2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

A generic diffusion-based approach for 3D human pose prediction in the wild

Saeed Saadatnejad[1], Ali Rasekh[2], Mohammadreza Mofayezi[2], Yasamin Medghalchi[2], Sara Rajabzadeh[2], Taylor Mordan[1] and Alexandre Alahi[1]

***Abstract*— Predicting 3D human poses in real-world scenarios, also known as human pose forecasting, is inevitably subject to noisy inputs arising from inaccurate 3D pose estimations and occlusions. To address these challenges, we propose a diffusion-based approach that can predict given noisy observations. We frame the prediction task as a denoising problem, where both observation and prediction are considered as a single sequence containing missing elements (whether in the observation or prediction horizon). All missing elements are treated as noise and denoised with our conditional diffusion model. To better handle long-term forecasting horizon, we present a temporal cascaded diffusion model. We demonstrate the benefits of our approach on four publicly available datasets (Human3.6M, HumanEva-I, AMASS, and 3DPW), outperforming the state-of-the-art. Additionally, we show that our framework is generic enough to improve any 3D pose prediction model as a pre-processing step to repair their inputs and a post-processing step to refine their outputs. The code is available online: https://github.com/vita-epfl/DePOSit.**

I. Introduction

Robots and humans are poised to work in close proximity. Yet, current technology struggles to read and anticipate the motion dynamics of humans. Predicting 3D human poses enables a safe co-existence between humans and robots, with direct applications in social robotics [13], autonomous navigation [38], assistive robotics [57], [59], and human-robot interaction [9], [30].

Predicting a sequence of future 3D poses of a person given a sequence of past observed ones, also referred to as human pose forecasting, is a challenging task since it must combine spatial and temporal reasoning to output multiple plausible outcomes. Previous models have yielded satisfactory results [39], [36], yet they fail to produce acceptable outcomes in noisy settings. Minor offsets from detection methods or partial occlusions of body parts can drastically impact the prediction accuracy.

Denoising Diffusion Probabilistic Models (DDPMs) [25] are one type of generative models that can denoise input signals iteratively. Motivated by this property, we propose a diffusion model that explicitly handles noisy data input so that it not only predicts accurate and in-distribution poses, but can also be used in the wild. As depicted in Figure 1, we construct a full sequence of observation and future frames where noise is placed in the missing observation elements and future poses. Our model denoises this sequence in several steps and produces the correct predictions. Naively predicting all future frames simultaneously results in inaccurate predictions in later frames. Hence, we propose a model comprised of two temporally-cascaded diffusion blocks. The first block predicts the short-term poses and repairs the noisy observations (if applicable), while the second block uses the output from the former as a condition to predict the long-term poses. We also leverage our model in a generic framework that can improve the performance of state-of-the-art prediction models in a black-box manner. To this end, we use our diffusion-based model as a pre-processing step to repair the observations providing pseudo-clean data for the prediction model to make more reliable predictions. Our model can then be used as a post-processing step to further refine these predictions.

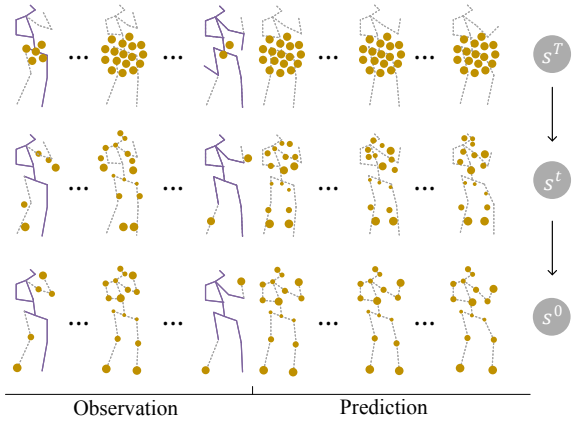

Fig. 1: Our proposed conditional diffusion model denoises the input sequence s^T over T steps by simultaneously 1) predicting poses for the future frames and 2) repairing the noisy observations in the case of partial occlusion (first column), missing whole frame (second column), or inaccurate observations (third column). The large yellow circles depict the Gaussian noise we consider for unavailable joints, which gradually become smaller and fit into the correct locations.

To summarize, our contributions are three-fold:

- We frame the 3D human pose prediction task as a denoising problem.
- We propose a two-stage diffusion model outperforming the state-of-the-art in both clean and noisy observation settings.
- We introduce a generic framework that leverages our model through pre-processing (repairing the input) and post-processing (refinement), which can enhance any pose prediction model.

