IEEE ROBOTICS AND AUTOMATION LETTERS, VOL. 9, NO. 5, MAY 2024

IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Toward Reliable Human Pose Forecasting With Uncertainty

Saeed Saadatnejad, Mehrshad Mirmohammadi, Matin Daghyani, Parham Saremi, Yashar Zoroofchi Benisi, Amirhossein Alimohammadi, Zahra Tehraninasab, Taylor Mordan, and Alexandre Alahi, *Member, IEEE*

***Abstract*—Recently, there has been an arms race of pose forecasting methods aimed at solving the spatio-temporal task of predicting a sequence of future 3D poses of a person given a sequence of past observed ones. However, the lack of unified benchmarks and limited uncertainty analysis have hindered progress in the field. To address this, we first develop an open-source library for human pose forecasting, including multiple models, supporting several datasets, and employing standardized evaluation metrics, with the aim of promoting research and moving toward a unified and consistent evaluation.Second, we devise two types of uncertainty in the problem to increase performance and convey better trust: 1) we propose a method for modeling aleatoric uncertainty by using uncertainty priors to inject knowledge about the pattern of uncertainty. This focuses the capacity of the model in the direction of more meaningful supervision while reducing the number of learned parameters and improving stability; 2) we introduce a novel approach for quantifying the epistemic uncertainty of any model through clustering and measuring the entropy of its assignments. Our experiments demonstrate up to 25% improvements in forecasting at short horizons, with no loss on longer horizons on Human3.6 M, AMSS, and 3DPW datasets, and better performance in uncertainty estimation. The code is available online.**

***Index Terms*—Computer vision for automation, human-centered robotics, human-robot collaboration, uncertainty.**

I. Introduction

HUMAN pose forecasting consists in predicting a sequence of future 3D poses of a person, given a sequence of past observed ones. It has attracted significant attention in recent years due to the critical applications in autonomous driving [42], human-robot collaboration [10], [51], robot navigation [8], and healthcare [53]. The field is now witnessing an arms race of forecasting models using different architectures that have shown increasing performances [29], [31], [46].

Manuscript received 31 July 2023; accepted 12 February 2024. Date of publication 6 March 2024; date of current version 2 April 2024. This letter was recommended for publication by Associate Editor 'YZ' Yezhou Yang and Editor M. Vincze upon evaluation of the reviewers' comments. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie Grant 754354 and in part by SNSF Sinergia Fund. *(Mehrshad Mirmohammadi, Matin Daghyani, Parham Saremi, Yashar Zoroofchi Benisi, Amirhossein Alimohammadi, and Zahra Tehraninasab contributed equally to this work.) (Corresponding author: Saeed Saadatnejad.)*

The authors are with VITA Laboratory, EPFL, 1015 Lausanne, Switzerland (e-mail: Saeed.Saadatnejad@epfl.ch).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3374188, provided by the authors.

Project page: https://github.com/vita-epfl/UnPOSed.

Digital Object Identifier 10.1109/LRA.2024.3374188

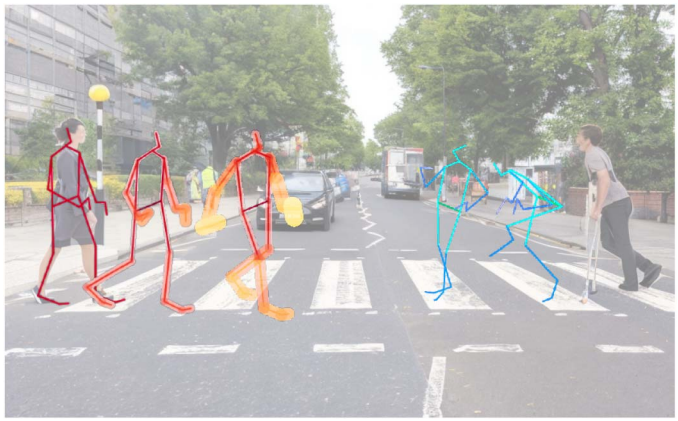


Fig. 1. We propose to model two kinds of uncertainty. 1) Aleatoric uncertainty, highlighting the inherent temporal evolution with lighter colors and thicker bones over time, illustrated by the left person. 2) Epistemic uncertainty, to detect non valid, out-of-distribution forecast poses due to unseen scenarios in training, exemplified by the right person.

Forecasting human poses is a difficult task with multiple challenges to solve: it mixes both spatial and temporal reasoning, with a huge variability in scenarios; and human behavior is difficult to predict, as it changes in dynamic and multi-modal ways to react to its environment. To guarantee safe interactions with humans, robots should not only predict human motions, but also identify scenarios in which they are uncertain [4], [22], [26], [49], and act accordingly. As an example, Fig. 1 illustrates a pose forecasting scenario in autonomous driving context. Without an uncertainty measure, all the forecast poses are considered valid. However, uncertainty measures can detect unconfident outputs, so be treated with more caution. Researchers have shown the benefits of estimating uncertainty for classification [26], [49] and regression tasks [4], [22], but how to apply it to pose forecasting is not yet studied.

In this letter, we present two solutions to capture the uncertainty of pose forecasting models from two important perspectives. The first one deals with the aleatoric uncertainty, i.e., the irreducible intrinsic uncertainty in the data. We reformulate the pose forecasting objective function to capture the aleatoric uncertainty. To reduce the number of learned parameters and improve stability, we introduce uncertainty priors based on our knowledge about the uncertainty, e.g., that the uncertainty increases with time. We then train the forecasting model with the new objective function. This allows the model to focus its capacity to learn forecasting at shorter time horizons, where uncertainty is lower and learning is more meaningful, compared to

longer ones that are intrinsically harder and uncertain to forecast. We apply our proposed uncertainty method to several models from the literature and evaluate on three well-known datasets (Human3.6M [17], AMASS [30], 3DPW [52]) and achieve up to 25% improvements in forecasting at short horizons, with no loss on longer horizons.

