Towards Accurate 3D Human Motion Prediction from Incomplete Observations

Qiongjie Cui Huaijiang Sun*
Nanjing University of Science and Technology, Nanjing, PR China
cuiqiongjie@njust.edu.cn, sunhuaijiang@njust.edu.cn

Abstract

Predicting accurate and realistic future human poses from historically observed sequences is a fundamental task in the intersection of computer vision, graphics, and artificial intelligence. Recently, continuous efforts have been devoted to addressing this issue, which has achieved remarkable progress. However, the existing work is seriously limited by complete observation, that is, once the historical motion sequence is incomplete (with missing values), it can only produce unexpected predictions or even deformities. Furthermore, due to inevitable reasons such as occlusion and the lack of equipment precision, the incompleteness of motion data occurs frequently, which hinders the practical application of current algorithms.

In this work, we first notice this challenging problem, i.e., how to generate high-fidelity human motion predictions from incomplete observations. To solve it, we propose a novel multi-task graph convolutional network (MT-GCN). Specifically, the model involves two branches, in which the primary task is to focus on forecasting future 3D human actions accurately, while the auxiliary one is to repair the missing value of the incomplete observation. Both of them are integrated into a unified framework to share the spatio-temporal representation, which improves the final performance of each collaboratively. On three large-scale datasets, for various data missing scenarios in the real world, extensive experiments demonstrate that our approach is consistently superior to the state-of-the-art methods in which the missing values from incomplete observations are not explicitly analyzed.

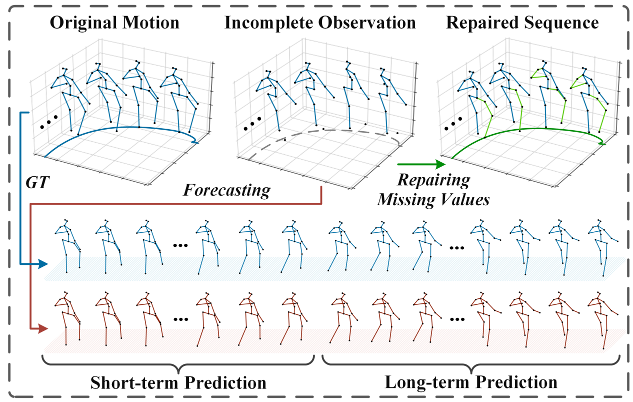

Figure 1. **Example results.** In the upper part, the middle is an incomplete observation (single leg or arm is missing) of the original motion. Our approach focuses particularly on generating the predicted poses directly from the incomplete data while repairing the missing value incidentally.

1. Introduction

3D human motion prediction has present considerable potential in many computer vision applications, such as human behavior understanding, machine intelligence, and autonomous driving [51, 31, 6, 42, 5, 41, 47]. For instance, robots in our daily life plan their actions in advance to perform seamless human-machine interaction by accurately anticipating the human actions [19, 30, 32].

Recently, due to its increasing significance, this fascinating topic has been extensively investigated by various emerging technologies [16, 33, 23, 58, 2]. Researchers typically regard it as a sequence-to-sequence (seq2seq) generation task and then resort to RNNs to speculate the next plausible human movement from the historical observation [26, 21]. Current approaches have attempted to exploit GCNs to effectively access the topological relationship of 3D human skeleton for predicting future human motion [41, 13, 34]. These solutions fully analyze the temporal and spatial correlation of human motion sequences.

Although encouraging progress has been achieved, from the actual scene of human motion prediction, we suggest that the existing literature ignores an essential aspect, *i.e.*, the incompleteness of historical observations has not been considered. Stated in a different way, state-of-the-art approaches [21, 34, 10, 13] are over-sensitive to the missing items of the observed data that are very common in real-world scenarios [16, 46, 22, 8]. For example, due to the mutual occlusion of joints or the occlusion of objects in the environment, the sensor measurement frequently involves missing values, as shown in Figure 1. Even for professional motion capture (MoCap) devices, the incompleteness of the

*Corresponding author

raw motion data is also inevitable [55, 12, 38]. Current predictive algorithms never consider the realistic scenario of incomplete historical observations, which may yield unexpected or even distorted predictions, leading to the failure of the human motion prediction task.

To investigate this new issue, we develop a novel multi-task graph convolutional network (**MT-GCN**), which simultaneously considers two supervised learning tasks, *i.e*., predicting human actions and repairing the incomplete observation. Specifically, MT-GCN mainly includes three modules, including a shared context encoder (SCE), a sequence repairing module (SRM), and a human action predictor (HAP). From temporal and spatial perspectives, the SCE resorts to the GCN [7, 28] and temporal convolutional networks (*i.e*., TCNs) [4] to extract the context code of 3D skeleton sequences. In back-propagation, this shared context is supervised by both HAP and SRM. For SRM, in addition to GCNs, it is also embedded with a temporal self-attention mechanism to select the most related information from the whole sequence to repair the corrupted pose [3, 57]. This strategy can also be regarded as an alternative to RNNs or TCNs to capture the temporal pattern. For HAP, we propose a multi-head graph attention network (GAT) to aggregate information from neighboring nodes, to bring a richer topological representation and stable training [54]. Besides, we design a non-autoregressive pipeline to generate each predicted frame independently, thus avoiding error propagation over the time dimension. Meanwhile, inspired by neural machine translation (NMT) [14, 48], position embedding is introduced into the HAP to ensure continuity of the predicted sequence. Finally, the above modules are jointly optimized in a unified framework to improve the prediction performance from the incomplete sequence.

The major contributions are threefold: (1) To best our knowledge, this is the first research that explicitly focuses on predicting human motion when the observed poses involve missing values; (2) We propose a multi-task learning framework to consider both tasks of repairing the corrupted observation and predicting future human actions; (3) On three large-scale benchmarks, our model achieves the state-of-the-art (SoTA) performance against the existing work.

2. Related Work

Human Motion Prediction. With the availability of large-scale MoCap datasets [25, 1, 50], typical methods resort to RNNs to treat human motion prediction as a seq2seq learning problem [19, 47, 22, 8, 9]. In [16], researchers first introduce RNNs to address the human motion prediction problem, in which two models are proposed, *i.e*., 3-layer long short-term memory (LSTM-3LR) and encoder-recurrent-decoder (ERD). Jain *et al*. [26] develop a structural RNN to consider the tree structure of human kinematics. However, these two methods frequently encounter a significant discontinuity between the first predicted frame and the last observed frame. Martinez *et al*. [42] alleviate this limitation with a residual single-layer GRU model. Ghosh *et al*. [20] construct two-level processing to help generate the planned motion trajectory. Liu *et al*. [37] introduce a hierarchical recursive method combined with a Lie algebra. In [10], the authors consider the influence of the environment on human action and then employ RNNs to predict future motions. Despite promising results, due to the unavoidable error accumulation, the variants of RNNs are prone to converge to an undesired mean pose.

Currently, state-of-the-art approaches utilize GCNs to predict future human movements [58, 10, 36, 40]. Mao *et al*. [41] first introduce an unconstrained graph to represent the human skeleton sequence. To explicitly leverage the topological relationship of human joints, Cui *et al*. [13] propose a dynamic GCN to consider the connections of both adjacent joints and geometrically separated ones. Li *et al*. [34] develop a multi-scale GCN model to comprehensively extract the rich connections of the human body.

All of the aforementioned methods formulate human motion prediction from a simple aspect, which is not applicable to actual situations where the observation involves missing values. Our work fills this gap.

