Meta-Auxiliary Learning for Adaptive Human Pose Prediction

Qiongjie Cui[1], Huaijiang Sun[1]*, Jianfeng Lu[1], Bin Li[2], Weiqing Li[1]

[1]Nanjing University of Science and Technology
[2]Tianjin AiForward Science and Technology Co., Ltd., China
cuiqiongjie@njust.edu.cn, sunhuaijiang@njust.edu.cn
lujf@njust.edu.cn, libin@aiforward.com, li_weiqing@njust.edu.cn

Abstract

Predicting high-fidelity future human poses, from a historically observed sequence, is crucial for intelligent robots to interact with humans. Deep end-to-end learning approaches, which typically train a generic pre-trained model on external datasets and then directly apply it to all test samples, emerge as the dominant solution to solve this issue. Despite encouraging progress, they remain non-optimal, as the unique properties (*e.g.*, motion style, rhythm) of a specific sequence cannot be adapted. More generally, once encountering out-of-distributions, the predicted poses tend to be unreliable. Motivated by this observation, we propose a novel test-time adaptation framework that leverages two self-supervised auxiliary tasks to help the primary forecasting network adapt to the test sequence. In the testing phase, our model can adjust the model parameters by several gradient updates to improve the generation quality. However, due to catastrophic forgetting, both auxiliary tasks typically have a low ability to automatically present the desired positive incentives for the final prediction performance. For this reason, we also propose a meta-auxiliary learning scheme for better adaptation. Extensive experiments show that the proposed approach achieves higher accuracy and more realistic visualization.

Introduction

Human pose forecasting, accurately predicting how a person will move in the near future, is a fundamental task in computer vision, which has enormous potential in machine intelligence, and human-robot interaction (Gui et al. 2018a; Wang et al. 2021; Liu et al. 2021; Sofianos et al. 2021).

Over the past few years, extensive literature has sprung up exploring this fascinating topic, with deep-learning based end-to-end approaches proving increasingly popular (Li et al. 2020a; Gui et al. 2018b; Li et al. 2018, 2021). Researchers typically train on external large-scale datasets (Ionescu et al. 2014) to achieve a generic pre-trained model, which is then indiscriminately applied to all test sequences with the same set of network weights in the inference stage (Jain et al. 2016; Wei Mao 2021; Dang et al. 2021). These approaches have extensively investigated this issue from various perspectives, emerging as the mainstream solutions (Guo et al. 2022; Cui et al. 2021b).

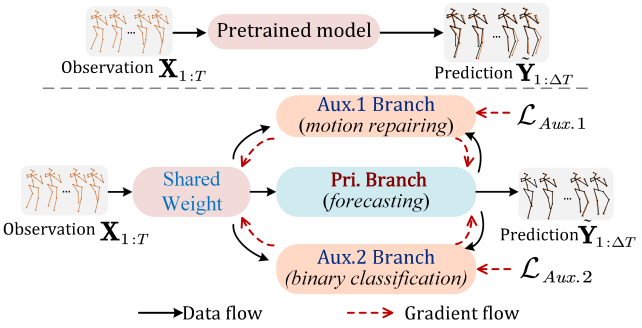

Figure 1: Comparison of a classic deep end-to-end model (top) with our approach (bottom) at test-time. Given a sequence $\mathbf{X}_{1:T}$, typical approaches utilize the pre-trained model obtained from large-scale datasets to generate the prediction $\tilde{\mathbf{Y}}_{1:\Delta T}$, which is sub-optimal, as the internal information within the specific sample is ignored. By contrast, our model learns to adapt to the unique properties of the test sample via several meta-learning steps.

While the empirical result is encouraging, it is not optimal. In real-world applications, the inherent defects of the existing DNN-based models (Ruiz et al. 2018; Gopalakrishnan et al. 2019; Barsoum et al. 2018; Kundu et al. 2019; Aliakbarian et al. 2021) cannot be overlooked, among which the major one is that the features learned from external datasets can hardly cover the unique attributes within a specific sequence, including motion style and rhythm, etc. In this case, the model tends to focus on the dominant distribution in training, while failing to take account of the unique patterns of new test sequence (out-of-distribution), therefore, unreliable results may be yielded. This inability to adapt to the internal properties of a given sequence hinders the realistic application of the predictive algorithms (Yuan et al. 2020; Mao et al. 2020; Shu et al. 2021).

To solve it, we propose a novel test-time adaptation (TTA) approach. Concretely, our model falls into auxiliary learning, where the network consists of one primary task and two self-supervised auxiliary ones. The primary task (Pri.) focuses on mapping historical observations to the predicted poses. The auxiliary task-1 (Aux.1) is a simple binary classifier to distinguish whether the input sequence is a scrambled

*indicates corresponding author

counterpart of the observation. As a contrast, some joints of the observed sequence are randomly removed to construct the corrupted sequence, and then Aux.2 aims to repair these missing joints. The Pri., Aux.1, and Aux.2 share most of the parameters, and are jointly trained to achieve a base model. Then, in the testing phase, Aux.1 and Aux.2 behave as a regularization to further update the shared weights to enhance the generalization for specific sequences.

Intuitively, the auxiliary task provides rich semantic cues to fine-tune the model parameters (Chi et al. 2021; Varsavsky et al. 2020). However, empirical results show that, a rough update of the base model may lead to the criticized *negative transfer*, as the invalid message may be exchanged (Xiao et al. 2018; Vafaeikia et al. 2020). To solve it, we design a Gate Sharing Unit (GSU), which learns to control the relative intensity of message transmission among tasks in both training and testing, to pass the favorable information, while hindering the redundant or even incorrect ones.

Even so, there is a legacy problem: how to ensure that the Pri. branch obtains better-adapted parameters to ensure the forecasting performance of specific sequences. For this purpose, inspired by MXML (Liu et al. 2019; Chi et al. 2021), we integrate meta-learning into auxiliary learning to form meta-auxiliary learning. Our meta-objective is to optimize the whole network via meta-auxiliary learning so that the Pri. branch can better adapt to test sequences. Note that we call the pair composed of the observed and the future poses the 'task' in the meta-learning nomenclature. Moreover, for each observed sequence, the adapted parameters are different, and its specific motion patterns can be generalized.

Methodologically, to capture the spatio-temporal pattern of skeleton data, we introduce two virtual relay nodes into the sparse transformer, to form the Spatial Sparse-Relay Transformer (SS-RT) and Temporal Sparse-Relay Transformer (TS-RT) (Aksan et al. 2021; Cai et al. 2020). The relay nodes are capable of receiving information from all human joints along with spatial and temporal aspects, to extract the global spatio-temporal correlations. With the sparse transformer and relay-nodes update, the newly designed SS-RT and TS-RT explicitly consider the human topology and temporal smoothness of motion sequences, as well as long-term correlations in space and time.

