Temporal Constrained Feasible Subspace Learning for Human Pose Forecasting

Gaoang Wang[1,2] and **Mingli Song**[2]
[1]Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, China
[2]College of Computer Science and Technology, Zhejiang University, China
gaoangwang@intl.zju.edu.cn, brooksong@zju.edu.cn

Abstract

Human pose forecasting is a sequential modeling task that aims to predict future poses from historical motions. Most existing approaches focus on the spatial-temporal neural network model design for learning movement patterns to reduce prediction errors. However, they usually do not strictly follow the temporal constraints in the inference stage. Even though a small Mean Per Joint Position Error (MPJPE) is achieved, some of the predicted poses are not temporal feasible solutions, which disobeys the continuity of the body movement. In this paper, we consider the temporal constrained feasible solutions for human pose forecasting, where the predicted poses of input historical poses are guaranteed to obey the temporal constraints strictly in the inference stage. Rather than direct supervision of the prediction in the original pose space, a temporal constrained subspace is explicitly learned and then followed by an inverse transformation to obtain the final predictions. We evaluate the proposed method on large-scale benchmarks, including Human3.6M, AMASS, and 3DPW. State-of-the-art performance has been achieved with the temporal constrained feasible solutions.

1 Introduction

Human pose forecasting is a sequential modeling task that aims to predict future poses from historical motions. This task has received increasing attention in numerous applications, such as autonomous driving [Paden *et al.*, 2016; Mangalam *et al.*, 2020], healthcare [Troje, 2002], teleoperations [Rubagotti *et al.*, 2019], and collaborative robots [Koppula and Saxena, 2013; Unhelkar *et al.*, 2018]. Unlike human pose estimation task [Gu *et al.*, 2019; Gu *et al.*, 2021; Li *et al.*, 2022b; Zhang *et al.*, 2022; Chai *et al.*, 2023] that predicts the pose for observed frames, pose forecasting focuses on future pose estimation. Most of the existing approaches focus on the spatial-temporal neural network model design [Li *et al.*, 2018; Mao *et al.*, 2019; Mao *et al.*, 2020; Sofianos *et al.*, 2021; Li *et al.*, 2022a; Bouazizi *et al.*, 2022; Xu *et al.*, 2022] for learning the movement patterns to reduce prediction errors. Though such data-driven deep learning models can approximate very complex functions and achieve outstanding performance with large-scale datasets, they suffer from less interpretability than conventional simple machine learning models. Unexpected output may occur when the testing data drifts from the training data distribution. Even though a small Mean Per Joint Position Error (MPJPE) is achieved, some of the predicted poses are not temporal feasible solutions. The drawback of commonly used methods is shown in Figure 1. Even though commonly used methods employ temporal constraints in the **training** stage, the forecasting model may still generate infeasible poses in the prediction since the constraints are NOT employed in the **inference** stage. For example, since a human cannot make an abrupt movement, there should be constraints on the upper bound of the change of the joint velocities. Such temporal constraints should be satisfied not only in the training stage but also in the inference stage for any unseen testing data. However, the temporal constrained feasible solutions are seldom discussed in the literature for the pose forecasting task.

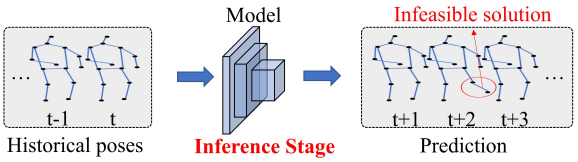

Figure 1: Drawback of commonly used methods in the inference stage. Even though commonly used methods employ temporal constraints in the **training** stage, the forecasting model may still generate infeasible poses in the prediction since the constraints are NOT employed in the **inference** stage.

Physical constraints can alleviate such issues for data-driven deep learning models. By bringing laws from the physical world and general knowledge from humans, predictions and decisions with physical constraints become more trustworthy than those made from purely data-driven models. Some existing approaches take into account physical constraints in deep learning models, typically by adapting ideas from constrained optimization [Amos and Kolter, 2017], training algorithms/loss functions with regularization [Diligenti *et al.*, 2017; Berk *et al.*, 2017], or correcting the model output with iterative projections [Yang *et al.*, 2020; Detassis *et al.*, 2020]. However, most existing approaches

use approximation rather than searching for feasible solutions that strictly comply with the constraints. Besides, iterative adaptation may be required in the inference stage, which takes extra computational costs. Furthermore, compared with unconstrained models, constrained models usually find suboptimal solutions, resulting in degradation in performance.

In this paper, we tackle the human pose forecasting problem with the consideration of temporal constrained feasible solutions, where the change of velocities of human joints should follow the continuity property of the physical movement. To deal with this problem, we propose a temporal constrained feasible subspace learning framework for human pose forecasting. Specifically, rather than direct supervision of the prediction in the original pose space, the temporal constrained subspace is explicitly learned by exploiting simple projection functions, such as rectified linear unit (ReLU) and exponential linear unit (ELU). Then the backward transformation can be applied to obtain the final human pose prediction. Since the constraints are guaranteed in both training and inference stages, no extra steps of iterative projections or optimization are required in the inference. Moreover, unlike most existing approaches that usually sacrifice accuracy to satisfy physical constraints, our proposed method can further reduce the prediction error with constrained feasible solutions. Our framework is shown in Figure 2. The key contributions are summarized as follows:

- We apply temporal constraints in the human pose forecasting task and provide feasible solutions for both training and inference stages. Unlike most existing works, the temporal feasible solutions of our proposed method can be obtained without any iteration, adaptation, optimization, or approximation in the inference.
- We propose a novel subspace learning framework for the temporal constraints. Rather than direct supervision of the prediction in the original pose space, a temporal constrained subspace is explicitly learned, followed by an inverse transformation to obtain the final predictions.
- With the STS-GCN [Sofianos *et al.*, 2021] as the encoder backbone network, we achieve state-of-the-art (SOTA) performance on the Human3.6M [Ionescu *et al.*, 2013], AMASS [Mahmood *et al.*, 2019], and 3DPW [von Marcard *et al.*, 2018] datasets in the human pose forecasting task with temporal feasible solutions.

