IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 June 2024, accepted 23 July 2024, date of publication 29 July 2024, date of current version 7 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434695

SURVEY

Human Motion Prediction: Assessing Direct and Geometry-Aware Approaches in 3D Space

SARMAD IDREES[1], (Graduate Student Member, IEEE), JIWOO KIM[2], JONGEUN CHOI[1], (Member, IEEE), AND SEOKMAN SOHN[3]

[1]School of Mechanical Engineering, Yonsei University, Seodaemun-gu, Seoul 03722, South Korea
[2]School of Electrical and Electronic Engineering, Yonsei University, Seodaemun-gu, Seoul 03722, South Korea
[3]Power Generation Laboratory, Korea Electric Power Research Institute, Yuseong-gu, Daejeon 34056, South Korea

Corresponding author: Jongeun Choi (jongeunchoi@yonsei.ac.kr)

This work was supported in part by Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant funded by Korea Government through MOTIE (Development of Work Safety Management Platform in Power Plant) under Grant 20206610100290, in part by the National Research Foundation of Korea (NRF) Grant funded by Korea Government through MSIT under Grant RS-2023-00221762, and in part by Korea Institute of Science and Technology (KIST) Intramural Research Grants under Grant 2E31570.

ABSTRACT Predicting 3D human motion is a complex task, owing to the unpredictable nature of human movements. The influx of deep learning innovations and the availability of extensive datasets have intensified research interest in this field. This survey provides an exhaustive review of human motion prediction algorithms and categorizes them according to their core architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Convolutional Networks (GCNs), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Transformers, and Equivariant Neural Networks (ENNs). Our key contribution is a systematic presentation of the latest prediction methodologies, classified into direct and geometry-aware modeling. We begin with the problem formulation of human motion prediction, explore assorted techniques, and discuss data representation, accompanied by a list of accessible datasets. We also identify and analyze the ongoing challenges and limitations of the current algorithms, offering insights into potential future developments in this domain.

INDEX TERMS Human motion prediction, deep learning, neural network, equivariant models.

I. INTRODUCTION

In computer vision, human motion prediction is characterized as the anticipation of future human poses based on the observation of historical frames [7]. Understanding and predicting human motion is vital for improving safety and optimizing performance across diverse sectors. For example, in autonomous driving, precise prediction of pedestrians' and vehicles' future movements can mitigate collision risks and enhance safety [8]. Furthermore, in industrial settings where collaborative robots operate alongside humans, forecasting human motion and intentions is essential to detect safety hazards and avert accidents in facilities such as factories, power plants, and construction sites [4]. Additionally, in Virtual Reality (VR) and Augmented Reality (AR) environments where users interact with digital elements, predicting user movements and intentions is essential to provide a seamless, immersive, and safe virtual experience [9], [10].

The interest of the research community in human motion prediction has grown due to advancements in deep learning algorithms, enhanced computational resources, and the availability of extensive motion-capture datasets [11], [12], [13], [14]. Predicting 3D human motion is a complex task due to the stochastic nature of human movements, the large number of degrees of freedom in the human body, and the challenge of modeling temporal dependencies in human motion [7], [15]. The challenge is further compounded by the non-linear dynamics and interactions among multiple joints and muscles in 3D space, which are difficult to model accurately. Additionally, human motion can be influenced by various factors such as fatigue, emotion, and the external environment, making accurate prediction complex [16], [17]. These factors contribute to the inherent uncertainty and variability in human motion, thus making predicting future human motion a highly