The second one is about epistemic uncertainty which shows the model's lack of knowledge. To this end, we define a model-agnostic uncertainty metric to reflect the reliability and certainness of pose forecasting models in real-world scenarios, where the ground truth is absent for accuracy calculation. Unlike previous methods which require accessing model [12] (i.e., white-box methods) or are specific to certain models [49], our approach does not require access to the model (i.e., black-box approach) and is model-agnostic. Since there is no label for motions, we train a deep clustering network to learn the distribution of common poses and measure the dissimilarity between the predictions' embeddings and cluster centers. We achieve better performance in detecting out-of-distribution forecast poses using our epistemic uncertainty metric than other approaches from the literature.

Lastly, it is important to acknowledge that the field of pose forecasting is rapidly advancing, thanks to the significant interest from researchers and practitioners. However, this happens at the cost of unfair and non-unified evaluations. All current works use disparate metrics and dataset setups to report their results, leading to ambiguities and errors in interpretation. In an effort to mitigate these discrepancies, we release an open-source library for human pose forecasting named *UnPOSed*[1]. This includes our re-implementations of over 10 models, processing codes for 3 widely-used datasets and 6 metrics, all implemented and tested in a standardized way, in order to ease the implementation of new ideas and promote research in this field. To summarize our contributions:

- We propose a method for incorporating priors to estimate the aleatoric uncertainty in human pose forecasting and demonstrate its efficacy in improving several state-of-the-art models on multiple datasets;
- We propose a model-agnostic metric of quantifying epistemic uncertainty to evaluate models in unseen situations, outperforming previous methods;
- We develop and publicly release an open-source library for human pose forecasting.

II. Related Works

Human pose forecasting: While the literature has extensively examined the forecasting of a sequence of future center positions at a coarse-grained level [3], [27], [41] or a sequence of bounding boxes [6], [43], our focus in this work is on a more fine-grained forecasting i.e., pose. Additionally, we limit our focus to the observation sequence alone, rather than incorporating context information [15], social interactions [1], action class [7] or global movements [38]. Many approaches have been proposed for human pose forecasting, with some using feed-forward networks [25] and many others using Recurrent Neural Networks (RNNs) to capture temporal dependencies [20], [34]. To better capture spatial dependencies of body poses, Graph Convolutional Networks (GCNs) have been utilized [31], [33], along with separating temporal and spatial convolution blocks and using trainable adjacency matrices [46]. Attention-based approaches have also gained interest for modeling human motion, showing improvement with a spatio-temporal self-attention module [31]. More recently, forecasting in multiple stages [29] and a diffusion model with a transformer-based architecture [44] have been proposed.

We can categorize all previous works into stochastic and deterministic models. Stochastic models [2], [28], [32], [37], [44], [45], [59] can give diverse predictions but we mainly focus on deterministic models [9], [29], [31], [46] as they provide more accurate predictions which is crucial for robotics applications. Given the growing interest in this field, we believe that greater attention should be paid to uncertainty estimation in this task.

Uncertainty in pose forecasting: Knowing when a model does not know, i.e., uncertain, is important to improve trustworthiness and safety [35]. Traditionally, uncertainty in deep learning is divided into data (aleatoric) and model (epistemic) uncertainty [22]. The aleatoric originates from the intrinsic noise and inherent uncertainty of data and cannot be reduced by improving the model, while the epistemic uncertainty shows the model's weakness in recognizing the underlying structure of the data and can be reduced by enhancing the network architecture or increasing data. Many methods have been proposed to estimate and utilize these types of uncertainty in various tasks, including image classification [12], semantic segmentation [21], and natural language processing [56]. It has also been explored in pose estimation from images and videos [19], [23], visual navigation and trajectory forecasting tasks [16], [18] but not yet studied in human pose forecasting which includes spatio-temporal relationships modeling. We will show how modeling the uncertainty can improve accuracy.

Moreover, it is important to measure the epistemic uncertainty of models intended for real-world applications. Bayesian Neural Networks (BNNs) have conventionally been used to formulate uncertainty by defining probability distributions over the model parameters [36]. However, the intractability of these distributions has led to the development of alternative approaches to approximate Bayesian inference for uncertainty estimation. One widely used method is Variational Inference [5], [13], which is valued for its scalability. A notable example is Monte Carlo (MC) dropout [12], which involves applying dropout [47] at inference time to model the parameters of the network as a mixture of multivariate Gaussian distributions with small variances. However, those methods are not model-agnostic. Another approach, known as calibration [14], requires the model to provide probabilities, but deep neural networks have been shown to be poorly calibrated. One way to evaluate model reliability is by measuring the distance between a new sample and the training samples using a deep deterministic network, a technique proven effective in image classification [26], [49]. However, this approach measures the uncertainty for their own model and is not applicable to measuring the uncertainty of different models. In contrast, Deep Ensembles [24] can measure the uncertainty of different models by training multiple neural networks independently and averaging their outputs at inference time. Nevertheless, this method can be computationally expensive and slow. In this study, we concentrate on the model's output and define epistemic uncertainty as the extent to which the model's forecasts align with the training distribution, providing a black-box uncertainty measurement of pose forecasting models.

III. Aleatoric Uncertainty in Pose Forecasting

Pose forecasting models usually take as input a sequence x of 3D human poses with J joints in O observation time frames, and predict another sequence $\hat{y}$ of 3D poses to forecast its future

[1]https://github.com/vita-epfl/UnPOSed

y in the next T time frames. In addition to this, we want a model to estimate its aleatoric uncertainty u along with the predicted poses $\hat{y}$, to indicate how reliable these can be.