2 Related Works

Human pose forecasting Some works focus on motion modeling for human pose forecasting. For example, [Chiu *et al.*, 2019] propose a new action-agnostic method for short- and long-term human pose forecasting with triangular-prism RNN for modeling the hierarchical and multi-scale characteristics of human dynamics. [Mao *et al.*, 2019] propose a simple feed-forward deep network for motion prediction, which takes into account both temporal smoothness and spatial dependencies among human body joints. [Mao *et al.*, 2020] propose to extract motion attention to capture the similarity between the current motion context and the historical motion sub-sequences. [Sofianos *et al.*, 2021] propose a space-time-separable graph convolutional network for human pose forecasting. [Adeli *et al.*, 2021] propose a novel trajectory and pose dynamics method based on graph attention networks to model the human-human and human-object interactions both in the input space and the decoded future output space.

Some works combine reinforcement learning in the formulation. Specifically, [Wang *et al.*, 2019] propose a new reinforcement learning formulation for the problem of human pose prediction and develops an imitation learning algorithm for predicting future poses under this formulation through a combination of behavioral cloning and generative adversarial imitation learning. [Yuan and Kitani, 2019] propose the use of a proportional-derivative (PD) control-based policy learned via reinforcement learning to estimate and forecast 3D human poses from ego-centric videos.

Some works consider the action characteristic and scene context information in the setting. For example, [Diller *et al.*, 2020] propose the task of forecasting characteristic 3D poses: from a short sequence observation of a person, predict a future 3D pose of that person in a likely action-defining, characteristic pose. [Adeli *et al.*, 2020] consider incorporating both scene and social contexts as critical clues for the human motion and pose forecasting task.

And some works take account of embedding, constraints, and subtasks in the formulation. For example, [Mangalam *et al.*, 2020] tackle the problem of Human Locomotion Forecasting, a task for jointly predicting the spatial positions of several keypoints on the human body in the near future under an ego-centric setting and presents a method to disentangle the overall pedestrian motion into easier-to-learn subparts by utilizing a pose completion and a decomposition module. [Parsaeifard *et al.*, 2021] propose to learn decoupled representations for the global and local pose forecasting tasks. [Wang *et al.*, 2022] propose the velocity-to-velocity learning paradigm for human motion prediction, which attempts to directly build the sequence-to-sequence model in the velocity space. However, for most existing works, temporal constraints are usually treated as regularization terms in the learning for human pose forecasting. How to guarantee the temporal constrained feasible solutions is not discussed in such approaches.

Physically constrained learning There are some existing works that deal with physical constraints in deep neural networks. Specifically, [Yang *et al.*, 2020] propose a new family of neural networks to predict the behaviors of physical systems by learning their underpinning constraints. [Sangalli *et al.*, 2021] pose the training of deep neural networks for binary classification as a constrained optimization problem using an Augmented Lagrangian method (ALM). [Detassis *et al.*, 2020] use a decomposition scheme alternating master steps and learner steps on the constrained optimization problem. [Diligenti *et al.*, 2017] propose a unified approach to learning from constraints, which integrates the ability of classical machine learning techniques to learn from continuous feature-based representations with the ability of reasoning using higher-level semantic knowledge typical of statistical relational learning. [Amos and Kolter, 2017] present OptNet, a network architecture that integrates optimization

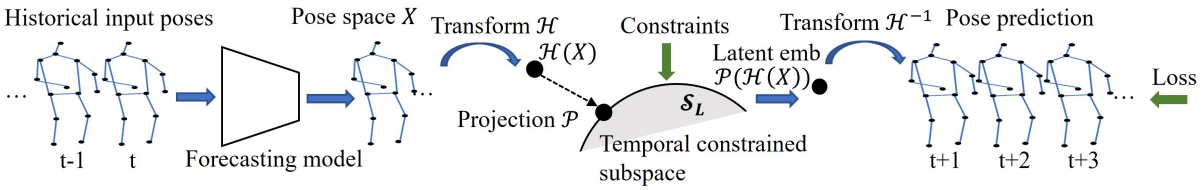

Figure 2: The framework of temporal feasible subspace learning for human pose forecasting. Given historical input poses, we first embed them to the target subspace. Then a projection operation is conducted in the subspace based on the temporal constraints, followed by a transformation from the subspace to the final pose space for future pose prediction.

problems as individual layers in larger end-to-end trainable deep networks. [Kotary *et al.*, 2021a] connect the variation of the training data to the ability of a model to approximate it and propose a method for producing solutions to optimization problems that are more amenable to supervised learning tasks. [Kotary *et al.*, 2021b] survey the recent attempts at leveraging machine learning to solve constrained optimization problems and focuses on surveying the work on integrating combinatorial solvers and optimization methods with machine learning architectures. [Huang *et al.*, 2021] propose the combinatorially efficient, equivariant, and constraint-aware Graph Mechanics Network (GMN), where the geometrical constraints are implicitly and naturally encoded in the forward kinematic. [Rubanova *et al.*, 2021] present a framework for constraint-based learned simulation, where a scalar constraint function is implemented as a graph neural network, and future predictions are computed by solving the optimization problem defined by the learned constraint. [Zhong *et al.*, 2021] introduce a differentiable contact model, which can capture contact mechanics and accommodate inequality constraints. However, most existing solutions usually require iterative projection steps in the inference stage, and feasible solutions are not strictly guaranteed.