Our contributions are multifaceted: (1) We develop a test-time adaptation approach that leverages meta-auxiliary learning to enable fast and effective adaptation to the specific information within test sequences. (2) Both motion repairing and binary classification are introduced as our self-auxiliary tasks. (3) To avoid the negative transfer across multi-tasks, the GSUs are designed to allow valid information to be passed easily among tasks, while preventing useless ones.

Related Work

Human Motion Forecasting. Nowadays, various GNN-based models are being developed to extract the semantic connectivity of the 3D skeleton sequence, with promising results (Mao et al. 2019; Cui et al. 2020; Li et al. 2020a; Dang et al. 2021; Cui et al. 2021b; Zhong et al. 2022). However, GCNs are capable only of gathering information from the local neighbor joints, and have a limited capacity to capture long-term relationships.

Currently, researchers attempt to exploit the Transformer to achieve the long-range correlation, whereas, it fails to consider the meaningful topology and temporal smoothness of motion sequences, and brings more computational cost (Mao et al. 2020; Aksan et al. 2021; Guo et al. 2022). In contrast, our approach, which includes a sparse transformer and virtual relay nodes, allows us to explicitly focus on the meaningful local structure and temporal continuity while still extracting long-term correlations.

Test-time Adaptation. To improve the generalization for diverse distributions, the test-time adaptation (TTA) scheme is recently proposed (Chi et al. 2021; Varsavsky et al. 2020; Hu et al. 2021; Shin et al. 2021). Typically, deep learning algorithms are trained on external datasets to produce a general model, and before making decisions, TTA resorts to auxiliary tasks to neatly fine-tune the weights according to the internal knowledge of test samples. Due to the utilization of both external and internal information, superior outcomes are achieved (He et al. 2021; Hao et al. 2021).

However, the existing test-time adaptation remains a key challenge, that is, the auxiliary task may send inaccurate or even incorrect messages to the primary task (Vafaeikia et al. 2020; Xiao et al. 2018). To address it, we elaborate a simple but effective gated sharing unit (GSU) that adaptively releases the important context while preventing others.

Meta-learning. Our work is related to meta-learning (learning to learn), particularly the model-agnostic version (MAML), which allows the pre-trained model to be adjusted to perform the fast adaptation of individual samples. (Liu et al. 2022) uses MAML for multi-domain single image dehazing, with the meta-objective of learning consistency across the losses of different tasks. Along with MAML, (Liu et al. 2019) presents the meta-auxiliary learning (MXML) framework, which generates labels for additional auxiliary tasks. Inspired by the MXML, (Chi et al. 2021) also achieves a fast adaptation to improve the performance of the primary deblurring operation for unseen images. Our approach, which draws inspiration from these publications in part, involves the following two changes: we design two auxiliary tasks to identify more effective semantics; our auxiliary tasks are self-supervised for automatic inference.

Proposed Approach

Suppose that $\mathbf{X}_{1:T} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_T]$ is an observed sequence over horizon T, where each $\boldsymbol{X}_t = [\boldsymbol{j}_1, \boldsymbol{j}_2, ..., \boldsymbol{j}_N] \in \mathbb{R}^{N \times D}$ records the 3D coordinate of N human joints in a frame. Current DNN-based models directly train a mapping from the observation to the future sequence, $\mathcal{M} : \mathbf{X}_{1:T} \rightarrow \mathbf{Y}_{1:\Delta T}$, with $\mathbf{Y}_{1:\Delta T} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{\Delta T}\}$.

In contrast, our approach incorporates the following developments. (1) Two self-auxiliary tasks are introduced, sharing the majority of model weights and allowing collaborative training alongside the primary forecasting one. (2) To avoid negative transfer across tasks, we build the GSU to prevent the passage of erroneous/incorrect messages. (3) We first train on large-scale datasets to achieve a base model, and the ultimate goal is to further optimize

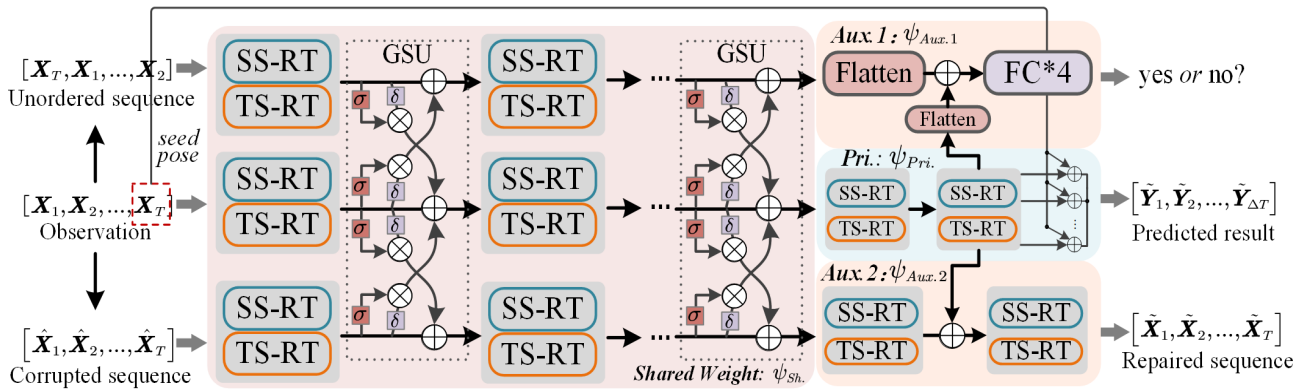

Figure 2: Illustration of our approach. It involves a primary (Pri.) task and two auxiliary (Aux.) ones, sharing most of the parameters w.r.t $\psi_{Sh.}$, except for the task-specific components w.r.t $\{\psi_{Pri.}, \psi_{Aux.1}, \psi_{Aux.2}\}$. The Pri. is concerned with mapping observations to the expected prediction. The objective of Aux.1 is to provide the correct label of the scrambled sequence, and Aux.2 is to repair the missing joints in the corrupted sequence, where both the scrambled and corrupted sequence are derived from the observation. With the proposed GSU, the valid contexts can be exchanged, while the invalid or incorrect ones are blocked. $\otimes$ is element-wise product and $\oplus$ is addition. The last observed frame is regarded as the seed pose (red rectangle).

it at test-time, to automatically adapt to the sample-specific properties, and then yield more realistic predicted results $\tilde{\mathbf{Y}}_{1:\Delta T} = \{\tilde{\boldsymbol{Y}}_1, \tilde{\boldsymbol{Y}}_2, ..., \tilde{\boldsymbol{Y}}_{\Delta T}\}$. (4) In practical studies, the naive updates might not bring desired improvements. To solve it, a meta-auxiliary learning framework is proposed, which learns the better-adapted parameters for the effective test-time adaptation of a specific sequence.