For this, we model the probability distribution of the error, i.e., the euclidean distance between ground truths y and forecasts $\hat{y}$, with an exponential distribution following [4]:

$$\|y - \hat{y}\|_2 \sim \text{Exp}(\alpha), \tag{1}$$

where α is the distribution parameter to be selected. Its log-likelihood therefore writes

$$\ln p(\|y - \hat{y}\|_2) = \ln \alpha - \alpha \|y - \hat{y}\|_2 . \tag{2}$$

We then define the aleatoric uncertainty as $u := -\ln \alpha$, and set it as a learnable parameter for the model. When training the model with maximum likelihood estimation, the loss function $\mathcal{L}$ to minimize is then given by

$$\mathcal{L}(y, \hat{y}, u) = -\ln p(\|y - \hat{y}\|_2) = e^{-u} \|y - \hat{y}\|_2 + u. \tag{3}$$

We consider pose forecasting as a multi-task learning problem with task-dependant uncertainty, i.e., independent of the input sequences x. There are several ways to define tasks in this manner, e.g., by separating them based on time frames, joints, actions (if the datasets provide them), or any other combination of them. In the following, we consider dividing tasks based on time and joints[2]. In this case, for each future time frame t and joint j, the model predicts an uncertainty estimate u_t^j associated with its 3D joint forecasts $\hat{y}_t^j$. This formulation yields the corresponding loss function:

$$\mathcal{L}_{total}(y, \hat{y}, u) = \sum_{\substack{t=1...T \\ j=1...J}} e^{-u_t^j} \left\| y_t^j - \hat{y}_t^j \right\|_2 + u_t^j, \tag{4}$$

where T refers to the number of prediction frames and J is the number of joints.

Since the loss function (4) weighs the error $\|y_t^j - \hat{y}_t^j\|_2$ based on the aleatoric uncertainty $e^{-u_t^j}$, it forces the model to focus its capacity to points with lower aleatoric uncertainty. In particular, we expect short time horizons to have lower uncertainty, and therefore to present better improvements than longer ones.

Unfortunately, learning all aleatoric uncertainty values u_t^j independently leads to an unstable training. To address this issue, we introduce uncertainty priors F, in order to inject knowledge about the aleatoric uncertainty pattern and stabilize the training. For this, we choose a family F of functions parameterized by a given number of parameters θ. Instead of learning all uncertainty values u_t^j independently, the model now only learns θ, which can be chosen to be of a smaller size so as to ease the training. With a learned θ^*, the uncertainty values u_t^j are obtained with the function $F(\theta^*)$:

$$u_t^j = F(\theta^*)(j, t). \tag{5}$$

It is noticeable that this framework generalizes the previous case (without prior) by setting F to yield a separate parameter for each uncertainty value:

$$u_t^j = \text{Id}(\theta^*)(j, t) = \theta_t^j. \tag{6}$$

[2]Extending the formulation to other task definitions is straightforward.

Intuitively, the more parameters F has, the more scenarios it can represent, but at the cost of stability. We, therefore, compare several choices for F, with variable numbers of learnable parameters as different trade-offs between ease of learning and representation power. We select three functions that constrain the temporal evolution of aleatoric uncertainty, independently for each joint. We select functions with a logarithmic shape due to the observed exponential pattern in error evolution over time. The first one, Sig_3, is a sigmoid function used to ensure that uncertainty only increases with time, and has three parameters per joint to control this pattern:

$$u_t^j = \text{Sig}_3(\theta)(j, t) = \frac{\theta_2^j}{1 + e^{-\theta_0^j (t - \theta_1^j)}}. \tag{7}$$

Then we leverage Sig_5, which is a generalized version of the sigmoid function [40] with 5 parameters per joint:

$$u_t^j = \text{Sig}_5(\theta)(j, t) = \theta_0^j + \frac{\theta_1^j}{1 + ab + (1 - a)c}, \tag{8}$$

where the terms a, b and c are defined by

$$a = \frac{1}{1 + e^{-\frac{2\,\theta_2^j\,\theta_4^j}{|\theta_2^j + \theta_4^j|}(\theta_3^j - t)}}, b = e^{\theta_2^j(\theta_3^j - t)}, c = e^{\theta_4^j(\theta_3^j - t)}. \tag{9}$$

Note that optimization with Sig_5 can converge to all uncertainty coefficients learnable with Sig_3, but also additional values, benefiting from its strictly larger output space.

We also compare with a more generic polynomial function Poly_d of degree d, which has $d + 1$ learnable parameters per joint and constrain the uncertainty less:

$$u_t^j = \text{Poly}_d(\theta)(j, t) = \theta_0^j + \theta_1^j t + \theta_2^j t^2 + ... + \theta_d^j t^d. \tag{10}$$

IV. Epistemic Uncertainty in Pose Forecasting

Now, we address the epistemic uncertainty to capture the model's uncertainty due to the lack of knowledge. We want to quantify the intuition that the models with predicted motions dissimilar to the training distribution in the latent representation are less reliable and, therefore, should be treated with caution. Notably, our aim in this section is not to improve accuracy but rather to measure uncertainties associated with pose forecasting models.

We improve upon existing literature of uncertainty quantification by introducing temporal modeling and clustering in epistemic uncertainty. Specifically, we employ an LSTM-based autoencoder (Fig. 2) due to its proficient capability to encode spatio-temporal dependencies and learn potent latent representations. We then rely on clustering on that space as there are no predefined motion classes.

In the next parts, we first explain how to estimate the number of motion clusters K and train the deep clustering. We then illustrate how to measure the epistemic uncertainty.