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

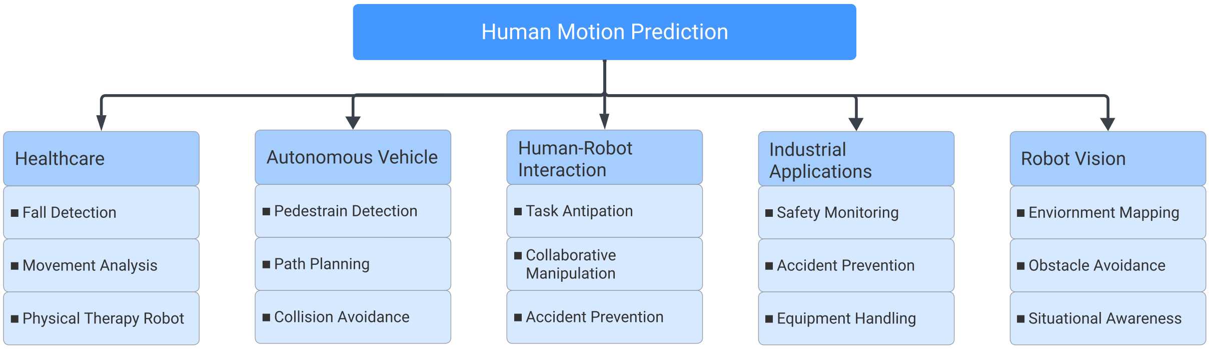


FIGURE 1. Related applications and research problems of human motion prediction task, in the field of healthcare [1], autonomous vehicle [2], human-robot interaction [3], industrial applications [4], and robot vision [5], [6].

challenging task and impeding the accuracy and reliability of 3D human motion prediction models.

Many diverse and interesting approaches have been proposed by the research community to address such challenges, including recurrent models [18], [19], [20], convolutional networks [21], [22], [23], [24], [25], [26], generative models [27], [28], [29], [30], [31], transformers [32], [33], [34]. However, additional refinements are needed to enhance the motion modeling strategy. Due to the stochastic nature of human motion, learning unnatural human behavior is always a challenging task. In this survey paper, we attempt to summarize the current challenges in motion prediction tasks, available proposed approaches to solve these challenges, and promising research directions for predicting plausible and semantically meaningful future motions.

In literature, anticipating the trajectories of road participants is closely linked with the task of predicting human motion. Despite both pursuits sharing a common objective, their ultimate outcomes diverge. In the task of trajectory prediction, our goal is to forecast the overall movement of the entire body, without considering individual body joint details. In contrast, within the domain of human motion prediction, our focus shifts toward estimating specific joint information for future frames. The scope of this survey is restricted to the exploration of human motion prediction methodologies exclusively.

While in-depth surveys such as those found in [2] focus on trajectory prediction methods, and overviews of human motion prediction are presented in [7] and [15], this survey paper contributes further by offering a systematic categorization of prediction algorithms. It specifically differentiates between direct and geometry-aware methodologies, thereby providing additional clarity and insight into the landscape of human motion prediction strategies not thoroughly explored in previous reviews. Finally, we also incorporate recently developed algorithms, reflecting current trends and future directions in this rapidly evolving field. The recent strides in equivariant group deep learning particularly highlight exciting possibilities for future research [35].

The remaining paper is organized as follows. Section II introduces the motion prediction problem formulation, our classification rules, and the human motion data representation techniques utilized in the literature. In Section III, available methods are introduced and categorized into relevant groups according to our classification rules. Section IV introduces the available datasets. Evaluation metrics and results are reported in Section V. Lastly, the limitations of current methods and possible future directions are discussed in Section VI, with the concluding remarks in Section VII.

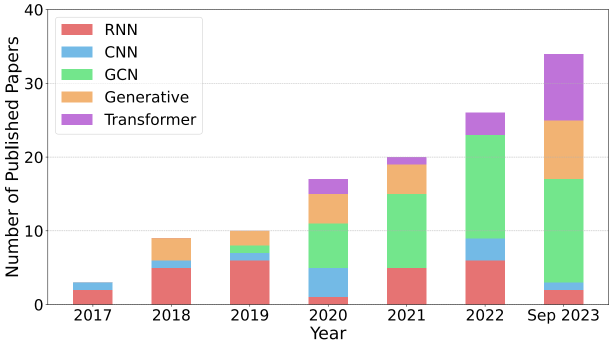


FIGURE 2. Published papers per year with respect to architecture utilized.

II. OVERVIEW

In this survey paper, our goal is to provide a comprehensive and in-depth overview of contemporary motion prediction algorithms. First, we explore the evolution of human motion prediction techniques over time. Next, we describe the formulation of the motion prediction problem. Lastly, we outline the criteria used for categorizing the methodologies within this field.

