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

DMMGAN: Diverse Multi Motion Prediction of 3D Human Joints using Attention-Based Generative Adversarial Network

Payam Nikdel, Mohammad Mahdavian, Mo Chen

***Abstract*— Human body motion prediction is a fundamental part of many human-robot applications. Despite the recent progress in the area, most studies predict human body motion relative to a fixed joint and only limit their model to predict one possible future motion, or both. However, due to the complex nature of human motion, a single prediction cannot adequately reflect the many possible movements one can make. Also, for any robotics application, prediction of the full human body motion including the absolute 3D trajectory – not just a 3D body pose relative to the hip joint – is needed. In this paper, we try to address these two shortcomings by proposing a transformer-based generative model for forecasting multiple diverse human motions. Our model generates *N* future possible body motions given the human motion history. This is achieved by first predicting the pose of the body relative to the hip joint as was done in prior work. Then, our proposed *Hip Prediction Module* predicts the trajectory of the hip position relative to a global reference frame for each predicted pose frame, an aspect of human body motion neglected by previous work. To obtain a set of diverse predicted motions, we introduce a similarity loss that penalizes the pairwise sample distance. Our system not only outperforms the state-of-the-art in human motion prediction, but also is able to predict a diverse set of future human body motions, including the hip trajectory.**

I. Introduction

An important ability of an intelligent system interacting with humans is to estimate plausible human body poses and trajectories in 3D space. With the advancement of artificial intelligence, there are multiple industrial applications for such algorithms in human-robot interactions (HRI) [1], autonomous driving [2] or visual surveillance [3]. Specifically, detailed human 3D body motion prediction plays a crucial role in many robotic applications, such as robot following ahead of a human [4, 5] or crowd navigation [6]. In the past decade, with the popularity of deep learning, sequence-to-sequence (seq2seq) prediction methods such as those involving Recurrent Neural Networks (RNN) [7] have shown promising results, and have become a viable alternative to conventional human motion prediction methods [8, 9]

In general, the 3D human motion prediction problem can be divided into *Human Pose Prediction* and *Human Trajectory Prediction*. *Pose* refers to a relative position of all body joints with respect to the hip joint and *Trajectory* refers to the hip joint path while the entire body moves in 3D space. For solving both problems, seq2seq models have been utilized successfully with room for improvement. Predicting a human future motion sequence can be defined as a probabilistic or deterministic problem [10]. In probabilistic methods, similar to how our brain performs, we predict multiple future motion sequences for an observed motion sequence. Arguably, the probabilistic approach is preferred in robotic applications as it provides more assurance by considering a set of possible scenarios. However, probabilistic methods may reduce the accuracy of each individual predicted sequence. Deep generative models such as generative adversarial networks (GANs) are one of the leading deep neural network architectures that help such methods achieve reasonable accuracy. Notably, DLow [11] is the state-of-the-art method that uses deep generative models and a novel sampling method for multi-future *pose* predictions. On the other hand, deterministic methods aim to predict one single sequence more accurately, while not considering the diverse and multi-modal nature of human behaviour which reduces their practicality in some robotic applications. Furthermore, works on human trajectory predictions are sometimes limited due to only considering the hip movements and ignoring other joints while making a prediction, even though the joints can provide valuable information about how the hip may move in space.

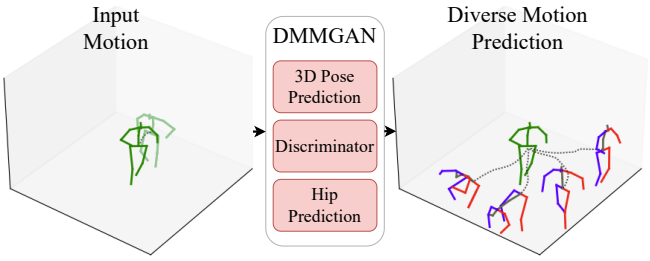

Fig. 1: Given a sequence of 3D human motions, our system generates a diverse set of future motions. The *3D pose prediction* module generates diverse 3d poses while *hip prediction* module estimates the human trajectory together forming a 3d human motion. The *discriminator* module distinguishes a real 3D human motion from a generated one.

In recent years, with the emergence of transformers [12], many works attempted to solve the human motion prediction problem by acquiring spatio-temporal autoregressive [13] or non-autoregressive transformers [14]. For instance, Aksan et al. [13] introduce a spatio-temporal-based transformer for 3D human motion prediction. It uses an autoregressive model that predicts human future *poses*. Gonzlez et al. [14] improved the transformer model's inference speed by making it non-autoregressive while trading off the accuracy of long-term predictions.

In this work, we combine the benefit of both probabilistic and deterministic methods to provide multiple accurate predictions for both 3D human trajectory and pose. We hope this opens doors to practical use in real robotic applications. To generate multiple future human motions, we

School of Computing Science, Simon Fraser University (SFU), Canada. {pnikdel, mmahdavi, mochen}@sfu.ca

This work received support from Terramera Inc., the Mitacs Accelerate Program, Amii, and the CIFAR Program. M. Mahdavian received support from the SFU Graduate Deans Entrance Scholarship.