A. Determining the Number of Motion Clusters

Determining K, the number of clusters, is essential since it corresponds to the diversity of motions in the training dataset. An optimal K, therefore, captures the diversity in the training dataset while also reducing the time complexity of our subsequent algorithms.

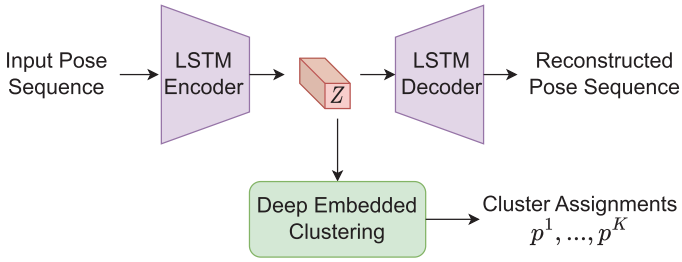


Fig. 2. Motion is encoded into a well-clustered representation space Z by our LSTM encoder-decoder. The probabilities of the cluster assignments are provided by our deep embedded clustering on that space to estimate the epistemic uncertainty.

We first train an LSTM auto-encoder (Fig. 2) to learn low dimensional embeddings Z by minimizing the reconstruction loss $\mathcal{L}_{recons}$ over the training dataset. We then follow DED [55] which uses t-SNE [50] to reduce Z to a 2-dimensional feature vector z'. Subsequently, local density ρ_i and delta δ_i for each data point are calculated:

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \quad \delta_i = \min_{j:\rho_j > \rho_i} d_{ij}, \tag{11}$$

where $\chi(.) = 1$ if $. < 0$ else $\chi(.) = 0$, d_{ij} is the distance between z'_i and z'_j, and d_c is the cut-off distance. We then define $\gamma_i = \rho_i.\delta_i$ similar to [54]. A larger γ_i corresponds to a greater likelihood of being chosen as a cluster center; however, the number of clusters still remains a hyperparameter. We fully automate it by defining r_i as the gap between two γ_i and γ_{i+1} values (where $\gamma_{i+1} < \gamma_i$):

$$r_i = \frac{\gamma_i}{\gamma_{i+1}}, i \in [1, N-1]. \tag{12}$$

We set $K = \text{argmax}(r_i)$ since γ_K represents the largest shift in likelihood of a sample being a cluster itself.

B. Deep Embedded Clustering

Having identified the number of clusters, we now learn the optimal deep clustering of our embedding. We initialize the cluster centers $\{\mu^k\}_{k=1}^K$ using the K-means algorithm on the feature space. We then minimize the clustering loss $\mathcal{L}_{cluster}$ as defined in DEC [57] jointly with the reconstruction loss in order to learn the latent representation as well as clustering. We incorporated the reconstruction loss into the loss function to act as a regularizer and prevent the collapse of the network parameters. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{cluster} + \lambda \mathcal{L}_{recons}, \tag{13}$$

where λ is the regularization coefficient. Finally, when the loss is converged, we fine-tune the trained network using the cross-entropy loss on the derived class labels in order to make clusters more compact.

C. Estimating Epistemic Uncertainty

Now, we estimate the epistemic uncertainty of a given forecasting model. Specifically, for each example, denote the probability of assignment to the kth cluster by p^k. The epistemic uncertainty is then calculated as follows:

$$EpU = \frac{1}{N} \sum_{i=1}^{N} \text{entropy}(p_i^1, \ldots, p_i^K), \tag{14}$$

where N is the size of the dataset. In other words, a model that does not generate outputs close to the motion clusters is considered uncertain.

V. Experiments

A. Datasets and Metrics

Human3.6M [17] contains 3.6 million body poses. It comprises 15 complex action categories, each one performed by seven actors individually. The validation set is subject-11, the test set is subject-5, and all the remaining five subjects are training samples. The original 3D pose skeletons in the dataset consist of 32 joints. Similar to previous works, we have 10/50 observation frames, 25 forecast frames down-sampled to 25 fps, with the subset of 22 joints to represent the human pose.

AMASS (The Archive of Motion Capture as Surface Shapes) [30] unifies 18 motion capture datasets totaling 13,944 motion sequences from 460 subjects performing a variety of actions. We use 50 observation frames down-sampled to 25 fps with 18 joints, similar to previous works.

3DPW (3D Poses in the Wild) [52] is the first dataset with accurate 3D poses in the wild. It contains 60 video sequences taken from a moving phone camera. Each pose is described as an 18-joint skeleton with 3D coordinates similar to AMASS dataset. We use the official instructions to obtain training, validation and test sets.

We measure the accuracy in terms of MPJPE (Mean Per Joint Position Error) in millimeters (mm) per frame and in terms of A-MPJPE as the average for all frames when needed. We also report EpU as defined in (14).

B. Baselines

We apply our approach to several recent methods that are open-source [29], [31], [46] and compare the performances of with and without the incorporation of our approach. Note that we follow their own training setup in which some use 10 frames of observation [9], [29], [46] and the rest 50 frames of observation [25], [31], [33], [34]. We report the results obtained from the pretrained model of deterministic STARS* [58] as documented on their GitHub page. We also consider *Zero-Vel*, a simple and competitive baseline [34], that forecasts all future poses by outputting the last observed pose.

Inspired by the common trend to treat sequences with Transformers, we have designed our own simple transformer-based architecture referred to as *ST-Trans*. We followed the best practices proposed in [48] and adapted their design elements to the task of pose forecasting. As depicted in Fig. 3, it is composed of several identical residual layers, each layer consists of a spatial and a temporal transformer encoder to learn the spatio-temporal dynamics of data utilizing the attention mechanism.