Motion Sequence Repairing. Researchers have attempted to repair the missing information in motion sequences based upon sparse representation [56, 15] or low-rank matrix completion [11, 55]. Compared with the statistical approach, RNN variants are also proposed to solve this issue [12, 35, 24]. However, these methods are not designed for human motion prediction, and accordingly, are unsuitable for predicting actions from incomplete observations. In [46], the authors consider human action prediction from the perspective of motion repairing. Particularly, a mask matrix is utilized to occlude the latter frames of a motion sequence, and then repairing these missing frames is transformed into predicting future human poses. Unfortunately, they still fail to consider the problem that the observed sequence is corrupted by missing joints.

Presumably, a trivial strategy to address this new paradigm consists of two stages: repairing the missing values, and then predicting actions from this repaired sequence. Although it seems to be more straightforward to handle two single-task separately, it ignores the internal relations between these two related problems. As shown in our experimental section, compared with this alternative solution, the proposed multi-task learning framework achieves more realistic results.

3. Proposed Approach

3.1. Problem Definition and Notations

Let $\mathbb{X}_{-T+1:0} = [\mathbf{X}_{-T+1}, ..., \mathbf{X}_{-1}, \mathbf{X}_0] \in \mathbb{R}^{J \times T \times 3}$ be the complete observation of historical poses, where each $\mathbf{X}$ in-

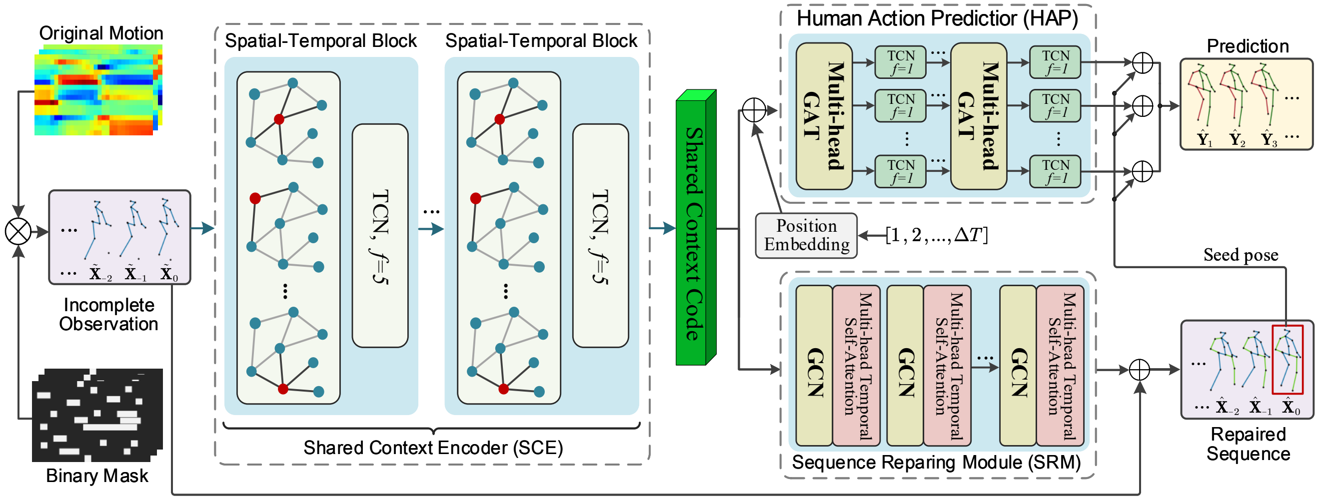

Figure 2. **Illustration of the proposed multi-task graph convolutional network (MT-GCN)**. It mainly consists of three modules. The first is the shared context encoder (SCE), which comprises multiple spatio-temporal blocks with residual connections for extracting a flexible context code of the input sequence. The second is the sequence repairing module (SRM), in which a temporal self-attention (TSA) is introduced to explicitly borrow information from the appropriate location to repair the missing value. The last is the human action predictor (HAP) that embeds a multi-head graph attention network (GAT) to effectively access the human skeleton and stabilize the training process. The aforementioned components are trained jointly to promote mutual cooperation and improve the final performance. $\oplus$ is addition operation, $\otimes$ indicates element-wise product, and $\Delta T = 25$. Finally, the overall model takes the incomplete observation $\tilde{\mathbb{X}} = \{\tilde{\mathbf{X}}_t\}_{t=-T+1}^{0}$ to produce the predicted poses $\hat{\mathbb{Y}} = \{\hat{\mathbf{Y}}_t\}_{t=1}^{\Delta T}$, and the auxiliary repaired sequence $\hat{\mathbb{X}} = \{\hat{\mathbf{X}}_t\}_{t=-T+1}^{0}$.

dicates the human pose represented by 3D coordinate with J joints. The actual future motion is formally expressed as $\mathbb{Y}_{1:\Delta T} = [\mathbf{Y}_1, ..., \mathbf{Y}_{\Delta T-1}, \mathbf{Y}_{\Delta T}] \in \mathbb{R}^{J\times\Delta T\times 3}$. Previous studies [13, 41, 5] are based upon the complete motion $\mathbb{X}_{-T+1:0}$ to learn an function $\mathcal{F} : \mathbb{X}_{-T+1:0} \to \hat{\mathbb{Y}}_{1:\Delta T}$ to make the prediction $\hat{\mathbb{Y}}_{1:\Delta T}$ as close as the ground truth $\mathbb{Y}_{1:\Delta T}$. These works ignore the situation of observations with missing values; hence, it may lead to the failure of the motion prediction task. We have noticed this limitation in the existing literature and are committed to solving it.

Suppose that $\mathbb{M} \in \{0, 1\}$ is a binary mask to set the missing/unobserved part to zero, $\otimes$ is the element-wise product. Our goal is, based on the incomplete observation $\tilde{\mathbb{X}}_{-T+1:0} = \mathbb{M} \otimes \mathbb{X}$, to train a unified mapping $\mathcal{F}$ to forecast the future human action $\hat{\mathbb{Y}}_{1:\Delta T}$, and incidentally, to obtain the repaired sequence $\hat{\mathbb{X}}_{-T+1:0}$:

$$\mathcal{F} : \tilde{\mathbb{X}}_{-T+1:0} \to \{\hat{\mathbb{X}}_{-T+1:0}, \hat{\mathbb{Y}}_{1:\Delta T}\}. \tag{1}$$

3.2. Multi-task Graph Convolutional Network

In this subsection, we illustrate the details of the MT-GCN from the following three components: Shared Context Encoder (SCE), Sequence Repairing Module (SRM), and Human Action Predictor (HAP), as shown in Figure 2.

3.2.1 Shared Context Encoder (SCE)

As a spatio-temporal time-series data, 3D skeleton sequence enjoys both spatial correlations of joints and temporal patterns among poses. Therefore, to extract a shared representation, we construct the SCE by stacking multiple spatio-temporal blocks composed of GCNs and TCNs.