Anticipating future human motion has been the subject of extensive research, involving various methodologies from the research community. Despite significant advancements in prediction algorithms and learning techniques, achieving accurate long-term predictions remains a formidable challenge. While there have been notable improvements in predictive accuracy, existing approaches primarily excel in short-term predictions by effectively capturing immediate movements and reactions [19], [36]. For long-term predictions, current methods perform better with repetitive actions, such as walking, but struggle with unpredictable actions such as posing for a picture due to error accumulation in later stages [23], [37], [38]. Maintaining accuracy over extended periods is still problematic due to the inherent complexity of human motion. To address these challenges, it is crucial to incorporate spatio-temporal correlations and human body movement constraints, ensuring the capture of essential features necessary for reliable long-term predictions.

Initially, the task of predicting human motion was approached using Recurrent Neural Networks (RNNs) due to their ability to model time-series data [18], [36]. These initial methods laid the groundwork for further refinements in the precision of 3D pose predictions. Later studies integrated the structural details of the human body to impose movement constraints, resulting in the hierarchical-RNN techniques that hierarchically interpret the local joint connections [20], [39]. Later research utilized the capabilities of Graph Convolutional Networks (GCNs) [40] to extract features from graph data, leading to significant advancements in learning human motion dynamics [24], [37], [41]. Some researchers have suggested that stochastic models tend to generalize to a mean pose for every motion, and have proposed the use of generative models with an adversarial training loss to learn a diverse distribution of motion [27], [42], [43], [44].

The implementation of Transformer networks [45] has leveraged their attention mechanisms and prowess in capturing long-range data dependencies. In [33], a full transformer network was employed for the task of predicting human motion, while Aksan et al. [32] utilized only the encoder layer to encode spatio-temporal features. The latter approach leverages the attention mechanism of transformer networks while constraining the complexity of the network. Furthermore, the MLP-Mixer architecture [46] has been explored for human motion prediction in studies like those by [47] and [48].

In addition, recent research on Equivariant Neural Networks (ENNs) [49] aims to leverage the group symmetry to enhance performance in motion prediction [35], [50], [51]. These diverse approaches highlight the ongoing evolution and innovation in methodologies addressing the challenges in human motion prediction.

A. PROBLEM FORMULATION

The task of predicting human motion revolves around anticipating future movements by analyzing past human motion data. Mathematically speaking, this can be visualized as a function that takes a series of historical human motion data points and then produces a prediction for the next data point or even an entire sequence of future data points. This prediction algorithm processes multiple human pose data from each historical frame, with its primary goal being to generate human pose predictions for a number of future frames.

To put it mathematically, let us denote a sequence of historical human poses as $X_{1:T} = x_1, x_2, \ldots, x_T$. Here, $x_t = j_1, j_2, \ldots, j_N$ symbolizes a single pose information at time t, comprising N distinct joints. Each of these joints is illustrated in a desired K dimensional pose representation format. For instance, when $K = 3$, it signifies joint position representation. Alternatively, $K = 9$ could refer to a 3×3 rotation matrix representation. These motion data representation techniques are defined later. The main objective here is to predict the poses for the upcoming L time steps, which translates to forecasting the sequence of future human poses $X_{T+1:T+L}$ after having observed the historical frames.

For motion prediction, two dominant strategies have gained traction: 1) sequence-to-sequence (seq2seq), and 2) auto-regressive. The seq2seq approach, pioneered by Sutskever et al. [52], allows models to produce an entire future sequence in one step. In contrast, auto-regressive models predict future frames sequentially, as described by Box et al. [53]. The latter uses the output from the previous prediction as input for the next, providing the model with a mechanism to adjust and refine future predictions. A recognized limitation of the seq2seq approach in various applications is its potential to generate static or repeated patterns, especially when forecasting beyond its typical training scope [52].

B. CATEGORIZATION CRITERIA

We categorize deep learning-based human motion prediction algorithms according to their prediction strategies. Over time, researchers have sought to understand the dynamics of human motion to anticipate near-future movements. The research community has proposed a plethora of prediction algorithms, with many designed to incorporate the structural constraints and movement limitations of the human body. For a coherent analysis, we divide these methods based on their prediction approach. Some techniques integrate the structure of the human body into their predictions, while others employ a feed-forward strategy without considering these structural constraints. In our examination of these prediction methods, we group the strategies into two primary categories: direct modeling and geometry-aware modeling.