C. Aleatoric Uncertainty

We first show the impact of aleatoric Uncertainty-Aware Loss (pUAL) with the prior Sig_5 to several models from the literature and our ST-Trans. Table I shows the overall results

TABLE I
COMPARISON OF OUR METHOD ON HUMAN3.6M [17] IN MPJPE (MM) AT DIFFERENT PREDICTION HORIZONS

Model	80 ms	160 ms	320 ms	400 ms	560 ms	720 ms	880 ms	1000 ms
Zero-Vel [34]	23.8	44.4	76.1	88.2	107.4	121.6	131.6	136.6
Res. Sup. [34]	25.0	46.2	77.0	88.3	106.3	119.4	130.0	136.6
ConvSeq2Seq [25]	16.6	33.3	61.4	72.7	90.7	104.7	116.7	124.2
LTD-50-25 [33]	12.2	25.4	50.7	61.5	79.6	93.6	105.2	112.4
MSR-GCN [9]	12.0	25.2	50.4	61.4	80.0	93.9	105.5	112.9
STARS* [58]	12.0	24.6	49.5	60.5	78.6	92.6	104.3	111.9
STS-GCN [46]	17.7	33.9	56.3	67.5	85.1	99.4	109.9	117.0
STS-GCN + pUAL (ours)	13.2	27.1	54.7	66.2	84.5	97.9	109.3	115.7
gain	**25.4 %**	20.1 %	2.8 %	1.9 %	0.7 %	1.5 %	0.5 %	1.1 %
HRI* [31]	12.7	26.1	51.5	62.6	80.8	95.1	106.8	113.8
HRI* + pUAL (ours)	11.6	25.3	51.2	62.2	80.1	93.7	105.0	112.1
gain	**8.7 %**	3.1 %	0.6 %	0.6 %	0.9 %	1.5 %	1.7 %	1.5 %
PGBIG [29]	10.3	22.6	46.6	57.5	76.3	90.9	102.7	110.0
PGBIG + pUAL (ours)	9.6	21.7	46.0	57.1	75.9	90.3	102.1	109.5
gain	**6.8 %**	4.0 %	1.3 %	0.7 %	0.5 %	0.7 %	0.6 %	0.5 %
ST-Trans	13.0	27.0	52.6	63.2	80.3	93.6	104.7	111.6
ST-Trans + pUAL (ours)	10.4	23.4	48.4	59.2	77.0	90.7	101.9	109.3
gain	**20.0 %**	13.3 %	8.0 %	6.3 %	4.1 %	3.1 %	2.7 %	2.1 %

+pUAL refers to models where aleatoric uncertainty is modeled.

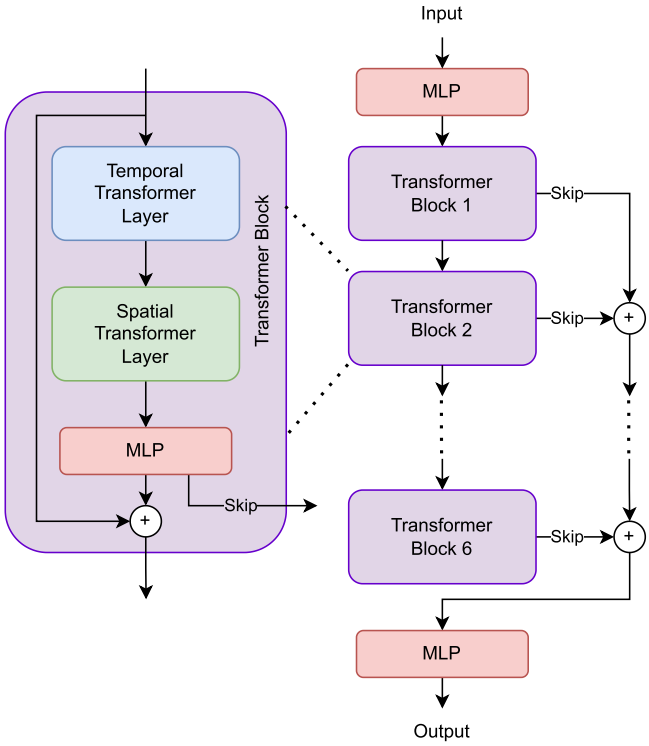


Fig. 3. ST-Trans consists of two MLP layers and six Transformer Blocks with skip connections. Each Transformer Block contains two cascaded temporal and spatial transformers to capture the spatio-temporal features of data.

on Human3.6M [17]. To have a fair evaluation between all models, we adapt HRI [31] to predict 25 frames in one step (denoted as HRI*). We observe that all methods get better results when taking aleatoric uncertainty into account during learning, therefore confirming the need for aleatoric uncertainty estimation. It is noticeable that pUAL gives better improvements for shorter prediction horizons, e.g., up to 25.4 % and 20.1 % for STS-GCN [46] at horizons of 80 ms and 160 ms, which

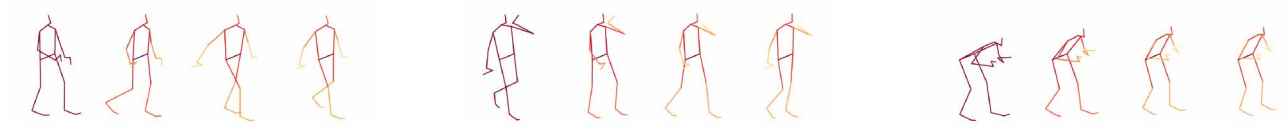
TABLE II
COMPARISON OF OUR PROPOSED METHOD ON AMASS [30] AND 3DPW [52] IN MPJPE (MM) AT DIFFERENT PREDICTION HORIZONS