[1]EPFL, Lausanne, Switzerland (e-mail: saeed.saadatnejad at epfl.ch)
[2]The research was conducted during an internship at EPFL

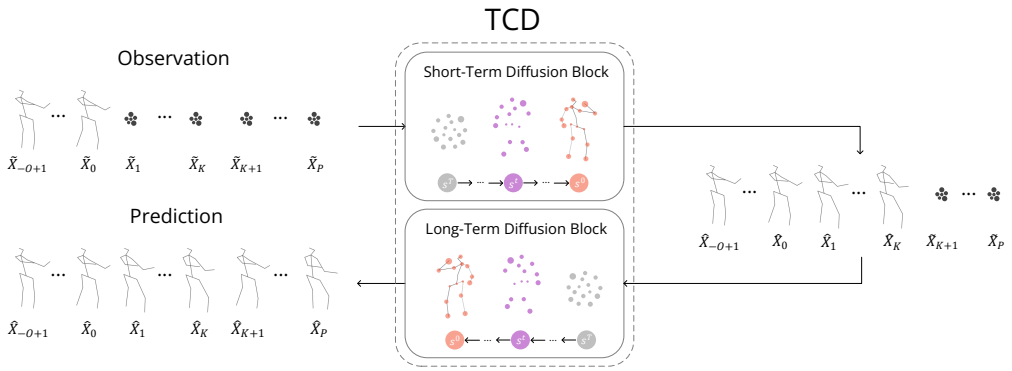

Fig. 2: Overview of our Temporal Cascaded Diffusion (TCD). The short-term diffusion block (top) takes the observed sequence padded with random noise and predicts short-term human poses in K frames. The predicted sequence along with the observation padded with random noise is given to the long-term diffusion block (bottom) to predict for all P frames.

II. Related Work

Predicting a sequence of future center positions at a coarse-grained level [29], [48], [5] or a sequence of bounding boxes [8], [49] have been extensively studied in the literature. However, in this work, we focus on a more fine-grained prediction, namely 3D pose. Recurrent Neural Networks (RNNs) have been widely used [20], [27], [42], [12], [21], [14] as they are capable of capturing the temporal dependencies in sequential data, and later networks with only feed-forward networks were introduced [31]. Subsequently, Graph Convolutional Networks (GCNs) were proposed to better capture the spatial dependencies of body poses [41], [17], [39], [33]. Separating temporal and spatial convolution blocks [36], and trainable adjacency matrices [54], [64] are among other proposed ideas. Attention-based approaches have recently gained interest for modeling human motion [43], [46] and showed a huge improvement with spatio-temporal self-attention module [39]. Our proposed model also incorporates attention. While various works have employed context information [11], [23], [15], social interactions [1] or action classes [2], [10] as conditions, this paper focuses on conditioning solely on the observation sequences.

Deterministic models [39], [36] offer satisfactory prediction accuracy, yet they lack the ability to generate diverse and multi-modal outputs compared to stochastic models [63], [4], [3], [52], [35], [40], [60]. In this category, Variational AutoEncoders (VAEs) have been widely adopted due to their strength in representation learning [45], [63], [4], [3]. Generative models, particularly diffusion models, have been recently utilized to model data distributions with remarkable results in image synthesis [18], [50], image repainting [34] and text-to-image generation [51], [47]. Recently, they have been used for time-series imputation [55], i.e., filling in missing elements. However, it was not explored for human motion. To the best of our knowledge, we are the first to propose a diffusion model for human pose prediction, which outperforms both stochastic and deterministic models.

Previous models perform poorly with partial noisy observations. A multi-task learning approach has been recently suggested in [16] to address this issue, by implicitly disregarding noise in the data. We provide detailed comparisons with [16], and show that explicitly denoising the input leads to a generalizable solution, and that our temporally-cascaded diffusion blocks better capture the spatio-temporal relationships in the poses. Furthermore, we present a generic framework that can be used to improve any existing state-of-the-art model in a black-box manner.

III. Method

In this section, we first describe the notations and conditional diffusion blocks, which are the fundamental elements of our model. We then present our model and finally introduce our generic framework.