Network Architecture

The network architecture consists of one primary branch and two self-supervised auxiliary ones, as seen in Figure 2. For convenience, the following uses subscripts to indicate the spatial indexes, and superscripts for temporal indexes.

Primary Branch. The Pri. is intended to predict future motions, where its backbone comprises SS-RT and ST-RT to extract the spatio-temporal correlation of motion sequences.

Spatial Sparse-Relay Transformer (SS-RT) is implemented to capture the spatial correlation. In contrast to the vanilla version (Vaswani et al. 2017), we use the spatial sparse transformer (SST) to explicitly consider the skeletal structure (Child et al. 2019). Moreover, we attach a virtual spatial-relay vertex, which utilizes a separate transformer, called spatial-relay transformer (SRT), to directly aggregate the global information in a frame, and distribute it to each one to consider the long-term correlation.

Let $\mathbf{c}^t = \{\boldsymbol{c}_1^t, \boldsymbol{c}_2^t, ..., \boldsymbol{c}_N^t\} \in \mathbb{R}^{N \times C_{in}}$ be the feature at t-th frame, and $\boldsymbol{c}_r^t$ be a spatial-relay vertex. For each node $\boldsymbol{c}_i^t$, we use 3 linear transformations to generate a query $\boldsymbol{q}_i \in \mathbb{R}^d$, a key $\boldsymbol{k}_i \in \mathbb{R}^d$ and a value $\boldsymbol{v}_i \in \mathbb{R}^d$. The SST is used to consider the natural connectivity of the human skeleton:

$$\boldsymbol{c}_i'^t = \sum softmax(\frac{\boldsymbol{q}_i \cdot \boldsymbol{k}_j}{\sqrt{d}})\boldsymbol{v}_j, j \in \{i, \mathcal{N}_i, r\}, \quad (1)$$

where $\mathcal{N}_i$ is the neighbors of ith joint, and r stands for the label of the spatial-relay vertex. Then, the meaningful inductive bias of the human skeleton is expressly considered.

In addition, we make use of SRT to capture the long-term spatial correlation:

$$\boldsymbol{c}_r'^t = \sum softmax(\frac{\boldsymbol{q}_r \cdot \boldsymbol{k}_j}{\sqrt{d}})\boldsymbol{v}_j, j \in \{r\} \cup \{j : 1 \le j \le N\}. \quad (2)$$

By stacking the SST and SRT, our SS-RT is formed, which is capable of extracting the intrinsic connections of human joints, and meanwhile, capturing the long-term spatial correlation at intra-frame. The resulting output of SS-RT can be formalized as: $\mathbf{c}'^t = \{\boldsymbol{c}_1'^t, \boldsymbol{c}_2'^t, ..., \boldsymbol{c}_N'^t\} \in \mathbb{R}^{N \times C_{out}}$.

Temporal Sparse-Relay Transformer (TS-RT) consists of a temporal sparse transformer (TST) for extracting the local inter-frame smoothness, and a temporal-relay transformer (TRT) for long-term temporal dependency. Let $\mathbf{c}_v = \{\boldsymbol{c}_v^1, \boldsymbol{c}_v^2, ..., \boldsymbol{c}_v^T\} \in \mathbb{R}^{T \times C_{in}}$ be the input hidden state, for $v \in N$, with $\boldsymbol{c}_v^i \in \mathbb{R}^{C_{in}}$, and $\boldsymbol{c}_v^r$ be the feature of temporal-relay node, 3 linear transformations are exploited to produce $\boldsymbol{q}_i \in \mathbb{R}^d$, $\boldsymbol{k}_i \in \mathbb{R}^d$ and $\boldsymbol{v}_i \in \mathbb{R}^d$. The TST is defined as:

$$\boldsymbol{c}_v'^i = \sum softmax(\frac{\boldsymbol{q}^i \cdot \boldsymbol{k}^j}{d})\boldsymbol{v}^j, j \in \{i, i-1, i+1, r\}. \quad (3)$$

Then, the temporal-relay node is updated with the TRT:

$$\boldsymbol{c}_v'^r = \sum softmax(\frac{\boldsymbol{q}^r \cdot \boldsymbol{k}^j}{d})\boldsymbol{v}^j, j \in \{r\} \cup \{j : 1 \le j \le T\}. \quad (4)$$

The TST and TRT are stacked to create the TS-RT, where the output feature is $\mathbf{c}'_v = \{\boldsymbol{c}_v'^1, \boldsymbol{c}_v'^2, ..., \boldsymbol{c}_v'^T\} \in \mathbb{R}^{T \times C_{out}}$.

With the TST and TRT, the TS-RT enables the consideration of both local and global temporal correlation, which is crucial for human motion prediction. In both SS-RT and TS-RT, we set $d = 64$, and in keeping with recent progress (Vaswani et al. 2017; Devlin et al. 2019), we exploit $H = 8$ independent heads to stabilize the training.

Finally, as illustrated in Figure 2, the Pri. branch is composed of 9 shared blocks and a task-specific one, each of which is formed by an SS-RT and a TS-RT.

Following the recent works (Mao et al. 2020; Cui et al. 2020), the combination of L_2 distance and bone length loss is exploited as the loss of the Pri.:

$$\mathcal{L}_{Pri.} = \|\mathbf{Y}_{1:\Delta T} - \tilde{\mathbf{Y}}_{1:\Delta T}\|_2 + \eta \mathcal{L}_B(\mathbf{Y}_{1:\Delta T}, \tilde{\mathbf{Y}}_{1:\Delta T}), \quad (5)$$

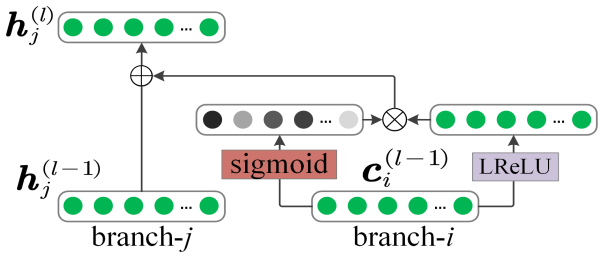

Figure 3: Illustration of the Gated Sharing Unit.

where $\mathcal{L}_B$ is the function to calculate the bone length difference of two motion sequences, and $\eta = 0.04$, as in (Xia et al. 2018; Cui et al. 2020). $\tilde{\mathbf{Y}}_{1:\Delta T}$ is the predicted poses, $\mathbf{Y}_{1:\Delta T}$ is the corresponding GT.

Auxiliary Branches. Our approach involves two self-supervised auxiliary branches: Aux.1 (binary classifier) and Aux.2 (motion repairing), as discussed below.