Let the bones between adjacent joints be edges, and we represent human body as an undirected graph, *i.e.*, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is vertex/joint set and $\mathcal{E} = \{e_{ij} | i, j \in 1, 2, ..., J\}$ is edge set. Each skeletal pose can be formulated as an adjacency matrix $\boldsymbol{A}$, where $\boldsymbol{A}_{ij} = 1$ if and only if i-th and j-th joints are connected (each joint connects with itself). Given the diagonal degree matrix $\boldsymbol{D}$ and the identity matrix $\boldsymbol{I}$, the following formula is used to extract the spatial relation of the human skeleton sequence:

$$\boldsymbol{H}^{l+1} = \sigma\left(\tilde{\boldsymbol{A}}\boldsymbol{H}^l\boldsymbol{W}^l\right), \tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}}, \tag{2}$$

where $\boldsymbol{W}^l \in \mathbb{R}^{C_{in}\times C_{out}}$ is the learnable weight, and σ is the *Mish* function [43]. $\boldsymbol{H}^l \in \mathbb{R}^{J\times C_{in}}$, $\boldsymbol{H}^{l+1} \in \mathbb{R}^{J\times C_{out}}$ are the input feature and the updated state at l-th layer, respectively. C_{in}, C_{out} are the channel number.

The latest studies show that TCN has an efficient expression for modeling time-series data [44, 4, 45, 18]. Following these progresses, the TCN (with same padding) is used to capture the temporal pattern of motion sequences.

Then, the SCE is composed of 7 GCN-TCN blocks with the channel numbers 64, 64, 128, 128, 256, 256, 512. Finally, the input 3-d joint is mapped to a 512-d shared context representation $SCE(\tilde{\mathbb{X}})$ for two downstream modules.

3.2.2 Sequence Repairing Module (SRM)

Usually, human poses are potentially related, even similar to, throughout the sequence. If the network is capable of leveraging the relevant context with the corrupted pose, it is of great benefit to repair the missing information. To this end, for SRM, we design a multi-head temporal self-attention (TSA) strategy to integrate heterogeneous contributions of different frames. In addition, TSA can also extract the temporal correlation without extra operations.

Let $\mathbf{h}^v = (\boldsymbol{h}_1^v, \boldsymbol{h}_2^v, ..., \boldsymbol{h}_T^v) \in \mathbb{R}^{T \times C_{in}}$ be the input feature of TSA with respect to the v-th spatial dimension, $\boldsymbol{h}_i^v \in \mathbb{R}^{C_{in}}$, for each of $v \in \mathcal{V}$. The result of TSA is a sequence $\mathbf{h}'^v = (\boldsymbol{h}_1'^v, \boldsymbol{h}_2'^v, ..., \boldsymbol{h}_T'^v)$ weighted according to its relevance in the sequence, where $\boldsymbol{h}_i'^v \in \mathbb{R}^{C_{out}}$ is associated with i-th frame. For each $\boldsymbol{h}_i^v$, we first use 3 learnable linear transformations to produce 3 different vectors: a query $\boldsymbol{q}_i^v \in R^{d_q}$, a key $\boldsymbol{k}_i^v \in R^{d_k}$ and a value $\boldsymbol{v}_i^v \in R^{d_v}$. Then, we use a dot product to obtain a weight for each pair $(\boldsymbol{h}_i^v, \boldsymbol{h}_j^v)$:

$$\alpha_{ij}^v = \boldsymbol{q}_i^v \cdot \boldsymbol{k}_j^v / \sqrt{d_k}, \; \forall v \in \mathcal{V}, \tag{3}$$

where $d_q = d_k = d_v = 64$. The score α_{ij}^v indicates how much the node v of j-th frame is relevant for the one of i-th frame. Then, the $\boldsymbol{h}_i'^v \in R^{d_v}$ can be obtained:

$$\boldsymbol{h}_i'^v = \sum\nolimits_j \text{softmax}\left(\alpha_{ij}^v\right) \boldsymbol{v}_j^v. \tag{4}$$

Similar to the vanilla Transformer [48], we use K independent TSA and then concat their output to enhance the representation. For each time step, we repeat the above operation to produce the attentive context $\mathbf{h}'^v = (\boldsymbol{h}_1'^v, \boldsymbol{h}_2'^v, ..., \boldsymbol{h}_T'^v) \in \mathbb{R}^{T \times C_{out}}$ of the node v, where the $C_{out} = K \cdot d_v$. Then, along the spatial dimension, the newly calculated feature is $\boldsymbol{H} = (\mathbf{h}'^1, \mathbf{h}'^2, ..., \mathbf{h}'^J) \in \mathbb{R}^{J \times T \times C_{out}}$ for the next layer.

The SRM involves 5 blocks, formed by a TSA and GCN layer, with same channel number $C_{out} = 512$, to help the SRM explicitly borrow information from related locations to effectively repair the missing value. With an additional linear layer, it decodes the shared code $SCE(\tilde{\mathbb{X}})$ into the original dimension. Finally, retaining the non-missing parts in the observation, the repaired sequence $\hat{\mathbb{X}}_{-T+1:0} = [\hat{\mathbf{X}}_{-T+1}, ..., \hat{\mathbf{X}}_{-1}, \hat{\mathbf{X}}_0]$ is obtained:

$$\hat{\mathbb{X}} = (1 - \mathbb{M}) \otimes SRM\big(SCE(\tilde{\mathbb{X}})\big) + \mathbb{M} \otimes \tilde{\mathbb{X}}. \tag{5}$$

3.2.3 Human Action Predictor (HAP)

As the main task of our multi-task learning framework, the HAP mainly involves three components: Multi-head Graph Attention Network (GAT), TCN, and Position Embedding.

Intuitively, all neighbors of joint v contribute unequally to its motion pattern. For example, during running, the movement of elbow joint is more driven by shoulder joint rather than wrist joint. To model this, we develop the GAT to explicitly consider the importance of the neighbors. Following the previous studies [49, 52, 59], with the hidden state of $\mathbf{h}^t = (\boldsymbol{h}_1^t, \boldsymbol{h}_2^t, ..., \boldsymbol{h}_J^t) \in \mathbb{R}^{J \times C_{in}}$, $\boldsymbol{h}_i^t \in \mathbb{R}^{C_{in}}$, for each $t \in \{1, 2, ..., T\}$, a single GAT layer can be defined as:

$$\beta_{ij}^t = \frac{\exp\left(\text{LReLU}\left(\boldsymbol{a}^T\left[\boldsymbol{W}\boldsymbol{h}_i^t, \boldsymbol{W}\boldsymbol{h}_j^t\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LReLU}\left(\boldsymbol{a}^T\left[\boldsymbol{W}\boldsymbol{h}_i^t, \boldsymbol{W}\boldsymbol{h}_k^t\right]\right)\right)}, \tag{6}$$

where β_{ij}^t is the attentive score of the vertex pair $(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t)$. $\mathcal{N}_i$ is the neighbors of i-th node in the graph, and $[\,,\,]$ represents a concatenation. $\boldsymbol{W} \in \mathbb{R}^{C_{out} \times C_{in}}$ and $\boldsymbol{a} \in \mathbb{R}^{2C_{out}}$ indicate the weight matrix of a linear transformation and a single-layer fully-connected network, respectively. LReLU ($\alpha = 0.2$) is the nonlinear activation.

The output of multi-head GAT for each node is obtained using the average computation of K independent GATs:

$$\mathbf{h}_i'^t = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in \mathcal{N}_i} \beta_{ij}^{tk} \boldsymbol{W}^k \boldsymbol{h}_i^t\right), \forall t \in \{1, ..., T\}. \tag{7}$$

Similarly, for each node, we repeat the GAT computation to obtain the output state $\mathbf{h}'^t = (\boldsymbol{h}_1'^t, \boldsymbol{h}_2'^t, ..., \boldsymbol{h}_J'^t) \in \mathbb{R}^{J \times C_{out}}$, with $\boldsymbol{h}_i'^t \in \mathbb{R}^{C_{out}}$ being the t-th vector in the sequence. σ is the $Mish$ function [43]. Then we apply it to each temporal dimension to produce the final result $\boldsymbol{H} = (\mathbf{h}'^1, \mathbf{h}'^2, ..., \mathbf{h}'^T) \in \mathbb{R}^{J \times T \times C_{out}}$.

