Optimizing Dance Training Programs Using Deep Learning: Exploring Motion Feedback Mechanisms Based on Pose Recognition and Prediction

Yuting Jiao*
Quanzhou Preschool Teachers College, Fujian China

Abstract—**Dance pose recognition and prediction is an important part of dance training and a challenging task in the field of artificial intelligence. Due to the diverse styles and significant variations in dance movements, conventional methods struggle to capture effective dance pose features for recognition. In this context, we have developed a dance pose recognition and prediction method based on deep learning. Given the characteristics of dance movements, such as complex human postures and dynamic movements, we proposed the MKFF-ST-GCN model, which integrates multi-kinematic feature fusion with ST-GCN. This model fully captures the dynamic information of dance movements by calculating the first and second-order kinematic features of keypoints and fuses the kinematic features using a multi-head attention mechanism. Additionally, to address dance pose prediction issues, we proposed the STGA-Net based on the spatial-temporal graph attention mechanism. This model improves the long-distance information modeling capability by calculating local and global graph attentions of dance poses, effectively solving the problem of dance pose prediction. To comprehensively evaluate the quality of the proposed methods in dance pose recognition and prediction, we conducted extensive experimental validations and comparisons with several common algorithms. The experimental results fully demonstrate the effectiveness of our methods in dance pose recognition and prediction. This study not only advances the technology of dance pose recognition and prediction but also provides valuable experience for the field.**

Keywords—Deep learning; pose recognition; pose prediction; dance training; graph convolutional network; attention mechanism

I. INTRODUCTION

Dance constitutes a significant expressive medium within the realm of modern art, serving as a pivotal channel for not only emotional expression but also the transmission of cultural heritage. Traditional dance training predominantly relies upon the professional abilities of instructors, with evaluative criteria heavily influenced by subjective experiences and judgments, which lacks objectivity and necessitates substantial resource and time investments. With the proliferation of dance training, an increasing number of practitioners are demanding a cost-effective, personalized learning experience equipped with real-time feedback, thus underscoring the imperative for the development of innovative dance training methodologies. In the recent era, advancements in artificial intelligence (AI) technology, particularly through deep learning techniques within the domain of computer vision, have been noteworthy [1]-[3]. These techniques are adept at extracting latent information from image data, and their application has been extensive in areas such as defect detection and facial recognition. The application of deep learning for human action recognition and prediction in dance training allows for the effective recognition and comprehension of dance movements. This not only markedly enhances the efficiency and quality of dance training but also opens novel avenues for the transformation of traditional dance training paradigms.

Integrating human pose recognition and prediction with dance training represents a multidimensional research domain where technology and art converge extensively. This approach seeks to leverage the sophisticated image recognition capabilities inherent in deep learning algorithms to capture and analyze the movements and postures of practitioners with precision, thereby achieving objective and scientific training feedback. Additionally, human posture prediction based on deep learning can integrate sequential image data to predict future movements, effectively preventing potential health risks during training. Therefore, accurately recognizing and predicting human dance movements holds significant theoretical and practical significance. It can reduce the subjectivity of teacher evaluations, help establish objective and personalized evaluation standards, and provide practitioners with more convenient and cost-effective dance training methods.

In the domain of dance education, accurately and effectively identifying and predicting sequences of dance movements presents a significant challenge [4]. This challenge stems from the complexity inherent in human posture, the diversity of dance styles, and the inherent uncertainty of movement execution. To address these issues, it is crucial to enhance the learning capabilities of algorithms to ensure better fitting and prediction accuracy for dance movements. To this end, we propose the integration of Graph Convolutional Neural Networks (GCNs) in the study of dance movement recognition and prediction. GCNs are particularly adept at capturing the complex spatial relationships between human postures within a dance context. Moreover, to augment the feature extraction capabilities of our model and to enhance focus on pivotal pose changes, we introduce a multi-granularity hierarchical adaptive attention mechanism. This mechanism is designed to dynamically adjust the focal points of attention at various layers within the network. It is foreseeable to capture the subtle details and nuances of key movements and transitions more accurately, thereby optimizing the process of dance teaching.

*Corresponding Author.

II. Literature Review

This article will collect existing work in the field of human pose recognition to highlight the shortcomings of current research.

A. Traditional Human Pose Recognition Methods

Over the past few years, methodologies based on pose recognition have been applied to various fields, such as sports and dance training. Early human pose recognition algorithms primarily utilized manually designed features to reflect the spatial and temporal information of human movements, which were then processed by advanced machine learning algorithms to obtain the corresponding recognition results. GIST and HOG (Histogram of Oriented Gradients) have been widely adopted as image feature descriptors. Kuehne et al. [5] examined the effectiveness of different feature extraction methods on various datasets. The study highlights that GIST features, which capture the background context, perform slightly better (60.0%) compared to HOG features (58.6%). Furthermore, it is believed to incorporate motion features in pose recognition. Techniques such as Motion Boundary Histogram, Histograms of Optical Flow, and dense trajectories were developed to capture motion information. Fan et al. [6] proposed an innovative method for improving the accuracy of human action recognition systems. Their approach integrates a dense sampling strategy that concentrates on motion boundaries with histograms of motion gradients to optimize the feature extraction process.