The Aux.1 is a typical binary classifier whose purpose is to identify whether the input counterpart of the observation sequence is disordered. In (Sun et al. 2020; Geirhos et al. 2018), the auxiliary branch is specified as a simple classification task, which improves the performance of the main task at test time. In image processing (Varsavsky et al. 2020; Shin et al. 2021; He et al. 2021), the image plane is rotated at a certain angle and the auxiliary task seeks to anticipate the rotation angle. Such a classification task has been proven to provide an effective context for primary image analysis. These solutions have demonstrated that the basic classifier is an effective auxiliary task to enhance the primary branch, which motivates us to design our auxiliary task as a simple self-supervised binary classifier. Let p be the correct label of the disordered sequence, and $\tilde{p}$ the predicted one. The loss function of Aux.1 can be denoted as:

$$\mathcal{L}_{Aux.1} = -\left(p \log \tilde{p} + (1-p)\log(1-\tilde{p})\right). \tag{6}$$

On the other hand, (Cui et al. 2021a) proposes to take repairing the missing joints as the auxiliary task of the predictor, which has generated a more realistic prediction result. Therefore, we design our Aux.2 as a motion repairing task. Specifically, we randomly remove some joints in the observed sequence, and the Aux.2 is exploited to impute these missing values. The loss function of Aux.2 is described as:

$$\mathcal{L}_{Aux.2} = \|\mathbf{X}_{1:T} - \tilde{\mathbf{X}}_{1:T}\|_2 + \mu \mathcal{L}_B(\mathbf{X}_{1:T}, \tilde{\mathbf{X}}_{1:T}). \tag{7}$$

We set $\mu = 0.04$. $\tilde{\mathbf{X}}_{1:T}$ is the repaired sequence, $\mathbf{X}_{1:T}$ is the underlying complete observation.

Compared with the Pri., the goals of our Aux. branches are rather straightforward and can be swiftly carried out during testing. Similar to the Pri., both the Aux.1 and Aux.2 include 9 shared blocks formed by SS-RT and TS-RT, as well as a task-specific portion, as shown in Figure 2.

Gated Sharing Units (GSUs)

Our network design fits into a multi-task framework in part, where redundant or even incorrect information may be exchanged among tasks, resulting in the criticized negative transfer. To mitigate it, we provide a simple but effective GSU to adaptively control the passage of features.

Given a feature $\boldsymbol{c}_i^{l-1} \in \mathbb{R}^{C_{out}}$ of branch-i at l-th layer, we first use a single FC layer to achieve the weight:

$$\boldsymbol{z}_{ij}^{l-1} = \sigma(\boldsymbol{W}_i^{l-1}\boldsymbol{c}_i^{l-1} + \boldsymbol{b}_i^{l-1}), \tag{8}$$

where σ denotes the sigmoid function, and $\boldsymbol{z}_{ij}^{l-1} \in \mathbb{R}^{C_{out}}$ is the intensity of the message passing form branch-i to branch-j. $\boldsymbol{W}_i^{l-1} \in \mathbb{R}^{C_{out}\times C_{out}}$ and $\boldsymbol{b}_i^{l-1} \in \mathbb{R}^{C_{out}}$ are the learnable weight and bias, respectively. Let $\boldsymbol{h}_j^{l-1} \in \mathbb{R}^{C_{out}}$ be the feature at the previous layer of branch-j, the output $\boldsymbol{h}_j^{l} \in \mathbb{R}^{C_{out}}$ is then calculated:

$$\boldsymbol{h}_j^{l} = \boldsymbol{z}_{ij}^{l-1} \otimes \delta(\boldsymbol{U}_i^{l-1}\boldsymbol{c}_i^{l-1} + \boldsymbol{e}_i^{l-1}) + \boldsymbol{h}_j^{l-1}, \tag{9}$$

where $\boldsymbol{U}_i^{l-1} \in \mathbb{R}^{C_{out}\times C_{out}}$ is the learnable weight, and $\boldsymbol{e}_i^{l-1} \in \mathbb{R}^{C_{out}}$ is the bias. $\otimes$ denotes the element-wise product, and δ is LeakyReLU with a slope of 0.2. Figure 3 presents an illustration of the proposed GSU.

Joint Training and Meta-auxiliary Learning

Joint Training. Since our model involves multiple branches, it can be directly trained, much as the multi-task learning solutions (Cui et al. 2021a; Li et al. 2020a; Chi et al. 2021). The overall objective is:

$$\mathcal{L} = \mathcal{L}_{Pri.} + \mathcal{L}_{Aux.1} + \mathcal{L}_{Aux.2}. \tag{10}$$

Once the training is complete on the external dataset, the pre-trained model, w.r.t. $\mathcal{M}_\psi$, is attained, which is regarded as the initialization of the meta-auxiliary learning.

Meta-auxiliary Learning. Due to the failure of exploiting the internal properties of test samples, the pre-trained model learned from Eq.10 has a low ability to adapt to the unseen data. We solve this problem by using the proposed meta-auxiliary learning to obtain the optimal parameter.

For each to use, we decompose the model parameters ψ into the shared weights $\psi_{Sh.}$ and task-specific ones $\{\psi_{Pri.}, \psi_{Aux.1}, \psi_{Aux.2}\}$ for each branch. To enable the model parameters to be customized according to the unique distribution of test samples, we propose to use meta-auxiliary learning to create the adapted parameters. Concretely, inspired by (Liu et al. 2019; Chi et al. 2021), our meta-auxiliary learning intends to learn the consistency of the parameters of our Aux. branches for the Pri. task, to ensure that the auxiliary tasks improve the performance of the Pri. task. In the inner loop of the meta-training phase, several gradient updates of the auxiliary losses are used to update the parameter of the whole network parameter ψ, thereby performing effective adaptation on a specific sample. Given a training pair $(\mathbf{X}_{1:T}^{(k)}, \tilde{\mathbf{X}}_{1:T}^{(k)})$, concerning the corrupted and repaired sequence, and their labels $(p^{(k)}, \tilde{p}^{(k)})$, it can be achieved:

$$\tilde{\psi}^{(k)} \leftarrow \psi - \alpha \nabla_\psi \left[\mathcal{L}_{Aux.1}(p^{(k)}, \tilde{p}^{(k)}) + \mathcal{L}_{Aux.2}(\mathbf{X}_{1:T}^{(k)}, \tilde{\mathbf{X}}_{1:T}^{(k)})\right] \tag{11}$$

where $\psi^{(k)} = \{\psi_{Sh.}^{(k)}, \psi_{Pri.}^{(k)}, \psi_{Aux.1}^{(k)}, \psi_{Aux.2}^{(k)}\}$ is the adapted parameter that is tailored by the specific observation. α is the learning rate of the adaptation procedure. We notice that both Aux. branches have the same gradient descent direction and are optimized concurrently to ensure a synergistic impact for the adaptation.