Typically, RNN-based models are based on previous predicted poses to forecast the next frame [42, 46, 23]. This autoregressive pipeline inevitably leads to the problem of error accumulation, even the convergence to the mean pose. To break through it, inspired by [53, 32], a TCN with filter size f=1 is used to forecast each frame independently. Our strategy bypasses the influence of previous frames on the current prediction, thus alleviating error accumulation.

One drawback of the above non-autoregressive scheme is that it cannot encode the temporal continuity of successive poses. To solve this problem, following the current progress in NMT [14, 48], we use position embedding to map each scalar index t to a vector in a supervised way, and then inject it into each time step of the input features of HAP. Considering two indexes t_1, and t_2, the closer they are, the more similar the positional vectors are, and vice versa. In this way, our non-autoregressive HAP clearly distinguishes the input context at different positions, thus explicitly ensuring the temporal continuity and the ordinal relation of the generated sequence. Then, each predicted frame $\hat{\mathbf{Y}}_t$ is independently computed as:

$$\hat{\mathbf{Y}}_t = \hat{\mathbf{X}}_0 + HAP\big(P(t), SCE(\tilde{\mathbb{X}})\big), \tag{8}$$

where $\hat{\mathbf{X}}_0$ is the last frame (seed pose) of the repaired sequence. P is the position embedding that transforms each index t into a vector. $SCE(\tilde{\mathbb{X}})$ is the shared code from the SCE. Finally, the HAP generates the smooth prediction $\hat{\mathbb{Y}}_{1:\Delta T} = [\hat{\mathbf{Y}}_1, ..., \hat{\mathbf{Y}}_{\Delta T-1}, \hat{\mathbf{Y}}_{\Delta T}]$ in parallel, in which each predicted frame is not affected by previous ones.

3.3. Training

Following previous work [34, 41, 12, 10], the model is trained to minimize L_2 distance and the bone length error.

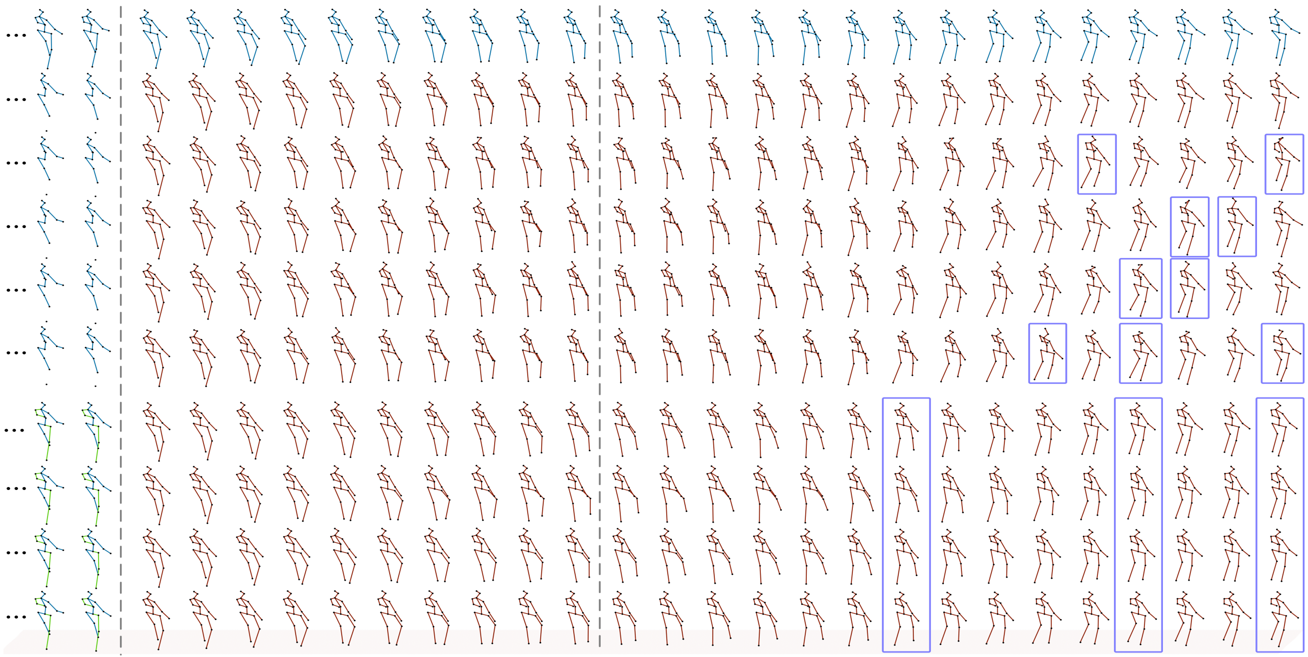

Figure 3. **Qualitative Comparison**. From top to bottom: Ground Truth (GT); and MT-GCN, STMIGAN [46], TrajGCN [41], LDRGCN [13], DMGNN [34], directly generated from incomplete observations; as well as R+STMIGAN [46], R+TrajGCN [41], R+LDRGCN [13], R+DMGNN [34], based upon the repaired sequence. As highlighted in the rectangle, patently unreasonable or abnormal predictions are exhibited. We observe that, even when the baselines are based on the repaired sequence, the proposed model still outperforms them.

The final objective function is then expressed as:

$$\begin{aligned}\mathcal{L} =&\lambda_P||\mathbb{Y}_{1:\Delta T} - \hat{\mathbb{Y}}_{1:\Delta T}||_2 + \lambda_{B_1}\mathcal{L}_B(\mathbb{Y}, \hat{\mathbb{Y}}) \\ &+ \lambda_R||\mathbb{X}_{-T+1:0} - \hat{\mathbb{X}}_{-T+1:0}\,||_2 + \lambda_{B_2}\mathcal{L}_B(\mathbb{X}, \hat{\mathbb{X}}),\end{aligned} \tag{9}$$

where the $\hat{\mathbb{X}}$ and $\hat{\mathbb{Y}}$ denote the repaired sequence and the prediction respectively, $\mathbb{X}$ and $\mathbb{Y}$ are the corresponding GT. The function of $\mathcal{L}_B$ is used to calculate the bone length difference of two motion sequences [13, 12]. In all experiments, we set $\lambda_P = 1$, $\lambda_{B_1} = 0.04$, $\lambda_R = 0.5$, $\lambda_{B_2} = 0.015$. Such a hyper-parameter setting brings several significant advantages: (1) Balancing the scale of each loss term; (2) Distinguishing the importance of two tasks; (3) Ensuring that HAP and SRM converge synchronously as much as possible to stabilize the training process.

3.4. Implementation Details

In our work, the 3D position-based sequence is used as the input and output. Compared with the action-specific model, we consider training the proposed MT-GCN under all action categories to achieve a general model.