1) DIRECT-MODELING

Given the complex and dynamic nature of human activities, decoding and forecasting human motion has always been a challenging task. One prevailing strategy in this domain is direct modeling. Such an approach primarily utilizes methods that predict motion without prior knowledge of the human physique. Instead, they primarily rely on deep learning paradigms to intuitively understand the physical layout and joint dynamics of the body directly from the input dataset, as highlighted by Martinez et al. [54]. This methodology effectively infers the connective and skeletal attributes of the body, eliminating the need for explicit structural human body information. Due to the simplified approach of modeling human motion without anatomical body priors, several challenges arise. These include the difficulty of accurately learning complex motion patterns from limited data, the challenge of generalizing across different types of

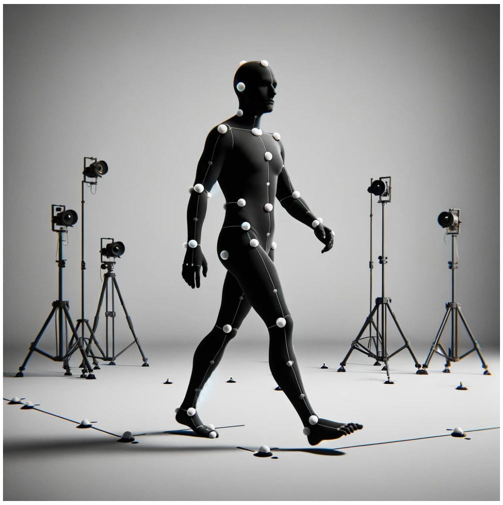


FIGURE 3. Example of collecting motion capture dataset. Multiple motion cameras are placed to capture human motions generated by an actor with tracking markers.

movements and scenarios, and the tendency to collapse into zero-velocity motion for extended horizon predictions. The research community has attempted to address these issues by incorporating spatio-temporal decoupled architectures to capture complex motion dynamics [32], [55], [56], by proposing efficient sampling strategies to build more generalized models [57], [58], and refining predicted motion to reduce errors in long-term forecasts [44], [59].

2) GEOMETRY-AWARE MODELING

Human body joints can be envisioned as nodes in a graph, interconnected by directed or undirected links representing the skeletal structure, as elaborated by Yan et al. [60]. Including prior knowledge about human anatomy and its joints is crucial in learning human motion representation. Although human activities inherently possess a degree of variability, understanding the fundamental anatomical constructs can greatly enhance the precision of motion prediction. By incorporating these anatomical details, prediction models can more precisely depict the complexities of human movement. However, increasing the number of joints in the model raises both its complexity and computational demands. Moreover, most models are trained with a fixed number of joints, limiting their predictive capability. Altering the number of joints can destabilize the model due to its reliance on the specific joint connections learned during training. Due to the tree-structured like body-joints representation, the GCN-based methods have the greater potential to decipher the motion dynamics, as further explained in Section III-B.

C. MOTION DATA REPRESENTATION

Motion capture (MoCap) technology is essential for recording the movement dynamics of individuals within a three-dimensional space. A widely employed method for acquiring such data is through marker-based motion capture systems. In this approach, markers are strategically placed on key points of the human body, and their movements are tracked using specialized cameras to capture their trajectories precisely (see Fig. 3).

The data thus obtained can be interpreted using various representation techniques. The selection of a suitable data representation is crucial for enabling the prediction model to more effectively grasp and interpret motion dynamics. Human motion data is typically rendered using four primary representation types:

1) Joint positions in the Cartesian plane
2) Axis-angle
3) Joint Rotation Matrix
4) Quaternion

1) JOINT POSITIONAL POSITIONS

In motion reconstruction and prediction, the Cartesian coordinate system is a favored method for representing human body joints in three dimensions. Mathematically, it can be expressed as $P = (x, y, z)$, where x, y, and z denote the position of joints in the 3D Cartesian plane. This representation is straightforward and provides a direct view of joint positions, facilitating easy visualization and qualitative analysis of the proposed method. However, it does not encapsulate the intricacies of joint orientation, which can be critical for certain applications. Additionally, this representation is prone to positional noise, leading to disruptions in the natural constraints of human body structure and movement. Even minor deviations in joint positions can accumulate over consecutive frames, resulting in significant errors in the overall body posture. Nevertheless, due to its simplicity and directness, the Cartesian representation remains a popular choice in various methodologies.