3 Proposed Method

3.1 Overview of Human Pose Forecasting

We use the joint-skeleton model to represent the human pose. Given the 3D coordinates of V joints for T_1 frames, we aim to predict the V body joints for the next T_2 future frames. Denote the 3D coordinate of joint v at frame t as $\boldsymbol{x}_{v,t} \in \mathbb{R}^3$. The motion history of human poses is denoted by $\boldsymbol{X}_p \in \mathbb{R}^{V \times T_1 \times 3}$ for all V joints in T_1 frames. We aim to learn a forecasting model $\mathcal{F}_{\boldsymbol{w}}$ parameterized by $\boldsymbol{w}$, *i.e.*,

$$\boldsymbol{X}_f = \mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p), \tag{1}$$

where $\boldsymbol{X}_f \in \mathbb{R}^{V \times T_2 \times 3}$ represent the predicted V joints in the future T_2 frames.

3.2 Temporal Constrained Subspace Learning

Usually, the learning of pose forecasting can be formulated as an optimization problem, *i.e.*,

$$\min_{\boldsymbol{w}} ||\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p) - \boldsymbol{X}_f^*||^2, \ \text{s.t.} \ \mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)) \leq \boldsymbol{0}, \tag{2}$$

where $\boldsymbol{X}_f^*$ is the ground truth of future poses in the training data, and $\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)) \leq \boldsymbol{0}$ represents the inequality constraint. The constraint can be temporal consistency for predicted poses. For example, the difference in motion velocities of joints in adjacent timestamps cannot exceed a certain threshold. The solution should be temporal feasible and follows the constraint. To take the constraint into consideration, we describe the conventional approaches and our proposed approach to address the problem.

Conventional approaches After combining the constraints into the training stage, the training loss of conventional approaches of pose forecasting can be formulated as follows,

$$\mathcal{L}_{\boldsymbol{w}} = ||\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p) - \boldsymbol{X}_f^*||^2 + \lambda \mathcal{R}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)), \tag{3}$$

In this formulation, the constraint $\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)) \leq \boldsymbol{0}$ is converted to a regularization term $\mathcal{R}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p))$ with the hyperparameter λ. For example, $\mathcal{R}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p))$ can be set as $[\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p))]_+$, where $[\cdot]_+$ clamps negative values to zeros.

After training, the inference stage can be represented as

$$\hat{\boldsymbol{X}}_f = \mathcal{F}_{\hat{\boldsymbol{w}}}(\boldsymbol{X}_p), \tag{4}$$

where $\hat{\boldsymbol{w}}$ are the learned model parameters, and $\hat{\boldsymbol{X}}_f$ are the predicted poses in the testing data. However, the constraint is not applied in the inference stage. Due to the gap between the training and testing set, the temporal feasible solution is **not guaranteed**. In other words, the constraint, $\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)) \leq \boldsymbol{0}$, may be not satisfied in the inference stage.

Our approach To address the common problem in the conventional approaches, we propose a temporal constrained subspace learning approach. Based on Eq. (3), we formulate our approach as follows,

$$\mathcal{L}_{\boldsymbol{w}} = ||\mathcal{H}^{-1}(\mathcal{P}(\mathcal{H}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)))) - \boldsymbol{X}_f^*||^2 + \lambda \mathcal{R}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)), \tag{5}$$

where the first term in the equation have three additional operations, *i.e.*, a subspace transformation $\mathcal{H}$, a projection $\mathcal{P}$, and an inverse transformation $\mathcal{H}^{-1}$. Specifically, the subspace transformation $\mathcal{H}$ transforms the pose coordinate space to the temporal subspace; the projection $\mathcal{P}$ projects the unconstrained subspace to the constrained subspace to ensure the feasibility; the inverse transformation $\mathcal{H}^{-1}$ maps from the temporal subspace back to the pose space. An illustration is shown in Figure 2.

Our proposed inference stage can be represented by

$$\hat{\boldsymbol{X}}_f = \mathcal{H}^{-1}(\mathcal{P}(\mathcal{H}(\mathcal{F}_{\hat{\boldsymbol{w}}}(\boldsymbol{X}_p)))). \tag{6}$$

Unlike the inference stage as shown in Eq. (4) of conventional learning approaches that do not take the temporal constraint

into consideration, our inference stage uses the projection $\mathcal{P}$ to ensure the temporal feasibility explicitly. As a result, the temporal feasibility is **always guaranteed** in the inference stage in our proposed method.

3.3 Approach Details

Subspace transformation We consider the temporal subspace as the change of velocity of joint coordinated between additional frames, denoted as $\mathcal{F}_{\boldsymbol{w}}^{''}$, *i.e.*, $\mathcal{F}_{\boldsymbol{w},t}^{''} = (\mathcal{F}_{\boldsymbol{w},t+1}(\boldsymbol{X}_p) - \mathcal{F}_{\boldsymbol{w},t}(\boldsymbol{X}_p)) - (\mathcal{F}_{\boldsymbol{w},t}(\boldsymbol{X}_p) - \mathcal{F}_{\boldsymbol{w},t-1}(\boldsymbol{X}_p))$. The subspace transformation $\mathcal{H}$ defined in Eq. (5) maps the original pose space $\mathcal{F}_{\boldsymbol{w}}$ to the temporal subspace $\mathcal{F}_{\boldsymbol{w}}^{''}$, which follows a linear transformation, *i.e.*,

$$\mathcal{F}_{\boldsymbol{w}}^{''} = \mathcal{H}\mathcal{F}_{\boldsymbol{w}}, \tag{7}$$

where $\mathcal{H}[t, t:(t+2)] = [1, -2, 1]$. Since $\mathcal{H}$ is with size $(T_2 - 2) \times T_2$, we concatenate the natural basis $\boldsymbol{e}_1 = [1, 0, ..., 0]^T$ and $\boldsymbol{e}_{T_2} = [0, ..., 0, 1]^T$ with $\mathcal{H}$ to form a full rank matrix $\tilde{\mathcal{H}} = [\boldsymbol{e}_1^T; \mathcal{H}; \boldsymbol{e}_T^T]$. For simplicity, we use $\mathcal{H}$ to represent $\tilde{\mathcal{H}}$. Then the temporal subspace transformation $\mathcal{H}$ is invertible, denoted as $\mathcal{H}^{-1}$.