As shown in Figure 2, our model is mainly composed of three modules: SCE, SRM, and HAP. The SCE is stacked with 7 residual spatial-temporal blocks, each of which is formed by a GCN and a TCN layer, with the channel number of 64, 64, 128, 128, 256, 256, 512. The filter size of TCNs is $f = 5$. The SRM and HAP contain 5 blocks with channel number 512, each of which follows an additional linear layer to map the output into the original dimension. In the SRM, the block is formed by a GCN and a multi-head TSA, while in the HAP, it is formed by a multi-head GAT and a TCN with a filter size $f = 1$. The head number of multi-head TSA and multi-head GAT is $K = 8$. In addition, for SRM, we use the skip-connection to connect the incomplete input and the repaired sequence, while for HAP, each predicted pose is added to the last repaired frame (seed pose) $\hat{\mathbf{X}}_0$. Then, a $Mish$ function is used as the activation [43]. The length of input and output is equal ($T = \Delta T = 25$). The position embedding module takes each index t as the input and returns its 512-d embedding from a learnable lookup table [17].

Throughout the model, each layer is followed by batch normalization, with dropout rate of 0.3. The mini-batch size is 64. We use Adam [27] to train the network, where the initial learning rate is 0.01, with a 0.98 decay every 2 epoch.

4. Experiments

4.1. Preliminaries

Dataset-1: H3.6M [25] is the largest benchmark for human action prediction, which involves 15 activity categories performed by 7 professional actors. Following the previous literature [34, 13, 33], the constant joints are removed so that each pose contained 17 joints (J=17). Then, all sequences are down-sampled to a frame rate of 25 frames per second (fps). Finally, the activities of subject-5 (S5) are used as the testing set, the S11 is the validation, and the remaining is the training samples.

	Walking					Eating					Smoking					Discussion					Directions				
Milliseconds (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
MT-GCN (Ours)	**11.5**	**18.8**	**34.1**	**41.7**	**60.4**	**9.1**	**17.2**	**35.2**	**42.3**	**74.9**	**9.0**	**15.7**	30.2	**39.7**	**70.8**	**10.8**	**22.7**	**53.3**	64.6	**115.7**	**8.4**	**23.2**	**42.7**	**56.7**	**108.5**
Residual sup. [42]	32.5	50.8	72.2	85.4	112.3	26.1	41.3	69.4	87.1	131.7	27.8	47.2	70.5	93.6	137.8	35.2	64.3	103.0	115.3	158.9	30.5	55.2	91.2	113.1	151.4
ConvSeqSeq [33]	26.5	46.2	67.1	77.4	108.2	20.7	34.8	64.3	82.7	113.4	18.7	45.3	65.6	87.4	107.7	26.2	49.4	87.5	107.1	150.1	20.3	41.5	72.4	87.7	142.6
TrajGCN [41]	17.5	29.0	49.8	57.7	102.4	14.1	26.6	48.7	61.5	127.1	16.0	24.3	50.2	76.4	98.4	17.2	34.1	78.4	88.3	138.8	12.5	33.6	60.3	78.0	129.1
LDRGCN [13]	15.3	27.7	50.5	56.6	97.0	13.2	23.4	53.7	60.3	117.8	13.4	20.8	51.9	72.1	97.2	15.6	30.2	69.1	87.4	131.8	14.2	32.5	57.4	74.5	125.7
DMGNN [34]	14.8	27.3	48.0	55.4	90.9	14.1	23.6	52.5	59.1	115.1	13.5	20.7	46.2	64.8	94.3	15.0	29.4	68.7	86.5	129.6	15.1	31.0	58.4	73.2	124.4
STMIGAN [46]	16.3	39.5	55.7	64.5	94.3	17.2	36.6	81.1	93.1	101.8	16.3	37.5	52.0	61.1	100.5	23.1	47.6	86.9	97.6	147.9	22.4	47.3	70.2	79.2	131.2
R+Residual sup. [42]	25.6	43.5	68.3	73.2	92.1	20.2	37.2	65.6	82.3	114.4	23.2	29.3	63.1	85.6	118.3	30.4	53.1	92.3	105.6	145.5	27.7	50.2	83.1	95.6	138.3
R+ConvSeqSeq [33]	20.2	37.3	61.2	68.3	88.5	15.6	29.4	54.4	70.3	96.7	13.1	35.2	56.4	69.8	89.7	20.2	42.3	73.5	87.2	133.8	18.4	36.3	65.6	80.9	121.6
R+TrajGCN [41]	13.4	25.1	43.4	48.1	67.3	9.2	19.3	38.1	46.3	83.6	10.3	20.3	38.2	51.6	80.2	14.5	27.9	59.2	69.5	120.1	11.4	24.4	50.6	65.8	119.3
R+LDRGCN[13]	12.4	22.2	42.1	46.6	65.3	10.1	19.4	41.8	44.2	81.6	11.2	17.2	35.9	48.3	77.4	13.9	24.4	56.5	65.7	117.1	10.9	23.2	45.4	61.7	117.4
R+DMGNN [34]	12.7	20.3	38.6	47.2	64.2	11.3	18.2	40.6	43.8	77.5	11.6	17.0	**34.4**	45.1	79.7	12.0	23.7	54.8	**64.4**	117.9	11.3	23.5	46.4	59.4	115.8
R+STMIGAN [46]	15.4	32.3	45.6	57.3	78.3	14.3	27.8	53.0	71.1	89.2	15.2	22.5	43.9	56.2	85.5	21.3	34.7	85.3	92.5	133.6	18.8	33.4	56.1	81.3	124.1

Table 1. **Comparisons of 3D error on five representative activities from H3.6M dataset.** The upper is the numerical result that is directly generated from the incomplete observation with missing values. In the lower part, the prefix 'R' means that the results are obtained from the repaired sequence. Note that for our MT-GCN, we only consider the challenging but practical solution of predicting human motion from the raw observation with missing information. The best result is highlighted in bold, and the second is underlined.

Dataset-2: We also report our experimental results on **CMU MoCap** [1]. Consistent with the previous work [21, 41, 42], the selected samples contain eight actions, with a total of about 86k poses. We use a similar test/training partition strategy as they published code. Notably, due to data limitations, the validation set is unavailable. Other preprocessing solutions are the same as the H3.6M dataset.

Dataset-3: 3DPW MoCap [50] is recently released human action analysis dataset. It involves more than 51k indoor or outdoor frames. For a fair comparison, we use the official training, testing and validation sets. A pose is represented as the 17-joint skeleton. Compared with the H3.6M and CMU MoCap, the frame rate of the 3DPW dataset is 30fps. Therefore, the input observation involves 30 frames, *i.e.*, $\hat{\mathbb{X}} \in \mathbb{R}^{J\times 30\times 3}$. Other configurations are consistent with those of the H3.6M and CMU MoCap.

Baselines. We compare the our MT-GCN with 5 representative approaches, *i.e.*, a RNN-based (Residual sup. [42]), a CNN-based (ConvSeq2Seq [33]), three GCN-based (TrajGCN [41], LDRGCN [13], DMGNN [34]), as well as STMIGAN [46]. For an unbiased comparison, the baseline models are retrained under incomplete observations, and the other experimental settings are consistent with their papers.

Evaluation Metric. We first animate the predicted pose for qualitative comparison. Then, following the previous work [41, 13, 33], we also provide 3D errors using Mean Per Joint Position Error (MPJPE) [25] in millimeter (mm).