Our approach strives to maximize the performance of the Pri. forecasting branch by adjusting the model parameters through the self-supervised auxiliary tasks. For this purpose, our meta-objective is formally denoted as:

$$\min_{\psi_{Sh.},\psi_{Pri.}} \sum_{k=1}^{K} \mathcal{L}_{Pri.}\left(\mathbf{Y}_{1:\Delta T}^{(k)}, \tilde{\mathbf{Y}}_{1:\Delta T}^{(k)}; \tilde{\psi}_{Sh.}^{(k)}, \tilde{\psi}_{Pri.}^{(k)}\right). \quad (12)$$

Here, $\mathcal{L}_{Pri.}$ is computed suing the pair $(\mathbf{X}_{1:T}^{(k)}, \mathbf{Y}_{1:\Delta T}^{(k)})$, while the optimization is over $\psi = \{\psi_{Sh.}, \psi_{Pri.}, \psi_{Aux.1}, \psi_{Aux.2}\}$ to achieve the updated parameter of the Pri. task. Eq.12 can be minimized using gradient descent algorithms:

$$\psi \leftarrow \psi - \beta \sum_{k=1}^{K} \nabla_{\psi} \mathcal{L}_{Pri.}\left(\mathbf{Y}_{1:\Delta T}^{(k)}, \tilde{\mathbf{Y}}_{1:\Delta T}^{(k)}; \tilde{\psi}_{Sh.}^{(k)}, \tilde{\psi}_{Pri.}^{(k)}\right), \quad (13)$$

where β is the meta-learning rate. The overall meta-auxiliary learning procedure is conducted in Algorithm.1, in which the parameters of the Pri. task are updated in the outer loop, and the auxiliary parameters are updated in the inner loop. Regarding the testing phase, Eq.11 is used to obtain the adapted parameters ψ, and then $\{\psi_{Sh.}, \psi_{Pri.}\}$ is used to improve the generalization ability of the primary task.

Implementation Details

As shown in Figure 2, our model includes a Pri. and two Aux. branches. The shared parts consist of 9 residual blocks, created by combining the outputs of SS-RT and TS-RT, and having the channel $C_{in} = C_{out} = 512$. In addition, the task-specific portions of the Pri. and Aux.2 are an additional block to map the feature into the original dimension. By contrast, the Aux.1 is a binary classifier, where its separate parts comprise a flatten layer, and 4 FC layers with channel numbers $256, 128, 64, 1$. Aux.1 takes a scrambled-order counterpart of the observation as the input, while for Aux.2, we randomly remove 20% of the joints from observations. Note that, the feature of the last layer of Pri., is directly connected to Aux.2, and passing through a flatten layer, is connected to Aux.1, so that the meta-auxiliary learning can update the whole parameters of the Pri. branch. We follow the current multi-task learning framework, and exploit the Adam optimizer to train our network, where the learning rate is initialized to 0.001, with a 0.98 decay every 2 epoch. The mini-batch size is 16. At the test-time adaptation, we fix the learning rate $\alpha = \beta = 2 \times 10^{-5}$, and 6 gradient descents of Eq.11 are performed. Finally, the fine-tuned parameters are acquired, allowing for the adaptation of the internal properties of a specific sequence to achieve a better prediction.

Algorithm 1: Meta-Auxiliary Training

Require: learning rates α, β. pre-trained parameter $\psi = \{\psi_{Sh.}, \psi_{Pri.}, \psi_{Aux.1}, \psi_{Aux.2}\}$
Ouput: meta-auxiliary learned parameter
1: initialize the model with the pre-trained parameter ψ
2: while *not converge* **do**
3: sample a training batch from the $\{\mathbf{X}_{1:T}^{(k)}, \mathbf{Y}_{1:\Delta T}^{(k)}\}_{k=1}^{K}$;
4: **for** *each k* **do**
5: evaluate the auxiliary losses $\mathcal{L}_{Aux.1}, \mathcal{L}_{Aux.2}$;
6: update the adapted parameter:
$\tilde{\psi}^{(k)} = \psi - \alpha \nabla_{\psi}[\mathcal{L}_{Aux.1}(p^{(k)}, \tilde{p}^{(k)}) + \mathcal{L}_{Aux.2}(\mathbf{X}_{1:T}^{(k)}, \tilde{\mathbf{X}}_{1:T}^{(k)})]$
7: **end**
8: evaluate the primary task and update:
$\psi \leftarrow \psi - \beta \sum_{k=1}^{K} \nabla_{\psi} \mathcal{L}_{Pri.}(\mathbf{Y}_{1:\Delta T}^{(k)}, \tilde{\mathbf{Y}}_{1:\Delta T}^{(k)}; \tilde{\psi}_{Sh.}^{(k)}, \tilde{\psi}_{Pri.}^{(k)})$
9: end

Experiments

Preliminaries

Dataset-1: H3.6M (Ionescu et al. 2014) involves 15 action categories performed by 7 professional human subjects (*S_1*, *S_5*, *S_6*, *S_7*, *S_8*, *S_9*, *S_11*). Each pose is represented as a 17-joint skeleton ($N = 17$), and the sequences are down-sampled to achieve 25 fps (Mao et al. 2019; Ma et al. 2022).

Dataset-2: We also select 8 action categories from **CMU MoCap**. The pre-processing solution is consistent with the H3.6M dataset. For both H3.6M and CMU MoCap, the proposed model is implemented where the length of the observed sequence is equal to the prediction ($T = \Delta T = 25$).

Baselines. To assess the effectiveness of the proposed approach, the following 5 state-of-the-art (SoTA) methods are selected as our baselines, including LTD (Mao et al. 2019), DMGNN (Li et al. 2020b), MSR (Dang et al. 2021), ST-Tr (Aksan et al. 2021), and PGBIG (Ma et al. 2022).

Metric. We test our model using the Mean Per Joint Position Error (MPJPE) in millimeters, in accordance with earlier work (Ma et al. 2022; Dang et al. 2021).

Experimental Setups. We use 3 alternative setups to analyze our model, as stated in Table 1. The prefix *S* indicates the *subject*, and *C* denotes the *category*. For fairness, we also apply the training/testing division in Table 1, but the hyper-parameters remain unchanged, to re-train the baselines.

Datasets	testing	training	purpose
H3.6M	(*i*) *S_5*	*S_1,S_6∼S_9,S_11*	general predictive ability
	(*ii*) *S_x*	other subjects	predictive ability on unseen subjects
	(*iii*) *C_x*	other categories	predictive ability on unseen categories
CMU MoCap	(*iv*) *C_x*	other categories	

Table 1: Experimental setups. As in the typical approaches, the setup-(*i*) is to evaluate the general predictive ability, while the setup-(*ii*)(*iii*)(*iv*) are newly designed to investigate the adaptability to out-of-distribution data.

