Few-shot Human Motion Prediction via Learning Novel Motion Dynamics

Chuanqi Zang[1,2*], **Mingtao Pei**[1] and **Yu Kong**[2]
[1]Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, China
[2]Golisano College of Computing and Information Sciences, Rochester Institute of Technology, USA
jnchuanqizang@gmail.com, peimt@bit.edu.cn, yu.kong@rit.edu

Abstract

Human motion prediction is a task where we anticipate future motion based on past observation. Previous approaches rely on the access to large datasets of skeleton data, and thus are difficult to be generalized to novel motion dynamics with limited training data. In our work, we propose a novel approach named Motion Prediction Network (MoPredNet) for few-short human motion prediction. MoPredNet can be adapted to predicting new motion dynamics using limited data, and it elegantly captures long-term dependency in motion dynamics. Specifically, MoPredNet dynamically selects the most informative poses in the streaming motion data as masked poses. In addition, MoPredNet improves its encoding capability of motion dynamics by adaptively learning spatio-temporal structure from the observed poses and masked poses. We also propose to adapt MoPredNet to novel motion dynamics based on accumulated motion experiences and limited novel motion dynamics data. Experimental results show that our method achieves better performance over state-of-the-art methods in motion prediction.

1 Introduction

Human motion prediction is the task of forecasting future human pose based on the observed pose data. It is one of the hallmarks of human intelligence, which has a wide range of applications such as autonomous driving [Behl *et al.*, 2017], motion simulation [Vondrak *et al.*, 2008], and human-robot interaction [Koppula and Saxena, 2015]. Taking human-robot interaction as an example, robots are supposed to greet us immediately once we are raising our hand to greet them, to help us take the cup once we are walking to get it, and to stop once their next motion may hurt others.

In contrast to human intelligence which acquires such a prediction capability from just a few experiences, existing methods for human motion prediction [Fragkiadaki *et al.*, 2015; Martinez *et al.*, 2017; Liu *et al.*, 2019] still rely on extensive training data in order to be adapted to novel motion dynamics. This limits their applications in few-shot task domain. In addition, even though these methods continuously improve the performance in motion prediction on public datasets, their prediction capability is limited in long-term prediction [Martinez *et al.*, 2017] mainly because they do not explicitly learn sub-motion patterns and do not receive explicit guidance from past related motion. Here, sub-motions represent the decomposition of complex motion, e.g., raising, holding, and lowering hands are sub-motions of eating.

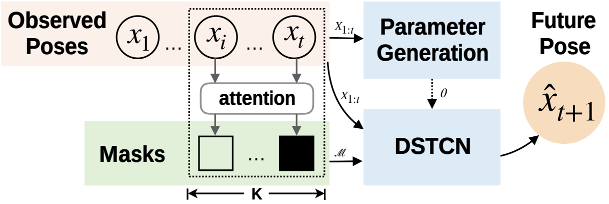

Figure 1: Overview of our few-shot human motion prediction architecture (MoPredNet). Future poses are predicted from observed poses. There are three modules in our architecture: DSTCN is used for motion generation. Sequence mask (of length K) is used to recall referable motion to get the explicit guidance for prediction. The parameter generation is used for quickly updating parameters of the DSTCN.

To deal with these issues, we propose a Motion Prediction Network (MoPredNet) that can dynamically select the most informative poses as prediction reference and adaptively learn spatio-temporal motion structure for long-term motion prediction. Besides, MoPredNet has the ability to predict new motion categories based on accumulated information and novel information from a few new samples. As shown in Fig. 1, our MoPredNet can be divided into three modules: Deformable Spatio-Temporal Convolution Network (DSTCN), a sequence mask module and a parameter generation module.

The DSTCN is proposed to adaptively model sub-motion dynamics and spatial correlation in the entire motion sequence to capture long-term dependency. In the spatial domain, human's skeletal joints are directly related to their parent joints but indirectly related to other joints (symmetry or regularity in motion). For traditional convolution, it is hard to model sub-motion patterns in a suitable starting and ending frames and flexibly capture spatial correlation based on mo-

[*] This work was completed while Chuanqi Zang was at Rochester Institute of Technology as a visiting student.

tion dynamics. The DSTCN can adaptively extend in the past time and establish selective connections with its neighboring joints. Besides, because the joint is embedded with high-dimensional representation compared to pixels, we present a local geometric extraction layer before the deformable spatio-temporal convolution layer.

The sequence mask module is used to recall a referable motion state from history, which provides explicit guidance for long-term motion prediction at specific moments. Many researchers extract the guidance information of motion dynamics from different cues, such as speed [Martinez *et al.*, 2017], incorporating derivative [Gopalakrishnan *et al.*, 2019], and smooth trajectory [Wei *et al.*, 2019]. However, a single physical property does not explicitly guide the architecture to generate the next moment pose as prediction errors accumulate. In this work, we use the attention mechanism to calculate the correlation scores between the past poses and the current pose, which are then used to pick key frames. This information combined with observed motion information helps the prediction module generate human-like motion.

