in Unity for capturing and driving Dunhuang dance postures, the experiment involved launching the Unity program, activating the camera, positioning the subject centrally, mimicking Dunhuang Feitian dance moves along with music, observing the 3D model's movements, and capturing the process at specific intervals. The 3D character model used is from Unity's Unity-chan resources. Figure 7 illustrates the real-time driving effect of the student avatar.

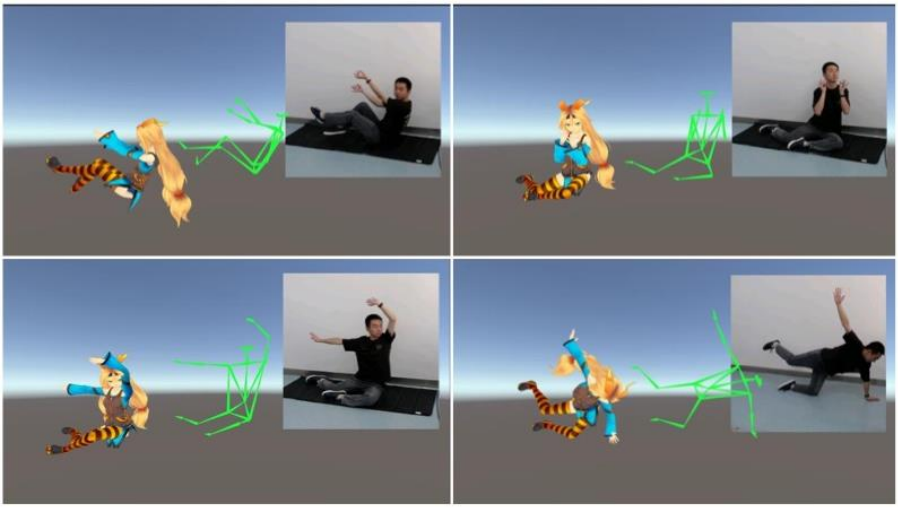

**Fig. 7.** Real-time 3D models driven by estimation data.

Observing the experiment, we noticed significant jitter in the skeleton during stillness and occasional disruptions or delays in consecutive movements. These issues can affect students' performance and learning quality in the virtual environment. To address this, noise reduction is considered using Kalman filtering for motion smoothing. The linear equation for the Kalman filter is represented by Formula 12.

$$\hat{x}_k = \hat{x}_{k-1} + K_k(\hat{z}_k - \hat{x}_{k-1}) \tag{12}$$

*The data fusion schematic diagram is illustrated in Figure 8.*

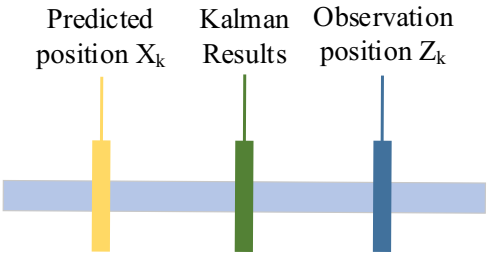


**Fig. 8.** Kalman data fusion.

The Kalman filter has a limitation: it tends to be influenced by the previous state position, causing predicted data to be close to the previous state position, especially when there's a significant gap between the previous and current positions. To address this, the paper proposes an improvement: using a convolutional neural network for joint

prediction and tracking the data changes in the xyz-axis directions of each joint sequentially. While this increases computation, it helps prevent jitter in predicted skeletons when the character is stationary.

To confirm the effectiveness of the enhancement, experiments will be conducted to observe the impact on character animation before and after improving the Kalman filter. It is crucial to maintain a frame rate of 24 frames per second throughout the experiment for stable graphics card performance and consistent state transitions.

The learning platform was initially operated using the unenhanced Kalman filter. Next, to address the patterns in human body movements, an improved Kalman filter is introduced for validation. The comparison of the two approaches is shown in Figure 9, with the left side depicting the original performance and the right side showcasing the enhanced results.