A. Problem Definition and Notations

Let $X = [X_{-O+1}, X_{-O+2}, \dots, X_0, X_1, \dots, X_P] \in \mathbb{R}^{(O+P)\times J\times 3}$ be a clean complete normalized sequence of human body poses with J joints in O frames of observation and P frames of future. Each joint consists of its 3D cartesian coordinates. The availability mask is a binary matrix $M \in \{0,1\}^{(O+P)\times J\times 3}$ where zero determines the parts of the sequence that are not observed due to occlusions or being from future timesteps. Note that the elements of M corresponding to P future frames are always zero. With this notation, the observed sequence $\tilde{X} = [\tilde{X}_{-O+1}, \tilde{X}_{-O+2}, \dots, \tilde{X}_0, \tilde{X}_1, \dots, \tilde{X}_P]$ is derived by applying the element-wise product of M into X and adding a Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ in non-masked area $\tilde{X} = M \odot X + (1 - M)\epsilon$. The model predicts $\hat{X} = [\hat{X}_{-O+1}, \hat{X}_{-O+2}, \dots, \hat{X}_0, \hat{X}_1, \dots, \hat{X}_P]$ and the objective is lowering $|\hat{X} - X| \odot (1 - M)$ given $\tilde{X}$.

B. Conditional Diffusion Blocks

We propose a conditional diffusion block, inspired by [55], which contains multiple residual layers. Each layer consists of two consecutive transformers with the same input and output shapes. The first (temporal) transformer is responsible for modeling the temporal behavior of data. Its output is then fed to the second (spatial) transformer to attend to the body pose within each frame.

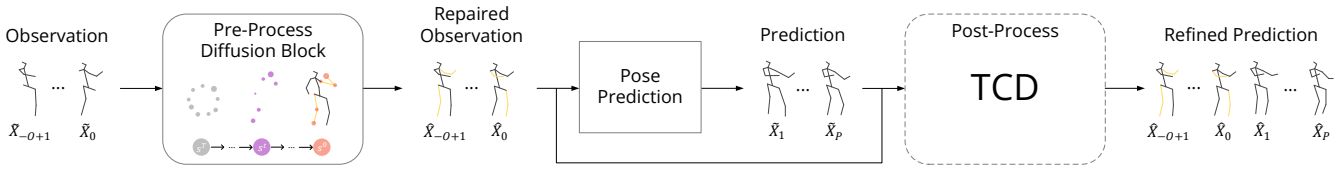

Fig. 3: An illustration of the pre-processing and post-processing framework. The pre-process diffusion block denoises the noisy observation sequence. The repaired observation is then given to a frozen predictor. The output of the predictor model is passed to TCD to perform the post-processing step and refine its predictions.

At training time, a Gaussian noise with zero mean and pre-defined variance is added to the input pose sequence s^0 to make a noisier version s^1. This process is repeated for T steps such that the output s^T will be close to a pure Gaussian noise in the non-masked area:

$$q(s^t|s^{t-1}) = M \odot s^{t-1} + (1-M) \odot \mathcal{N}(s^t; \sqrt{1-\beta^t} s^{t-1}, \beta^t \mathbf{I}), \quad (1)$$

where q denotes the forward process, and β^t is the variance of the noise in step t, determined using a scheduler. We use the cosine noise scheduler in our formulations, which was first introduced in [44]:

$$\beta^t = 1 - \frac{f(t)}{f(t-1)}, \quad f(t) = \cos^2\left(\frac{t/T + c}{1+c} \cdot \frac{\pi}{2}\right), \quad (2)$$

where c is a small offset and is set to 0.008 empirically. The cosine noise scheduler provides a smoother decrease in input quality than other popular schedulers, such as quadratic and linear [44], enabling more accurate learning of step noise variances in our problem. The network learns to reverse the diffusion process and retrieve the clean sequence by predicting the cumulative noise that is added to s^t as described in DDPM [25].

At inference time, the model begins with an incomplete and noisy input sequence s^T, where Gaussian noise is put in the non-masked area and observed data in the masked area. Subsequently, the model iteratively predicts the poses $s^{T-1}, \ldots, s^0$ through an iterative process by subtracting the additive noise learned during training from the output of the preceding step, until a clean output approximating the ground truth is obtained.

C. Temporal Cascaded Diffusion (TCD)

We illustrate our main model, which consists of a short-term and a long-term diffusion blocks, in Figure 2. The short-term block takes $\tilde{X}$ as input and predicts the first K frames of the future $[\hat{X}_1 \ldots \hat{X}_K]$, along with the observation frames $[\hat{X}_{-O+1} \ldots \hat{X}_0]$. The long-term block is tasked with predicting the remaining frames of the future $[\hat{X}_{K+1} \ldots \hat{X}_P]$, utilizing both the observation and the output of the short-term block. Note that during training, both blocks are trained using ground-truth input; however, at inference time, the average of five samples of the short-term block is supplied to the long-term block.