Additionally, the Scale-Invariant Feature Transform (SIFT) [7] is commonly utilized alongside the Histogram of Oriented Gradients (HOG). Zhang et al. [8] proposed a novel method to leverage the SIFT flow, which effectively captures the displacement of key points between video frames. This approach involves tracking key points that are invariant to scale changes across video frames, describing these key points with local appearance and motion descriptors like Histograms of Oriented Gradients (HOG).

Despite the progress made by the aforementioned methods, algorithms based on hand-crafted features and machine learning exhibit several notable limitations. These methods generally demonstrate poor generalizability and struggle to adapt to the diversity of movements in dance training. Additionally, the process of manually designing feature extraction is complex and time-consuming, requiring domain-specific expertise and often failing to capture critical information of dance movements comprehensively. These issues limit the effectiveness of traditional methods in meeting the high accuracy and generalizability requirements of dance training, making intelligent and adaptive human pose recognition technology still a challenge.

B. Deep Learning-based Human Pose Recognition Methods

Currently, deep learning-based methods dominate the field of human pose recognition. Compared to traditional approaches, deep learning methods utilize massive datasets to gather more accurate and comprehensive information about the subjects, thereby enhancing recognition accuracy and bolstering robustness against environmental variability. Additionally, these approaches offer better scalability and end-to-end inference capabilities, allowing for more comprehensive dance training systems by integrating with other advanced techniques.

Ng et al. [9] proposed to obtain the spatial representation of the human pose at each frame by 2D CNN and fused by a multi-layer Long-short Term Memory network (LSTM). Additionally, J. Donahue et al. [10] introduced a model that employs a two-layer LSTM, known as Long-term Recurrent Convolutional Networks (LRCN). Furthermore, Li et al. [11] proposed a human pose recognition method based on translation-scale invariant image mapping and multi-scale deep Convolutional Neural Networks (CNN). The joint positions were mapped to image space, and the multi-scale CNN was utilized to extract features and achieve recognition results.

Compared to RGB-based methods, skeleton-based models are widely adopted in human pose recognition. These models represent human skeletons as structured graphs, which are then processed by Graph Neural Networks (GNN). S. Yan et al. [12] introduced a Spatial-Temporal GCN (ST-GCN) for action recognition, which operates similarly to 3D convolutional networks but processes skeleton graphs, achieving an accuracy of 30.7% on the Kinetics-400 dataset. Meanwhile, Y. Song et al. [13] enhanced GCNs by incorporating a suite of advanced techniques, including batch normalization [14]. This led to the development of an Efficient-GCN, which not only delivers competitive performance on pose recognition but also requires less training time and offers greater explainability.

In recent years, the advancement of attention-based approaches in Natural Language Processing (NLP) has driven their integration with advanced techniques in the field of computer vision. G. Bertasius et al. [15] explored various configurations of spatial and temporal self-attention mechanisms and developed the TimeSformer. AGCN [16] integrates an attention mechanism into the Graph Convolutional Network (GCN) framework. It utilizes three forms of attention: spatial, temporal, and channel attention, which collectively enable AGCN to achieve superior accuracy scores. Similarly, C. Si et al. [17] introduced the Attention Enhanced Graph Convolutional LSTM Network (AGC-LSTM). In this model, temporal dynamics are handled by an LSTM, while spatial relationships are managed through a GCN augmented with attention mechanisms.

Overall, the application of deep learning in human pose recognition is advancing swiftly, with various deep learning models constantly expanding the capabilities of pose recognition technology. Despite encountering several obstacles, these models have proven highly effective in precise detection, analysis, and interpretation of complex human motions. With ongoing advancements in research and technology, we anticipate that future systems for human pose recognition will evolve to be more advanced, offering more accurate and varied pose analysis capabilities.

C. Research Gaps

Despite significant progress in the field of deep learning for human pose recognition and prediction, its application in dance training still has notable deficiencies. Key issues, such as the precision in capturing subtle movements, handling the correlation of long-distance movements, and the diversity of dance movements, have not been fully addressed. Therefore, future research should focus on the following areas:

1) Limitations of CNN models in dance pose recognition: The CNN-based methods for dance movement recognition have achieved certain results, but the complex spatial relationships between dance movements make it difficult for CNN methods that rely on RGB image inputs to effectively capture the dynamic relationships of dance movements. Additionally, CNN-based approaches heavily depend on the volume of training data, leading to weaker generalization capabilities when dealing with unfamiliar dance styles or new movements.

2) Lack of kinematic and multi-scale infomation in dance pose recognition: Currently, there is no unified standard for preprocessing dance data, leading to various studies employing their own methodologies. For instance, some methods focus on simple pose estimation while others may integrate complex motion analysis, but often these approaches do not fully capture the kinematic details such as the fluidity of motion and multi-scale movements. Even in advanced models that attempt to incorporate both static poses and dynamic sequences, there are challenges in accurately synchronizing small-scale movements with larger body movements during analysis. This absence of a standardized approach to capturing and analyzing kinematic and multi-scale information limits the effectiveness and interoperability of different dance pose recognition methods.