Model	AMASS				3DPW			
	160 ms	400 ms	720 ms	1000 ms	160 ms	400 ms	720 ms	1000 ms
Zero-Vel [34]	56.4	111.7	135.1	119.4	41.8	79.9	100.5	101.3
ConvSeq2Seq [25]	36.9	67.6	87.0	93.5	32.9	58.8	77.0	87.8
LTD-10-25 [33]	20.7	45.3	65.7	75.2	23.2	46.6	65.8	75.5
STS-GCN [46]	20.7	43.1	59.2	68.7	20.8	40.3	55.0	62.4
STS-GCN + pUAL	20.4	42.4	59.1	68.1	20.5	40.0	54.8	62.2
HRI [31]	20.7	42.0	58.6	67.2	22.8	45.0	62.9	72.5
HRI + pUAL	19.9	41.4	58.1	66.5	22.2	44.6	62.4	72.2
ST-Trans	21.3	42.5	58.3	66.6	24.5	47.4	64.6	73.8
ST-Trans + pUAL	18.3	39.7	56.5	66.7	22.3	45.7	63.6	73.2

+pUAL refers to models where aleatoric uncertainty is modeled. The models were trained on AMASS.

correspond to the less uncertain time frames, where pUAL focuses training more (smaller discount in the loss function, as seen in (4)). At the same time, adding pUAL does not degrade the performances at longer horizons. In the context of close human-robot interactions, this improved precision can significantly enhance the overall system performance. Examples of predicted 3D pose sequences using pUAL are depicted in Fig. 4, and show that the estimated uncertainty increases over time, with joints farther away from the body center associated with higher uncertainties. Moreover, we report the performances of the models on AMASS and 3DPW datasets in Table II. Again, we observe that modeling aleatoric uncertainty leads to more accurate predictions, especially at shorter horizons, with improvements up to 14.1 % on AMASS and up to 9.0 % on 3DPW for ST-Trans at a horizon of 160 ms.

We argue that modeling the aleatoric uncertainty leads to more stable training. In order to demonstrate this, we conduct five separate trainings of ST-Trans and present in Fig. 5 the average of the A-MPJPE values along with their respective standard deviations for each epoch. The plot highlights that the model with pUAL is more stable across runs, as indicated by a lower standard deviation. Moreover, we compute AP-MPJPE, which is the average pairwise distance of predicted motions in terms of MPJPE, and observe that it decreases from 24.2 mm



(a) Walking (b) Phoning (c) Taking Photo

Fig. 4. Qualitative forecast poses on Human3.6M [17] depicting different actions over time. For each action, time progresses from left to right. Higher aleatoric uncertainty is shown with a lighter color. Uncertainty of any bone is considered as its outer joint's uncertainty assuming the hip is the body center. We observe that the estimated uncertainty increases over time, with joints farther away from the body center associated with higher uncertainties.

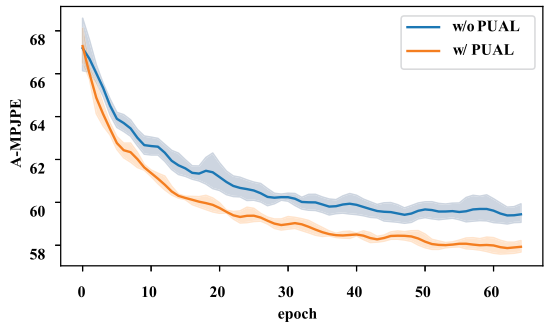


Fig. 5. A-MPJPE and its standard deviation in training epochs for 5 trained models. The model with pUAL has a lower standard deviation, meaning a more stable training.

TABLE III
COMPARISON OF DIFFERENT PRIORS FOR ALEATORIC UNCERTAINTY IN TERMS OF MPJPE (MM) AT 1 S ON HUMAN3.6M

Uncertainty prior (tasks)	Number of parameters	Standard deviation	ST-Trans	HRI*	STS-GCN
None	–	0.643	111.6	113.8	117.0
Id (T, J)	$25 \cdot 22$	0.557	**109.3**	114.6	115.8
Poly_9 (T, J)	$10 \cdot 22$	0.505	110.3	114.7	118.1
Sig_5 (T, J)	$5 \cdot 22$	**0.496**	**109.3**	**112.1**	**115.7**
Sig_3 (T, J)	$3 \cdot 22$	0.537	110.3	113.1	115.9
Sig_5 (T)	5	0.505	109.7	112.4	115.9

Lower standard deviation in training is better.

to 20.3 mm when pUAL loss is added, showing again lower standard deviation in the model's output.

So far, results have been reported using the Sig_5 uncertainty prior (8) to model the time and joint (T, J) aleatoric uncertainty. In Table III, we report the performances of other choices, and compare against using a single prior Sig_5 for all joints (only time dependency T) and other priors Sig_3, Poly_9. The results show again that taking aleatoric uncertainty into account with pUAL is beneficial and that a good choice of uncertainty prior is important. In particular, Sig_5 performs better than using no prior for all models. Using a prior can lead to similar aleatoric uncertainty than the unconstrained case, but with fewer learnable parameters and better stability.

D. Epistemic Uncertainty

Evaluating the quality of epistemic uncertainty is difficult due to the unavailability of ground truth annotations, yet important. Our goal is to identify instances where pose forecasting is not reliable, essentially making this a binary classification problem.

TABLE IV
AUROC, INFERENCE LATENCY (MS) AND THE NUMBER OF TRAINING RUNS FOR DIFFERENT EPISTEMIC UNCERTAINTY METHODS

Method	AUROC	Latency	Trainings
Deep-Ensemble-3	0.87	6.28	3
Deep-Ensemble-5	0.90	10.43	5
MC-Dropout-5	0.90	9.57	1
MC-Dropout-10	0.92	18.98	1
Ours	**0.95**	**6.23**	1

Selective classification is a widely used methodology to evaluate uncertainty quality, where a classifier has the option to refrain from classifying data points if its confidence level drops below a certain threshold [11]. In other words, if a pose forecasting model is trained on action A and evaluated on actions A and B, a reliable measure of epistemic uncertainty should effectively distinguish between these two sets of forecasts.