The parameter generation module is inspired by the human learning process. When facing a new motion category with a few samples, we analyze and predict motion using accumulated information from an external memory at first. Then we will pay attention to novel motion dynamics to acquire new knowledge. In the parameter generation module, we store the basic knowledge learned from large-scale dataset in the memory to instruct the model to generate human-like motion. Besides, we present a learner that can learn novel motion dynamics and quickly update the parameter of the DSTCN. Therefore, the DSTCN can predict motion by a new context vector with unique characteristics for a specific category.

2 Related Works

2.1 Human Motion Prediction

In early work, hand-craft features [Pavlovic *et al.*, 2001; Wang *et al.*, 2006; Taylor *et al.*, 2007] are proposed to model the motion patterns. However, human motion incorporates far more complicated changing, which is hard to be modeled in a fixed, manually designed model.

With the collection and publication of large-scale human motion datasets in recent years, many methods based on the deep neural network are proposed and achieve promising results. Recurrent Neural Network (RNN) based methods [Fragkiadaki *et al.*, 2015; Martinez *et al.*, 2017; Liu *et al.*, 2019] are popular in motion prediction and encode temporal structure by a hidden state. [Li *et al.*, 2018] presents a temporal convolution seq2seq Network with GAN to generate human-like motion. Recently, [Wei *et al.*, 2019] proposes a feed-forward network based on graph convolutional networks and improves the performance of fixed length motion prediction within 1000 ms.

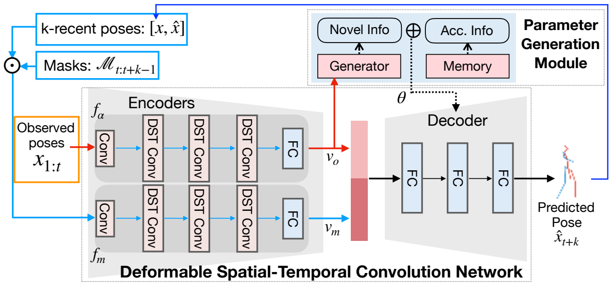

Figure 2: Pipeline of our framework. We take the generation of pose $\hat{x}_{t+k}$ as an example. $\hat{x}_{t+k}$ is predicted based on observed poses and generated poses $[x, \hat{x}]$. After masked by $\mathcal{M}$, k-recent poses and observed poses are input to the DSTCN. They are encoded by the encoders (observed encoder f_α and masked encoder f_m) of DSTCN, getting the context vector v_o and v_m respectively. The parameter generation module provides parameters θ to the Decoder through an external memory that shares the accumulated information and a parameter learner for capturing novel information. Then the concatenated context vectors (v_o and v_m) are used by the decoder to generate future pose. Note that the predicted pose $\hat{x}_{t+k}$ will be used to predict the pose $\hat{x}_{(t+k)+1}$ at the next moment.

The prediction capability of most current methods is limited to short-term predictions, mainly because they do not explicitly learn sub-motion patterns and do not dynamically select important motion segmentation as a reference in the past. Facing few-shot new motion categories, these methods, which have been trained with large amounts of annotated data, will quickly become overfitting. In this paper, we design a MoPredNet for few-shot human motion prediction.

2.2 Few-shot Learning

The goal of the few-shot learning is to learn new concepts with limited samples. Recent work on few-shot learning can be classified into three categories: metric-based [Koch, 2015; Vinyals *et al.*, 2016], gradient-based [Finn *et al.*, 2017], and memory-based [Ravi and Larochelle, 2016; Gidaris and Komodakis, 2018]. In few-shot human motion prediction , [Gui *et al.*, 2018] proposes to combine the model-agnostic meta-learning MAML [Finn *et al.*, 2017] and model regression networks (MRN) [Wang *et al.*, 2017] to jointly learn generic model initialization and adaptation strategy during the meta-training phase.

Unlike [Gui *et al.*, 2018] that slowly updates parameters, when facing novel features extracted from a new category, we directly update a specific model for this motion category from the external memory. Then instead of just using stored fixed parameters [Gidaris and Komodakis, 2018], we also generate target parameters by learning novel motion dynamics from new motion categories.

3 Approach

In human motion prediction task, we forecast future pose sequence based on the observed pose sequence. In this paper, the pose at each time step is defined by x (i.e., mocap representation for skeleton joints). For observed motion pose, the motion sequence from time 1 to t is expressed as $x_{1:t}$. The predicted motion sequence from $t+1$ to $t+T$ is expressed as $\hat{x}_{t+1:t+T}$, which is expected to be close to the ground truth $x_{t+1:t+T}$. To predict the correct motion of a new category from very few observed samples, as illustrated in Fig. 2, the proposed new architecture consists of three parts. DSTCN (Section 3.1) adaptively captures motion information from observed poses and mask poses (Section 3.2) through en coders and generates new pose through the decoder. In th few-shot motion prediction, the parameter generation modul (Section 3.3) generates new parameters by combining infor mation from the memory and generation module, enablin fast adaptation to new motion categories.