3) Limitations of deep learning models in dance pose prediction: While deep learning models like Graph Convolutional Networks (GCNs) have shown promise in understanding complex spatial relationships in dance movements, they face significant challenges in long-term motion prediction. Current GCN models excel in capturing the instantaneous relationships between body joints but struggle with generating or predicting extended sequences of movements. Due to the fundamental differences between static spatial data representation and the temporal dynamics of dance, GCNs cannot transition into predicting long-term dance sequences without modifications or integration with other temporal-focused models.

In summary, future research should focus on developing new models and techniques to address these challenges in dance pose recognition and prediction, to not only enhance the effectiveness of dance training but also improve the acceptance of participants.

III. Research on Dance Pose Recognition and Prediction Methods Based on Graph Convolutional Networks

In this section, a novel framework based on GCN is proposed for dance pose recognition. The preliminary knowledge of spatial-temporal GCN (ST-GCN) is first introduced, followed by the proposed multi-kinematic feature fusion-based spatial-temporal GCN (MKFF-ST-GCN). Finally, the spatial-tempotal graph attention-based network (STGA-Net) is designed to predict dance poses.

A. Spatial-Temporal Graph Convolutional Network

In the contemporary field of dance pose recognition, deep learning technologies such as Convolutional Neural Network (CNN) and Graph Convolutional Network (GCN) have begun to be explored for understanding human movements. With their advanced data processing capabilities, they exhibit notable representative potential.

Graph Convolutional Network (GCN) is a specialized type of neural network designed to operate directly on graphs instead of grid data such as images. Compared to traditional CNN, the convolutional operation in GCN is adapted to aggregate information from neighbors of nodes which can capture the spatial relationships within the graph. Additionally, the GCN models can leverage the node features and graph topology to generate powerful node embeddings that can be used for a variety of tasks, such as node classification, graph classification, and link prediction.

Considering the graph CNN model within one single frame, the N joint nodes can be expressed as V_t and the skeleton edges can be expressed as $E_S(\tau) = \{v_i^t v_j^t \mid t = \tau, (i, j) \in H\}$, where H represents the set of naturally connected human body joints. Assuming that the kernel size of the convolutional operator is K×K, and the number of channels of the input feature map f_{in} is c. The output feature map for a single channel can be expressed as

$$f_{out}(x) = \sum_{h=1}^{K}\sum_{w=1}^{K} f_{in}(\boldsymbol{p}(x,h,w)) \cdot \boldsymbol{w}(h,w) \tag{1}$$

where $\boldsymbol{p}: Z^2 \times Z^2 \to Z^2$ represents the sampling function that enumerates the neighbors of location x, and $\boldsymbol{w}: Z^2 \to \square^{c}$ denotes the weight function that provides the weight vector for computing the inner product with the sampled input features. Furthermore, the sampling function on the neighbor set of a node v_i^t can be defined as $B(v_i^t) = \{v_j^t \mid d(v_j^t, v_i^t) \le D\}$ and the weight function can be simplified by partitioning the neighbor set $B(v_i^t)$ of a joint node v_i^t into a fixed number of K subsets, where each subset has a numerical label. Then, the spatial graph convolution can be rewritten as

$$f_{out}(v_i^t) = \sum_{v_j^t \in B(v_i^t)} \frac{1}{Z_i^t(v_j^t)} f_{in}(\boldsymbol{p}(v_i^t, v_j^t)) \cdot \boldsymbol{w}(v_i^t, v_j^t) \tag{2}$$

In addition, the same joints across consecutive frames are connected to integrate temporal information, that can be expressed as

$$B(v_i^t) = \{v_j^q \mid d(v_j^t, v_i^t) \le K, |q - t| \le |\Gamma / 2|\} \tag{3}$$

The parameter Γ represents the temporal kernel size. Since the temporal axis is well-ordered, the label function l_{ST} can be expressed as

$$l_S^T(v_j^q) = l_i^t(v_j^t) + (q - t + \lfloor \Gamma / 2 \rfloor \times K) \tag{4}$$

Based on the above ST-GCN, this study introduced a novel multi-kinematic feature fusion (MKFF) module to comprehensively capture the movement information, thus addressing jitter and occlusion.

B. MKFF-ST-GCN for Dance Pose Recognition

In this section, we introduce the MKFF-ST-GCN (Multi-Kinematic Features Fusion-based Spatial-Temporal Graph Convolution Network), which is specially designed for dance pose recognition. Its network structure has been adjusted to suit the characteristics of dance movement data, as shown in Fig. 1.

Considering the jitter and occlusion during dance movements, relying on the structural information of human body keypoints cannot fulfill these requirements. As a result, this paper introduced MKFF module to leverage the keypoint kinematic features within a sliding window, obtaining from previous, current, and next frames' poses. Traditional ST-GCN capture keypoint information through complex network structure, which may not be sufficient to obtain the dynamic patterns in dance movements. By integrating the multi-kinematic features fusion module, the proposed method can effectively calculate the inherent kinematic features without increasing network complexity and extra parameters.