4.2. Result Analysis

Qualitative comparison. We visualize the character animation of each predicted pose on H3.6M dataset. For the competing methods, we utilize two different solutions: *First*, directly forecast future actions from incomplete observations (40% of the length of the left arm and right leg joint is invisible); *Second*, repair first using [12], and then generate the prediction based upon the repaired sequence; while for our MT-GCN, we only consider the former challenging but more practical solution. The generated results are shown in Figure 3, in which the vertical dashed lines separate the observation, the short-term prediction (400 ms) and the long-term prediction (1000 ms). We observe that once the observation involves missing values, the TrajGCN, LDRGCN, and DMGNN yield distorted results. We also simply modify the STMIGAN to accommodate the problem of predicting actions from incomplete observation. Although it has achieved specious visualization, with the increasing of the predictive horizon, the results are significantly different from GT. Moreover, the long-term prediction tends to converge to the mean pose. We suggest that a possible reason is that STMIGAN inevitably leads to the error accumulation. However, our MT-GCN explicitly considers the missing value in the observation, thus achieving remarkable improvements. Even if the baseline methods are based on the repaired sequence, they only achieve a slight progress. In contrast, our model directly infers from incomplete observations and obtains more accurate predictions that are almost indistinguishable from the GT.

Quantitative comparison. Table 1 shows the 3D error on five representative activities from the H3.6M dataset, which is evaluated directly from incomplete observations or the repaired sequence, respectively. The construction of incomplete observation is the same as the previous qualitative comparison part. Notably, the prefix 'R' means that the result is obtained from the repaired sequence. We observe that due to error accumulation, Residual sup. gradually obtains higher errors with the predicted range. ConvSeq2Seq is difficult to extract the structural relation, thus only achieving a lower accuracy. The GCN-based methods generate the sub-par result because they efficiently extract the spatio-temporal relationship of 3D skeleton sequences. Compared with these baselines, STMIGAN achieves better results under the incomplete observation because it solves the problem of human motion prediction from the perspective of repairing missing frames. In addition, the competitors usually produce a slightly better performance on the repaired sequence than on the incomplete observation. How-

	Greeting					Phoning					Posing					Purchase					Sitting				
Millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
MT-GCN (Ours)	**13.9**	**28.9**	**65.2**	**78.6**	**138.6**	**8.8**	**17.4**	**37.3**	**47.1**	**103.4**	**9.6**	**20.1**	**54.0**	**77.5**	**154.3**	**13.4**	**28.9**	67.9	**75.1**	**136.8**	**9.8**	**21.1**	48.5	**54.4**	**116.9**
R+TrajGCN [41]	18.2	39.4	75.1	92.8	145.0	11.1	23.3	45.4	57.7	111.3	14.7	31.5	64.5	89.7	173.8	18.5	39.5	70.3	86.1	152.2	11.2	26.5	54.1	67.4	124.3
R+LDRGCN[13]	16.2	33.3	72.7	86.6	143.1	10.4	22.6	41.4	52.1	110.5	12.5	26.3	62.1	89.2	169.5	17.1	38.0	**67.6**	80.4	149.9	12.5	28.7	50.5	61.1	120.8
R+DMGNN [34]	15.3	30.7	71.3	85.0	142.1	11.2	21.8	39.8	51.3	114.5	11.7	28.9	59.8	85.8	164.1	18.3	39.4	68.8	79.0	147.4	12.6	30.3	**47.7**	59.7	121.7
R+STMIGAN [34]	16.3	38.1	77.8	94.0	151.2	12.7	25.3	44.2	59.1	139.9	13.3	30.2	70.3	93.5	176.4	20.1	43.5	74.2	84.3	149.3	13.6	30.9	56.7	66.4	131.8

	Sitting down					Taking photo					Waiting					Walking dog					Walking together				
Millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
MT-GCN (Ours)	**13.7**	**29.6**	**57.3**	**70.8**	**140.2**	**8.9**	**18.0**	**39.5**	**48.1**	**117.3**	**8.4**	**18.7**	**42.9**	**52.7**	**108.8**	**21.1**	**41.1**	**77.0**	**90.6**	**145.6**	**8.4**	**21.3**	**33.7**	**44.2**	**69.0**
R+TrajGCN [41]	17.9	36.3	61.6	78.5	153.1	10.5	22.4	49.4	60.1	128.7	10.9	23.0	53.1	72.8	120.5	23.5	46.4	83.6	97.3	160.5	11.2	22.1	40.3	48.7	76.4
R+LDRGCN[13]	16.4	34.3	60.2	76.1	149.0	10.2	23.3	47.7	57.7	127.7	10.1	21.7	51.4	59.3	117.9	24.6	45.8	80.4	95.1	157.6	11.4	20.6	38.5	47.1	75.2
R+DMGNN [34]	15.5	33.4	59.2	76.7	147.9	9.8	24.9	46.7	55.9	124.3	10.5	21.5	50.3	54.5	115.4	25.0	43.9	79.2	94.6	156.7	10.3	22.1	35.7	45.2	74.7
R+STMIGAN [34]	16.8	36.7	60.2	83.4	158.5	11.3	25.4	49.6	65.0	137.1	12.9	27.0	57.7	70.1	130.9	30.2	58.4	92.5	109.7	176.3	13.6	21.4	43.1	52.5	80.2

Table 2. 3D error comparisons on the remaining 10 actions of H3.6M dataset. The results of our MT-GCN are directly from the incomplete observation, while others are generated from the repaired sequence.

ever, the price of this improvement is the addition of an additional sequence repairing procedure. Our MT-GCN generates superior results only based on the incomplete observation, which is more practical and efficient. The numerical results on the remaining 10 activities are shown in Table 2, and the conclusions are consistent with the above.

Different scenarios of missing value. We report the predicted 3D error (1000 ms) on different types of missing values on H3.6M dataset. Except for the data missing scenario, other experimental configurations are the same as before. As shown in Table 3, under various types of missing values, the results of our MT-GCN are reliable.

Scenario	Joint Random Missing	Structured Missing	Random Missing
MT-GCN (Ours)	**110.7**	**115.2**	**112.3**
TrajGCN [41]	144.3	163.2	139.1
LDRGCN [13]	135.7	149.3	139.3
DMGCN [34]	133.0	146.3	138.2
STMIGAN [46]	138.1	144.6	135.9
R+TrajGCN [41]	123.6	131.5	122.6
R+LDRGCN [13]	120.7	127.1	120.4
R+DMGCN [34]	117.1	126.5	118.5
R+STMIGAN [46]	132.6	128.2	135.2

Table 3. **Predicted 3D errors on different types of missing values. Joint Random Missing:** 40% of the right leg is randomly missing. **Structured Missing:** 40% of the length of the right leg joint is continuously missing. **Random Missing:** 30% of the random entries in the whole sequence is missing.

Results on CMU and 3DPW MoCap. As with the evaluation of H3.6M, we also investigate the quantitative 3D error on the CMU and 3DPW datasets. For baseline methods, we predict human motion based on two different historical sequences, including a.) raw incomplete observation; b.) the repaired observation after filling the missing value using the model in [12]. As shown in Table 4 and Table 5, our MT-GCN still achieves better performance than all competitors, which coincides with the conclusion on H3.6M dataset.

Limit testing. Because the person is occluded by a pillar, the whole pose may be invisible to the sensor for a period. To simulate this challenging situation, we remove continuous frames with different lengths from the observation to evaluate the 3D error (1000 ms). From Table 6, with the increase of the missing number, our model still yields reliable results, which evidences our superiority again.

4.3. Repairing Missing Values

We select several sequence repairing algorithms to verify our model in repairing missing values. From Table 7, we observe that MT-GCN achieves higher accuracy in terms of filling the missing value. We suggest that, with two supervised tasks, the SRM additionally utilizes a complete knowledge from the future poses to repair missing values, thus achieving a better repaired results. This also reflects from the side why our method is capable of generating high-fidelity predictions from incomplete observations.