3.1 Deformable Spatio-Temporal Convolution

To adaptively model sub-motion dynamics and spatial co relation in the entire motion sequence, inspired by [Dai e *al.*, 2017], we propose a Deformable Spatio-Temporal Con volution Network (DSTCN) based on Temporal Convolutio Network (TCN) [Bai *et al.*, 2018]. Same as [Li *et al.*, 2018 the observed skeletons data are represented in a 2D forma temporal sequence domain t and spatial skeleton domain s.

Compared with TCN, our DSTCN can adaptively model sub-motions, and can flexibly capture spatio-temporal structure. We show the first two layers of DSTCN in Fig. 3. The feature $F_{l+1} \in \mathbb{R}^{n_s \times n_t}$ (feature map size: n) of location P_0 in $l+1$-th layer includes not only original receptive field O information in l layer but also its neighbor's information. The selection of neighbor information for motion modeling is guided by the offset of receptive field ΔP. Then feature $F_{l+1}(P_0)$ convolves $(*)$ this deformable information by a 2D convolution kernel $K_l \in \mathbb{R}^{m_s \times m_t}$ (kernel size: m). We formulate the deformable convolution process as follows:

$$F_{l+1}(P_0) = \sum_{P_{ts} \in O} K_l * F_l(P_0 + P_{ts} + \Delta P), \tag{1}$$

where $P_{ts} = (P_t, P_s)$. P_t is the temporal dimension and P_s is the spatial dimension. Due to the limitation of the fixed kernel, TCN does not have the offset ΔP. Unlike the original application of deformable convolution using image data [Dai *et al.*, 2017], the DSTCN is designed to fit the temporal data, where the information from the previous moment cannot be leaked to the next moment.

In our work, we use the nonlinear activation function (ReLU) to limit the direction of convolution expansion. We also adaptively expand the receptive field in space to capture the correlation of joints that are not directly connected, e.g., the symmetry of arms. We have the spatio-temporal offset:

$$\Delta P = (-ReLU(\Delta P_t), \Delta P_s). \tag{2}$$

However, high-dimensional 3D skeleton data are different from the explicit relationship of adjacent pixels. Adjacent pixels are not coaxial, so it is difficult to extract spatial correlation directly. Therefore, reasonable deformation expansion should occur on the same geometric pixels. In this work, we present a local geometric extraction layer before deformable spatio-temporal convolution layer to encode coordinates data $\Delta P_s = P_{joint}$.

The extension of convolution to the past is beneficial for the modeling of periodic and aperiodic sub-motions. The captured motion dynamics in base motion categories can be easily generalized to new motion categories.

3.2 Sequence Mask

Correlated motion sequence is selected in past sequences based on attention mechanism to provide explicit motion information guidance. As shown in Fig. 2, the observed pose and masked pose (masked by $\mathcal{M}$) are encoded by observed encoder f_α and masked encoder f_m respectively. At each time step, by combining observed motion context vector v_o and masked motion context vector v_m, a reasonable future motion sequence can be predicted by decoder function f_β. The generation of $\hat{x}_{t+k}$ can be presented by

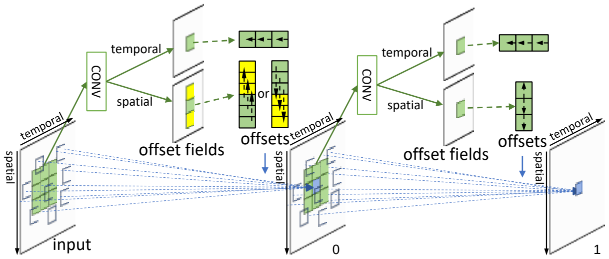

Figure 3: Illustration of Deformable Spatio-Temporal Convolution. The input is human motion data with spatio-temporal information. A convolutional layer extracts the temporal and spatial offset fields. In the temporal domain, the offset field is restricted to the past direction. In the spatial domain, the offset field is encouraged to expand to both two directions. To effectively extract the spatial correlation, from the input to the feature map F_0, the offset is limited to the joint level (green coordinate to yellow coordinate).