Constraint and projection The temporal constraint $\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p))$ is set as $\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)) = |\mathcal{F}_{\boldsymbol{w}}^{''}| - (\max(\boldsymbol{X}_p^{''}) + \delta_{ub})$, where $\max(\boldsymbol{X}_p^{''})$ is the maximum change of velocity in the historical poses and δ_{ub} is the pre-defined upper bound. The constraint assumes that the predicted poses in the temporal subspace should not exceed $\max(\boldsymbol{X}_p^{''}) + \delta_{ub}$. For simplicity, we use $\tilde{\delta}_{ub}$ to represent $\max(\boldsymbol{X}_p^{''}) + \delta_{ub}$.

To explicitly follow the constraint in the temporal subspace, we use projection $\mathcal{P}$ that maps from unconstrained space to the constrained space. Note that we focus on the temporal feasible solutions, *i.e.*, $|\mathcal{F}_{\boldsymbol{w}}^{''}| - (\max(\boldsymbol{X}_p^{''}) + \delta_{ub}) \leq \mathbf{0}$. Then the projection function can be explicitly defined as

$$\mathcal{P}(\mathcal{F}_{\boldsymbol{w}}^{''}) = \left[2\tilde{\delta}_{ub} - [\tilde{\delta}_{ub} - \mathcal{F}_{\boldsymbol{w}}^{''}]_+\right]_+ - \tilde{\delta}_{ub}. \tag{8}$$

The function can be easily implemented with the rectified linear unit (ReLU). We denote this projection as $\mathcal{P}_{\text{relu}}$. Other appropriate projection functions can also be applied. Discussions are provided in the Experiments Section.

4 Experiments

4.1 Datasets and Metrics

Human3.6M [Ionescu *et al.*, 2013] It is a large-scale dataset consisting of 3.6 million 3D human poses and corresponding images. It includes 7 actors performing 15 different actions like *Walking*, *Eating* and *Phoning*. Following the current literature [Mao *et al.*, 2020; Mao *et al.*, 2019; Martinez *et al.*, 2017], we use subject 11 (S11) for validation, the subject 5 (S5) for testing, and all the rest of the subjects for training.

AMASS [Mahmood *et al.*, 2019] The Archive of Motion Capture as Surface Shapes (AMASS) dataset has been recently proposed with 18 existing MoCap datasets. Following [Mao *et al.*, 2020; Sofianos *et al.*, 2021], we take 13 datasets from AMASS in the experiment, with 8 datasets for training, 4 for validation and 1 for testing. Then we use the SMPL [Loper *et al.*, 2015] parameterization for the human skeleton and joint rotation angle to represent the human pose based on the shape vector. 3D Human poses are obtained by applying forward kinematics.

3DPW [von Marcard *et al.*, 2018] The dataset consists of in-the-wild video sequences and 3D human poses captured by a moving camera. The dataset includes both indoor and outdoor actions. In total, it contains 51,000 frames captured at 30Hz, divided into 60 video sequences.

Evaluation metrics Following the benchmark protocols, we adopt the Mean Per Joint Position Error (MPJPE) metric for evaluation. It quantifies the error of the 3D coordinate predictions in mm. The MPJPE is defined as follows,

$$\text{MPJPE} = \frac{1}{VT}\sum_{v=1}^{V}\sum_{t=1}^{T}\|\hat{\boldsymbol{x}}_{v,t} - \boldsymbol{x}_{v,t}^*\|_2, \tag{9}$$

where $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}^*$ are predictions and ground truth joint coordinates, respectively. In addition, we also propose a novel metric, namely infeasible rate (IR), to measure the percentage of predicted joints that do not satisfy the temporal constraint $\mathcal{C}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)) \leq \mathbf{0}$.

4.2 Implementation Details

Model architecture We adopt the Space-Time-Separable Graph Convolutional Network (STS-GCN) [Sofianos *et al.*, 2021] for human pose forecasting, which is one of the SOTA methods for spatial-temporal graph embedding. Details of STS-GCN can be found in [Sofianos *et al.*, 2021]. We use 4 STS-GCN layers in the encoding, which only differ in the number of channels: from 3 (the input 3D coordinates x, y, z), to 64, then 32, 64, and finally 3, by means of the projection matrices. At each layer we adopt batch normalization [Ioffe and Szegedy, 2015] and residual connections.

Training details We use Pytorch for training the neural networks and use ADAM [Kingma and Ba, 2014] as the optimizer. The learning rate is set to 0.01 and decayed by a factor of 0.1 every 5 epochs after the 20th epoch. The batch size is set to 256. The maximum epoch is set to 50. The constraint L is set to 50. One NVIDIA RTX 3090 GPU is used for training.

4.3 Comparison with State-of-the-Art Methods

Human3.6M The MPJPE error in mm on the Human3.6M dataset is shown in Table 1. Following [Sofianos *et al.*, 2021], we estimate the human pose forecasting for 720, 880 and $1,000$ milliseconds. To show the effectiveness of the model, we also report the performance from ConvSeq2Seq [Li *et al.*, 2018], LTD [Mao *et al.*, 2019], RNN-GCN [Mao *et al.*, 2020] and STS-GCN [Sofianos *et al.*, 2021]. Meanwhile, to verify the effectiveness of our proposed constrained learning approach, we also compare with SBR [Diligenti *et al.*, 2017], INP [Yang *et al.*, 2020] and MT [Detassis *et al.*, 2020]. For the compared constrained methods, they have degradation in the prediction of joint positions, which is a common phenomenon for constrained problems. Since we can learn

the constrained feasible solutions in an end-to-end manner with backpropagation, our proposed method can further reduce the prediction error, demonstrating the effectiveness of our method. Except for the superior performance on the prediction error, the proposed method has two additional benefits. First, unlike SBR and INP, which may not strictly comply with the constraints, our method strictly follows the constraints in the inference. Second, unlike INP and MT, our method does not need extra iterative steps in the inference stage, showing the efficiency of our method.

