17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Data Augmentation for Human Motion Prediction

Takahiro Maeda Norimichi Ukita
Toyota Technological Institute
Nagoya, Japan
{sd19445, ukita}@toyota-ti.ac.jp

Abstract

Human motion prediction is seldom deployed to real-world tasks due to difficulty in collecting a huge amount of motion data. We propose two motion data augmentation approaches using Variational AutoEncoder (VAE) and Inverse Kinematics (IK). Our VAE-based generative model with adversarial training and sampling near samples generates various motions even with insufficient original motion data. Our IK-based augmentation scheme allows us to semi-automatically generate a variety of motions. Furthermore, we correct unrealistic artifacts in the augmented motions. As a result, our method outperforms previous noise-based motion augmentation methods.

1 Introduction

Human motion prediction, which forecasts future body poses based on past poses, is a key technique for human-robot interaction [1–5], autonomous driving [6], VR/AR applications [7], performance capture [8, 9], etc. These applications are still limited because of the lack of motion data, which results in low prediction accuracy. This is because the acquisition of motion data requires a vast amount of costs such as the motion capture equipment and post-processing such as denoising.

Data Augmentation (DA) is useful for alleviating the data insufficiency [10–12]. However, as far as we know, there is only one less-effective augmentation method proposed for human motion prediction [13]. This paper presents two new augmentation approaches using Variational AutoEncoder (VAE) [14] and Inverse Kinematics (IK), which are shown in "Motion Augmentation" in Fig. 1. Although most of our augmented motions are physically plausible, we observed some of them have unrealistic artifacts. These artifacts are corrected with imitation learning and physics simulation ("Motion Correction" in Fig. 1) in our proposed method.

Our contributions are as follows:

1. **VAE-based human-motion augmentation:** Our generative model with adversarial training and sampling near samples can generate plausible motions even with insufficient motion data.
2. **IK-based human-motion augmentation:** This requires less effort just deciding the target sampling space and keyframe rather than annotating IK targets for all frames.

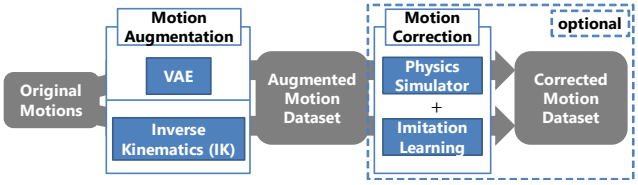

Figure 1. Overview of our proposed motion data augmentation. Original motions are augmented independently using VAE and IK. Both VAE and IK are superior to previous augmentation approaches using additive noise. Optionally, we also propose motion correction where the augmented motions are modified to be physically plausible using imitation learning and a physics simulator.

3. **Motion correction for physical reality:** Our method is designed to improve the prediction accuracy by maintaining the physical reality of augmented motions.

2 Related Work

Motion augmentation: Fragkiadaki *et al.* [13] proposed to corrupt input motions with zero-mean Gaussian noise. This simple additive noise might lose contexts represented in the input motion.

Data augmentation with GAN: In image classification tasks, generative models such as GAN are used for data augmentation [15–18]. This approach can be applied to other tasks including prediction.

Inverse Kinematics (IK): IK modifies the pose of the whole body so that key points in the body reach their target positions. A motion can be also modified by providing the target positions in all frames of the motion [19]. Although IK can greatly modify each pose and potentially be useful for motion augmentation, it is impractical to annotate the target positions in all frames included in a training dataset.

Motion synthesis with physics simulation: For better accuracy and reliability, motion prediction models should output physically plausible motions. Recent deep reinforcement learning enables a physically simulated character to imitate diverse motions [20–23]. Such physics simulation might improve the quality of augmented motions.