$$\hat{x}_{t+k} = f_\beta\big(f_\alpha(x_{1:t};\theta_\alpha), f_m(\mathcal{M} \odot [x, \hat{x}];\theta_m);\theta_\beta\big), \tag{3}$$

where θ_α, θ_m, θ_β are parameters of deep networks. $\odot$ denotes Hadamard product. $[x, \hat{x}]$ is the observed and generated motion before time $t+k$. The temporal mask $\mathcal{M}$ in this paper is binary which depends on the correlation score a_i between past poses and the current pose. The process of selection can be given by

$$\mathcal{M}_i = \begin{cases} 0, & a_i \le \delta \\ 1, & a_i > \delta \end{cases}, \tag{4}$$

where a_i is defined as

$$a_i = \frac{exp\left(x_{t+k}{}^{\intercal} x_{t+k-i}\right)}{\sum_{i=0}^{K} exp\left(x_{t+k}{}^{\intercal} x_{t+k-i}\right)}. \tag{5}$$

Here, δ and K are hyper-parameters that determine the number of reference frames. In summary, different from previous work [Martinez *et al.*, 2017; Gopalakrishnan *et al.*, 2019] in which a single physical property is selected, we enhance correlated motion sequence encoding to provide explicit guidance for long-term motion prediction.

3.3 Few-shot Learning

Fig. 4 summarizes our new memory-based parameter generation module for the few-shot motion prediction. Following related few-shot work [Gidaris and Komodakis, 2018; Sun *et al.*, 2019], we divide large human motion dataset $\mathcal{D}$ into two subcategory datasets: base categories split $\mathcal{D}_{base} = \{\mathcal{D}_i\}_{c=1}^{\mathcal{C}_b}$ of $\mathcal{C}_b$ categories and novel categories split $\mathcal{D}_{novel} = \{\mathcal{D}_i\}_{c=1}^{\mathcal{C}_n}$ of $\mathcal{C}_n$ categories. In few-shot motion prediction, the training data for a new category are limited. Therefore, we sample a few prediction task $\mathcal{T}^{(tr)} = \bigcup_{c=1}^{\mathcal{C}_b} \{x_{c,s}\}_{s=1}^{S}$ of S

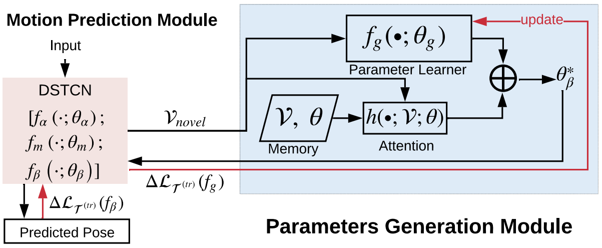
Motion Prediction Module
Input
DSTCN
$[f_\alpha(\cdot;\theta_\alpha);$
$f_m(\cdot;\theta_m);$
$f_\beta(\cdot;\theta_\beta)]$
$\mathcal{V}_{novel}$
$\Delta\mathcal{L}_{\mathcal{T}^{(tr)}}(f_g)$
$\Delta\mathcal{L}_{\mathcal{T}^{(tr)}}(f_\beta)$
Predicted Pose
$f_g(\bullet;\theta_g)$
Parameter Learner
update
θ_β^*
$\mathcal{V},\ \theta$
Memory
$h(\bullet;\mathcal{V};\theta)$
Attention
Parameters Generation Module

Figure 4: Illustration of the Parameter Generation Module. Motion Prediction Module provides the context vector v_{novel} of new motion category to Parameters Generation Module and gets new parameters ${\theta_\beta}^*$ from it. In the Parameter Generation Module, accumulated knowledge is stored in the memory, including context vectors $\mathcal{V}$ of base category and corresponding parameters θ. The novel information in v_{novel} which is non-existence in based context vectors, is decoded by the Parameter Learner to generate target parameters. It combines with the base parameter computed by attention mechanism to generate new parameters. $\nabla\mathcal{L}_{\mathcal{T}^{(tr)}}(f_\beta)$ and $\nabla\mathcal{L}_{\mathcal{T}^{(tr)}}(f_g)$ denote backpropagation (red line) for DSTCN and the Parameter Learner, respectively.

samples from $\mathcal{D}_{base}$ and $\mathcal{T}^{(ts)} = \bigcup_{c=1}^{\mathcal{C}_n} \{x_{c,s}\}_{s=1}^{S'}$ of S' samples from $\mathcal{D}_{novel}$ for meta-train and meta-test separately.