AMASS Similar to Human3.6M, we conduct experiments on AMASS for all the comparison methods. The results are shown in the left part of Table 2. We have achieved competitive results compared with the baseline methods like ConvSeq2Seq, LTD, RNN-GCN, and STS-GCN. Though the improvement is not significant, our solution can generate temporal constrained feasible solutions without iterative steps. Besides, the proposed method also outperforms the constraint-based approaches such as SBR, INP and MT, demonstrating the effectiveness of our subspace learning strategy.

3DPW To test the generalizability of our proposed method, we use AMASS dataset for training and the 3DPW dataset for testing, as shown in the right part of Table 2. Since AMASS is an extremely large-scale dataset that contains many sub-datasets, the generalizability among different approaches does not have a significant difference. With our temporal constrained feasible subspace learning, we can still achieve competitive results compared with the baseline methods like ConvSeq2Seq, LTD, RNN-GCN, and STS-GCN. Similarly, the proposed method also outperforms the other constraint-based approaches, demonstrating the generalizability of our proposed method.

Comparison on infeasible rate (IR) We show the percentage of infeasible solutions that do not satisfy the temporal constraint on Human3.6M testing set in Table 3. Since our proposed method strictly follows the temporal constraint in the inference stage, the IR keeps 0% in all experiments. Compared with the baseline method STS-GCN, our method is over 3% better on 1000 ms forecasting.

Comparison on the predicted distributions We show the distributions of predicted $\boldsymbol{X}_f^{''}$ and the ground truth $\boldsymbol{X}_f^{''}$ on Human3.6M testing set in Figure 3. The top sub-figure shows the comparison between distributions of the ground truth and STS-GCN predictions. The bottom sub-figure shows the comparison between distributions of the ground truth and our constrained predictions. It is obvious that our constrained predictions are much closer to the ground truth distribution.

4.4 Qualitative Results

Qualitative results of the predicted poses We show some qualitative results of the predicted pose on Human3.6M in Figure 4. Each subfigure represents the predicted pose along with the time. Black, red and blue colors represent the ground truth pose, the predicted pose by STS-GCN, and the predicted pose by our method. From the results, the generated poses by our proposed method are closer to the ground truth compared with the baseline method, demonstrating the effectiveness of our method.

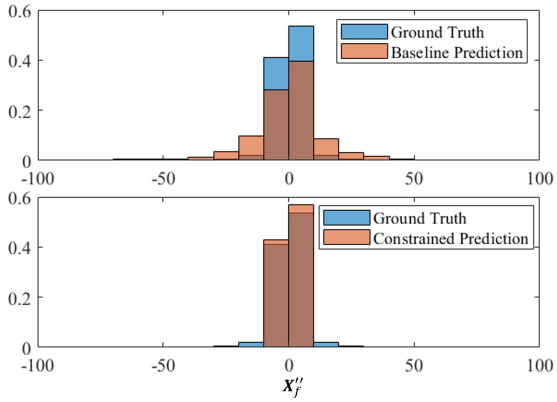

Figure 3: Top: the comparison between $\boldsymbol{X}_f^{''}$ distributions of the ground truth and STS-GCN predictions on the Human3.6M testing set. Bottom: the comparison between $\boldsymbol{X}_f^{''}$ distributions of the ground truth and our constrained predictions the on Human3.6M testing set.

Examples of the temporal constraints on joints To verify the effectiveness of our temporal constrained solutions, we visualize the change of velocities of each joint of the predicted poses in Figure 5. We calculate the change of velocity, *i.e.*, $\boldsymbol{X}_f^{''}$, on the predicted pose along x, y, z and then draw arrows on the joints, where the arrow direction represents the direction of $\boldsymbol{X}_f^{''}$ and the arrow length represents the magnitude of $\boldsymbol{X}_f^{''}$. The first and second row shows pose examples predicted by STS-GCN and our method, respectively. Obviously, the predictions by STS-GCN have many abrupt changes that do not obey the temporal constrained rules. In contrast to STS-GCN, our method can generate poses with only subtle changes in the velocities, which are closer to real-world movement, showing the effectiveness of our constrained solution.

4.5 Ablation Study

Regularization loss functions As defined in Eq. (5), we consider several regularization terms $\mathcal{R}_{\boldsymbol{w}}$ in the training. The regularization terms include the MSE of the change of velocities with respect to the ground truth, $\mathcal{L}_v$; and the regularization on the change of velocities, $\mathcal{L}_r$. These regularization losses are defined as follows,

$$\begin{aligned} \mathcal{L}_v &= \|\mathcal{P}(\mathcal{H}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p))) - \boldsymbol{X}_f^{*''}\|^2, \\ \mathcal{L}_r &= [\|\mathcal{P}(\mathcal{H}(\mathcal{F}_{\boldsymbol{w}}(\boldsymbol{X}_p)))\| - \tilde{\delta}_{ub}]_+. \end{aligned} \tag{10}$$

Note that the prediction in the inference stage is defined as $\hat{\boldsymbol{X}}_f = \mathcal{H}^{-1}(\mathcal{P}(\mathcal{H}(\mathcal{F}_{\hat{\boldsymbol{w}}}(\boldsymbol{X}_p))))$, where the temporal feasibility is not affected by the regularization terms. The results with different combinations of the regularization terms training on the Human3.6M dataset are shown in Table 4. The weight λ in Eq. (5) is set to 1 by default. For each column in the table, ✓ mark represents the adopted regularization term in training. For the rows of $\mathcal{P}_{\text{relu}}$ and $\mathcal{P}_{\text{elu}}$, the best performance is shown in bold. For both the projection functions $\mathcal{P}_{\text{relu}}$ and $\mathcal{P}_{\text{elu}}$, smaller errors are achieved when $\mathcal{L}_v$ and $\mathcal{L}_r$