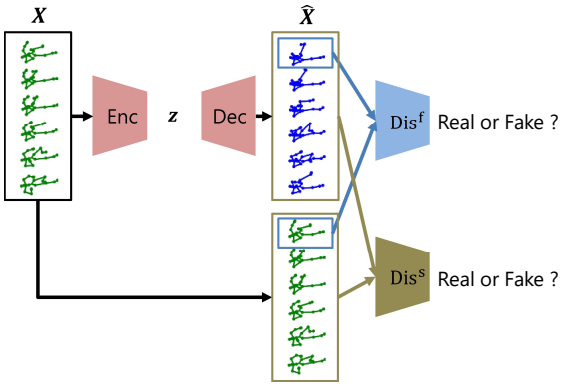

Figure 2. Proposed VAE-based network architecture with adversarial training:

3 Proposed Motion Augmentation

We propose two independent augmentation approaches with VAE and IK. Furthermore, the motion correction can be applied optionally to augmented motions. We also temporally expand and contract motion sequences in the range from 10% longer to 10% shorter as temporal data augmentation. The details are explained in the following subsections.

3.1 DA with VAE

Although GAN is widely used as a generative model, we found that, for motion prediction, GAN often produces only static motions where all poses are almost identical due to data insufficiency and GAN's training instability (i.e., mode collapse). Instead, we propose a VAE-based model. Our VAE-based model successfully generates diverse within-class motions with adversarial training and sampling near samples.

Adversarial training: Our proposed network is shown in Fig. 2. We validated that VAE with adversarial training can generate more realistic motions than those of the vanilla VAE. The encoder outputs a latent variable $\boldsymbol{z}$ from an input motion $\boldsymbol{X} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ where each $\boldsymbol{x}_t$ denotes a pose vector in t-th frame. The decoder reconstructs the motion $\widehat{\boldsymbol{X}}$ from $\boldsymbol{z}$. Frame-wise and sequence-wise discriminators (denoted by Dis^f and Dis^s, respectively, in Fig. 2) discriminate $\boldsymbol{X}$ from $\widehat{\boldsymbol{X}}$ for improving $\widehat{\boldsymbol{X}}$

Sampling near samples in the latent space: It is not easy to determine the appropriate dimension of the latent space in general. In the vanilla VAE, the latent representation $\boldsymbol{z}$ is sampled from the normal distribution with zero mean and unit variance $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. While the high-dimensional representation leads to the sparsity of training data, which makes it difficult to sample realistic data, the low-dimensional one generates inaccurate data.

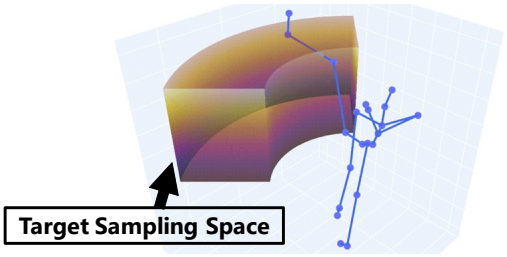

Figure 3. Sampling space of target points for the action class kick. The target space is a fan-shaped one to which a foot end-effector may reach.

To solve the aforementioned problem, we propose a method for sampling from only regions that are appropriately represented by training data in the high-dimensional latent space. In this method, each motion in the training data is encoded into mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$. Given $\overline{\boldsymbol{\mu}}$ and $\overline{\boldsymbol{\sigma}^2}$ that are respectively the mean of $\boldsymbol{\mu}$ and the mean of $\boldsymbol{\sigma}^2$ over randomly-sampled n motions, the latent representation $\boldsymbol{z}$ is drawn from $\mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}^2})$, and $\boldsymbol{z}$ is fed into the decoder for generating $\boldsymbol{X}_{\text{aug}}$, as expressed by Eqs. (1) and (2). In our experiments, $n = 2$.

$$\boldsymbol{X}_{\text{aug}} = \text{Dec}(\boldsymbol{z}) \tag{1}$$

$$\boldsymbol{z} \sim \mathcal{N}(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\sigma}^2}) \tag{2}$$

3.2 DA with IK

The proposed IK-based motion editing needs target positions in all frames of a motion. To achieve this semi-automatically, we present a new effortless IK-based augmentation method that requires a user only to provide a target sampling space $\mathbb{P}$ and a keyframe $\boldsymbol{x}_{\text{key}}$ on each class. Examples of a kick class are shown in Figs. 3 and 4. The user determines the target sampling space as shown in Fig. 3. Then, a keyframe where a kicking foot reaches the farthest position from the body is given with the key pose.