In our work, we train a model suitable for a new category of motion prediction in three phases: a pre-training phase, a meta-training phase, and a meta-test phase. In the pre-training phase, to simulate the accumulation of basic human knowledge, we pre-train our motion prediction module on a large dataset with base categories $\mathcal{D}_{base}$. Pose encoder (including whole past sequence encoder $f_\alpha(\cdot;\theta_\alpha)$, masked sequence encoder $f_m(\cdot;\theta_m)$), and pose decoder $f_\beta(\cdot;\theta_\beta)$ are trained by the loss function $\mathcal{L}_\mathcal{D}$:

$$[\theta_\alpha;\theta_m;\theta_\beta] \leftarrow [\theta_\alpha;\theta_m;\theta_\beta] - \gamma_1 \nabla\mathcal{L}_\mathcal{D}([f_\alpha;f_m;f_\beta]), \quad (6)$$

where $\bigtriangledown\mathcal{L}_\mathcal{D}$ is computed by the L2 loss ℓ_2 on frames:

$$\nabla\mathcal{L}_\mathcal{D}([f_\alpha;f_m;f_\beta]) = \frac{1}{\mathcal{D}} \sum_{x\in\mathcal{D}} \ell_2(x_{t+1:t+T}, \hat{x}_{t+1:t+T}), \quad (7)$$

and γ_1 denotes the learning rate. In addition to pre-training general parameters $[\theta_\alpha;\theta_m;\theta_\beta]$ of the model for all motion categories, the parameters θ_β of the prediction decoder are also kept in an external memory with its n-dimensional context vector $v = f_\alpha(\cdot;\theta_\alpha) \in \mathbb{R}^n$ (including general parameters and specific category parameters). So, similar to memory networks [Sukhbaatar *et al.*, 2015], the external memory stores key-value pairs: $(\mathcal{V},\theta)$, where $\mathcal{V} = \{v_c\}_{c=1}^{\mathcal{C}_b+1}$ and $\theta = \{\theta_{\beta_c}\}_{c=1}^{\mathcal{C}_b+1}$.

The second phase is the meta-training phase, as shown in Fig. 4. To quickly adapt the motion prediction module to new motion categories, we use the parameter generation module to generate decoder parameters θ_β instead of optimizing them slowly [Gui *et al.*, 2018]. A crucial difference between our network and previous work [Gidaris and Komodakis, 2018] is that our proposed parameter generation module not only references a credible parameter set θ in the memory but also relies on **novel** information in a new motion category. The parameters in the memory have the ability to extract shared features in the motion but cannot capture unique information for new motion. Therefore, we propose a parameter learner to extract novel motion dynamics and generate a novel weight for $f_g(\cdot;\theta_g)$, which enables the motion prediction module decode new motion. Combining the base parameters and novel parameters, the p-dimensional target parameters ${\theta_\beta}^* \in \mathbb{R}^p$ are computed as:

$${\theta_\beta}^* = f_g(v_{novel};\theta_g) + h(v_{novel},\mathcal{V},\theta), \quad (8)$$

where the context vector v_{novel} is encoded by observed encoder $f_\alpha(x_{t+1:t+T};\theta_\alpha)$ and is the same as v_o in base categories. $v_{novel} \in \mathbb{R}^n$ denotes novel feature as a query. $h(\cdot,\cdot,\cdot)$ computes the content-based attention scores by normalized cosine similarity and used to weight the basic parameters. Therefore, the Motion Prediction Module will use the updated parameters $[\theta_\alpha;\theta_m;{\theta_\beta}^*]$ of the DSTCN to generate future pose.

For parameter update in the meta-training phase, parameters $[\theta_\alpha;\theta_m]$ of the pose encoder in DSTCN are frozen, and the learnable generated parameter ${\theta_\beta}^*$ of the pose decoder in DSTCN is optimized by gradient descent:

$${\theta_\beta}^*_{upd} = {\theta_\beta}^*_{ass} - \gamma_2 \nabla\mathcal{L}_{\mathcal{T}^{(tr)}}([f_\alpha;f_m;f_\beta]), \quad (9)$$

where ${\theta_\beta}^*_{ass}$ denotes the assigned parameters before the optimization and ${\theta_\beta}^*_{upd}$ denotes the updated parameters after the optimization. ${\theta_\beta}^*_{ass}$ and ${\theta_\beta}^*_{upd}$ are used for guiding the optimization of θ_g:

$$\theta_g \leftarrow \theta_g - \gamma_3 \nabla\mathcal{L}_{\mathcal{T}^{(tr)}}(f_g), \quad (10)$$

where loss function $\mathcal{L}_{\mathcal{T}^{(tr)}}$ is computed by the L2 loss between ${\theta_\beta}^*_{ass}$ and ${\theta_\beta}^*_{upd}$:

$$\nabla\mathcal{L}_{\mathcal{T}^{(tr)}}(f_g) = \frac{1}{\mathcal{T}^{(tr)}} \sum_{\mathcal{T}^{(tr)}} \ell_2\left({\theta_\beta}^*_{ass}, {\theta_\beta}^*_{upd}\right). \quad (11)$$

In the meta-test phase, we train the parameter generation module on the support data of a new prediction task $\mathcal{T}^{(ts)}$. With the motion of the new category as input, the parameter generation module is adjusted by new training data. To avoid the problem of "catastrophic forgetting" [Lopez-Paz and Ranzato, 2017], the encoder's parameters of the motion generation module are frozen.