Cascading two diffusion models improves overall and particularly long-term forecasting due to the division of the complex task. In other words, the short-term prediction block focuses on predicting a limited number of frames, and thanks to its accurate short-term predictions, the long-term prediction block acquires more data, thus allowing it to focus its capacity on longer horizons.

D. Pre-processing and Post-processing

Given a frozen pose prediction model, we can enhance its performance through pre-processing by repairing its input sequence, and through post-processing by refining its outputs. This framework is illustrated in Figure 3.

a) Pre-Processing: Since most of the existing pose prediction models are unable to handle noisy observations, we present a simpler version of our model that serves as a pre-processing step for denoising the observations only. This module takes the noisy observation sequence $[\tilde{X}_{-O+1}, \tilde{X}_{-O+2}, \ldots, \tilde{X}_0]$ as input and outputs a repaired sequence $[\hat{X}_{-O+1}, \hat{X}_{-O+2}, \ldots, \hat{X}_0]$. The architecture of this model is similar to TCD, yet predicting within a single stage, with both the input and output sequences containing O frames. Our precise repair strategies allow any pose prediction models trained on complete datasets to predict reasonable poses in noisy input conditions.

b) Post-Processing: Furthermore, we want to improve the prediction results of existing models. We feed the results of any black-box pose prediction model $[\tilde{X}_1, \ldots, \tilde{X}_P]$ concatenated with repaired observation $[\hat{X}_{-O+1}, \hat{X}_{-O+2}, \ldots, \hat{X}_0]$ as the input to our TCD and retrain it to predict better. The initial prediction acts as the starting point that is gradually shifted toward the real distribution by our post-processing.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluate the performance of all approaches on four widely-used 3D pose prediction datasets:

Human3.6M [26] is the largest benchmark dataset for human motion analysis, comprising 3.6 million body poses. It consists of 15 complex action categories, each performed by seven actors individually. The training set comprises five subjects, and the validation and test sets comprise two different subjects. We train our models on all action classes concurrently. The original 3D pose skeletons in the dataset consist of 32 joints, but different subsets of joints have been used in previous works to represent human poses. To ensure a fair and comprehensive comparison, we consider three different settings for the dataset as follows:

Model	Human3.6M [26]				HumanEva-I [53]	
	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	ADE ↓	FDE ↓
Pose-Knows [58]	461	560	522	569	269	296
MT-VAE [61]	457	595	716	883	345	403
HP-GAN [6]	858	867	847	858	772	749
BoM [7]	448	533	514	544	271	279
GMVAE [19]	461	555	524	566	305	345
DeLiGAN [22]	483	534	520	545	306	322
DSF [62]	493	592	550	599	273	290
DLow [63]	425	518	495	531	251	268
Motron [52]	375	488	–	–	–	–
Multi-Objective [35]	414	516	–	–	228	236
GSPS [40]	389	496	476	525	233	244
STARS [60]	358	445	442	471	217	241
TCD (ours)	**356**	**396**	**463**	**445**	**199**	**215**

TABLE I: Comparison with stochastic models on Human3.6M [26] Setting-A and HumanEva-I [53] at a horizon of 2s.

- **Setting-A**: 25 observation frames, 100 prediction frames at 50 frames per second (fps), with the subset of 17 joints to represent the human pose;
- **Setting-B**: 50 observation frames, 25 prediction frames down-sampled to 25 fps, with the subset of 22 joints to represent the human pose;
- **Setting-C**: 25 observation frames, 25 prediction frames down-sampled to 25 fps, with the subset of 17 joints to represent the human pose.

AMASS (Archive of Motion capture As Surface Shapes) [37] is a recently published human motion dataset that combines 18 motion capture datasets, totaling 13,944 motion sequences from 460 subjects performing various actions. We use 50 observation frames down-sampled to 25 fps with 18 joints, as in previous studies.

3DPW (3D Poses in the Wild) [56] is the first dataset with accurate 3D poses in the wild. It contains 60 video sequences and each pose is described with an 18-joint skeleton, similar to the AMASS dataset. We use the official instructions to obtain training, validation, and test sets.