Given the sampling space and the keyframe, the IK target position $\boldsymbol{p}_{\text{key}} \in \mathbb{P}$ for the keyframe is randomly sampled. Target positions $\boldsymbol{p}_t$ for all frames are determined by propagating the differences between $\boldsymbol{p}_{\text{key}}$ and the end-effector positions at 0-th and T-th frames backward and forward, respectively, in a linearly-decreasing manner, as shown in Fig. 4 and expresed as follows:

$$\boldsymbol{p}_{\text{diff}} = \left\{ \boldsymbol{p}_{\text{key}} - \text{FK}(\hat{\boldsymbol{x}}_{\text{key}}, j) \right\} \tag{3}$$

$$\boldsymbol{p}_t = \text{FK}(\hat{\boldsymbol{x}}_t, j) + \boldsymbol{p}_{\text{diff}} \cdot f(t_{\text{key}}, t) \tag{4}$$

$$f(t_{\text{key}}, t) = \begin{cases} \dfrac{t}{t_{\text{key}}} & \text{if } t \leq t_{\text{key}} \\ \dfrac{n-t}{n-t_{\text{key}}} & \text{if } t > t_{\text{key}} \end{cases} \tag{5}$$

$$\boldsymbol{X}^{\text{np}} = \{\text{IK}(\hat{\boldsymbol{x}}_1, j, \boldsymbol{p}_1), \ldots, \text{IK}(\hat{\boldsymbol{x}}_T, j, \boldsymbol{p}_T)\}, \tag{6}$$

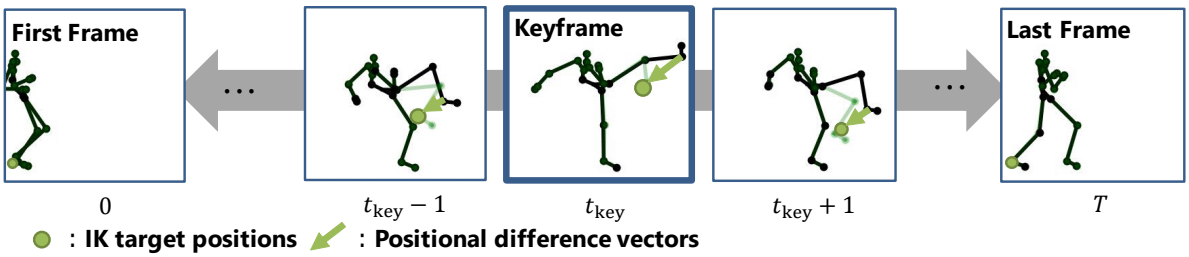

Figure 4. Overview of our sequential IK scheme. Given a keyframe and the body pose in the keyframe, body poses in all other frames are automatically determined by IK.