Comparison with State-of-the-arts on H3.6M

The existing predictors are normally tested on the actions of *S_5* and trained on the other subjects. However, our key observation is that, the motion patterns of different individuals tend to be distinct; therefore, this distribution-shift deteriorates the performance of deep pre-trained models. As a comparison, at test-time, our model is able to be further optimized by meta-auxiliary learning, to achieve a better result. Consistent with the previous work (Mao et al. 2019; Cui et al. 2020), we first use the setup-(*i*) (in Table 1) to evaluate the general predictive ability of our model. Table 2 reports the comparison of 3 representative activities. We observe that, our result tends to be better in almost all scenarios, which reveals that the dynamic characteristics of *S_5* are potentially distinguishing from other subjects, and our model can adapt to them.

Intuitively, due to unique height and body proportion,

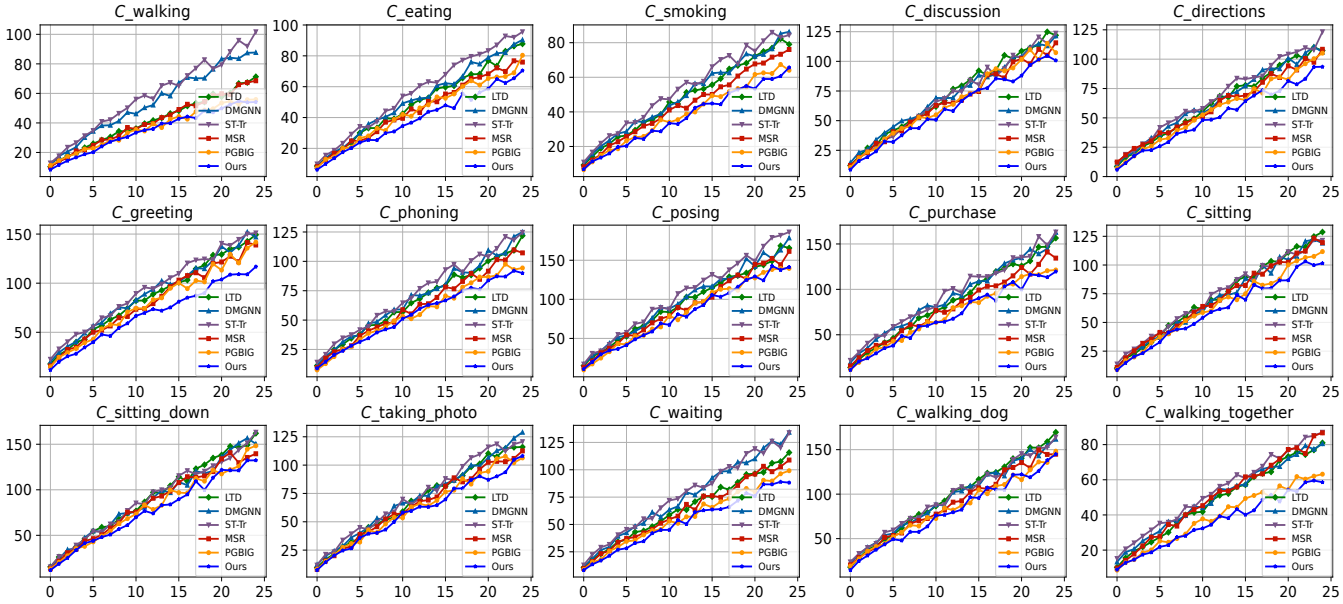

Figure 4: Comparison of each unseen action category *C_x* from the H3.6M dataset. We observe that, at test-time, our approach is able to be fine-tuned for a specific category *C_x* to adapt to its internal properties, thus achieving higher prediction accuracy.

	walking			eating			smoking		
ms	160	400	1000	160	400	1000	160	400	1000
LTD	23.0	46.1	59.8	16.9	40.7	77.8	16.2	38.9	72.6
DMGNN	30.7	65.2	95.8	21.4	43.9	86.7	17.6	40.3	72.2
ST-Tr	27.4	67.3	103.2	24.5	47.0	84.2	17.9	42.8	79.6
MSR	22.7	45.2	63.0	17.1	40.4	77.1	16.3	38.2	71.6
PGBIG	19.8	40.3	56.4	**15.1**	38.1	76.0	14.1	34.7	69.5
Ours	**18.9**	**38.1**	**54.3**	15.4	**36.7**	**73.6**	**13.5**	**32.0**	**67.5**

Table 2: MPJPE comparisons on 3 activities from the H3.6M dataset, where the experimental design follows the conventions of predictive algorithms (*S_5* is used for testing while the other is used for training). It reveals that the dynamic characteristics of *S_5* are slightly different from those of other subjects, and our approach is able to adapt to them.

even for the same category, the motion properties (*e.g.*, styles or rhythms) of different subjects are potentially inconsistent. To further investigate the adaptation ability of different subjects, the experimental setup-(*ii*) is used. Concretely, we fine-tune the base model under the test actions of a specific subject-x (*S_x*), where the base model is learned from the others. Table 3 provides the average MPJPE of the end predicted pose (1000ms) of different unseen subjects. From the results, we observe that our model produces better predictions against the baseline models. It implies that the dynamic characteristics of different humans indeed involve distinct motion attributes. Moreover, our approach exploits the external large dataset, and meanwhile, can be tailored based on the internal information of test sequences via meta-auxiliary learning, to consistently yield a superior result for unseen subjects.

Due to the diversity and uncertainty, human action involves unenumerable categories. Typically, the training dataset falls short of covering all action types. In practical applications, existing deep end-to-end algorithms face a major challenge, that is, once encountering the unseen category at test-time, their performance tends to decline sharply. However, our model is able to further optimize the base model learned from large datasets, to adapt to the unique attributes of a new action category. To verify it, we exploit the experimental setup-(*iii*). Specifically, our approach and the baselines are evaluated under each specific category *C_x* respectively, while the training is conducted on the remaining ones. From Figure 4, we observe that our model brings superior results in all scenarios for such out-of-distribution data of unprecedented categories. It evidences that our model is indeed capable of adapting to the characteristics of unseen action categories.

Subjects	*S_1*	*S_6*	*S_7*	*S_8*	*S_9*	*S_11*
LTD	115.4	132.8	133.7	120.1	123.8	124.3
DMGNN	122.5	139.3	131.0	125.2	134.7	120.2
ST-Tr	133.6	147.5	134.2	128.0	140.2	124.5
MSR	115.7	131.0	123.1	116.5	118.8	116.2
PGBIG	113.2	127.3	124.4	118.3	114.6	112.0
Ours	**107.0**	**123.2**	**118.7**	**113.5**	**109.7**	**110.2**

Table 3: Average MPJPE of a total of 15 activities at the end predicted pose (1000ms), evaluated on each unseen subject.

Also, Figure 4 illustrates two qualitative comparisons between the proposed model and SoTA PGBIG (Ma et al. 2022), for the greeting activity of the unseen subject-11 and unseen action category-phoning.