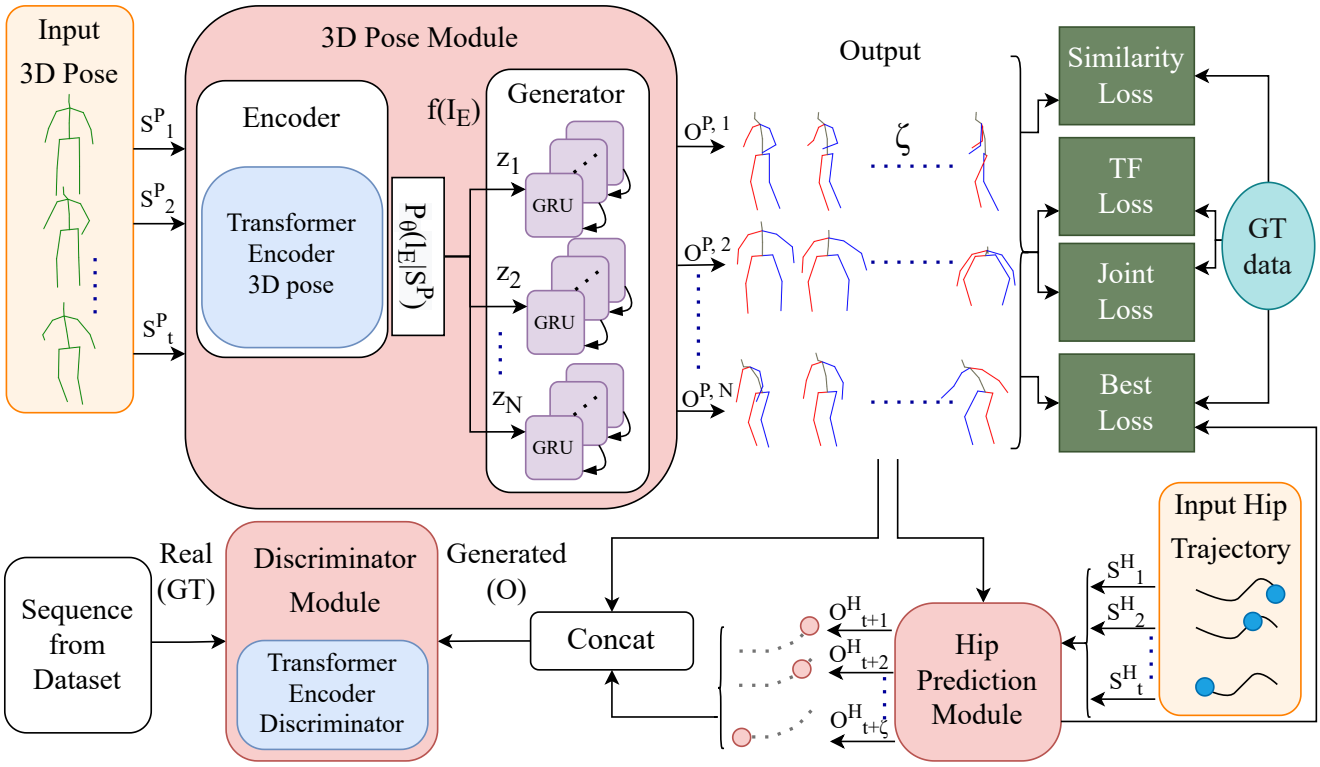

Fig. 2: System overview: Given a sequence of 3D human motion, our method generates N future sequences of human 3D motion using a discriminator and four loss functions. Our system consists of three main parts. The first part is predicting the human 3D pose (*3D Pose module*) by receiving a history of the human 3D pose. The second part is the *Hip Prediction* module (more details in Fig. 3) which predicts the future position of the hip joint for each of the predicted human 3D poses. Finally, the discriminator module learns the distribution of the Human 3.6M dataset by distinguishing between generated and real data. The system uses the discriminator loss to generate sequences similar to the dataset distribution while using four supervised loss functions to promote accuracy and diversity. See Fig. 3 for Transformer Encoder architecture.

use a conditional generative adversarial network (CGAN) with a transformer-based encoder for better encoding of the observed sequence. At the end, a GRU combined with a GAN provides multiple future predictions autoregressively.

The contributions of this paper are as follows:

- We propose a novel deep generative architecture involving transformer-based encoders to predict a diverse set of possible human body motions.
- We provide a real-time solution for diverse 3D human motion prediction, including both pose and trajectory prediction, which can potentially be more suitably used for robotics and autonomous car applications.
- In addition to providing both pose and trajectory predictions, our work achieves better accuracy compared to the state-of-the-art models in standard evaluation metrics.

II. Related Work

The human pose prediction problem is divided into probabilistic and deterministic approaches. Early deterministic approaches use RNN modules for making predictions [7, 15]–[17]. In recent years, Graph Convolutions Networks (GCN) [18, 19] and Spatio-Temporal [13, 19, 20] methods attempt to improve the predictions by better learning the spatial and temporal dependencies between the joints. More recently, transformers [12] parallelized the training process which improved the accuracy and speed of the predictions [13, 14].