Our parameter generation module can quickly update parameters of motion prediction module for new motion categories. The memory module helps generate model parameters for human-like motion, and the parameter learner can help generate model parameters for decoding novel motion.

3.4 Schedule Sampling

There are many training strategies for human motion prediction, including open-loop prediction with noise [Fragkiadaki *et al.*, 2015], closed-loop prediction [Martinez *et al.*, 2017] and fixed sampling [Li *et al.*, 2018]. All these training strategies focus on the challenging error accumulation problem, which ignores the model's convergence efficiency and convergence effect. Inspired by [Bengio *et al.*, 2015], we use

		walking					eating					smoking					discussion				
Methods		80	320	400	560	1000	80	320	400	560	1000	80	320	400	560	1000	80	320	400	560	1000
PAML [Gui *et al.*, 2018]		0.35	0.76	0.82	0.80	0.83	0.36	0.65	0.70	0.71	0.79	0.39	0.81	1.01	1.03	**1.01**	0.41	1.01	1.02	1.09	**1.12**
Baseline	Single	0.32	0.58	0.59	1.09	1.08	0.30	0.58	0.69	0.94	1.53	0.26	0.76	0.89	1.19	2.35	0.44	1.02	1.07	1.33	2.21
	Transfer	0.24	0.44	0.45	0.85	0.87	0.31	0.51	0.57	0.80	0.87	0.23	0.67	0.76	1.02	2.16	0.43	0.99	1.08	1.39	2.54
Abalation	Memory (attention)	0.20	0.44	0.44	0.77	0.85	0.32	0.54	0.59	0.82	0.97	0.24	0.64	0.74	1.02	2.17	0.42	0.95	1.00	1.26	2.22
MoPredNet(Ours)		**0.19**	**0.43**	**0.44**	**0.75**	**0.83**	**0.30**	**0.45**	**0.47**	**0.63**	**0.73**	**0.21**	**0.53**	**0.59**	**0.78**	1.88	**0.41**	**0.94**	**0.90**	**1.06**	1.17

Table 1: Mean Angle Error of different methods on Human 3.6M dataset for few-shot motion prediction task.

scheduled sampling to balance the convergence and generalization during the training phase. Let f denote the entire prediction function, including encoding function (f_α, f_β) and decoding function (f_θ). We have

$$\hat{x}_{t+k+1} = f\left(\left[x_{0:t+k-1},\ \varphi \cdot\ \hat{x}_{t+k} + (1-\varphi)\cdot x_{t+k}\right]\right), \tag{12}$$

where φ decays after each series of iterations I (or epoch), like the learning rate:

$$\varphi_I = 1 - s \cdot \left(1 - \varphi_{I-1}\right). \tag{13}$$

Here, s determines the decay speed of the function, with the range of $[0, 1]$. Hence, in the early stage of training, the network can quickly converge relying on the guidance of ground truth. As the training proceeds, the network can generate plausible motion based on motion model and dynamics. The network then reduces the sampling rate to focus on error accumulation and improves its generalization ability.

4 Experiments

We evaluate the effectiveness of our MoPredNet for few-shot motion prediction on two popular human motion datasets:1) Human 3.6M dataset [Ionescu *et al.*, 2013] and 2) CMU MOCAP. We follow [Martinez *et al.*, 2017] on Human 3.6M dataset, where the sequence of subject 5 is selected for test (same in meta-learning and meta-test) and others are selected for training. Besides few-shot motion prediction, we also report the performance of the sequence mask and DSTCN module on long-term prediction using Human 3.6M dataset. Similar to [Li *et al.*, 2018], 8 representative motions in CMU MOCAP are used and split into training and test set.

4.1 Experimental Setup

In our Motion Prediction Network (MoPredNet), pose encoder consists of whole past sequence encoder and masked motion sequence encoder with the longest mask K set as 20 and mask threshold γ set as 0.0666. They share the parameters of the convolution layer in the encoder, including 1 convolution layer with 1×3 convolutions for joints coding and 3 deformation convolution layers with 2×3 convolutions, followed by leaky ReLU nonlinearity. The feature channels of each convolution layer are set as 16, 64, 128, 256 respectively. After that, the feature is encoded to a 512-dimensional vector by a fully-connected layer. The pose decoder network contains three fully-connected layers with the size of 512, 128, and 54, respectively. Leaky ReLU action function and drop out are both set as 0.5 in the first two layers. For the parameter learner in the parameter generation network, it contains two fully-connected layers. We adopt the ADAM optimizer with the initial learning rate set as $\gamma_1 = \gamma_2 = 1e-4$, $\gamma_3 = 1e-8$. The initial sampling rate and decayed rate are set as 0.8 and 0.7, respectively.

In order to fairly evaluate previous work, we report experimental results by implementing their released code with their original setting. The input window is set as 50 frames (2s), and the output window is set to 25 frames (1s) for training (GCNs' output window is set to 100 frames as it required). It is noteworthy to mention that both in few-shot motion prediction and long-term motion prediction, we do not use any supervised label for motion prediction.