Results on CMU MoCap. We also evaluate the predictive ability on unseen categories from the CMU MoCap using the setup-(*iv*). From Figure 5, the results show that our model

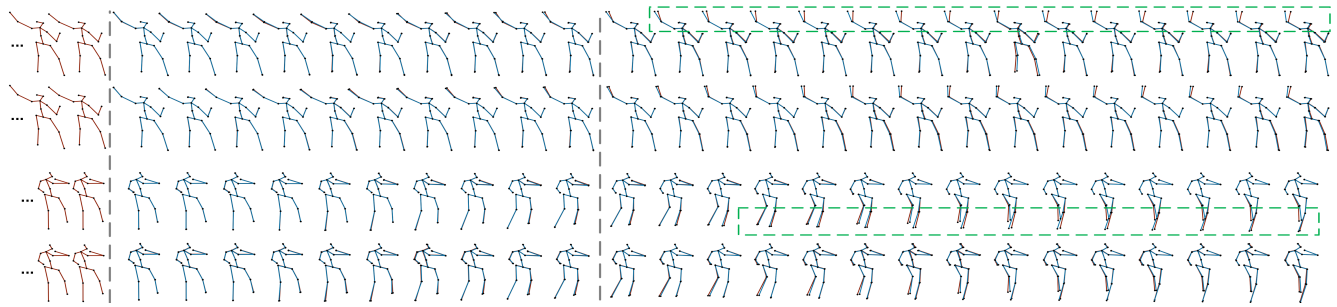

Figure 5: Qualitative comparison on the greeting activity of the unseen subject-11 (top) and unseen category-phoning (bottom). In each sub-figure, the first row is the PGBIG (Ma et al. 2022), followed by our result, where the blue pose refers to the prediction. We observe that, our predicted poses are closer to the GT, as it is tailored according to the specific sequence.

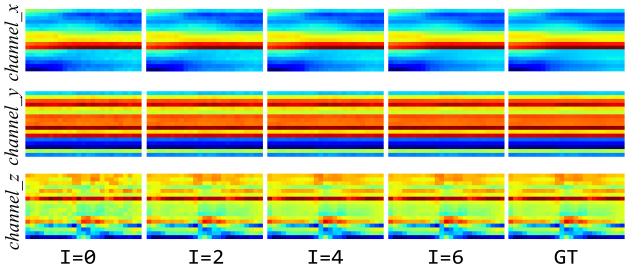

Figure 6: Results of the unfolding TTA with different steps of gradient descents $I = \{0, 2, 4, 6\}$ on *C_smoking*. With the iteration, the coarse result tend to be close to the GT.

substantially outperforms the baselines.

	basket	*bas-sig*	*dir-trc*	*jump*	*run*	*soccer*	*walk*	*washwin*
LTD	109.0	75.3	121.2	142.4	65.7	115.3	49.0	83.1
DMGNN	145.6	72.7	130.3	163.1	73.4	121.9	52.1	89.8
ST-Tr	150.2	77.1	131.0	153.2	76.8	130.4	60.3	95.6
MSR	101.1	66.6	117.4	138.6	56.3	110.4	45.2	74.9
PGBIG	94.9	61.9	113.0	134.8	57.1	105.1	41.4	74.7
Ours	**87.4**	**53.1**	**97.3**	**120.4**	**46.2**	**93.2**	**35.8**	**62.5**

Table 4: Average MPJPE of the end predicted pose (1000ms) of each unseen category *C_x* from the CMU MoCap dataset.

Progressive results. At test time, with several gradient updates, our model allows us to learn to adapt to the internal properties of the test sequence. To better explain it, we show the progressive result by unfolding the TTA procedure after each gradient descent. The inference is run on smoking, and the training is run on the other categories. Figure 6 presents the 3 channels (*i.e.*, x, y, z axes) of these intermediate results by the heat map, with more red denoting larger, and bluer, smaller values. We see that as the iteration goes on, the result gradually tends to be closer to the GT.

Ablation Studies

Here, the following ablation experiments are conducted. We adapt our approach to each action category *C_x* and take the average as the result, as in the setup-(*iii*).

Millisecond (ms)	160	400	1000
w/o GSUs	23.1	60.3	109.3
w/ GSUs	**21.6**	**55.2**	**104.1**

Table 5: Effects of the proposed GSU.

Aux.1	*Aux.2*	160	400	1000
✓	×	25.2	62.7	111.0
×	✓	22.6	58.5	108.4
✓	✓	**21.6**	**55.2**	**104.1**

	160	400	1000
$I=0$	23.2	58.1	109.5
$I=5$	22.1	56.0	106.2
$\boldsymbol{I=6}$	**21.6**	**55.2**	**104.1**
$I=7$	22.0	**54.9**	106.7

Table 6: Effects of Aux. branches (left), and different number of gradient descents (right). $I = 0$ is the base model.

***w/* GSUs *v.s. w/o* GSUs.** Intuitively, the GSU facilitates the transfer of useful information. It is confirmed in Table 5.

Impact of Aux. branches. Both the Aux.1 and Aux.2 branches behave as the complement to the Pri. task. To verify the effectiveness of the Aux. branches, we analyze the effects to the Pri. of Aux.1 and Aux.2 on retaining one of them. As shown in Table 6 (left), when Aux.1 and Aux.2 are introduced concurrently, a better result is achieved.

Number of gradient descents. Here, we provide the impact of the maximum number of gradient updates $I = \{0, 5, 6, 7\}$ at test-time adaptation. From Table 6 (right), we observe that, overall, the larger I obtains smaller errors. When $I = 5$, the best result is yielded, and a larger value brings no benefits.

Conclusion

In this work, we have introduced a test-time adaptation model for human motion forecasting. It uses meta-auxiliary learning to ensure that the update of auxiliary tasks can bring superior adaptability and better performance to the main task on specific samples. At test-time, it resorts to meta-auxiliary learning to ensure that the updates of both auxiliary tasks can bring better adaptation and higher performance to the primary task on specific sequences. Extensive experiments show that our model consistently outperforms the SoTA approaches on unseen subjects and categories.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62176125), in part by the Jiangsu Funding Program for Excellent Postdoctoral Talent (2022ZB269), in part by the Natural Science Foundation of Jiangsu Province (BK20220939), and in part by the China Postdoctoral Science Foundation (2022M721629).

References

Aksan, E.; Kaufmann, M.; Cao, P.; and Hilliges, O. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, 565–574. IEEE.

Aliakbarian, S.; Saleh, F.; Petersson, L.; Gould, S.; and Salzmann, M. 2021. Contextually plausible and diverse 3d human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 11333–11342.

Barsoum, E.; Kender, J.; and Liu, Z. 2018. HP-GAN: Probabilistic 3D Human Motion Prediction via GAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1418–1427.