On the other hand, probabilistic approaches gained popularity with the development of GANs. These methods [21]–[24] usually use CGANs or Conditional Variational Autoencoders (CVAEs). As one of the state of the art methods, DLow [11] generates a diverse set of samples from a pretrained deep generative model. The authors train a mapping function that samples diversely using a pretrained CVAE. To diversify the samples, they train a set of learnable mapping functions with correlated latent space that use an energy-based formulation based on pairwise sample distance. We use DLOW as one of our baselines in Section V. Also, Yan et al. [25] developed a Motion Transformation Variational Autoencoder (MT-VAE) to generate multiple diverse and plausible motion sequences for facial and full body motion from an observed sequence. More recently, Agand et al. [26] developed a probabilistic and optimal approach for human navigational intent inference. All these algorithms make predictions from a human 3D pose sequence. There are a few works that perform pose prediction directly from video [27], but they can be less accurate. In this work, we assume that the human motion (pose and trajectory) are available as the

input to our model, since for a real-world application we can simply get these 3D motion using hardware such as the ZED2 camera[1].

To predict the human motion for a robotics application we need both the 3D pose and trajectory. Human trajectory prediction has been studied and implemented using RNNs [28, 29] and transformers [30]. But there are very few prior works that attempted to combine the human pose and trajectory predictions [31]. This combination can improve each individual prediction as the two parts are interdependent.

III. Problem Setup

Our framework predicts a diverse set of human motions. The input is a sequence of 3D body motion $S = \{S_{t-\alpha}, S_{t-\alpha+1}, ..., S_t\}$ of the past human's skeleton movements capture up to the current time t where $S_i \in \mathbb{R}^{51}$ represents the 3D positions of 17 human joints at time i. The outputs of our system are N possible sequences of future 3D human motion $O_i^\gamma = \{O_{t+1}^\gamma, ..., O_{t+\zeta}^\gamma\}$ where $\gamma \in 1, ..., N$ is the sequence number and ζ is the forecast duration. We divide the human 3D motion into two parts so that $S_i = (S_i^H, S_i^P)$ and $O_i = (O_i^H, O_i^P)$. The position of the hip joint is denoted by S^H and O^H for input and output hip trajectories. The relative positions of all joints with respect to the hip joint (called 3D pose, or just *pose*), denoted by S^P and O^P for input and output 3D pose sequences.

IV. Method

The overall framework of our system is summarized in Fig. 2. Our method learns to generate valid and rich human motions by leveraging the Human 3.6M dataset [32]. It divides the prediction of human 3D motion into predicting the joints motion relative to the hip joint (3D pose) and predicting the 3D position of the hip joint in the global frame for each predicted 3D pose (human trajectory). We estimate the human trajectory by considering both the predicted 3D pose and the trajectory history.

Specifically, we design our model to benefit from both paired and unpaired data by introducing four supervised losses and a discriminator loss respectively. Here, given a sequence of 3D motion $\{S_{t-\alpha}, \ldots, S_t\}$, a transformer encoder learns representation of the input in a latent space. Then, a generator uses this latent representation to output N future motions. To train our system, we use 5 losses. The *Best Loss* finds the best match between all the outputs and the ground truth data. The *Teacher Forcing Loss* improves the final prediction by randomly feeding ground truth instead of the model prediction in the decoding phase. Similar to the Best Loss, the Teacher Forcing Loss only applies for the output that matches the most closely with the ground truth. The *Similarity Loss* promotes diversity by penalizing the pairwise distance between the N generated sequences, and lastly, we use the *Joint Loss* to encourage joint length constraints. We combine these losses with the *Discriminator Loss* to generate plausible sequences matching the Human 3.6M dataset [32].

[1]https://www.stereolabs.com/zed-2/

A. Model Architecture

Our model consists of three main modules, the first module is the *3D pose module*, which generates N sequences of human 3D pose (relative to the hip joint). The second module is the *Hip Prediction module*, which predicts the trajectory of the hip joint in the global frame for each predicted human 3D pose. Finally, the last module is the *Discriminator module*, which learns the distribution of the dataset by distinguishing between the real and generated 3D sequences of human's motion.

1) 3D Pose Module: The 3D Pose module consists of two parts, as shown in Fig. 2. The first part is the encoder. Given a sequence of human 3D pose S^P, it outputs a latent representation l that encodes the past motion $P_\theta(l_E|S^P)$. Our encoder network uses a Transformer architecture, as shown in Fig. 3, to learn meaningful information over a sequence of 3D poses, similar to the model introduced by Vaswani et al. [12].

The second part is the generator. It forecasts N sequences of human 3D pose $O^{P,1}, \ldots, O^{P,N}$ given the past latent representation l_E. Instead of using random variables as the input of the generator to forecast the future, we design it to learn a mapping from the latent representation to N priors $z = f(l_E)$. Then it initializes N generator networks with Gated Recurrent Units (GRU) [33], each of which forecasts a sequence of future 3D pose based on their prior $P_{\phi n}(O_n^P|z_n), n \in \{1, ..., N\}$.