(Left) (Right)

**Fig. 9.** Comparison of Kalman Filter Smoothing Effects.

When the arm is rapidly lowered, the following observations can be made:

(1) Kalman filtering improvement before. When the arm swings down rapidly, there is a significant delay in the corresponding movement of the virtual character's arm in three dimensions. As a result, the virtual character's arm tends to be closer to the position of the previous state. Adjusting the Kalman's k value to reduce the predicted value would cause the result to lean toward the observed value, thus reducing the filtering and noise reduction effect.

(2) Kalman filtering improvement after. When the arm undergoes the same acceleration motion, due to considering the current state along with the previous three states, we can determine the motion in the x-axis, y-axis, and z-axis directions. This helps to infer the current state's position more accurately by separately calculating the three axes, and the effect is significantly better than the Kalman filtering improvement before.

## 4 Implementation of One-on-One Virtual Avatar Teaching

Using Unity3D as the main platform, the development of the two-dimensional and three-dimensional presentation modules is integrated with posture estimation and Dunhuang dance motion capture in the teaching scenario. The presentation module is a

crucial component that facilitates the visualization and interaction among teaching platform features, scenes, and animations. Leveraging Unity's visual programming approach, code is written to implement menu options, button functionalities, progress bar features, perspective changes, and various UI elements required for the platform. Through UI controls, interactive "one-on-one" learning is achieved between the teacher and student in the virtual environment. Figures 10-11 showcase the effects of the Dunhuang dance teaching scenario.

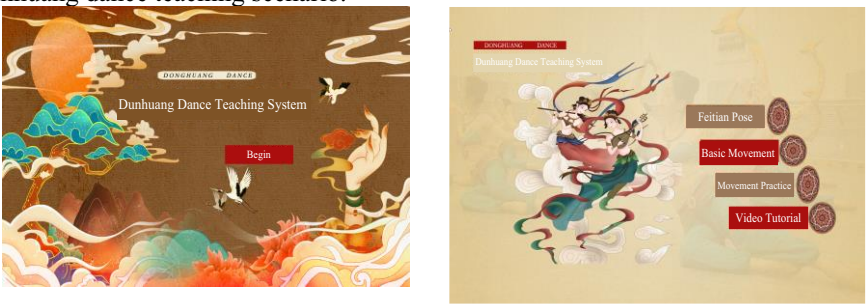


**Fig. 10.** Interface Design.

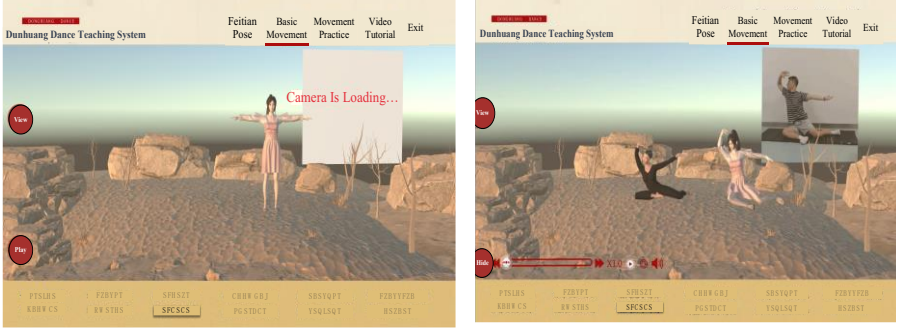


**Fig. 11.** "one-on-one" virtual teaching scenarios.

## 5 Conclusion

This research focuses on diverse applications for preserving and transmitting cultural art, specifically in the context of Dunhuang dance teaching. The study integrates established motion capture and pose estimation technologies, creating a comprehensive database of Dunhuang dance movement data. Utilizing the Unity platform, a novel one-on-one teaching approach is implemented, where the teacher's movements are driven by motion capture data, and the student's movements are guided by single-camera pose estimation. This innovative system offers a unique learning experience and opens up new possibilities for Dunhuang dance teaching.