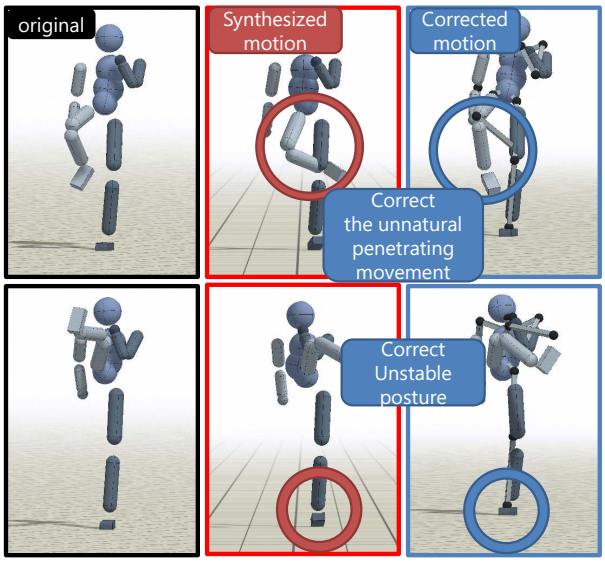

Figure 5. Examples of our motion correction. Upper: Foot penetrate each others. Lower: Unstable pose.

where j is an end-effector, $\boldsymbol{p} = FK(\boldsymbol{x}, j)$ is a forward kinematics function and $\widehat{\boldsymbol{x}} = IK(\boldsymbol{x}, j, \boldsymbol{p})$ is an inverse kinematics function. We apply IK with $\boldsymbol{p}_t$ to all frames and obtain an augmented motion $\boldsymbol{X}_{\text{aug}}$ where the end-effector smoothly reaches $\boldsymbol{p}_{\text{key}}$.

3.3 Motion Correction with Imitation Learning using Physics Simulation

Although most augmented motions generated by our method are physically realistic, some of them are not. For example, footskating by VAE, mutual penetrations between body parts by IK, and unstable poses by both VAE and IK are observed empirically. Our motion correction method enhances the physical reality of these augmented motions as shown in Fig. 5.

Peng *et al.* [20] proposed an imitation learning scheme that allows a physically simulated character to mimic various motions. Given a goal motion (e.g., motion measured by a motion capture system), imitation learning learns a policy that modifies the pose of the character at $t+1$ from its body status at t so that the sequence of the modified poses gets close to the goal motion. Then, to control the character toward the modified pose at each moment under physical constraints, a PD controller suggests physical torques given to the character at t. While the motion is modified for compensating the difference between the body properties of the goal motion and the character in DeepMimic [20], the goal motion is already physically plausible because it is measured by a motion capture system. On the other hand, we apply this imitation learning using physics simulation to correct physically implausible motions produced by our method. To cope with this more challenging problem, our imitation learning scheme employs Residual Force Control (RFC) [23] which maintains the physical stability (e.g., fall-prevention) by applying additional external forces to the character.

Table 1. Quantitative evaluation of augmented motions.

	Min DTW dist.	MMD
Original Motions	2.92	-
GAN	2.90	4.28
VAE	2.73	1.94
VAE + adv training	2.71	1.37
VAE + sampl. near samples	2.72	0.85
VAE + both (proposed)	**2.70**	**0.20**

While DeepMimic using RFC allows us to generate physically stable motions, the convergence of DeepMimic is usually more than one day for correcting one motion with 10 CPU threads. This is a critical problem when we augment a large number of motions. The dominant cost in this convergence time is on learning the policy network. While the character pose at $t+1$ modified by the policy network inevitably differs from the goal motion for satisfying physical constraints, the residual between these modified and goal poses is usually small. Based on this observation, we change the policy network so that it provides the residual between the modified and goal poses instead of the modified

Table 2. Quantitative results of the effectiveness of proposed data augmentation on human motion prediction.