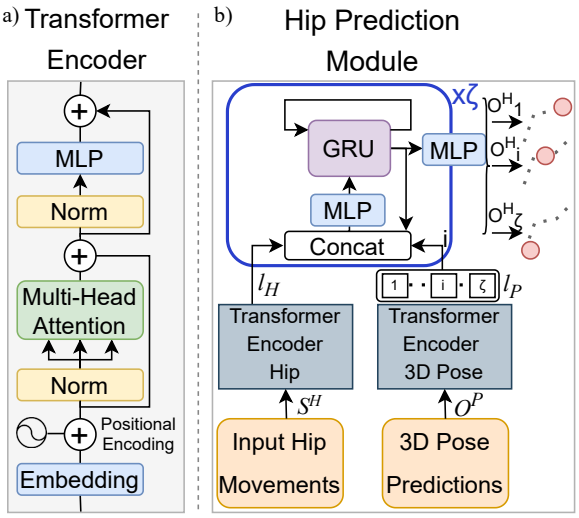

Fig. 3: a) The Transformer Encoder [12] and b) the *Hip Prediction* module architectures. The *Hip Prediction* module, estimates the hip joint positions of each predicted 3D pose by receiving the history of the hip movements and the motion predicted by the *3D Pose* module.

2) Hip Prediction Module: The second module is the *hip prediction* module. Given the 3D pose predictions O^P and the trajectory history S^H, it estimates the position of the hip for each predicted 3D pose.

Fig. 3 shows the architecture of the *Hip Prediction* module. It uses two Transformer encoders. The first one learns a representation l_H from the observed hip movements S^H and the second one learns a representation l_P from a predicted

3D pose sequence. If the transformer embedding has σ dimensions and the input has a length of α frames and we predicted a 3D pose sequence with ζ frames, the output of S^H has $\alpha\sigma$ and S^P has $\zeta\sigma$ dimensions. The GRU gets the concatenation of $l_{P,i}$, l_H and the previous output of the GRU as its input, to predict the position of the hip at time i for $i = 1, ..., \zeta$.

3) Discriminator Module: The last module is the discriminator. Here we use a Transformer-based Encoder architecture shown in Fig. 3a. The input of the discriminator is the full human 3D motion, consisting of the hip trajectory and the 3D pose trajectory. The discriminator needs to distinguish between the real and the generated data (Fig. 2).

B. Model Training

During training, we exploit paired data by introducing four supervised losses to promote the diversity and accuracy of the predictions. We also benefit from unpaired data by using a discriminator that learns to distinguish between the real and generated data. In the following we use the ground truth, GT, term to refer to the paired data, GT^P and GT^H to refer to the ground truth paired 3D pose and hip trajectory respectively.

1) Discriminator Loss: We implement the discriminator loss based on the Wasserstein Generative Adversarial Network (WGAN) [34]. To make the training more stable we used the Gradient Penalty (GP) version of the WGAN. If f is the discriminator network, the GP WGAN critic's loss function is defined as follows:

$$\mathcal{L}_{cWGAN} = \mathop{\mathbb{E}}_{O\sim P_g}[f(O)] - \mathop{\mathbb{E}}_{GT\sim P_r}[f(GT)] \tag{1}$$

$$\mathcal{L}_{cGP} = \mathcal{L}_{WGAN} + \lambda \mathop{\mathbb{E}}_{\bar{x}\sim P_{\bar{x}}}[(||\nabla_{\bar{x}} f(\bar{x})||_2 - 1)^2] \tag{2}$$

where (1) is the original critic loss function of WGAN method and the last term of (2) is the gradient penalty term. Consider a line connecting real (P_r) to generated (P_g) distributions. $P_{\bar{x}}$ is the distribution of these samples and λ is the weight of the gradient penalty.

The second part of the discriminator loss function is the generator objective. The objective of the generator is to minimize the distance between P_g and P_r by maximizing the expectation of the generated samples:

$$\mathcal{L}_g = -\mathop{\mathbb{E}}_{O\sim P_g}[f(O)] \tag{3}$$

2) Best Loss: Given a sequence of human's 3D motion, our model predicts multiple forecasts of future motions. Using the discriminator loss, these forecasts would be similar to the distribution of the dataset. The Best Loss minimizes the distance between the closest prediction and the GT data using mean squared error (MSE). The Best Loss is defined as follows:

$$\mathcal{L}_{best} = \sum_{T=t+1}^{t+\zeta} MSE(O_T^{\Gamma}, GT_T) \tag{4}$$

$$\text{where } \Gamma = \mathop{\arg\min}_{\gamma=1,...,N} \sum_{T=t+1}^{t+\zeta} D(O_T^{P,\gamma}, GT_T^P) \tag{5}$$

$$\text{and } D(O_t^{\Gamma}, GT_t) = \sum_{T=t+1}^{t+\zeta} \sum_{j=1}^{17} d(O_{t,j}^{\Gamma}, GT_{t,j}) \tag{6}$$

Here, D is the distance between two 3D motion predictions and d is the Euclidean distance between two joints.