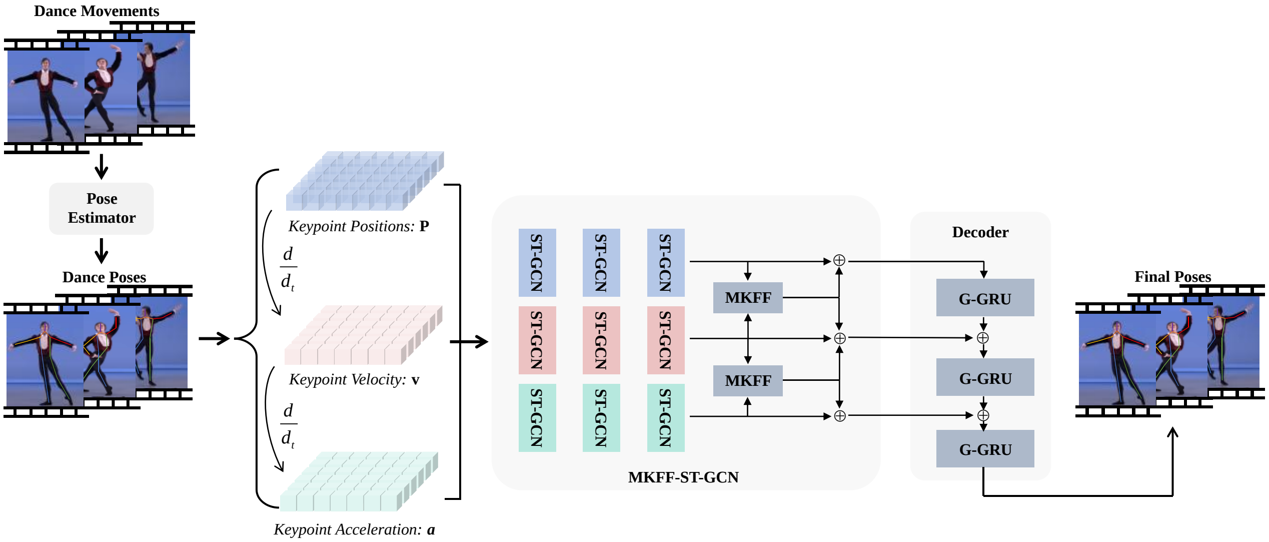

Fig. 1. Network structure diagram based on MKFF-ST-GCN dance pose recognition.

The input dance poses are firstly computed by off-the-shelf pose estimatore, denotaed as $\boldsymbol{P} \in \mathbb{R}^{T/N\times(K\cdot D)}$. By leveraging the consecutive poses's coordinates at each frame, the keypoint flow can be expressed as

$$\bar{\boldsymbol{P}}_t = (\boldsymbol{P}_t + \boldsymbol{P}_{t+d_t} + \boldsymbol{P}_{t-d_t})/3 \tag{5}$$

where d_t denotes an interval from the previous and next poses to the current poses. Based on the above keypoint flow, the keypoint velocity and acceleration can be expressed as

$$\begin{aligned} v_t &= (\boldsymbol{P}_t - \boldsymbol{P}_{t-d_t})/d_t \\ a &= (v_t - v_{t-d_t})/d_t \end{aligned} \tag{6}$$

Dance movements often have rich spatio-temporal relationships and subtle dynamic changes, making it difficult for a single feature extraction structure to fully capture these characteristics. To address this issue, we propose a multi-kinematic feature fusion (MKFF) module based on a multi-head attention mechanism. This module is designed to facilitate information exchange among different motion features, enhance the perception of complex dance movement structures and details, and improve the capture of dance motion features. By conducting in-depth analysis and reasonable fusion of dynamic features of keypoints, our method effectively strengthens the relationships between key frames, accurately identifies various types of dance movements, and provides strong technical support for dance training and performance.

As shown in Fig. 2, the MKFF module consists of three feature fusion block followed by an activate function and dropout layer, each block includes several critical parts, such as Layer normalization, multi-head self-attention, and MLP block.

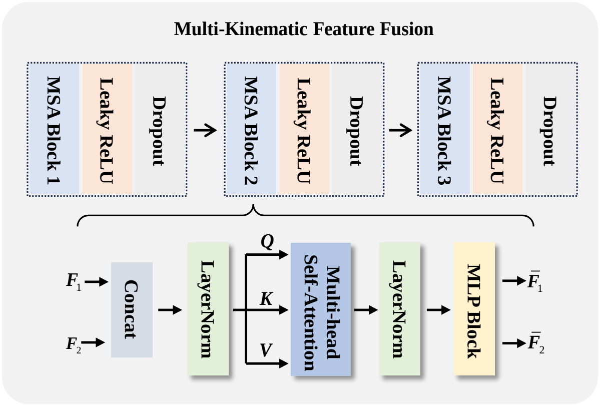

Fig. 2. Schematic diagram of the multi-kinematic feature structure fusion module.