The strategy of using single-camera pose estimation for learning addresses dance movement education, providing practical solutions for various industries. Additionally, the joint real-time driving of virtual human movements shows promise for application

in augmented reality and virtual reality scenarios. The outcomes of this research can contribute significantly to Dunhuang cultural promotion, dance dissemination, virtual teaching, and the digitization of intangible heritage, marking a meaningful advancement in Dunhuang dance preservation and protection.

## Acknowledgement

Supported by National Natural Science Foundation of China (62341209), Supported by the Fundamental Research Funds for the Central Universities (31920230054), Northwest Minzu University Educational Teaching Reform Research Project (2022XJJG-105).

## References

1. Gao, J. R. 2011. Training Course of Dunhuang Dance. Shanghai Music Publishing House Co., Ltd. Shanghai, China.
2. Li, S. B. 2019. Analysis of the Formation of Dance Language from Image Materials in Dunhuang Murals. MSc Thesis, Beijing Dance Academy.
3. Gao, Y. H. 2021. Research on the Creation of Dunhuang-Style Dance Animation Based on Motion Capture Technology. MSc Thesis, Lanzhou Jiaotong University.
4. Hu, Y. R. 2022. Design and Implementation of a Demonstration System for Dunhuang Dance Teaching Based on Motion Capture Technology. MSc Thesis, Northwest Minzu University.
5. Zhu, Y. C, Li, Y. Z, Tian, F. 2022. Design and Implementation of Kinect Based Dance Teaching System. J. Industrial Control Computer, 35(04), 107-109.
6. Takala T M, Hirao Y, Morikawa H, et al. 2020. Martial arts training in virtual reality with full-body tracking and physically simulated opponents. In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), IEEE, 858-858.
7. Liu, J., Zheng, Y., Wang, K., et al. 2020. A real-time interactive tai chi learning system based on VR and motion capture technology. J. Procedia Computer Science, 174, 712-719.
8. Papandreou G, Zhu T, Chen L C, et al. 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European conference on computer vision (ECCV), 269-286.
9. Rong Y, Shiratori T, Joo H. 2020. Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. J. arXiv preprint arXiv:2008.08324.
10. Theobalt C, Casas D, Mehta D, et al. (2017) VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. Retrieved November 11, 2023, from DOI: 10. 1145/ 3072959. 3073596
11. Ren W, Ma O, Ji H, et al. 2020. Human posture recognition using a hybrid of fuzzy logic and machine learning approaches. In IEEE Access, 8, 135628-135639.
12. Nirme J, Haake M, Gulz A, et al. 2020. Motion capture-based animated characters for the study of speech–gesture integration. J. Behavior Research Methods, 52(3), 1339-1354.
13. Mikić I, Trivedi M, Hunter E, et al. 2002. Human body model acquisition and motion capture using voxel data. In International Conference on Articulated Motion and Deformable Objects. Springer, Berlin, Heidelberg, 104-118.

14. Mehta, D., Sridhar, S., Sotnychenko, O., et al. 2017. Vnect: Real-time 3D human pose estimation with a single RGB camera. J. ACM Transactions on Graphics (TOG), 36(4), 1-14.
15. Wu, Z. F. 2009. Research and Implementation of Digitalization of Puppetry Based on Motion Capture Technology. MSc Thesis, University of Electronic Science and Technology of China.
16. Guo, Y. 2020. Research on Interactive Teaching System for Ansai Waist Drum Based on Motion Capture Technology. MSc Thesis, Xi'an University of Technology.
17. Covaci A, Postelnicu C C, Panfir A N, et al. 2012. A virtual reality simulator for basketball free-throw skills development. In Doctoral Conference on Computing, Electrical and Industrial Systems. Springer, Berlin, Heidelberg, 105-112.
18. Chan, J. C. P., Leung, H., Tang, J. K. T., et al. 2010. A virtual reality dance training system using motion capture technology. J. IEEE Transactions on Learning Technologies, 4(2), 187-195.
19. Wang, L. C. 2016. Research on Dance Pose Analysis and Teaching Methods Based on Motion Capture Technology. MSc Thesis, Liaoning Normal University.