Cai, Y.; Huang, L.; Wang, Y.; Cham, T.-J.; Cai, J.; Yuan, J.; Liu, J.; Yang, X.; Zhu, Y.; Shen, X.; et al. 2020. Learning Progressive Joint Propagation for Human Motion Prediction. In *European Conference on Computer Vision (ECCV)*, 226–242. Springer.

Chi, Z.; Wang, Y.; Yu, Y.; and Tang, J. 2021. Test-Time Fast Adaptation for Dynamic Scene Deblurring via Meta-Auxiliary Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9137–9146.

Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*.

Cui, Q.; and Sun, H. 2021a. Towards Accurate 3D Human Motion Prediction From Incomplete Observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4801–4810.

Cui, Q.; Sun, H.; and Yang, F. 2020. Learning Dynamic Relationships for 3D Human Motion Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6519–6527.

Cui, S. H. S. X., Qiongjie; and Kong, Y. 2021b. Deep Human Dynamics Prior. In *ACM MM*, 1418–1427.

Dang, L.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 11467–11476.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Geirhos, R.; Temme, C. R.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

Gopalakrishnan, A.; Mali, A.; Kifer, D.; Giles, C. L.; and Ororbia, A. 2019. A Neural Temporal Model for Human Motion Prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gui, L.-Y.; Wang, Y.-X.; Liang, X.; and Moura, J. M. F. 2018a. Adversarial Geometry-Aware Human Motion Prediction. In *European Conference on Computer Vision (ECCV)*, 786–803.

Gui, L.-Y.; Wang, Y.-X.; Ramanan, D.; and Moura, J. M. F. 2018b. Few-Shot Human Motion Prediction via Meta-learning. In *European Conference on Computer Vision (ECCV)*, 432–450.

Guo, W.; Bie, X.; Alameda-Pineda, X.; and Moreno-Noguer, F. 2022. Multi-Person Extreme Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13053–13064.

Hao, M.; Li, Y.; Di, Z.; Gundavarapu, N. B.; and Wang, X. 2021. Test-Time Personalization with a Transformer for Human Pose Estimation. In *Advances in neural information processing systems (NeurIPS)*.

He, Y.; Carass, A.; Zuo, L.; Dewey, B. E.; and Prince, J. L. 2021. Autoencoder Based Self-supervised Test-Time Adaptation for Medical Image Analysis. *Medical Image Analysis*, 102136.

Hu, M.; Song, T.; Gu, Y.; Luo, X.; Chen, J.; Chen, Y.; Zhang, Y.; and Zhang, S. 2021. Fully Test-Time Adaptation for Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 251–260. Springer.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. 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.

Jain, A.; Zamir, A. R.; Savarese, S.; and Saxena, A. 2016. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5308–5317.

Kundu, J. N.; Gor, M.; and Babu, R. V. 2019. BiHMP-GAN: Bidirectional 3D Human Motion Prediction GAN. In *AAAI*, volume 33, 8553–8560.

Li, B.; Tian, J.; Zhang, Z.; Feng, H.; and Li, X. 2020a. Multitask Non-Autoregressive Model for Human Motion Prediction. *IEEE Transactions on Image Processing*.

Li, C.; Zhang, Z.; Sun Lee, W.; and Hee Lee, G. 2018. Convolutional Sequence to Sequence Model for Human Dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5226–5234.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. Symbiotic Graph Neural Networks for 3D Skeleton-based Human Action Recognition and Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; and Tian, Q. 2020b. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), 214–223.

Liu, H.; Wu, Z.; Li, L.; Salehkalaibar, S.; Chen, J.; and Wang, K. 2022. Towards Multi-Domain Single Image Dehazing via Test-Time Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5831–5840.

Liu, S.; Davison, A.; and Johns, E. 2019. Self-supervised generalisation with meta auxiliary learning. *Advances in Neural Information Processing Systems*, 32.

Liu, Z.; Lyu, K.; Wu, S.; Chen, H.; Hao, Y.; and Ji, S. 2021. Aggregated Multi-GANs for Controlled 3D Human Motion Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2225–2232.

Ma, T.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2022. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6437–6446.

Mao, W.; Liu, M.; and Salzmann, M. 2020. History Repeats Itself: Human Motion Prediction via Motion Attention. In *European Conference on Computer Vision (ECCV)*, 474–489.

Mao, W.; Liu, M.; Salzmann, M.; and Li, H. 2019. Learning Trajectory Dependencies for Human Motion Prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 9489–9497.

Ruiz, A. H.; Gall, J.; and Moreno-Noguer, F. 2018. Human Motion Prediction via Spatio-Temporal Inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7134–7143.

Shin, C.; Kim, T.; Lee, S.; and Leey, S. 2021. Test-Time Adaptation for Out-Of-Distributed Image Inpainting. In *IEEE International Conference on Image Processing (ICIP)*.

Shu, X.; Zhang, L.; Qi, G.-J.; Liu, W.; and Tang, J. 2021. Spatiotemporal Co-attention Recurrent Neural Networks for Human-skeleton Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sofianos, T.; Sampieri, A.; Franco, L.; and Galasso, F. 2021. Space-Time-Separable Graph Convolutional Network for Pose Forecasting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 11209–11218.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248. PMLR.

Vafaeikia, P.; Namdar, K.; and Khalvati, F. 2020. A Brief Review of Deep Multi-task Learning and Auxiliary Task Learning. *arXiv preprint arXiv:2007.01126*.

Varsavsky, T.; Orbes-Arteaga, M.; Sudre, C. H.; Graham, M. S.; Nachev, P.; and Cardoso, M. J. 2020. Test-time Unsupervised Domain Adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 428–436. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *NeurIPS*, 5998–6008.

Wang, H.; Dong, J.; Cheng, B.; and Feng, J. 2021. PVRED: A Position-Velocity Recurrent Encoder-Decoder for Human Motion Prediction. *IEEE Transactions on Image Processing*, 30: 6096–6106.

Wei Mao, M. S., Miaomiao Liu. 2021. Generating Smooth Pose Sequences for Diverse Human Motion Prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Xia, G.; Sun, H.; Chen, B.; Liu, Q.; Feng, L.; Zhang, G.; and Hang, R. 2018. Nonlinear Low-Rank Matrix Completion for Human Motion Recovery. *IEEE Transactions on Image Processing*, 27: 3011–3024.

Xiao, L.; Zhang, H.; and Chen, W. 2018. Gated Multi-Task Network for Text Classification. *NAACL HLT 2018*, 726–731.

Yuan, Y.; and Kitani, K. 2020. Dlow: Diversifying Latent Flows for Diverse Human Motion Prediction. In *European Conference on Computer Vision (ECCV)*, 346–364.

Zhong, C.; Hu, L.; Zhang, Z.; Ye, Y.; and Xia, S. 2022. Spatio-Temporal Gating-Adjacency GCN for Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6447–6456.