Actions	Walking			Eating			Smoking			Discussion		
msec	720	880	1000	720	880	1000	720	880	1000	720	880	1000
ConvSeq2Seq	77.2	80.9	82.3	72.8	81.8	87.1	69.4	77.2	81.7	112.9	123.0	129.3
LTD	54.4	57.4	60.3	62.6	71.3	75.8	59.3	67.1	72.1	103.9	113.6	118.5
RNN-GCN	52.1	55.5	58.1	61.4	70.6	75.5	56.6	64.4	69.5	102.2	113.2	119.8
STS-GCN	**45.0**	**48.0**	**51.8**	**40.2**	**46.2**	**52.4**	**39.6**	**45.4**	**50.0**	**63.6**	**72.3**	**78.8**
SBR	48.1	52.3	56.3	41.6	49.9	55.7	40.7	46.9	52.8	66.8	75.7	81.3
INP	54.1	56.6	57.3	50.6	53.7	58.1	47.8	51.6	53.9	71.1	78.2	83.8
MT	46.0	49.3	52.6	41.4	47.3	54.1	40.5	46.2	51.0	64.5	73.1	79.4
Ours (STS-GCN + TCSL)	**44.4**	**48.1**	**50.7**	**39.1**	**45.4**	**49.7**	**38.1**	**44.0**	**47.6**	**63.5**	**72.1**	**76.9**

Actions	Directions			Greeting			Phoning			Posing		
msec	720	880	1000	720	880	1000	720	880	1000	720	880	1000
ConvSeq2Seq	99.8	109.9	115.8	130.7	142.7	147.3	92.1	105.5	114.0	148.8	171.8	187.4
LTD	88.1	99.4	105.5	119.7	132.1	136.8	83.6	96.8	105.1	137.8	160.8	174.8
RNN-GCN	88.2	100.1	106.5	118.4	132.7	138.8	82.9	96.5	105.0	136.8	161.4	178.2
STS-GCN	**56.5**	**64.5**	**71.0**	**76.3**	**85.5**	**91.6**	**51.1**	**59.3**	**66.1**	**79.2**	**94.5**	**106.4**
SBR	59.8	69.5	74.9	79.9	89.5	95.5	52.5	62.7	69.2	84.2	102.0	111.4
INP	66.0	72.7	77.7	83.1	91.4	98.0	60.2	65.0	70.3	88.9	100.8	110.2
MT	57.3	65.0	71.4	77.2	86.4	92.3	52.1	59.9	67.1	79.6	**95.0**	106.0
Ours (STS-GCN + TCSL)	**55.2**	**63.9**	**68.7**	**75.8**	**85.4**	**91.7**	**50.1**	**58.4**	**63.5**	**79.0**	95.3	**102.9**

Actions	Purchases			Sitting			Sitting Down			Taking Photo		
msec	720	880	1000	720	880	1000	720	880	1000	720	880	1000
ConvSeq2Seq	129.1	143.1	151.5	98.8	112.4	120.7	125.1	139.8	150.3	102.4	117.7	128.1
LTD	114.9	127.1	134.9	96.2	110.3	118.7	118.2	133.1	143.8	93.5	108.4	118.8
RNN-GCN	110.9	125.0	134.2	93.1	107.0	115.9	116.1	132.1	143.6	90.1	105.5	115.9
STS-GCN	**74.9**	**86.2**	**93.5**	**57.0**	**67.4**	**75.2**	**73.9**	**86.2**	**94.3**	**57.4**	**67.2**	**76.9**
SBR	79.6	90.3	96.4	58.3	67.9	76.1	76.7	90.3	98.4	61.6	74.8	80.4
INP	83.2	91.7	97.9	65.8	72.0	76.0	85.3	93.3	96.2	69.8	75.5	81.0
MT	75.3	86.1	93.6	57.8	67.6	75.9	74.3	86.1	93.9	57.6	67.1	75.6
Ours (STS-GCN + TCSL)	**74.7**	**85.2**	**90.9**	**56.2**	**66.2**	**71.4**	**73.6**	**84.5**	**91.5**	**56.1**	**65.2**	**72.5**

Actions	Waiting			Walking Dog			Walking Together			**Average**		
msec	720	880	1000	720	880	1000	720	880	1000	720	880	1000
ConvSeq2Seq	100.3	110.7	117.7	133.8	151.1	162.4	77.7	82.9	87.4	104.7	116.7	124.2
LTD	90.6	101.1	108.3	120.3	136.3	146.4	60.3	63.1	65.7	93.6	105.2	112.4
RNN-GCN	89.0	100.3	108.2	120.6	135.9	146.9	57.8	62.0	64.9	91.8	104.1	112.1
STS-GCN	**56.8**	**66.1**	**72.0**	**85.7**	**96.2**	**102.6**	**44.0**	**48.2**	**51.1**	**60.1**	**68.9**	**75.6**
SBR	59.3	69.1	75.6	91.2	100.3	106.3	47.2	52.6	57.2	63.2	72.9	79.2
INP	64.9	72.6	76.8	93.6	101.3	107.0	53.1	55.2	58.9	69.2	75.5	80.2
MT	57.8	66.6	72.2	86.4	96.6	103.3	44.9	**48.7**	**51.3**	60.8	69.4	76.0
Ours (STS-GCN + TCSL)	**55.3**	**64.3**	**70.3**	**86.2**	**93.9**	**100.7**	**43.9**	48.9	**51.3**	**59.4**	**68.1**	**73.4**

Table 1: MPJPE error in mm for prediction of 3D joint positions on Human3.6M. **TCSL** is short for temporal constrained subspace learning. We report results for the future 720, 880, and 1000 milliseconds, respectively. The first four methods are commonly used pose forecasting methods, and the last four methods are constrained learning methods.

are both used as the regularization, which shows the regularization in the constrained subspace can improve the learning performance.