We assess the performance of our epistemic uncertainty estimation using selective classification, and measure how well actions "sitting" and "sitting down" can be separated from actions "walking" and "walking together", all from the test set of Human3.6M, based solely on the predicted uncertainty of the model. The forecasting model and clustering are trained on Human3.6M walking-related actions, and we anticipate low uncertainty values for those actions and high uncertainty values for sitting-related actions, i.e., not encountered and significantly distinct actions. During the assessment, we compute uncertainty scores for both actions and measure the classification results for a range of thresholds. Similar to prior research [39], we utilize the AUROC metric, where a higher score is desirable and a value of 1 indicates that all walking-related data points possess lower uncertainty than all sitting-related data points. In Table IV, we present our findings and compare them to alternative approaches, where our proposed method demonstrates higher AUROC. The full ROC curve is in Fig. 6. Note that our approach is model-agnostic in contrast to MC-Dropout.

Another feature of our approach is computational efficiency, which is attributed to its ability to compute in a single forward path. This is in contrast to MC-Dropout and Ensemble methods. We provide a comparison of the average inference latency, measured in milliseconds, between our method and other approaches in Table IV. Our approach shows lower latency and only requires one training. Notably, the performance gap between our approach and other methods may increase when using more computationally expensive forecasting models.

We conducted another experiment to showcase our metric's effectiveness in out-of-distribution (OOD) motions. By shuffling the frames' order or joints in each pose sequence of the test set, we generated OOD data. The EpU on the original test set is

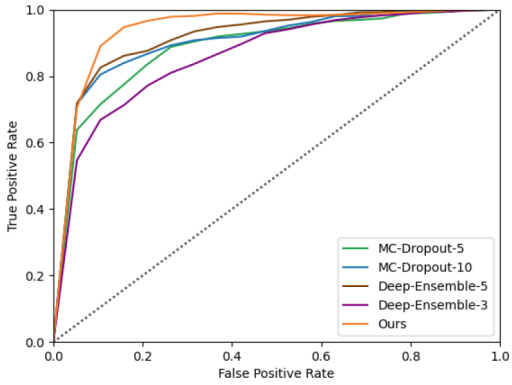


Fig. 6. ROC curve for a model trained on walking-related actions and tested on both walking-related and sitting-related actions. The objective is to distinguish between these sets by utilizing uncertainty estimates.

TABLE V
COMPARISON OF DIFFERENT MODELS IN TERMS OF A-MPJPE AND EPU ON AMASS AND 3DPW DATASETS

	AMASS		3DPW	
Model	EpU	A-MPJPE	EpU	A-MPJPE
Zero-Vel [34]	0.449	85.72	0.566	64.44
HRI [31]	0.351	43.76	0.463	43.62
STS-GCN [46]	0.332	45.49	0.455	42.60
ST-Trans + pUAL	0.336	35.86	0.439	40.02

The clustering and forecasting models were trained on AMASS.

0.085, while for shuffled joints, it was observed to be 1.53 due to the lack of correspondence with in-distribution (ID) poses. Furthermore, a high EpU value of 2.18 was obtained for shuffled frames, highlighting the importance of frame order in generating an ID motion. The full table of performances in all actions can be found on our webpage.

Additionally, we report the forecasting models' performances in Table V in terms of A-MPJPE, along with the epistemic uncertainties EpU associated with their predictions, on both the AMASS and 3DPW datasets. Note that the forecasting models and the clustering method were trained on the AMASS dataset. Higher uncertainties were recorded on 3DPW as an unseen dataset while prediction errors were lower. It underscores the reliability of our uncertainty quantification approach and suggests that relying solely on a model's prediction errors may not provide a comprehensive assessment.

VI. CONCLUSION

In this letter, we focused on modeling the uncertainty of human pose forecasting. We suggested a method for modeling aleatoric uncertainty of pose forecasting models that could make state-of-the-art models uncertainty-aware and improve their performances. We showed the effect of uncertainty priors to inject knowledge about the pattern of uncertainty. Moreover, we measured the epistemic uncertainty of pose forecasting models by clustering poses into motion clusters, which enables us to evaluate the trustworthiness of victim models. We made an open-source library of human pose forecasting with several models, datasets, and metrics to move toward a unified and fair evaluation. We hope that the findings and the library will pave the way to more uncertainty-aware pose forecasting models.

ACKNOWLEDGMENT

The authors extend their sincere gratitude to Armin Saadat and Nima Fathi for their invaluable contributions in the project's initial phase and library development. Special thanks also go to Mohamad Asadi, Ali Rasekh, Megh Shukla, and Mohammadhossein Bahari for their helpful input.

REFERENCES

[1] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and H. Rezatofighi, "Socially and contextually aware human motion and pose forecasting," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 6033–6040, Oct. 2020.

[2] S. Aliakbarian, F. S. Saleh, M. Salzmann, L. Petersson, and S. Gould, "A stochastic conditioning scheme for diverse human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5222–5231.

[3] M. Bahari, S. Saadatnejad, A. Rahimi, M. Shaverdikondori, S.-M. Moosavi-Dezfooli, and A. Alahi, "Vehicle trajectory prediction works, but not everywhere," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17102–17112.

[4] L. Bertoni, S. Kreiss, and A. Alahi, "MonoLoco: Monocular 3D pedestrian localization and uncertainty estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6860–6870.

[5] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[6] S. Bouhsain, S. Saadatnejad, and A. Alahi, "Pedestrian intention prediction: A multi-task perspective," 2020, *arXiv:2010.10270*.