action class	punch				kick				walk			
timesteps[ms]	100	200	300	400	100	200	300	400	100	200	300	400
No aug	1.31	1.87	2.18	2.33	1.08	1.68	2.05	2.26	0.52	0.88	1.11	1.23
Noise	1.31	1.90	2.23	2.35	1.06	1.65	2.02	2.25	0.52	0.87	1.09	1.21
VAE	1.29	1.82	**2.15**	**2.30**	1.06	1.63	1.97	2.17	0.52	0.88	1.10	1.22
IK	1.24	1.80	2.20	2.42	0.96	1.39	1.60	1.73	**0.47**	0.77	**0.94**	**1.05**
VAE & IK	**1.21**	**1.75**	2.19	2.45	**0.95**	**1.38**	**1.59**	**1.71**	**0.47**	**0.76**	**0.94**	1.06
VAE corrected	1.30	1.86	2.20	2.38	1.03	1.60	1.95	2.14	0.53	0.89	1.12	1.25
IK corrected	1.35	1.99	2.39	2.61	1.06	1.66	2.01	2.20	0.53	0.91	1.15	1.30
VAE & IK corrected	1.35	2.00	2.37	2.61	1.03	1.65	1.98	2.17	0.54	0.95	1.21	1.37

pose itself. This residual learning is much easier than learning arbitrary modified poses. The convergence time is around one-third (∼8 hours) compared to the original imitation learning using RFC.

4 Experiments

Our experiments consist of two parts: (1) Ablation study on our VAE-based method (Sec. 3.1). The effects of several components of our proposed method are validated in terms of physical and contextual closeness relationships. (2) Performance evaluation in motion prediction with different data augmentations.

Dataset: Our experiments were conducted on HDM05 Motion Database [24]. HDM05 is a relatively small but challenging dataset with dynamic motions compared to other common benchmarks such as Human3.6M [25]. We tested our method with 5 fold cross-validation so that our models were trained on motions from 4 subjects and tested on those from 1 subject. Motions of *punch*, *kick* and *walk* action classes were resampled to 30Hz and used for the experiments. We augmented the train set to 10 times larger by VAE or IK methods.

Implementation Details: All encoders, decoders, and discriminators consist of 256-D LSTM cells and MLP layers. The dimension of a latent space is 128 for VAE. A noise dimension for GAN is also 128. We used the Adam optimizer to train models for 10,000 epochs.

4.1 Motion Augmentation by VAE

The effectiveness of our VAE-based motion augmentation is validated by ablation. For comparison, a GAN-based method is also evaluated.

Metrics: The quality of generated motions is evaluated with two metrics: the minimum Dynamic Time Warping (DTW) [26] distance and the Maximum Mean Discrepancy (MMD) [27]. The minimum DTW distance is calculated so that the augmented motion closest to each test motion is found from all augmented motions. For DTW, frame-wise distances are calculated based on the Euclidean distance in the Euler angle. The mean over all test motions is shown in Table 1. MMD is a distance between two distributions. Table 1 shows MMD between the test set and the augmented set. Both metrics evaluate whether the augmented motions deviate from the domain of real observed motions. The lower score is better in both metrics.

Results: Table 1 shows the proposed VAE-based method with adversarial training and sampling near samples gets the best performance in both minimum DTW distance and MMD. Meanwhile, a GAN-based method increases the minimum DTW distance and gets the highest MMD probably because the training dataset is too small for GAN to learn various patterns.

4.2 Motion Prediction with DA

Prediction Model and Metrics: We use the SOTA human motion prediction model [28] to evaluate the effectiveness of our motion DA method. We follow the standard evaluation protocol used in [13, 29], and report the Euclidean distance between the predicted and ground-truth joint angles in Euler angle representation.

Results: Table. 2 shows quantitative results for human motion prediction with all data augmentations. The prediction errors are shown on 4 timesteps (100, 200, 300, 400ms) and 3 action classes (punch, kick, walk). In most cases, the combination of our proposed data augmentations achieved the lowest prediction error. However, the motion correction couldn't improve the prediction accuracy because the corrected motions have fewer motion patterns than the augmented motions. We conclude that physical reality has less importance than variation in terms of accuracy.

5 Conclusion

In this work, we presented two new human motion augmentation approaches using VAE and IK. The motion correction method is also proposed to fix unrealistic artifacts of augmented motions. Experiments showed that our augmentation outperformed previous methods but the motion correction couldn't improve prediction accuracy. However, animations or humanoid robots can utilize our physically plausible motions. We find our approaches encouraging the applications of human motion by cutting the data acquisition costs. Our future work will focus on a new data augmentation approach and fully automatic setting of IK targets.