Projection functions Many projection functions can be applied as long as they are differentiable in the training stage. Except for the projection function described in Eq. (8), we also consider the function constrained by the exponential linear unit (ELU), *i.e.*,

$$\mathcal{P}_{\text{elu}}(\mathcal{F}_{\boldsymbol{w}}^{''}) = \text{ELU}\left(2\tilde{\delta}_{ub} - \text{ELU}(\tilde{\delta}_{ub} - \mathcal{F}_{\boldsymbol{w}}^{''}; \alpha); \alpha\right) - \tilde{\delta}_{ub}, \tag{11}$$

where $\text{ELU}(x;\alpha) = \max(0,x) + \min(0, \alpha(\exp(x)-1))$ and α is set to 1 by default. To verify the effectiveness of different projection functions, *i.e.*, $\mathcal{P}_{\text{relu}}$ and $\mathcal{P}_{\text{elu}}$, we show the results in Table 4, along with the combination of different regularization losses. The results that $\mathcal{P}_{\text{elu}}$ achieves worse performance

Actions	AMASS-BMLrub			3DPW		
msec	720	880	1000	720	880	1000
ConvSeq2Seq	87.0	91.5	93.5	77.0	83.6	87.8
LTD	65.7	71.3	75.2	65.8	71.5	75.5
RNN-GCN	58.6	63.4	67.2	63.6	69.7	73.7
STS-GCN	**38.1**	**42.7**	**45.5**	**35.7**	**39.6**	**42.3**
SBR	38.7	43.7	53.8	36.3	40.7	48.9
INP	37.9	43.2	45.8	35.8	40.3	42.9
MT	39.5	44.2	46.4	45.4	46.6	51.9
Ours	**37.5**	**42.6**	**45.4**	**35.5**	**40.1**	**42.5**

Table 2: Left: average MPJPE in mm over the BMLrub test sequences. Right: average MPJPE in mm, testing the generalizability on 3DPW of models trained on AMASS.

msec	720	880	1000
STS-GCN	1.55	2.10	3.22
SBR	1.17	0.92	0.78
INP	9.67	10.74	9.67
MT	1.24	1.30	2.25
Ours	**0.00**	**0.00**	**0.00**

Table 3: Infeasible rate (%) on Human3.6M dataset.

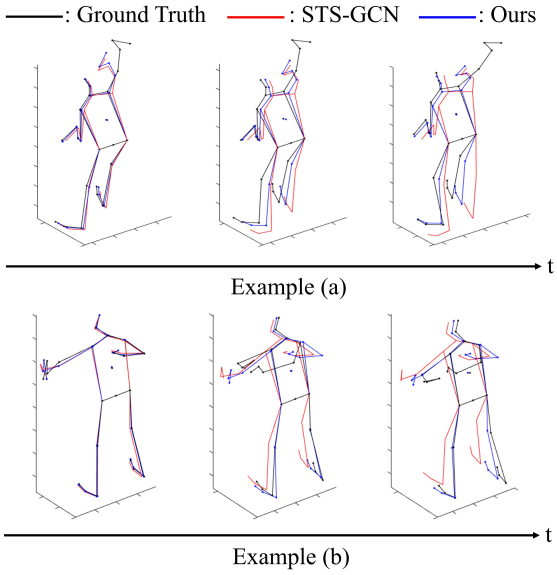

Figure 4: Visualization examples on the Human3.6M dataset. Each subfigure represents the predicted pose along with the time. Black, red and blue colors represent the ground truth pose, predicted pose by STS-GCN, and predicted pose by our method. Obviously, our method produces smaller errors.

than $\mathcal{P}_{\text{relu}}$ for each loss combination are reasonable. When the target output of $\mathcal{P}_{\text{elu}}$ is touching the bound of the constraint, it will push the input to positive infinity or negative infinity in the back-propagation, causing instability in training and degradation on the performance.

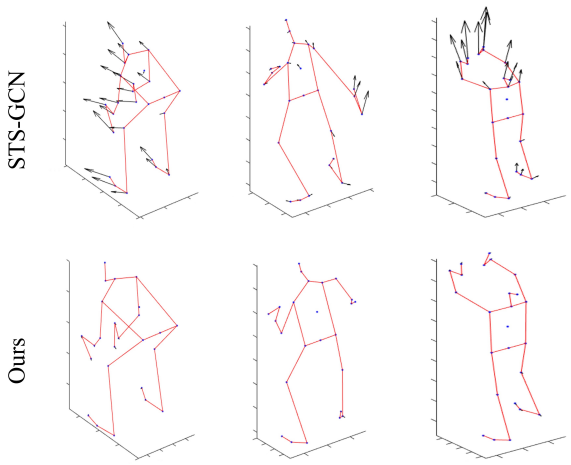

Figure 5: The visualization of the change of velocities for the joints. The first and second row shows examples of predicted poses by STS-GCN and our method, respectively. We calculate the change of velocity, *i.e.*, $\boldsymbol{X}_f^{''}$, on the predicted pose along x, y, z and then draw arrows on the joints, where the arrow direction represents the direction of $\boldsymbol{X}_f^{''}$ and the arrow length represents the magnitude of $\boldsymbol{X}_f^{''}$. Compared with STS-GCN, ours can generate only a subtle change of velocities, showing the effectiveness of our constrained solution.

Regularization Loss		Projection	
$\mathcal{L}_v$	$\mathcal{L}_r$	$\mathcal{P}_{\text{relu}}$	$\mathcal{P}_{\text{elu}}$
		78.2	78.6
✓		74.3	75.5
✓	✓	**73.4**	**73.8**

Table 4: MPJPE error in mm for prediction of 3D joint positions on Human3.6M with variant regularization terms and projection functions for training. The best performance is shown in bold.