First, the input multi-kinematic features are concatenated in keypoint dimernsion to obtain the initial embedding features, which can be expressed as

$$Z_0 = \text{concat}(\boldsymbol{F}_1; \boldsymbol{F}_2) W_o \tag{7}$$

where $W_0 \in \mathbb{R}^{(K \cdot 2D) \times C}$ is a linear projection matrix. Then, the initial embedding features are mapped to Q_0, K_0, V_0 representing queries, keys, and values, which can be expressed as

$$\begin{aligned} Q_t &= (Z_{t-1}) W_t + E_{pos,t} \\ K_t &= (Z_{t-1}) W_t + E_{pos,t} \\ V_t &= (Z_{t-1}) W_t + E_{pos,t} \end{aligned} \tag{8}$$

where $E_{pos,t}$ represents the positional embedding at t-th block. The multi-head self-attention is adopted to automatically calculate value map thus capturing the distribution of large and small dance movements. Besides, a layer normalization module is added before every multi-head self-attention and multi-layer perceptron (MLP) block. The entire process can be expressed as

$$\begin{aligned} \bar{Z}_t &= \text{MSA}(\text{LN}(Z_t)) + Z_t \\ \hat{Z}_t &= \text{MLP}(\text{LN}(\bar{Z}_t)) + \bar{Z}_t \\ MSA(Q_t, K_t, V_t) &= \text{soft}\max\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t \end{aligned} \tag{9}$$

The fused multi-kinematic features are then proceeded by the decoder module to obtain the final dance poses. The proposed model is trained by a weighted loss with objective of minimizing the weighted L_1 norm between prediction and ground truth joint positions, which is defined as

$$L_w = \frac{1}{N_j} \sum_{j=1}^{N_j} v_j \left\| G_j - P_j \right\| + \frac{\lambda}{N_k} \sum_{k=1}^{N_k} v_k \left\| G_k - P_k \right\| \tag{10}$$

where N_k, G_j, P_j, v_j represents the number of top-k keypoints, ground truth, predicition, and visibility of joint j, respectivley. The initial term of the loss function targets errors across all keypoints, whereas the subsequent term specifically addresses errors associated with the top-k keypoints. With this carefully designed network structure, MKFF-ST-GCN can recognize dance poses with high accuracy and precision, providing a powerful tool for automated dance training.

C. STGA-Net for Dance Pose Prediction

In dance training, the smoothness and continuity of movements greatly affect the quality of the dance. Therefore, recognizing the current dance posture and predicting the next frame of dance movement is crucial. Effective prediction of future dance movements can help dancers understand the transitions between movements, accelerate the learning process, and correct potential errors or dangerous movements, significantly impacting the efficiency and effectiveness of dance teaching and training. To address these issues, this study introduced a novel spatial-temporal graph attention mechanism (STGA-Net) to predict the dance poses based on the above recognition result.

As shown in Fig. 3, the proposed graph attention block includes the global graph attention layer and the local graph attention layer. The local spatial attention module is designed to model the hierarchical and symmetrical structure of dance poses providing fine-grained dance movements. The global spatial attention module is introduced to adaptively extract global semantic information to better understand the spatial characteristic and the relationship between consecutive dance movements. Furthermore, the proposed local and global spatial graph attention module holds significant advantages in dance movement prediction. Traditional GCN models have a limited receptive field, making it difficult to capture subtle interactions between keypoints. Dance training often involves rapid and minor changes in movement patterns, which poses challenges for traditional GCNs. The proposed local and global spatial graph attention mechanism optimizes the flow of information, enabling adaptation to a variety of dance styles and enhancing the accuracy and robustness of predictions. This is crucial for understanding complex dance sequences and predicting dancers' movement trajectories in advance, thereby significantly improving the quality of dance training and performance.

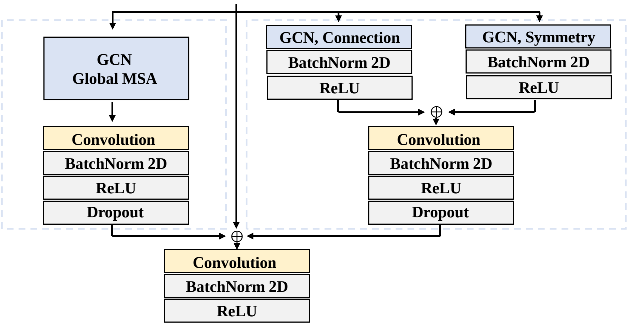

Fig. 3. Schematic diagram of graph attention module.