3) Teacher Forcing Loss: After calculating the predicted sequence that matches with the GT, the *Teacher Forcing* (TF) loss is calculated by randomly using the next frame from the GT instead of the last prediction in the GRU (Fig. 2 Generator). The TF loss can be especially useful in reducing the final displacement error as the model can learn to predict the next frames by using a combination of the GT and its own predictions [35].

4) Similarity Loss: We define the *Similarity* loss to increase the variety of the model predictions. We first find the distance between each pair of the predicted human 3D pose. Then select the two predictions, Γ_1 and Γ_2, with the shortest distance.

$$\Gamma_1, \Gamma_2 = \mathop{\arg\min}_{\substack{\gamma_1\in\{1,...,N\},\\ \gamma_2\in\{1,...,N\}\setminus\gamma_1}} \sum_{T=t+1}^{t+\zeta} D(O_T^{P,\gamma_1}, O_T^{P,\gamma_2}) \tag{7}$$

We can define the distance of each two joints of Γ_1 and Γ_2 by:

$$distJoints_j = \sum_{T=t+1}^{t+\zeta} d(O_{T,j}^{P,\Gamma_1}, O_{T,j}^{P,\Gamma_2}) \tag{8}$$

Then we apply the negative of MSE to the joints that exceed the average *Similarity loss* threshold of ϵ. We can define the $Similarityloss$ as follows:

$$\mathcal{L}_{similarity} = -\frac{1}{16}\sum_{j=0}^{16} distPenalize_j^2, \text{ where} \tag{9}$$

$$distPenalize_j = \begin{cases} 0 & \text{if } distJoints_j < \epsilon \\ distJoints_j & \text{otherwise} \end{cases} \tag{10}$$

To make the training more stable we use the *Similarity loss* only during the first *M* steps of the training.

5) Joint Loss: As human's bone length stay the same, joint Loss works as a regularizer that helps the model by forcing it to keep the bone length similar over time. If V is the set of vertices of a graph representing all human joints and E is the edges of this graph representing all human bones, then the joint loss is defined as follows:

$$\mathcal{L}_{joint} = \sum_{(i,j)\in E} \sum_{\gamma=1}^{N} MSE(J_{i,j}^{P,\gamma}, J_{i,j}^{P,GT}) \tag{11}$$

$$\text{where } J_{i,j}^{P,\gamma} = \frac{1}{\zeta}\sum_{T=t+1}^{t+\zeta} (d(O_{T,i}^{P,\gamma}, O_{T,j}^{P,\gamma})), \tag{12}$$

$$J_{i,j}^{P,GT} = \frac{1}{\zeta}\sum_{T=t+1}^{t+\zeta} (d(GT_{T,i}^{P}, GT_{T,j}^{P,})) \tag{13}$$

C. Data Prepossessing

To improve the model prediction and avoid over-fitting, we convert each 3D position in a sequence of human motion to a relative coordinate system based on the position of the hip joint at the time t. We also normalize each skeleton 3D pose ($\mu = 0, \sigma = 1$).

D. Dataset

For our experiments and training, we use the Human 3.6M dataset [32]. Human 3.6M is a large dataset with 7 actors[2]. For each actor, there are 15 actions that are recorded using a high-speed motion capture system at 50 Hz. Similar to DLow [11], we use 17 joints skeleton and train on actors S1, S5, S6, S7 and S8 while testing on S9 and S11. For future prediction, our model observes 0.5 seconds sequence of human's body motion to forecast the next 2 seconds.

V. Experiments and Results

Our method is specifically designed to forecast 3D motions that are suitable for autonomous car or robotics applications. It can predict the human 3D pose (position of joints relative to the hip joint) while predicting their trajectory (hip joint) separately. Most of the previous works only predict the human 3D pose without the human's hip trajectory.

Here we designed two experiments. The first one evaluates our 3D pose prediction without the trajectory prediction module. Then in the second experiment, we evaluate our full system. For both experiments, we used the same model (DMMGAN). Our model can run at 10 frames per second (FPS) on a GeForce 1080 GPU. Since most robotics applications require the observation to come with a frequency of fewer than 10 FPS, we train our model and the baselines using the Human3.6M [32] at 10 FPS. For DLow and our methods, we predict 10 sequences per observation ($N = 10$).

To evaluate our model versus the baselines we measure the accuracy and diversity using the following metrics (we are using the evaluation metrics similar to [11, 36]):

1) Average Pairwise Distance (APD): Evaluates diversity among the predictions. We calculate the APD by averaging the pairwise distance between all pairs of 3D pose samples between the predictions. The APD is calculated as $\frac{1}{N\times(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}||O_i^P - O_j^P||$.

2) Average Displacement Error (ADE): Mean squared distance between the ground-truth and the closest prediction. We define the ADE for both the 3D pose and the hip trajectory movements. We first calculate the closest prediction index, Γ, using the 3D pose predictions by: $\Gamma = \arg\min_{\gamma=1,\dots,N}\sum_{T=t+1}^{t+\zeta} D(O_T^{P,\gamma}, GT_T^P)$. Then use this index to calculate the ADE for both the 3D pose and the trajectory: $ADE_p = \sum_{T=t+1}^{t+\zeta} D(O^{P,\Gamma_T}, GT_T^P)$ and $ADE_h = \sum_{T=t+1}^{t+\zeta} D(O^{H,\Gamma_T}, GT_T^H)$.