4.4. Robustness to Noise

The captured motion data are often damaged by noise [29, 24, 39]; however, the existing work seldom considers it. We add Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the observed data and then randomly remove 50% of the leg joints. Then, based on this severely corrupted observation, the different methods are evaluated. As shown in Table 8, our model performs better than those non-multitask learning frameworks.

5. Ablation Studies

Here, we analyze the effect of several essential components on predictive performance on H3.6M dataset.

We first investigate the impact of **(1) different definitions of human skeleton**, including a.) the undirected graph in this work, b.) the directed graph from parent joint to child joint, c.) the reverse graph from child joint to parent joint, as well as d.) an unconstrained adjacency matrix to adaptively learn the topological relation. From Table 9, we observe that the undirected graph shows better performance, which implies that, for predicting human actions from incomplete observations, it is necessary to consider both the positive and reverse correlation of adjacent joints.

To verify the relevance of two branches (SRM, HAP), we separately analyze the results of sequence repairing and motion prediction when **(2) one of them is reserved**. From Table 10, when considering these two branches jointly, it achieves better results than the single one. This evidences

	Basketball					Basketball signal					Drecting traffic					Jumping				
Millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
MT-GCN (Ours)	**13.2**	**22.1**	**41.6**	**54.2**	**102.3**	4.0	**6.3**	**12.0**	**14.3**	**50.5**	**6.6**	17.1	**26.1**	**37.7**	**137.5**	**13.1**	**30.9**	67.2	**90.7**	**150.6**
R+TrajGCN [41]	15.3	27.3	51.7	63.1	112.4	**3.9**	7.1	13.4	17.8	59.6	7.2	**17.0**	33.4	41.9	153.7	17.4	33.5	67.8	93.5	166.2
R+LDRGCN[13]	14.3	26.5	48.7	60.7	111.1	4.2	7.9	13.2	16.7	54.8	7.0	17.5	30.5	39.4	150.4	16.8	32.7	66.6	92.1	158.5
R+DMGNN [34]	15.1	24.9	50.1	57.4	108.2	4.4	7.4	12.9	16.2	52.5	7.1	17.9	29.5	39.5	147.2	15.6	32.6	**65.7**	93.2	157.7
	Running					Soccer					Walking					Wash window				
Millisecond (ms)	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000	80	160	320	400	1000
MT-GCN (Ours)	**16.4**	**19.8**	**24.3**	**34.2**	**49.1**	**11.3**	**20.9**	**41.1**	**50.5**	**94.9**	**7.5**	**10.8**	**17.3**	**20.3**	**37.6**	**6.0**	**12.5**	**26.5**	**38.4**	**56.7**
R+TrajGCN [41]	19.4	27.8	31.5	40.1	60.1	14.4	26.4	47.8	62.2	101.2	9.8	15.7	27.1	28.3	40.6	7.4	15.5	30.3	48.7	65.3
R+LDRGCN[13]	18.7	25.4	28.0	38.4	52.8	13.8	24.2	43.6	54.9	99.0	8.9	13.0	23.3	24.7	39.4	8.2	13.3	29.1	42.1	63.9
R+DMGNN [34]	17.4	23.6	27.0	37.1	54.5	13.3	23.4	43.7	53.4	102.6	8.2	12.1	23.7	26.1	39.0	8.3	13.8	28.2	42.6	61.2

Table 4. **Comparisons of 3D error on 8 activities of CMU MoCap dataset.** Our model is evaluated on incomplete observation, while the baselines are based on the repaired sequence; even so, the proposed MT-GCN also achieves better results.

Millisecond (ms)	200	400	600	800	1000
MT-GCN (Ours)	**39.3**	**60.8**	**90.3**	**106.1**	**123.5**
R+TrajGCN [41]	47.1	73.0	103.7	126.6	146.7
R+LDRGCN [13]	42.3	65.5	97.4	115.6	134.2

Table 5. **Mean 3D error on whole testing set of 3DPW dataset.**

Number of Missing Frames	1	3	5	7
MT-GCN (Ours)	109.4	111.9	122.0	139.1

Table 6. **Limit Testing** for missing frames of different numbers.

Missing Time (ms)	80	160	320	400	480	560
MT-GCN (Ours)	**9.2**	**12.3**	**14.5**	**19.4**	**21.9**	**26.6**
BAN [12]	10.2	14.6	18.7	22.7	26.1	30.3
NonLinear MC [55]	15.3	20.3	25.1	33.2	37.3	42.7
STMIGAN [46]	12.2	13.5	17.2	22.4	25.5	31.4

Table 7. **Sequence Repairing Results,** which evaluates the L2 distance between the repaired observation and the real one when both left arm and right leg with different lengths are missing.

Models	MT-GCN	R+TrajGCN	R+LDRGCN	R+DMGNN
$\sigma = 25mm$	**114.3**	127.1	126.4	124.4
$\sigma = 50mm$	**119.7**	135.0	133.6	132.7

Table 8. **Robustness to noise.** Predicted 3D error (1000 ms) when the incomplete observation is attached to a Gaussian noise.

Graph Type	**Undirected**	Directed	Reverse	Unconstrained
MT-GCN (Ours)	**112.0**	126.4	131.2	117.5

Table 9. **Top:** Effects of various definitions of human body; **Bottom:** Effects of the number of heads in multi-head GAT. The results show the predicted 3D error of 1000 ms on H3.6M dataset.

		Sequence repairing				Motion prediction			
SRM	*HAP*	10%	20%	30%	40%	10%	20%	30%	40%
✓	×	9.1	14.4	20.6	26.9	-	-	-	-
×	✓	-	-	-	-	110.2	117.3	121.4	126.5
✓	✓	**8.6**	**13.7**	**18.7**	**24.5**	**109.4**	**110.5**	**112.3**	**114.4**

Table 10. **The repaired and predicted result** at 1000 ms with different random missing ratio, using SRM, HAP, or the both.

Seed Pose	80	160	320	400	1000
w/o	11.7	23.8	49.2	61.5	114.5
w/	**11.0**	**22.8**	**47.9**	**58.9**	**110.7**

Table 11. **Effects of the seed pose on each predicted pose**.

Filter Size	80	160	320	400	1000
3	11.4	24.7	50.6	62.6	114.6
5	11.1	**22.8**	**47.9**	**58.9**	**110.7**
7	**11.0**	23.1	49.5	60.6	115.7

Table 12. **Effects of different filter size of TCNs.**

that the human motion prediction and repairing incomplete observations are related tasks, and considering the both can improve their respective performance.

The last repaired frame $\hat{\mathbf{X}}_0$, as a seed pose, is added to each predicted frame. To verify its effectiveness, we investigate **(3) the impact of the seed pose** on the predictive performance. Besides, we also analyze the effect of **(4) different filter sizes of TCNs**. These results are shown in Table 11 and Table 12. We observe that the proposed components indeed facilitate the final generation.

Notably, Table 9, 11, and 12 are evaluated on the condition of 40% length of left arm and right leg are missing.