References

[1] Hongyi Liu and Lihui Wang. Human motion prediction for human-robot collaboration. *Journal of Manufacturing Systems*, 44:287–294, 2017.

[2] Wansoo Kim, Jinoh Lee, Luka Peternel, Nikolaos G. Tsagarakis, and Arash Ajoudani. Anticipatory robot assistance for the prevention of human static joint overloading in human-robot collaboration. *IEEE Robotics Autom. Lett.*, 3(1):68–75, 2018.

[3] Luka Peternel, Wansoo Kim, Jan Babič, and Arash Ajoudani. Towards ergonomic control of human-robot co-manipulation and handover. In *IEEE ICHR*, 2017.

[4] Marta Lorenzini, Wansoo Kim, Elena De Momi, and Arash Ajoudani. A synergistic approach to the real-time estimation of the feet ground reaction forces and centers of pressure in humans with application to human-robot collaboration. *IEEE Robotics Autom. Lett.*, 3(4): 3654–3661, 2018.

[5] Stefano Tortora, Stefano Michieletto, Francesca Stival, and Emanuele Menegatti. Fast human motion prediction for human-robot collaboration with wearable interface. In *IEEE CIS*, pages 457–462, 2019.

[6] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, pages 6120–6127, 2019.

[7] Xueshi Hou and Sujit Dey. Motion prediction and pre-rendering at the edge to enable ultra-low latency mobile 6dof experiences. *IEEE Open J. Commun. Soc.*, 1:1674–1690, 2020.

[8] Norimichi Ukita, Michiro Hirai, and Masatsugu Kidode. Complex volume and pose tracking with probabilistic dynamical models and visual hull constraints. In *ICCV*, 2009.

[9] Norimichi Ukita and Takeo Kanade. Gaussian process motion graph models for smooth transitions among multiple actions. *Comput. Vis. Image Underst.*, 116(4):500–509, 2012.

[10] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1106–1114, 2012.

[12] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *Arxiv preprint*, 1712.04621, 2017.

[13] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015.

[14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[15] Antreas Antoniou, Amos J. Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *Arxiv preprint*, 1711.04340, 2017.

[16] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger N. Gunn, Alexander Hammers, David Alexander Dickie, Maria del C. Valdés Hernández, Joanna M. Wardlaw, and Daniel Rueckert. GAN augmentation: Augmenting training data using generative adversarial networks. *Arxiv preprint*, 1810.10863, 2018.

[17] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and A. Cristiano I. Malossi. BAGAN: data augmentation with balancing GAN. *Arxiv preprint*, 1803.09655, 2018.

[18] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle J. Palmer, and Ian D. Reid. A bayesian data augmentation approach for learning deep models. In *NeurIPS*, pages 2797–2806, 2017.

[19] Michael Gleicher. Motion path editing. In *SI3D*, pages 195–202, 2001.

[20] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: example-guided deep reinforcement learning of physics-based character skills. *ACM Trans. Graph.*, 37(4):143:1–143:14, 2018.

[21] Kevin Bergamin, Simon Clavet, Daniel Holden, and James Richard Forbes. Drecon: data-driven responsive control of physics-based characters. *ACM Trans. Graph.*, 38(6):206:1–206:11, 2019.

[22] Seunghwan Lee, Moon Seok Park, Kyoung-Min Lee, and Jehee Lee. Scalable muscle-actuated human simulation and control. *ACM Trans. Graph.*, 38(4):73:1–73:13, 2019.

[23] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *NeurIPS*, 2020.

[24] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE PAMI*, 36(7):1325–1339, jul 2014.

[26] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[27] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[28] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *IEEE ICCV*, pages 9488–9496, 2019.

[29] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE CVPR*, pages 4674–4683, 2017.