HumanEva-I [53] includes three subjects captured at 60 fps. Each person has 15 body joints. We remove the global translation and use the official train/test split of the dataset. The prediction horizon is 60 frames (1 second) given 15 observed frames (0.25 seconds), similar to [40].

2) Other Implementation Details: We train our models using the Adam optimizer [28], with a batch size of 32 and a learning rate of 0.001. The learning rate is decayed by a factor of 0.1 at 75% and 90% of the total epochs. Our model consists of 12 layers of residual blocks and 50 diffusion steps by default. In TCD, the length of short-term prediction K is set to 20% of the total prediction length P. Each transformer has 64 channels and 8 attention heads.

3) Evaluation Metrics: We measure the Displacement Error (DE), in millimeters (mm), over all joints in a frame. Then, we report the Average Displacement Error (ADE), which is the average DE across all prediction frames, and/or the Final Displacement Error (FDE), which is the DE in the final predicted frame. We also report the multi-modal versions of ADE (MMADE) and FDE (MMFDE), following [40]. We additionally report ADE for the missing joints of the observation frames in the repairing task (r-ADE).

B. Baselines

We compare our model with several recent methods, including stochastic [40], [63], [52], [35], [60] and deterministic approaches [42], [31], [41], [39], [36], [54], [64] when possible. Note that Some methods are not open-source and have different settings than ours. We also include *Zero-Vel* as a competitive baseline. *Zero-Vel* is a simple model that predicts the last observed pose for all future frames.

C. Comparisons with the State of the Art

We separate our experiments into three different settings: we first compare to other stochastic approaches, then to deterministic ones, and finally evaluate on noisy scenarios, with missing or noisy observation data.

1) Comparisons with Stochastic Approaches: We evaluate our model on two datasets, Human3.6M [26] Setting-A and HumanEva-I [53], and compare it with other stochastic approaches in Table I. Each model is sampled 50 times given each observation sequence. TCD (ours) clearly outperforms all previous works in terms of accuracy of the best sample (as measured by ADE and FDE) and multiple samples (as measured by MMADE and MMFDE).

2) Comparisons with Deterministic Approaches: We then compare our model to deterministic approaches on Human3.6M [26] Setting-B, tabulated in Table II. To compare with deterministic models, our model is sampled five times, and the best sample is considered. Our proposed model surpassed previous works in the short-term and with a marked margin in the long-term, thanks to our two-stage prediction strategy. The detailed results of our model's performance on all categories of Human3.6M, along with comparisons with models that are not reported in standard settings, can be found in the appendix. We have also included the results of two previous state-of-the-art models that have been post-processed by our generic framework at the bottom of Table II. Note that as the input data is complete, we only add post-processing (TCD) to their outputs. The improvements from our framework are non-negligible and can even beat our original model. Our two-stage prediction reveals a more

Model	80ms	320ms	560ms	720ms	880ms	1000ms
Zero-Vel	23.8	76.0	107.4	121.6	131.6	136.6
Res. Sup. [42]	25.0	77.0	106.3	119.4	130.0	136.6
ConvSeq2Seq [31]	16.6	61.4	90.7	104.7	116.7	124.2
LTD-50-25 [41]	12.2	50.7	79.6	93.6	105.2	112.4
HRI [39]	10.4	47.1	77.3	91.8	104.1	112.1
PGBIG [36]	10.3	**46.6**	76.3	90.9	102.6	110.0
TCD (ours)	**9.9**	48.8	**73.7**	**84.0**	**94.3**	**103.3**
HRI [39] + TCD (ours)	10.3	47.3	72.9	83.8	94.0	102.9
PGBIG [36] + TCD (ours)	10.2	46.1	72.4	83.6	93.9	102.8

TABLE II: Comparison with deterministic models on Human3.6M [26] Setting-B in FDE (mm) at different horizons.

	AMASS [37]				3DPW [56]			
Model	560ms	720ms	880ms	1000ms	560ms	720ms	880ms	1000ms
Zero-Vel	130.1	135.0	127.2	119.4	93.8	100.4	102.0	101.2
convSeq2Seq [31]	79.0	87.0	91.5	93.5	69.4	77.0	83.6	87.8
LTD-10-25 [41]	57.2	65.7	71.3	75.2	57.9	65.8	71.5	75.5
HRI [39]	51.7	58.6	63.4	67.2	56.0	63.6	69.7	73.7
TCD (ours)	**49.8**	**54.5**	**60.1**	**66.7**	**55.4**	**61.6**	**67.9**	**73.4**

TABLE III: Comparison with deterministic models on AMASS [37] and 3DPW [56] in FDE (mm) at long horizons.