The skeleton 2D poses are first defined as a graph $\Omega = \{V, E\}$, where V represents the set of nodes and E represents edges. The node features are defined as $\boldsymbol{X} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N \mid \vec{x}_i \in \mathbb{R}^C\}$ which has C channels. Then, the output features of graph convolutional layer can be expressed as

$$\boldsymbol{X}^{l+1} = \text{soft}\max(\boldsymbol{M} \odot \tilde{\boldsymbol{A}}) \boldsymbol{X}^l W \tag{11}$$

where W is a learnable matrix for channel transformation, M represents a learnable mask matrix, $\tilde{\boldsymbol{A}}$ represents the connections between joints. By designing $\tilde{\boldsymbol{A}}$, this study introduces two different kinds of spatial graph attention, a symmetric matrix to encode the symmetrical counterpart and am adjacency matrix to encode the connections for distal joints. Besides, the global attention mechanism aims to encode the relationship across disconnected joints, thus addressing depth ambiguities and occlusions. The global attention mechanism can be expressed as

$$\boldsymbol{X}^{l+1} = concat(\boldsymbol{B}_k + \boldsymbol{C}_k) \boldsymbol{X}^l W_k \tag{12}$$

where $\boldsymbol{B}_k$ represents an adaptive global adjacency matrix, $\boldsymbol{C}_k$ signifies a learnable global adjacency matrix, and W_k is a transformed matrix. In addition, the temporal information is captured by the temporal dilated convolutional block. To save the spatial information across dance poses, the original convolutional block is replaced by 2D convolutions with $k \times 1$ kernel size. After encoding the keypoints features of dance poses, the latent features are processed by a G-GRU decoder to predict future dance poses, as shown in Fig. 4.

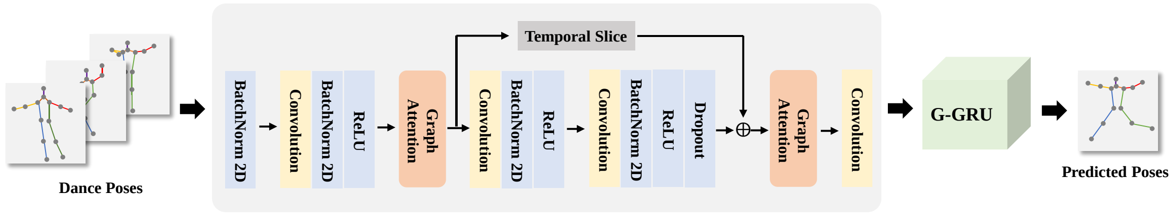

Fig. 4. Schematic diagram of STGA-Net for dance pose predictio.

IV. Case Verification

This section will validate the effectiveness of the proposed method based on a self-made experimental dataset.

A. Experimental Environment

The hardware environment for the experiments in this chapter is shown in Table I:

TABLE I. Experimental Software and Hardware Environment Table

CPU	Intel(R) Core(TM) i5-13400F
GPU	NVIDIA GeForce RTX 4070
Memory	12.0 Gb
Operating System	Ubuntu 18.04
CUDA	11.1
Main Frameworks	Pytorch1.10.0
Main Programming Language	Python 3.8

To explore dance pose recognition and prediction, this study has collected dancing videos including ballet, street dance, and modern dance. The collected datasets dataset contains 30 dance video clips, totalling about 150,000 frames. From these, approximately 8,000 images have been selected for dance motion recognition and prediction. Each dance movement image is manually annotated with 14 key points, including the head, shoulders, elbows, wrists, etc., to support precise capture and analysis of dance movements, and the annotation information is stored in JSON format. Additionally, data augmentation techniques have been applied to the constructed dance motion dataset, including random rotation, scaling, flipping, adding random noise, and Gaussian blur, to improve the model's generalization ability. The dataset is divided into 70% training set, 15% validation set, and 15% test set.

The experimental section designed several different comparisons to comprehensively evaluate the performance of the proposed model in dance movement recognition and prediction. It compares with two widely used dance movement recognition models, HBRNN [18] and HRNet [19]. Specific aspects include: dance movement recognition results under different dance styles, accuracy of dance movement keypoint recognition, prediction results of different dance styles, and ablation experiments on the proposed graph attention mechanism. The model parameters used for validation were pre-trained on the Kinetics dataset and fine-tuned on the dataset constructed in this paper. During the training process, the stochastic gradient descent algorithm was used, with an initial learning rate of 0.01, a linear learning rate decay strategy with a decay weight of 0.95, a batch size set at 16, and a total of 80 training epochs.

B. Experimental Results

To accurately and effectively recognize dance movements, this study constructed the MKFF-ST-GCN model based on the experimental setup mentioned above. The model training was conducted in a supervised manner, and the error variation curve during the training process is shown in Fig. 5. It can be observed that after pre-training, the network model achieved good results on the dataset used in this paper, with the error curve quickly decreasing and gradually stabilizing. Furthermore, we presented the qualitative recognition results of dance movements in various styles such as contemporary, ballet, and street dance, as shown in Fig. 6. It is evident that the proposed MKFF-ST-GCN demonstrated the highest recognition accuracy and the ability to capture complex postures in all dance styles. Benefiting from the MKFF module, the proposed model was able to accurately locate key points, showing significant improvements over HBRNN, and performed better in handling complex movements, significantly outperforming the other two algorithms in the dance movement recognition task.