3) Final Displacement Error (FDE): Mean squared distance between the final ground-truth and the closest final prediction. Similar to ADE, we first calculate the closest final prediction index by $\beth = \arg\min_{\gamma=1,\dots,N} D(O_{\gamma,t+\zeta}, GT_{t+\zeta})$. Then we calculate the FDE for both the 3D pose and the trajectory: $FDE_p = D(O_{t+\zeta}^{P,\beth}, GT_{t+\zeta}^P)$ and $FDE_h = D(O_{t+\zeta}^{H,\beth}, GT_{t+\zeta}^H)$.

4) Multi-modal ADE (MADE): To evaluate our system's ability to generate multi-modal predictions, we used the multi-modal version of ADE [11, 36]. The MADE uses multi-modal GT future motions by grouping similar past motions.

5) Multi-modal FDE (MFDE): Similar to MADE, The MFDE is the multi-modal version of FDE [11, 36].

Approach	APD ↑	ADE (m) ↓	FDE (m) ↓	MADE (m) ↓	MFDE (m) ↓
DMMGAN (Ours)	**5.81**	**0.44**	**0.52**	**0.54**	**0.60**
DLow	5.53	0.48	0.61	0.55	0.63
STPOTR	NA	0.50	0.75	NA	NA

TABLE I: Comparison of our systems versus two baselines for the 3D Pose experiment.

A. 3D Pose Experiment

In the first experiment, we evaluate our 3D Pose generation module. Here, we compare our method against two baselines. The first one is DLow [11], the state-of-the-art in diverse human 3D pose forecasting which outperforms all the currently known methods to the best of our knowledge. The authors of DLow [11] provide detailed comparisons to several other methods, which we will omit in this paper for brevity. The second baseline is STPOTR [31], a more recent method that also focuses on 3D human motion prediction for robotics applications. STPOTR predicts only one future motion so we cannot use it for multi-modal evaluation.

Table I shows the results of this experiment. Our method outperforms both of the baselines and achieves the highest diversity while keeping both ADE and FDE lowest. Our method also has the highest coverage of the multi-modal ground-truth (MADE and MFDE). Also, we visually evaluate our method against DLow, in Fig. 4, we visualize the 10 end poses of our predictions versus the DLow for 2 random samples. In both methods, we can see a comparable accuracy against the ground-truth data (GT). Although the diversity of our method is close to DLow, closer examination of Seq 1 shows that our method predicted sitting down, crouching, lying down, walking left and right, while DLow has qualitatively less diverse samples.

B. Full 3D Motion Experiment

The second experiment evaluates our full system. In order to compare our system with a state-of-the-art diverse 3D motion model, we repurposed and retrained DLow [11] to forecast the human's trajectory by adding the hip joint to the joints **Adapted DLow** predicts. We also compare our system with STPOTR [31], which is one of the few papers that provides full 3D motion (pose and hip) prediction. We also include two variations of our models as an ablation study. The first model is **MMGAN** which is our full system trained without the *similarity loss* and the second one is called **HipOnly** which is our *Hip Prediction* module without the 3D pose prediction inputs. The HipOnly model evaluates the impact of the predicted 3D pose data on the accuracy of the trajectory estimation. (Fig. 3b without the right 3D pose Transformer encoder).

Based on the result of this experiment (Table II), our method outperforms the baselines by achieving the highest diversity while keeping the ADE and FDE lowest. In Fig. 5, we compare our prediction versus Adapted DLow and the ground-truth (GT) qualitatively [3]. In these examples, Adapted

[2]There are 4 other actors without ground truth data

[3]Please refer to https://youtu.be/osJuFbtJsMg for more examples.