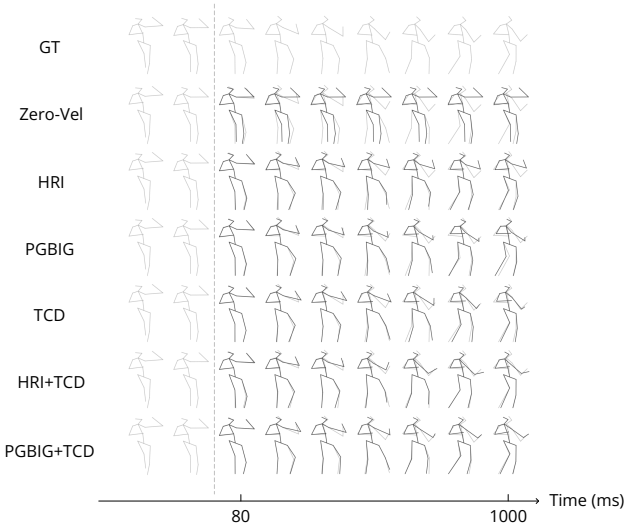

Fig. 4: Qualitative results on Human3.6M [26] Setting-B. The left part of each row shows the input observation, while the right part displays the predicted poses superimposed on the ground truth.

pronounced benefit for longer horizons, which suggests that starting with a better initial guess can better shift the pose sequence toward the real distribution.

Substantial long-term improvement can be observed in AMASS [37] and 3DPW [56] as well. Similar to previous works, we train our model on AMASS and measure the FDE on both datasets. The comparison with models reporting in this setting is in Table III. Note that for faster training, $K = 0$ was considered in this experiment.

Qualitative results on Human3.6M are shown in Figure 4. Predictions from our model are displayed along with predictions from several baselines and are superimposed on the ground-truth poses for direct comparison. Our model has successfully learned the data distribution, resulting in accurate and realistic poses; for instance, the hand movement is natural when the feet move while HRI has fixed hands and PGBIG has a momentum that avoids large hand movements. Moreover, post-processing can be used to further refine the predicted pose and shift it toward the ground truth.

3) Comparisons on Noisy Observation Data: We now examine the performance of models in the realistic scenario of noisy observations, since occlusions and noise are commonly seen in practice. To simulate occlusions, we remove 40% of the left arm and right leg from the observations of Human3.6M Setting-B, both during training and evaluation. The results in the top half of Table IV show that the state-of-the-art models perform inadequately when the observation is noisy, whereas our model achieves results close to those of the clean input observation. Our pre-processing module repairs the observation sequences before feeding to the state-of-the-art models and Zero-Vel, resulting in significant improvements in forecasting performance. MT-GCN [16] was designed to provide accurate predictions in incomplete observations. We compared our model to it and some other prior models and present the results on Human3.6M Setting-C in the first column of Table V. Our model achieved a remarkable improvement of 33.2mm in FDE at 1s horizon (30% improvement) over MT-GCN. It should be noted that the models in the upper part of the table received repaired sequences using MT-GCN's own preprocessing, while the rest received noisy sequences.

We analyzed the performance of our model in several occlusion patterns masks M that are applied to input data:

- Random Leg, Arm Occlusions: leg and arm joints are randomly occluded with a probability of 40%;
- Structured Joint Occlusions: 40% of the right leg joints for consecutive frames are missing;
- Missing Frames: 20% of the consecutive frames are missing;
- Gaussian Noise: Gaussian noise with a standard devia-

Model	80ms	320ms	560ms	720ms	880ms	1000ms
Zero-Vel	84.9	138.2	169.9	184.2	193.7	198.2
HRI [39]	65.2	104.5	130.0	141.6	151.1	157.1
PGBIG [36]	67.0	107.1	132.1	143.5	152.9	158.8
TCD (ours)	**11.2**	**51.3**	**75.4**	**85.4**	**95.4**	**104.5**
Pre(ours) + Zero-Vel	24.1	76.3	107.6	121.7	131.7	136.7
Pre(ours) + HRI [39]	11.4	48.6	78.3	92.7	105.0	112.8
Pre(ours) + PGBIG [36]	11.1	**47.9**	77.2	91.7	103.5	110.8
Pre(ours) + TCD (ours)	**10.8**	49.9	**74.4**	**84.9**	**95.1**	**104.2**

TABLE IV: Comparison on noisy observation data and pre-processed observation data (Pre(ours)+) on Human3.6M [26] Setting-B in FDE (mm) at different horizons.