2) AXIS-ANGLE REPRESENTATION

The axis-angle representation characterizes the orientation of a 3D object using a unit vector for the rotation axis and an angle for the rotation magnitude, represented as $\vec{r} = \theta \cdot \hat{u}$, where θ is the rotation angle and $\hat{u}$ is the rotation axis unit vector. This representation offers an advantage by avoiding the gimbal lock problem, a challenge inherent in techniques such as Euler angles or quaternions. Despite its strengths, it faces a limitation in that multiple axis-angle pairs can denote the same orientation, leading to potential inconsistencies in rotation representation. Resolving these ambiguities and determining the accurate orientation could introduce additional computational overhead. The axis-angle representation is highly utilized in 3D modeling and animation due to its ability to circumvent gimbal lock issues, such as [61] and [62].

3) JOINT ROTATION MATRIX

This method uses a rotation matrix to represent motion data for each joint, capturing the movement with rotation values within a 3 × 3 matrix. This matrix encapsulates nine rotation values for each joint, offering a comprehensive dataset. The detailed nature of this representation provides prediction algorithms with abundant information, fostering improved accuracy and enhanced generalization. As indicated by researchers in [32], the rotation matrix approach often exhibits superior performance, especially when contrasted with methods like the axis-angle representation. However, the complex nature of this matrix might escalate computational demands, especially when simultaneously handling an array of joints.

4) QUATERNION

The quaternion representation offers a concise and effective means to depict rotations in 3D space. Mathematically, a quaternion, q, is given by

$$q = a + bi + cj + dk, \tag{1}$$

where the components corresponding to $bi+cj+dk$ define the imaginary parts, while the coefficient a is the real component. In computer graphics and robotics, quaternions are frequently favored over Euler angles and rotation matrices due to their robustness and efficiency in portraying 3D rotations. Specifically, they sidestep issues such as gimbal lock which can pose challenges in those fields. Pavllo et al. [19] advocate for the quaternion approach in learning human dynamics. This methodology proves especially valuable in arenas like virtual reality, gaming, and biomechanical simulations where precise human movement forecast is crucial [63].

III. METHOD CHARACTERIZATION

In this section, we organize previous motion prediction methods according to the criteria outlined in Section II. These methods are first grouped by their modeling strategy and then further categorized by the primary architecture employed in the prediction framework. Notably, some architectures blend multiple approaches. For example, they might combine sequential neural networks with generative models to leverage the unique advantages of each. In these cases, we classify the methods based on their distinctive architectural combination. It is crucial to understand these classifications as they offer insights into the evolution and diversity of motion prediction techniques.

A. DIRECT MODELING

1) RNN-BASED METHODS

RNNs have demonstrated remarkable effectiveness with time-series data, leading to their widespread use in human motion prediction. The unique architecture of RNNs allows them to maintain an updated hidden state at each time step, enabling them to remember previous inputs, handle sequences of varying lengths, and identify temporal relationships. Early work involving RNNs primarily adopted the direct modeling strategy.

The research by Fragkiadaki et al. [36] is a foundational contribution to human motion prediction. They leveraged the strengths of RNNs to capture the repetitive nature inherent in human movements. Their approach utilized the encoder-decoder network to intricately understand human motion patterns. Given the integration of Long Short-Term Memory (LSTM) layers between the feedforward encoder and decoder sections, this framework received the fitting name of Encoder Recurrent Decoder (ERD). In the training phase, they minimize the Gaussian Mixture Model (GMM) negative log-likelihood, which is defined as:

$$\mathcal{L}(x) = -\sum_{t=1}^{T} \log \Pr(x_{t+1}|y_t), \tag{2}$$

where x_t is the input sequence and y_t is a decoder's output, at time step t. However, a recurring challenge with these methodologies is their inclination to stray when predicting over extended periods. This arises from their strategy of refining motion representation by focusing on short-term prediction errors. As a result, anomalies like the 'sliding-foot' phenomenon and instances of zero-velocity collapses are often observed in long-term predictions, highlighting the need for more comprehensive solutions in such contexts.