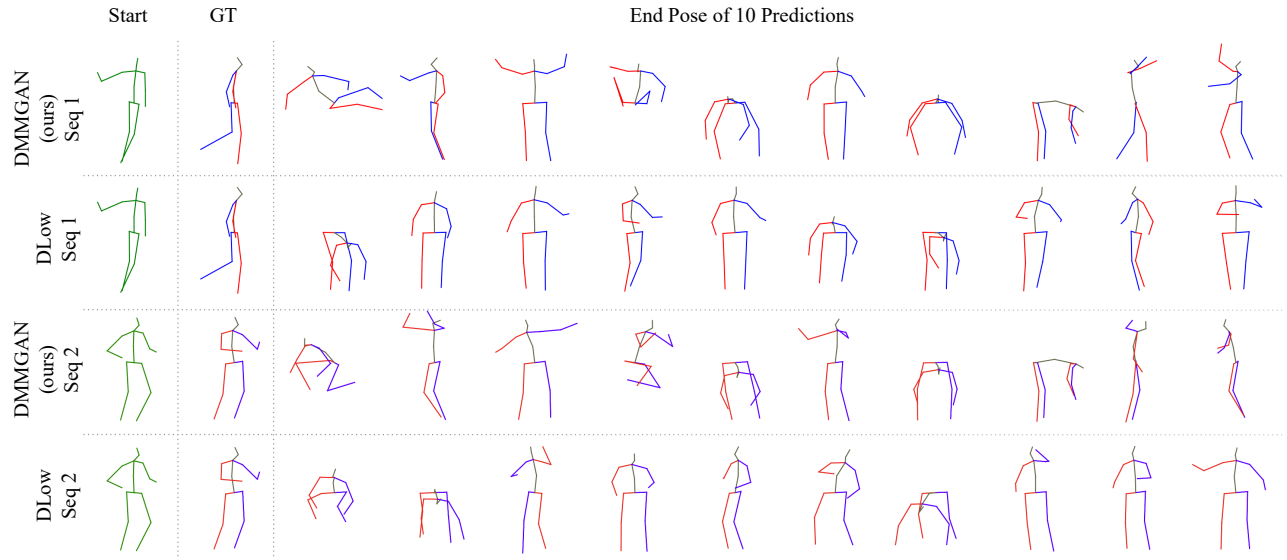

Fig. 4: Qualitative results of 3D pose predictions comparing our method, DMMGAN, to DLow in terms of diversity.

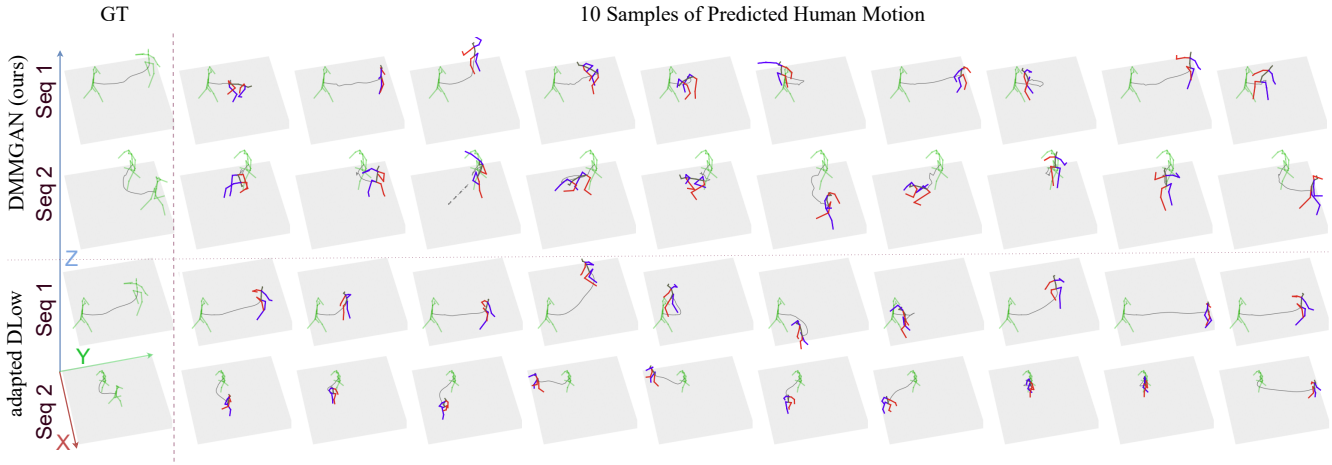

Fig. 5: Qualitative results of 3D motion predictions comparing our method, to DLow in terms of diversity.

Approach	APD ↑	ADE (m) ↓		FDE (m) ↓		MADE (m) ↓		MFDE (m) ↓	
		Pose	Trajectory	Pose	Trajectory	Pose	Trajectory	Pose	Trajectory
Adapted DLow	5.55	0.483	0.195	0.621	0.457	0.563	0.306	0.649	0.553
STPOTR	NA	0.507	0.139	0.758	0.277	NA	NA	NA	NA
ours:									
DMMGAN (ours)	**5.81**	0.443	0.122	0.520	0.228	**0.540**	**0.192**	**0.597**	**0.342**
MMGAN	2.01	**0.422**	**0.104**	**0.494**	**0.190**	0.589	0.198	0.665	0.360
HipOnly	NA	NA	0.156	NA	0.306	NA	NA	NA	NA

TABLE II: Comparison of our systems versus two baselines for the full 3D motion experiment.

DLow predicted only walking movement while DMMGAN could capture more diverse motions.

The *HipOnly* model achieved a higher FDE and ADE compared to our model, which shows the benefit of using an attention-based 3D pose generator during trajectory forecasting. The results also highlight the impact of the *similarity loss* on the diversity of the predicted 3D motions. Our model without the similarity loss, MMGAN, achieved APD of 2 versus 5.8 for our full system. It is interesting to note that by removing the *similarity loss*, the model achieves a lower ADE and FDE with the cost of less diverse predictions.

VI. CONCLUSION

We proposed DMMGAN, a novel method to predict diverse human motions. DMMGAN combined a generative adversarial network with Transformer based encoders to generate both the trajectory and the 3D pose of human motions.

Our implementation outperformed the previous state of the art in diverse human 3D pose prediction while also predicting the human's trajectory.

REFERENCES

[1] G. Ferrer and A. Sanfeliu, "Bayesian human motion intentionality prediction in urban environments," *Pattern Recognition Letters*, vol. 44, pp. 134–140, 2014.