Model	Random Leg, Arm Occlusions	Structured Joint Occlusions	Missing Frames	Gaussian Noise $\sigma = 25$	Gaussian Noise $\sigma = 50$
R+TrajGCN [41]	121.1	131.5	–	127.1	135.0
R+LDRGCN [17]	118.7	127.1	–	126.4	133.6
R+DMGCN [32]	117.6	126.5	–	124.4	132.7
R+STMIGAN [24]	129.5	128.2	–	–	–
MT-GCN [16]	110.7	114.5	122.0	114.3	119.7
TCD (ours)	**77.5**	**77.2**	**80.5**	**81.9**	**84.9**

TABLE V: Comparison on noisy observation data on Human3.6M [26] Setting-C in FDE (mm) at a horizon of 1s. The upper part of the table contains models that received repaired sequences (R+), while the lower part contains models that received noisy sequences.

Model	Train and Test Missing Ratio 10%	20%	30%	40%
MT-GCN [16]	109.4 / 8.6	110.5 / 13.7	112.3 / 18.7	114.4 / 24.5
TCD (ours)	**77.1 / 2.2**	**77.2 / 2.3**	**77.6 / 2.6**	**79.1 / 2.9**

TABLE VI: Results of motion prediction and sequence repairing on Human3.6M [26] Setting-C with varying amounts of randomly occluded joints in input data in FDE (mm) at a horizon of 1s / r-ADE (mm) of missing elements.

tion of $\sigma = 25$ or $\sigma = 50$ is added to the coordinates of the joints, and 50% of the leg joints are randomly occluded.

The results of training and evaluating our model on these observation patterns, in FDE at a prediction horizon of 1 second on Human3.6M Setting-C, are presented in Table V. Our model outperformed previous works in different patterns of occlusions and noises in input that can occur in the real world. Furthermore, we observed that missing 5 consecutive frames is more challenging than missing a part of the body in 10 consecutive frames, as the network can recover the latter with spatial information.

To have a thorough comparison with MT-GCN, we trained four models by varying the percentage of joints randomly removed from the pose observation sequence. The performance of sequence repairing (r-ADE of the occluded observation sequence) and motion prediction (FDE at 1-second horizon) is presented in Table VI. Our model exhibited a negligible error of 2.9mm in repairing with up to 40% of all joints missing, whereas MT-GCN exhibited an error of 24.5mm. Indeed, our model achieved more than 31% lower FDE compared to MT-GCN in forecasting.

D. Ablations Studies

Here, we investigate different design choices of the network and report ADE on Human3.6M [26] Setting-B. For faster training, only a fifth of the dataset was utilized in this section. The full model yielded an ADE of 63.3mm. When predicting in one stage, without any subdivisions, the ADE increased to 65.5mm due to erroneous predictions in longer time frames. Conversely, when predicting in three stages, i.e., 20%, 20%, and 60%, the performance dropped to 66.9mm, as cascading multiple stochastic processes leads to either random outcomes or a lack of diversity. This illustrates the efficacy of two-stage prediction. Another important factor is the length of short-term prediction. In our experiments, a prediction of $P = 25$ frames was made with $K = 5$. A lower $K = 2$ reduced the benefits of two-stage prediction (ADE of 65.1mm). On the other hand, a higher $K = 10$ made short-term prediction more difficult, leading to an increased ADE of 66.6mm.

We tested a quadratic scheduler instead of our cosine scheduler and it increased ADE by 1mm. Our full model employed 12 residual layers in its diffusion blocks; however, decreasing this number to 4 resulted in a decrease in performance by 3mm. We refrained from utilizing more than 12 residual layers due to the considerable negative influence on the sampling time. Moreover, we conducted several experiments on the architecture of the transformers and found that spatial transformer and time transformer both facilitated the learning of spatio-temporal features of the pose sequence. Eliminating either of these resulted in an ADE of 74.5mm and 261.1mm, respectively.

V. Conclusion

In this work, we proposed a denoising diffusion model for 3D human pose prediction suitable for noisy input observations occurring in the wild. Our model predicted future poses in two stages (short-term and long-term) to better capture human motion dynamics, achieved superior performance compared to the state-of-the-art on four datasets, including both clean and noisy input settings. We then leveraged it to create a generic framework that is easily applicable to any existing predictor in a black box manner in two steps: pre-processing to repair the observations and post-processing to refine the predicted poses. We have applied it to several previous predictors and enhanced their predictions. The high computational complexity of diffusion models is a well-known challenge, and future studies may explore ways to accelerate the model's performance without sacrificing accuracy.