4.2 Evaluation on Human 3.6M Dataset

We first evaluate our MoPredNet architecture for few-shot human motion prediction on Human 3.6M dataset. Following previous work [Gui *et al.*, 2018], we sample 6 motion sequences (5 sequences for training, 1 sequence for test) from each of the 11 categories (except walking, eating, smoking and discussion) as base prediction task $\mathcal{T}^{(tr)}$ during meta-training phase. The remaining categories of motion are used as novel prediction task $\mathcal{T}^{(ts)}$ with its small training set and test set like $\mathcal{T}^{(tr)}$. The performance of our meta-learning architecture is shown in Table 1. Obviously, our method outperforms PAML [Gui *et al.*, 2018] almost all the time in all novel motion categories, which proves that encoding novel motion dynamics is beneficial for few-shot motion prediction.

We construct two baseline methods (MoPredNet without parameter generator module): Single means training a single model with few samples on every specific motion category, and Transfer means training one model for all basic motion categories with large data and fine-tuning new motion categories with few samples. Compared with baseline methods, we can see that our meta-learning method improves the performance of our MoPredNet on novel motion categories.

We also apply the DSTCN and the sequence mask module to long-term motion prediction (4000 milliseconds) [Gopalakrishnan *et al.*, 2019] and compare with the state-of-the-art methods. Table 3 shows the Mean Angle Error (MAE) results on the Euler angles on the Human 3.6M dataset. From the experiment results, we can see that our architecture achieves the best prediction results. Especially at 560ms, 1000ms and 2000ms, our errors are 0.07, 0.08 and 0.1 lower than HMR results respectively.

The qualitative results are illustrated in Fig. 5, which shows the long-term human motion prediction over 4000ms for two actions: walking and walking together. We can see that poses predicted by our MoPredNet are closed to ground truth in short-term prediction and more like human motion. Take walking as an example, our generated pose sequence is consistent with ground truth before 1 second. After that, our MoPredNet hardly holds the same motion frequency with ground

		running					soccer					basketball					washwindow				
Methods		80	320	400	560	1000	80	320	400	560	1000	80	320	400	560	1000	80	320	400	560	1000
Baseline	Single	0.37	0.95	1.14	1.53	1.58	0.23	0.78	0.95	1.18	1.70	0.35	1.06	1.26	1.66	2.80	0.35	0.99	1.19	1.44	1.75
	Transfer	**0.26**	0.68	0.75	0.78	1.27	0.21	0.67	0.81	0.95	1.60	0.28	0.98	1.26	1.63	2.73	**0.34**	0.89	1.09	1.32	1.60
MoPredNet(Ours)		**0.26**	**0.61**	**0.65**	**0.62**	**1.02**	**0.16**	**0.47**	**0.60**	**0.78**	**1.36**	**0.27**	**0.71**	**0.94**	**1.30**	**2.32**	**0.34**	**0.87**	**1.06**	**1.27**	**1.50**

Table 2: Mean Angle Error of our method and baselines on CMU dataset for few-shot motion prediction task.

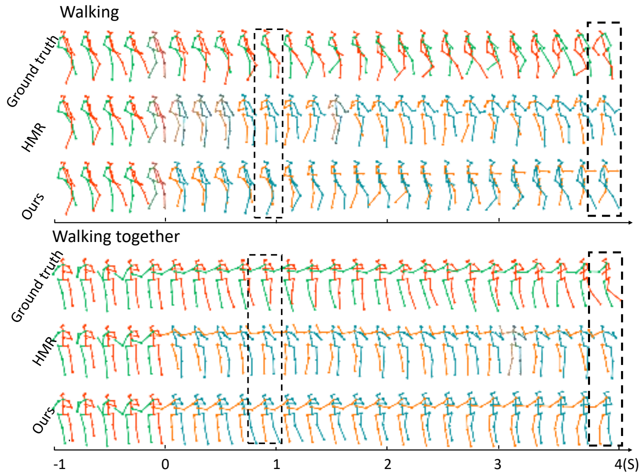

Figure 5: Qualitative results on Human 3.6M dataset.

truth, but can still capture the motion direction changes.

4.3 Evaluation on CMU MOCAP Dataset

Using similar architecture settings for Human 3.6M, we report our results on the CMU MOCAP dataset in terms of angle errors in Table 2. The CMU MOCAP dataset is divided into base categories and novel categories. The base categories include walking, jumping, basketball signal, and directing traffic. The novel categories include running, soccer, basketball, and wash window. The experiment results show that our method outperforms other baselines in all the new categories.