[2] M. Gulzar, Y. Muhammad, and N. Muhammad, "A survey on motion prediction of pedestrians and vehicles for autonomous driving," *IEEE Access*, 2021.

[3] K. M. Rashid and A. H. Behzadan, "Enhancing motion trajectory prediction for site safety by incorporating attitude toward risk," *Computing in Civil Engineering 2017*, pp. 425–433, 2017.

[4] P. Nikdel, R. Vaughan, and M. Chen, "LBGP: Learning Based Goal Planning for Autonomous Following in Front," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 3140–3146.

[5] G. P. Moustris and C. S. Tzafestas, "Assistive front-following control of an intelligent robotic rollator based on a modified dynamic window planner," in *Biomedical Robotics and Biomechatronics (BioRob), 2016 6th IEEE Int. Conf.* IEEE, June 2016, pp. 588–593.

[6] C. Chen, S. Hu, P. Nikdel, G. Mori, and M. Savva, "Relational graph learning for crowd navigation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 007–10 013.

[7] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2891–2900.

[8] A. M. Lehrmann, P. V. Gehler, and S. Nowozin, "Efficient nonlinear markov models for human motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1314–1321.

[9] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 15–48, 2012.

[10] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3d human motion prediction: A survey," *Neurocomputing*, vol. 489, pp. 345–365, 2022.

[11] Y. Yuan and K. Kitani, "Dlow: Diversifying latent flows for diverse human motion prediction," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–364.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

[14] A. Martínez-González, M. Villamizar, and J.-M. Odobez, "Pose transformers (POTR): Human motion prediction with non-autoregressive transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2276–2284.

[15] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE International conference on Computer Vision*, 2015, pp. 4346–4354.

[16] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Ji, S. Lu, and L. Cheng, "Investigating pose representations and motion contexts modeling for 3D motion prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[17] Z. Zhang, J. Rhim, A. Lim, and M. Chen, "A Multimodal and Hybrid Framework for Human Navigational Intent Inference," in *International Conference on Intelligent Robots and Systems*, 2021.

[18] Y. Bin, Z.-M. Chen, X.-S. Wei, X. Chen, C. Gao, and N. Sang, "Structure-aware human pose estimation with graph convolutional networks," *Pattern Recognition*, vol. 106, p. 107410, 2020.

[19] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 7760–7775, 2021.

[20] O. Medjaouri and K. Desai, "HR-STAN: High-Resolution Spatio-Temporal Attention Network for 3D Human Motion Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2540–2549.

[21] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proceedings of the IEEE International conference on Computer Vision*, 2017, pp. 3332–3341.

[22] X. Lin and M. R. Amer, "Human motion modeling using DVGANS," *arXiv preprint arXiv:1804.10652*, 2018.

[23] Barsoum, Emad and Kender, John and Liu, Zicheng, "HP-GAN: Probabilistic 3D Human Motion Prediction via GAN," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1499–149 909.

[24] S. Aliakbarian, F. S. Saleh, M. Salzmann, L. Petersson, and S. Gould, "A stochastic conditioning scheme for diverse human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5223–5232.

[25] X. Yan, A. Rastogi, R. Villegas, K. Sunkavalli, E. Shechtman, S. Hadap, E. Yumer, and H. Lee, "Mt-vae: Learning motion transformations to generate multimodal human dynamics," in *Proceedings of the European conference on Computer Vision (ECCV)*, 2018, pp. 265–281.

[26] P. Agand, M. Taherahmadi, A. Lim, and M. Chen, "Human navigational intent inference with probabilistic and optimal approaches," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8562–8568.

[27] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik, "Predicting 3d human dynamics from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7114–7123.

[28] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 961–971.

[29] C. Song, Z. Chen, X. Qi, B. Zhao, Y. Hu, S. Liu, and J. Zhang, "Human trajectory prediction for automatic guided vehicle with recurrent neural network," *The Journal of Engineering*, vol. 2018, no. 16, pp. 1574–1578, 2018.

[30] L. Achaji, T. Barry, T. Fouqueray, J. Moreau, F. Aioun, and F. Charpillet, "Pretr: Spatio-temporal non-autoregressive trajectory prediction transformer," *arXiv preprint arXiv:2203.09293*, 2022.

[31] M. Mahdavian, P. Nikdel, M. TaherAhmadi, and M. Chen, "STPOTR: Simultaneous Human Trajectory and Pose Prediction Using a Non-Autoregressive Transformer for Robot Following Ahead," in *IEEE International Conference on Robotics and Automation (ICRA), 2023 IEEE Int. Conf.*, 2023.

[32] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[33] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," *CoRR*, vol. abs/1409.1259, 2014. [Online]. Available: http://arxiv.org/abs/1409.1259

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223. [Online]. Available: https://proceedings.mlr.press/v70/arjovsky17a.html

[35] R. J. Williams and D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[36] Y. Yuan and K. M. Kitani, "Diverse Trajectory Forecasting with Determinantal Point Processes," in *ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=ryxnY3NYPS