6. Conclusion

In this work, we explore a new problem, namely, predicting future accurate human motions from historically incomplete sequences. Moreover, we also propose a novel multi-task graph convolutional network (MT-GCN) to solve it. Our approach jointly considers two supervised tasks of repairing missing values in the observed sequence and predicting human actions, rather than dealing with them separately. Compared with traditional algorithms which produce unreasonable or even abnormal results under incomplete observations, the proposed model achieves higher-quality and more realistic predictions, even if the baseline methods are based on the repaired sequence. In addition, on several large-scale human motion benchmarks, our MT-GCN surpasses the state-of-the-art approaches in various scenarios of joint missing. Therefore, we reasonably conclude that the proposed model is more convenient for the practical application of human motion prediction.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NO. 61772272), in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX20_0283, and in part by the Project of Science and Technology of Jiangsu Province of China under Grant BE2017031.

References

[1] CMU Graphics Lab: Carnegie-Mellon Motion Capture (Mocap) Database, 2003.

[2] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A Stochastic Conditioning Scheme for Diverse Human Motion Prediction. In *CVPR*, pages 5223–5232, 2020.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *CoRR*, abs/1803.01271, 2018.

[5] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *CVPR*, pages 1418–1427, 2018.

[6] M. Brand and Aaron Hertzmann. Style machines. In *SIGGRAPH '00*, 2000.

[7] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[8] Judith Bütepage, Michael J. Black, Danica Kragic, and Hedvig Kjellström. Deep Representation Learning for Human Motion Prediction and Classification. In *CVPR*, pages 1591–1599, 2017.

[9] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-Agnostic Human Pose Forecasting. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432, 2019.

[10] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, pages 6992–7001, 2020.

[11] Qiongjie Cui, Beijia Chen, and Huaijiang Sun. Nonlocal Low-rank Regularization for Human Motion Recovery based on Similarity Analysis. *Information Sciences*, 493:57–74, 2019.

[12] Qiongjie Cui, Huaijiang Sun, Yupeng Li, and Yue Kong. A Deep Bi-directional Attention Network for Human Motion Recovery. In *IJCAI*, pages 701–707, 2019.

[13] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning Dynamic Relationships for 3D Human Motion Prediction. In *CVPR*, pages 6519–6527, 2020.

[14] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 2019.

[15] Yinfu Feng, Mingming Ji, Jun Xiao, Xiaosong Yang, Jian J Zhang, Yueting Zhuang, and Xuelong Li. Mining Spatial-temporal Patterns and Structural Sparsity for Human Motion Data Denoising. *IEEE transactions on cybernetics*, 45(12):2693–2706, 2014.

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent Network Models for Human Dynamics. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 4346–4354, 2015.

[17] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. Convolutional Sequence to Sequence Learning. In *ICML*, 2017.

[18] Yue Geng, Lingling Su, Yunhong Jia, and Ce Han. Seismic Events Prediction Using Deep Temporal Convolution Networks. volume 2019. Hindawi, 2019.

[19] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning Human Motion Models for Long-Term Predictions. *2017 International Conference on 3D Vision (3DV)*, pages 458–466, 2017.

[20] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, C. Lee Giles, and Alexander Ororbia. A Neural Temporal Model for Human Motion Prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial Geometry-Aware Human Motion Prediction. In *ECCV*, pages 786–803, 2018.

[22] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José M. F. Moura. Few-Shot Human Motion Prediction via Meta-learning. In *ECCV*, pages 432–450, 2018.

[23] Xiao Guo and Jongmoo Choi. Human Motion Prediction via Learning Local Structure Representations and Temporal Dependencies. In *AAAI*, pages 2580–2587, 2019.

[24] Daniel Holden. Robust solving of optical motion capture data by denoising. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018.

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36:1325–1339, 2014.

[26] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *CVPR*, pages 5308–5317, 2016.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[28] Thomas N Kipf and Max Welling. Semi-supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*, 2016.

[29] Taras Kucherenko, J. Beskow, and Hedvig Kjellstrom. A Neural Network Approach to Missing Marker Reconstruction in Human Motion Capture. *arXiv: Learning*, 2018.

[30] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. In *AAAI*, volume 33, pages 8553–8560, 2019.

[31] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. Efficient Nonlinear Markov Models for Human Motion. In *CVPR*, pages 1314–1321, 2014.

[32] Bin Li, Jian Tian, Zhongfei Zhang, Hailin Feng, and Xi Li. Multitask non-autoregressive model for human motion prediction. *IEEE Transactions on Image Processing*, 2020.

[33] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional Sequence to Sequence Model for Human Dynamics. In *CVPR*, pages 5226–5234, 2018.

[34] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *CVPR*, pages 214–223, 2020.

[35] Shujie Li, Yang Zhou, Haisheng Zhu, Wenjun Xie, Yang Zhao, and Xiaoping Liu. Bidirectional Recurrent Autoencoder for 3D Skeleton Motion Data Refinement. *Computers & Graphics*, 81:92–103, 2019.

[36] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning Lane Graph Representations for Motion Forecasting. In *ECCV*, pages 541–556, 2020.

[37] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Minghua Tang, Shijian Lu, Richard Zimmermann, and Li Chen Cheng. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *CVPR*, pages 10004–10012, 2019.

[38] Suhas Lohit, Rushil Anirudh, and Pavan Turaga. Recovering trajectories of unmarked joints in 3d human actions using latent space optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2342–2351, 2021.

[39] Utkarsh Mall, G Roshan Lal, Siddhartha Chaudhuri, and Parag Chaudhuri. A deep recurrent framework for cleaning motion capture data. *arXiv preprint arXiv:1712.03380*, 2017.

[40] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History Repeats Itself: Human Motion Prediction via Motion Attention. In *ECCV*, pages 474–489, 2020.

[41] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning Trajectory Dependencies for Human Motion Prediction. In *ICCV*, pages 9489–9497, 2019.

[42] Julieta Martinez, Michael J Black, and Javier Romero. On Human Motion Prediction using Recurrent Neural Networks. In *CVPR*, pages 2891–2900, 2017.

[43] Diganta Misra. Mish: A Self Regularized Non-Monotonic Neural Activation Function. In *BMVC*, 2020.

[44] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *SSW*, 2016.

[45] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. In *CVPR*, pages 7753–7762, 2019.

[46] Alejandro Hernandez Ruiz, Juergen Gall, and Francesc Moreno-Noguer. Human Motion Prediction via Spatio-Temporal Inpainting. In *CVPR*, pages 7134–7143, 2018.

[47] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamics. In *IJCAI*, 2018.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NeurIPS*, pages 5998–6008, 2017.

[49] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[50] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering Accurate 3D Human Pose in The Wild Using IMUS and a Moving Camera. In *ECCV*, pages 601–617, 2018.

[51] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian Process Dynamical Models for Human Motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:283–298, 2008.

[52] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The World Wide Web Conference*, pages 2022–2032, 2019.

[53] Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Non-autoregressive machine translation with auxiliary regularization. In *AAAI*, volume 33, pages 5377–5384, 2019.

[54] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[55] Guiyu Xia, Huaijiang Sun, Beijia Chen, Qingshan Liu, Lei Feng, Guoqing Zhang, and Renlong Hang. Nonlinear Low-Rank Matrix Completion for Human Motion Recovery. *IEEE Transactions on Image Processing*, 27:3011–3024, 2018.

[56] Jun Xiao, Yinfu Feng, and Wenyuan Hu. Predicting Missing Markers in Human Motion Capture Using L1-sparse Representation. *Computer Animation and Virtual Worlds*, 22(2-3):221–228, 2011.

[57] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[58] Ye Yuan and Kris Kitani. Dlow: Diversifying Latent Flows for Diverse Human Motion Prediction. In *ECCV*, pages 346–364, 2020.

[59] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.