ACKNOWLEDGMENT

The authors would like to thank Mohammadhossein Bahari and Bastien Van Delft for their helpful comments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754354, and SNSF Sinergia Fund.

REFERENCES

[1] 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.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. *International Conference on 3D Vision (3DV)*, 2021.

[3] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. Contextually plausible and diverse 3d human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11333–11342, 2021.

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5223–5232, 2020.

[5] Mohammadhossein Bahari, Saeed Saadatnejad, Ahmad Rahimi, Mohammad Shaverdikondori, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Vehicle trajectory prediction works, but not everywhere. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Emad Barsoum, John R. Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1499–149909, 2018.

[7] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2018.

[8] Smail Bouhsain, Saeed Saadatnejad, and Alexandre Alahi. Pedestrian intention prediction: A multi-task perspective. In *European Association for Research in Transportation (hEART)*, 2020.

[9] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4563–4570, 2018.

[10] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, Xiaohui Shen, Ding Liu, and Nadia Magnenat Thalmann. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11645–11655, October 2021.

[11] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020.

[12] Yu-Wei Chao, Jimei Yang, Brian Price, Scott Cohen, and Jia Deng. Forecasting human dynamics from static images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 548–556, 2017.

[13] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

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

[15] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001, 2020.

[16] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4801–4810, 2021.

[17] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6519–6527, 2020.

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (Neurips)*, 34:8780–8794, 2021.

[19] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648, 2016.

[20] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4346–4354, 2015.

[21] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[22] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4941–4949, 2017.

[23] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11374–11384, October 2021.

[24] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[26] 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, jul 2014.

[27] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[29] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[30] Przemyslaw A Lasota and Julie A Shah. A multiple-predictor approach to human motion prediction. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2300–2307, 2017.

[31] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5226–5234, 2018.

[32] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020.

[33] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. Motion prediction using trajectory cues. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13299–13308, October 2021.

[34] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022.

[35] Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, and Chiho Choi. Multi-objective diverse human motion prediction with knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8161–8171, June 2022.

[36] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6437–6446, June 2022.

[37] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

[39] 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.

[40] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth

pose sequences for diverse human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13309–13318, 2021.

[41] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[42] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2891–2900, 2017.

[43] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2276–2284, October 2021.

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021.

[45] Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2294–2303, 2021.

[46] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10985–10995, October 2021.

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[48] Saeed Saadatnejad, Mohammadhossein Bahari, Pedram Khorsandi, Mohammad Saneian, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Are socially-aware trajectory prediction models really socially-aware? *Transportation Research Part C: Emerging Technologies*, 2022.

[49] Saeed Saadatnejad, Yi Zhou Ju, and Alexandre Alahi. Pedestrian 3d bounding box prediction. In *hEART*, 2022.

[50] Saeed Saadatnejad, Siyuan Li, Taylor Mordan, and Alexandre Alahi. A shared representation for photorealistic driving simulators. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[52] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6457–6466, June 2022.

[53] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, Mar. 2010.

[54] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11209–11218, October 2021.

[55] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems (Neurips)*, 34:24804–24816, 2021.

[56] 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 *European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[57] Fabien B Wagner, Jean-Baptiste Mignardot, Camille G Le Goff-Mignardot, Robin Demesmaeker, Salif Komi, Marco Capogrosso, Andreas Rowald, Ismael Seáñez, Miroslav Caban, Elvira Pirondini, et al. Targeted neurotechnology restores walking in humans with spinal cord injury. *Nature*, 563(7729):65–71, 2018.

[58] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3332–3341, 2017.

[59] Nikolaus Wenger, Eduardo Martin Moraud, Jerome Gandar, Pavel Musienko, Marco Capogrosso, Laetitia Baud, Camille G Le Goff, Quentin Barraud, Natalia Pavlova, Nadia Dominici, et al. Spatiotemporal neuromodulation therapies engaging muscle synergies improve motor control after spinal cord injury. *Nature medicine*, 22(2):138–145, 2016.

[60] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *European Conference on Computer Vision (ECCV)*, 2022.

[61] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *European Conference on Computer Vision (ECCV)*, pages 265–281, 2018.

[62] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019.

[63] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020.

[64] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6447–6456, June 2022.