In addition, to further demonstrate the effectiveness of the proposed method in recognizing dance movements, ballet dance data is used as an example, employing the Percentage of Correct Keypoints as the evaluation metric, with results shown in Tables II and III. Specifically, Table II displays the recognition accuracy for keypoints such as the head, shoulders, wrists, and knees under a threshold of 0.2, where MKFF-ST-GCN performs the best across all keypoints. Table III further tests more stringent thresholds (0.1 and 0.05), and the results show that MKFF-ST-GCN continues to lead under high precision requirements, followed by HRNet, while HBRNN performs relatively poorly. These findings indicate that MKFF-ST-GCN has significant advantages in handling complex dance movements and high-precision keypoint localization, making it suitable for high-demand dance movement recognition tasks.

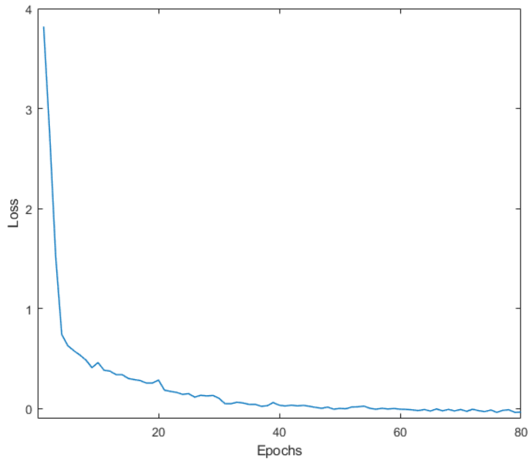

Fig. 5. Model loss rate curve.

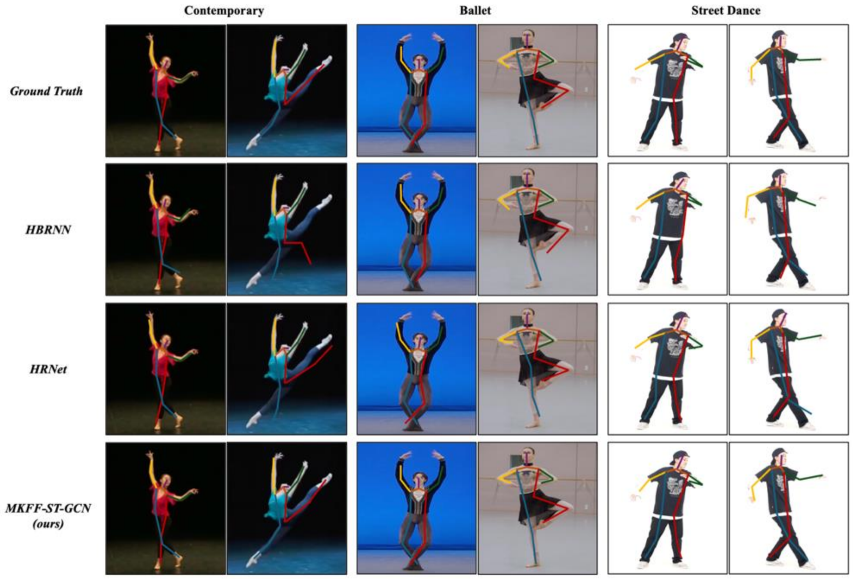

Fig. 6. Qualitative comparison of dance pose recognition among various dance styles.

TABLE II. Quantitative Comparison of Dance Pose Recognition Among Different Keypoints

Method	PCK@0.2			
	Head	Shoulder	Wrist	Knee
HBRNN	86.4	82.9	83.3	88.1
HRNet	90.2	88.3	87.5	91.1
MKFF-ST-GCN	**93.5**	**90.1**	**90.7**	**92.5**

TABLE III. Quantitative Comparison of Dance Pose Recognition Under Different Thresholds

Method	PCK@0.1	PCK@0.05
	Avg.	Avg.
HBRNN	81.2	79.6
HRNet	84.3	81.1
MKFF-ST-GCN	**86.2**	**83.9**

The aforementioned results adequately demonstrate that the proposed MKFF-ST-GCN can effectively extract latent features of dance movements, capture subtle motion changes, and accurately identify dance movements of different styles. Furthermore, this section verifies the performance of the proposed STGA-Net in predicting dance movements with two widely adopted methods, DMGNN [20] and T-GCN [21]. Extensive experiments were conducted for different dance styles and various time intervals, with the results presented in the Tables IV to VI.