5 Conclusion

In this paper, we propose a temporal constrained feasible subspace learning for the human pose forecasting task. The temporal constrained subspace is explicitly formulated with projection operations. Then the final prediction is obtained by an inverse transformation. The proposed method can guarantee feasible solutions, and no iterative steps are required in the inference stage. The efficiency and effectiveness are verified on the Human3.6M, AMASS, and 3DPW datasets. Ablation studies show the importance of different projection functions and regularization terms. We hope this research can open up more exploration of real-world constrained problems.

Limitations and future work Though good performance has been achieved for the proposed feasible solutions, there are a few limitations. First, we only consider the temporal constraints in the current work. For human pose models, there are also spatial constraints among different joints. Second, only human pose forecasting problems are considered in the current work. We plan to consider spatial constraints and extend our method to other tasks in future work.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant (2022ZD0162000), the National Natural Science Foundation of China (62106219), and the Fundamental Research Funds for the Central Universities (226-2023-00045, 2021FZZX001-23, 226-2023-00048).

References

[Adeli *et al.*, 2020] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5(4):6033–6040, 2020.

[Adeli *et al.*, 2021] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021.

[Amos and Kolter, 2017] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[Berk *et al.*, 2017] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[Bouazizi *et al.*, 2022] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 791–798. International Joint Conferences on Artificial Intelligence Organization, 7 2022.

[Chai *et al.*, 2023] Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. *arXiv preprint arXiv:2303.16456*, 2023.

[Chiu *et al.*, 2019] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

[Detassis *et al.*, 2020] Fabrizio Detassis, Michele Lombardi, and Michela Milano. Teaching the old dog new tricks: supervised learning with constraints. In *NeHuAI@ ECAI*, pages 44–51, 2020.

[Diligenti *et al.*, 2017] Michelangelo Diligenti, Marco Gori, and Claudio Sacca. Semantic-based regularization for learning and inference. *Artificial Intelligence*, 244:143–165, 2017.

[Diller *et al.*, 2020] Christian Diller, Thomas Funkhouser, and Angela Dai. Forecasting characteristic 3d poses of human actions. *arXiv preprint arXiv:2011.15079*, 2020.

[Gu *et al.*, 2019] Renshu Gu, Gaoang Wang, Zhongyu Jiang, and Jenq-Neng Hwang. Multi-person hierarchical 3d pose estimation in natural videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):4245–4257, 2019.

[Gu *et al.*, 2021] Renshu Gu, Gaoang Wang, and Jenq-Neng Hwang. Exploring severe occlusion: Multi-person 3d pose estimation with gated convolution. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8243–8250. IEEE, 2021.

[Huang *et al.*, 2021] Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. Equivariant graph mechanics networks with constraints. In *International Conference on Learning Representations*, 2021.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[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.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Koppula and Saxena, 2013] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities for reactive robotic response. In *IROS*, page 2071. Tokyo, 2013.

[Kotary *et al.*, 2021a] James Kotary, Ferdinando Fioretto, and Pascal Van Hentenryck. Learning hard optimization problems: A data generation perspective. *Advances in Neural Information Processing Systems*, 34, 2021.

[Kotary *et al.*, 2021b] James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378*, 2021.

[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.

[Li *et al.*, 2022a] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European Conference on Computer Vision*, pages 18–36. Springer, 2022.

[Li *et al.*, 2022b] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 2022.

[Loper *et al.*, 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[Mahmood *et al.*, 2019] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[Mangalam *et al.*, 2020] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2784–2793, 2020.

[Mao *et al.*, 2019] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[Mao *et al.*, 2020] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020.

[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.

[Paden *et al.*, 2016] Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles*, 1(1):33–55, 2016.

[Parsaeifard *et al.*, 2021] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2303, 2021.

[Rubagotti *et al.*, 2019] Matteo Rubagotti, Tasbolat Taunyazov, Bukeikhan Omarali, and Almas Shintemirov. Semi-autonomous robot teleoperation with obstacle avoidance via model predictive control. *IEEE Robotics and Automation Letters*, 4(3):2746–2753, 2019.

[Rubanova *et al.*, 2021] Yulia Rubanova, Alvaro Sanchez-Gonzalez, Tobias Pfaff, and Peter Battaglia. Constraint-based graph network simulator. *arXiv preprint arXiv:2112.09161*, 2021.

[Sangalli *et al.*, 2021] Sara Sangalli, Ertunc Erdil, Andeas Hötker, Olivio Donati, and Ender Konukoglu. Constrained optimization to train neural networks on critical and under-represented classes. *Advances in Neural Information Processing Systems*, 34, 2021.

[Sofianos *et al.*, 2021] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021.

[Troje, 2002] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2–2, 2002.

[Unhelkar *et al.*, 2018] Vaibhav V Unhelkar, Przemyslaw A Lasota, Quirin Tyroller, Rares-Darius Buhai, Laurie Marceau, Barbara Deml, and Julie A Shah. Human-aware robotic assistant for collaborative assembly: Integrating human motion prediction with planning in time. *IEEE Robotics and Automation Letters*, 3(3):2394–2401, 2018.

[von Marcard *et al.*, 2018] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[Wang *et al.*, 2019] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7124–7133, 2019.

[Wang *et al.*, 2022] Hongsong Wang, Liang Wang, Jiashi Feng, and Daquan Zhou. Velocity-to-velocity human motion forecasting. *Pattern Recognition*, 124:108424, 2022.

[Xu *et al.*, 2022] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 251–269. Springer, 2022.

[Yang *et al.*, 2020] Shuqi Yang, Xingzhe He, and Bo Zhu. Learning physical constraints with neural projections. *Advances in Neural Information Processing Systems*, 33:5178–5189, 2020.

[Yuan and Kitani, 2019] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10082–10092, 2019.

[Zhang *et al.*, 2022] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022.

[Zhong *et al.*, 2021] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Extending lagrangian and hamiltonian neural networks with differentiable contact models. *Advances in Neural Information Processing Systems*, 34, 2021.