4.4 Ablation Study

We use the parameter learner to update the parameters of decoder according to novel motion dynamics beyond the memory combination. We remove it from our MoPredNet to verify its effectiveness as shown in Table 1. The results show that using the parameter learner provides a significant boost in performance. We also show the influence of the sequence mask module, the DSTCN module, and the local geometric extraction layer in Table 3. The def.cnn in Table 3 indicates DSTCN without local geometric extraction layer before the deformable convolution layer. Results show that the sequence mask, the DSTCN and the local geometric extraction layer all contribute to the improvement of the performance.

In the training stage, we use schedule sampling to cope with the error accumulation problem. We compare the performance in Human 3.6M dataset using schedule sampling with open-loop prediction, closed-loop prediction, and fixed sampling in Table 4. We can see that using schedule sampling can converge to a better local optimal solution.

Methods		80	240	400	560	1000	2000	4000
ERD [Fragkiadaki *et al.*, 2015]		1.21	1.44	1.65	1.80	2.18	2.56	2.96
LSTM-3LR [Fragkiadaki *et al.*, 2015]		1.39	1.53	1.72	1.89	2.33	2.79	3.29
Res-GRU [Martinez *et al.*, 2017]		0.39	0.95	1.28	1.49	1.91	2.44	3.01
Zero-velocity [Martinez *et al.*, 2017]		0.40	0.90	1.21	1.42	1.85	2.21	2.55
Conv seq2seq [Li *et al.*, 2018]		0.38	0.87	1.15	1.35	1.77	2.17	2.51
HMR [Liu *et al.*, 2019]		0.38	0.83	1.14	1.37	1.80	2.14	2.46
GCNs [Wei *et al.*, 2019]		0.40	0.89	1.18	1.36	1.76	2.22	2.69
Ablation	None	0.40	0.90	1.21	1.41	1.85	2.22	2.60
	Mask	0.40	0.89	1.19	1.39	1.81	2.15	2.52
	def. cnn	0.39	0.87	1.16	1.36	1.81	2.24	2.65
	DSTCN	0.39	0.85	1.13	1.33	1.75	2.10	2.45
MoPredNet(Ours)		**0.38**	0.84	**1.11**	**1.30**	**1.72**	**2.04**	**2.43**

Table 3: Mean Angle Error of different methods on Human 3.6M dataset for short-term and long-term motion prediction tasks.

	80	240	400	560	1000	2000	4000
Close L	0.40	0.87	1.13	1.32	1.75	2.13	2.87
Open L	0.40	0.88	1.17	1.37	1.80	2.18	2.53
Fixed S	0.39	**0.84**	1.12	1.32	1.74	2.09	2.47
Schedule S	**0.38**	**0.84**	**1.11**	**1.30**	**1.72**	**2.04**	**2.43**

Table 4: Mean Angle Error of different sampling methods on Human 3.6M dataset.

5 Conclusion

In this paper, we propose MoPredNet for few-shot motion prediction. In addition to the base knowledge accumulated through base motion categories, MoPredNet can capture novel motion dynamics from a few new motion samples. This improves its performance in new motion categories. MoPredNet can select the correlated motion frames of past motion to provide explicit guidance for long-term pose prediction. In addition, the MoPredNet can adaptively capture past motion spatio-temporal structure to improve its encoding capability. Experiment results demonstrate that our method achieves state-of-the-art performance in human motion prediction.

References

[Bai *et al.*, 2018] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[Behl *et al.*, 2017] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Bengio *et al.*, 2015] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[Fragkiadaki *et al.*, 2015] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[Gidaris and Komodakis, 2018] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018.

[Gopalakrishnan *et al.*, 2019] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12116–12125, 2019.

[Gui *et al.*, 2018] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 432–450, 2018.

[Ionescu *et al.*, 2013] 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*, 36(7):1325–1339, 2013.

[Koch, 2015] Gregory Koch. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning*, 2015.

[Koppula and Saxena, 2015] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[Li *et al.*, 2018] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018.

[Liu *et al.*, 2019] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019.

[Lopez-Paz and Ranzato, 2017] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476, 2017.

[Martinez *et al.*, 2017] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

[Pavlovic *et al.*, 2001] Vladimir Pavlovic, James M Rehg, and John MacCormick. Learning switching linear models of human motion. In *Advances in neural information processing systems*, pages 981–987, 2001.

[Ravi and Larochelle, 2016] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2016.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[Sun *et al.*, 2019] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[Taylor *et al.*, 2007] Graham W Taylor, Geoffrey E Hinton, and Sam T Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352, 2007.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[Vondrak *et al.*, 2008] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Physical simulation for probabilistic motion tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[Wang *et al.*, 2006] Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. In *Advances in neural information processing systems*, pages 1441–1448, 2006.

[Wang *et al.*, 2017] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[Wei *et al.*, 2019] Mao Wei, Liu Miaomiao, Salzemann Mathieu, and Li Hongdong. Learning trajectory dependencies for human motion prediction. In *International Conference on Computer Vision*, 2019.