[7] Y. Cai et al., "A unified 3D human motion synthesis model via conditional variational auto-encoder," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11625–11635.

[8] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, "Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 6015–6022.

[9] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11447–11456.

[10] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: Reading the intentions of humans and robots," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 4132–4139, Oct. 2018.

[11] R. El-Yaniv et al., "On the foundations of noise-free selective classification," *J. Mach. Learn. Res.*, vol. 11, pp. 1605–1641, 2010.

[12] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[13] A. Graves, "Practical variational inference for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2348–2356.

[14] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1321–1330.

[15] M. Hassan et al., "Stochastic scene-aware motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11354–11364.

[16] X. Huang, S. G. McGill, B. C. Williams, L. Fletcher, and G. Rosman, "Uncertainty-aware driver trajectory prediction at urban intersections," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 9718–9724.

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[18] B. Ivanovic, Y. Lin, S. Shrivastava, P. Chakravarty, and M. Pavone, "Propagating state uncertainty through trajectory forecasting," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 2351–2358.

[19] T. M. Iversen, A. G. Buch, and D. Kraft, "Prediction of ICP pose uncertainties using monte carlo simulation with synthetic depth images," in *Proc. Int. Conf. Intell. Robots Syst.*, 2017, pp. 4640–4647.

[20] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5308–5317.

[21] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *Proc. Brit. Mach. Vis. Conf. Assoc.*, 2017.

[22] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5580–5590.

[23] J. Kundu, S. Seth, P. YM, V. Jampani, A. Chakraborty, and R. Babu, "Uncertainty-aware adaptation for self-supervised 3d human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 20416–20427.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6405–6416.

[25] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee, "Convolutional sequence to sequence model for human dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5226–5234.

[26] J. Liu, Z. Lin, S. Padhy, D. Tran, T. B. Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 7498–7512.

[27] Y. Liu, A. Rahimi, P.-C. Luan, F. Rajič, and A. Alahi, "Sim-to-real causal transfer: A metric learning approach to causally-aware interaction representations," 2023, *arXiv:2312.04540*.

[28] H. Ma, J. Li, R. Hosseini, M. Tomizuka, and C. Choi, "Multi-objective diverse human motion prediction with knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8151–8161.

[29] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6427–6436.

[30] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 5441–5450.

[31] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 474–489.

[32] W. Mao, M. Liu, and M. Salzmann, "Generating smooth pose sequences for diverse human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13289–13298.

[33] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9488–9496.

[34] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4674–4683.

[35] R. T. McAllister et al., "Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning," in *Proc. Int. Joint Conferences Artif. Intell.*, 2017, pp. 4745–4753.

[36] R. M. Neal, *Bayesian Learning for Neural Networks*. Berlin, Germany: Springer, 2012.

[37] P. Nikdel, M. Mahdavian, and M. Chen, "DMMGAN: Diverse multi motion prediction of 3D human joints using attention-based generative adversarial network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 9938–9944.

[38] B. Parsaeifard, S. Saadatnejad, Y. Liu, T. Mordan, and A. Alahi, "Learning decoupled representations for human pose forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 2294–2303.

[39] J. Ren et al., "Likelihood ratios for out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 14707–14718.

[40] J. H. Ricketts and G. A. Head, "A five-parameter logistic equation for investigating asymmetry of curvature in baroreflex studies," *Amer. J. Physiol.-Regulatory, Integrative Comput. Physiol.*, vol. 277, pp. R441–R454, 1999.

[41] S. Saadatnejad, M. Bahari, P. Khorsandi, M. Saneian, S.-M. Moosavi-Dezfooli, and A. Alahi, "Are socially-aware trajectory prediction models really socially-aware?," *Transp. Res. Part C: Emerg. Technol.*, vol. 141, 2022, Art. no. 103705.

[42] S. Saadatnejad, Y. Gao, K. Messaoud, and A. Alahi, "Social-transmotion: Promptable human trajectory prediction," 2024, *arXiv:2312.16168*.

[43] S. Saadatnejad, Y. Z. Ju, and A. Alahi, "Pedestrian 3D bounding box prediction," 2022, *arXiv:2206.14195*.

[44] S. Saadatnejad et al., "A generic diffusion-based approach for 3D human pose prediction in the wild," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 8246–8253.

[45] T. Salzmann, M. Pavone, and M. Ryll, "Motron: Multimodal probabilistic human motion forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6447–6456.

[46] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11189–11198.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[48] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24804–24816.

[49] J. V. Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9690–9700.

[50] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[51] L. Vianello, J.-B. Mouret, E. Dalin, A. Aubry, and S. Ivaldi, "Human posture prediction during physical human-robot interaction," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 6046–6053, Jul. 2021.

[52] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 601–617.

[53] F. B. Wagner et al., "Targeted neurotechnology restores walking in humans with spinal cord injury," *Nature*, vol. 563, no. 7729, pp. 65–71, Nov. 2018.

[54] J. Wang, Y. Zhang, and X. Lan, "Automatic cluster number selection by finding density peaks," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 13–18.

[55] Y. Wang, Z. Shi, X. Guo, X. Liu, E. Zhu, and J. Yin, "Deep embedding for determining the number of clusters," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 8173–8174.

[56] Y. Xiao and W. Y. Wang, "Quantifying uncertainties in natural language processing tasks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7322–7329.

[57] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[58] S. Xu, Y.-X. Wang, and L.-Y. Gui, "Diverse human motion prediction guided by multi-level spatial-temporal anchors," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 251–269.

[59] Y. Yuan and K. Kitani, "DLow: Diversifying latent flows for diverse human motion prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 346–364.