TABLE IV. Quantitative Comparison of Ballet Pose Prediction

Method	Ballet		
	80	160	320
DMGNN	0.88	1.10	1.39
T-GCN	0.45	0.62	**0.88**
STGA-Net	**0.24**	**0.40**	0.90

TABLE V. QUANTITATIVE COMPARISON OF CONTEMPORARY DANCE POSE PREDICTION

Method	**Contemporary Dance**		
	80	160	320
DMGNN	0.31	0.67	0.90
T-GCN	0.39	0.44	0.81
STGA-Net	**0.19**	**0.36**	**0.58**

TABLE VI. QUANTITATIVE COMPARISON OF STREET DANCE POSE PREDICTION

Method	**Street Dance**		
	80	160	320
DMGNN	0.39	0.80	1.32
T-GCN	0.41	0.76	1.09
STGA-Net	**0.22**	**0.60**	**0.92**

The above results showcase the performance of STGA-Net in predicting dance movements in ballet, contemporary dance, and street dance, using the Mean Absolute Error (MAE) between predicted and actual keypoints as the evaluation metric. The results indicate that across all tested dance styles and prediction intervals (80ms, 160ms, 320ms), STGA-Net consistently demonstrated superior accuracy compared to the other two methods (DMGNN and T-GCN). Specifically, for ballet, STGA-Net achieved MAE values of 0.24, 0.40, and 0.90 at prediction intervals of 80ms, 160ms, and 320ms respectively, significantly outperforming DMGNN and T-GCN. In contemporary dance, STGA-Net also performed the best across all intervals, with MAE values of 0.19, 0.36, and 0.58 respectively. In the prediction of street dance, STGA-Net maintained its lead with MAE values of 0.22, 0.60, and 0.92 at the 80ms, 160ms, and 320ms intervals, respectively. These results demonstrate that STGA-Net can effectively reduce prediction errors in dance movement prediction tasks across different dance styles and time intervals, showcasing its potential and practicality in the field of motion prediction.

V. CONCLUSION

This study aims to optimize the dance training process using artificial intelligence technology, implementing dance movement recognition and prediction based on deep learning algorithms, which greatly promotes the application of AI technology in the field of dance training. To extract the dynamic and complex features contained in dance movements, a MKFF-GCN model based on the multi-kinematic features fusion has been developed. This model supplements the dynamic information missing in structured data by calculating the kinematic information of keypoints and effectively fuses these features using a multi-head attention mechanism. Additionally, to accurately predict dance poses and avoid accidents during training, this paper proposes a STGA-Net based on the spatial-temporal graph attention mechanism, which can effectively model the local and global relationships between dance movement keypoints, enhancing the capability to extract long-distance information. Experiments show that compared to previous algorithms, the proposed models can effectively recognize and predict dance poses, significantly improving the efficiency of dance training. It must be acknowledged that there is still considerable room for improvement in the real-time recognition performance of the proposed models. In the future, we will continue to optimize the model structure to enhance real-time recognition capabilities, further aiding the dance training process.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[2] P. Zhou, B. Gao, S. Wang, and T. Chai, “Identification of Abnormal Conditions for Fused Magnesium Melting Process Based on Deep Learning and Multisource Information Fusion,” IEEE Trans. Ind. Electron., vol. 69, no. 3, pp. 3017–3026, Mar. 2022.

[3] G. Lan, Y. Wu, F. Hu, and Q. Hao, “Vision-Based Human Pose Estimation via Deep Learning: A Survey,” IEEE Trans. Human-Mach. Syst., vol. 53, no. 1, pp. 253–268, Feb. 2023.

[4] A. Bera, M. Nasipuri, O. Krejcar, and D. Bhattacharjee, “Fine-Grained Sports, Yoga, and Dance Postures Recognition: A Benchmark Analysis,” IEEE Trans. Instrum. Meas., vol. 72, pp. 1–13, 2023.

[5] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” in 2011 International Conference on Computer Vision, Barcelona, Spain: IEEE, Nov. 2011, pp. 2556–2563.

[6] M. Fan, Q. Han, X. Zhang, Y. Liu, H. Chen, and Y. Hu, “Human Action Recognition Based on Dense Sampling of Motion Boundary and Histogram of Motion Gradient,” in 2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS), May 2018, pp. 1033–1038.

[7] Z. Wang, Z. Wang, H. Liu, and Z. Huo, “Scale - invariant feature matching based on pairs of feature points,” IET Computer Vision, vol. 9, no. 6, pp. 789-796, Dec. 2015.

[8] J.-T. Zhang, A.-C. Tsoi, and S.-L. Lo, “Scale Invariant Feature Transform Flow trajectory approach with applications to human action recognition,” in 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China: IEEE, Jul. 2014, pp. 1197–1204.

[9] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 4694–4702.

[10] J. Donahue et al., “Long-term recurrent convolutional networks for visual recognition and description,” in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 2625–263.

[11] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He, “Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN,” in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, Hong Kong: IEEE, Jul. 2017, pp. 601–604.

[12] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition,” AAAI, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12328.

[13] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, “Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition,” in Proceedings of the 28th ACM International Conference on Multimedia, Oct. 2020, pp. 1625–1633.

[14] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in International conference on machine learning, 2015, pp:448-456.

[15] G. Bertasius, H. Wang, and L. Torresani, “Is Space-Time Attention All You Need for Video Understanding,” in International conference on machine learning, 2021.

[16] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks,” IEEE Trans. on Image Process., vol. 29, pp. 9532–9545, 2020.

[17] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, “An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action

Recognition," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019.

[18] Yong Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 1110–1118.

[19] J. Wang et al., "Deep High-Resolution Representation Learning for Visual Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[20] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020.

[21] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning Trajectory Dependencies for Human Motion Prediction," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 9488–9496.