Martinez et al. [18] observed that a constant pose predictor outperformed existing methods. They approached the task by focusing on learning velocities rather than joint angles. To accommodate random noise, they used the predicted pose of the network as input for the next predictions. Along with proposing the RNN-based architecture, they introduced a simple zero-velocity baseline to qualitatively assess subsequent research in the same predictive domain.

A notable challenge with RNN-based approaches is their tendency to generate motion sequences that appear less natural and are prone to artifact predictions. In response, Gui et al. [64] designed an encoder-decoder predictor enhanced with fidelity and continuity discriminators, ensuring smoother and more coherent motion predictions. They argued that relying solely on Euclidean loss does not effectively capture the subtle details of human anatomy, which can result in convergence to a generic pose. To remedy this, they integrated a geodesic loss, aiming to bridge the gap between predicted and true motions. This strategy was further reinforced by two adversarial losses from discriminators, leading to more realistic and diverse human motion forecasts.

In a related observation, Tang et al. [61] pointed out that predicting motion for every joint might be redundant. As an illustration, during a 'phoning' gesture, certain back joints remain unchanged. To address this inefficiency, they implemented a Modified Highway Unit (MHU) to filter out static joints. Moreover, they argued that past research methodologies frequently utilized mean-squared error on joint values, which inadvertently pushed models toward a generalized pose. Therefore, contrary to previous studies, the authors train the network with the gram matrix instead of individual joint values. Given a history sequence $\{x_{t'}\}_{t'=1}^{T'}$ and the ground truth sequence $\{x_t\}_{t=1}^{T}$, the model aimed to predict

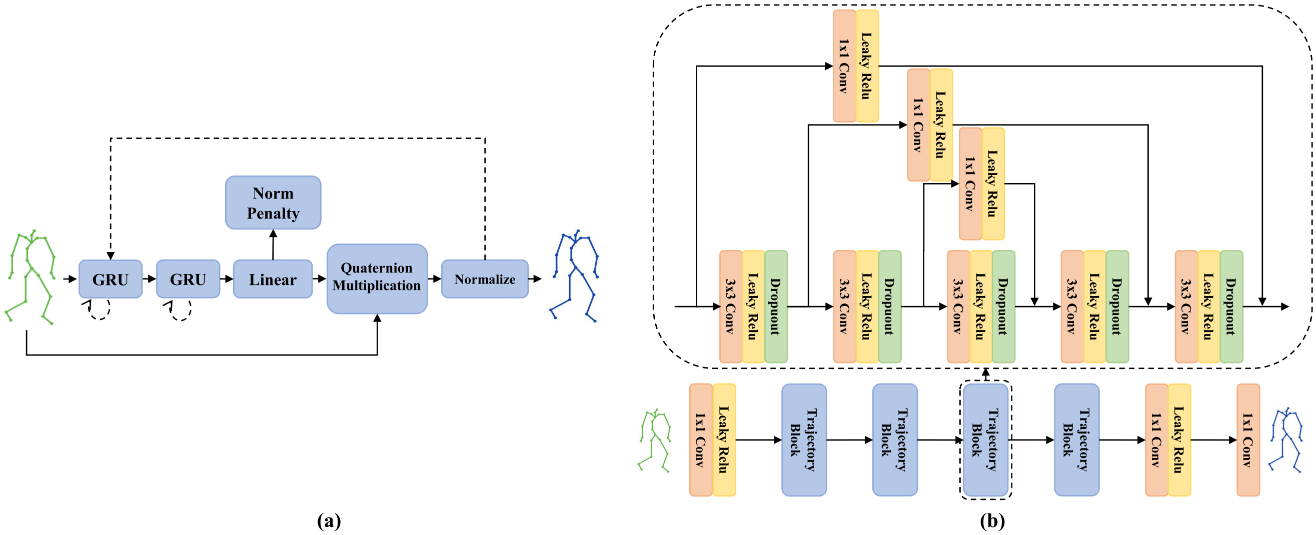


FIGURE 4. (a): The outline of the short-term generation model by Pavllo et al. [19]. The rotations of the current joints and the previous states are computed by Gated Recurrent Units (GRUs) to predict the future pose in an autoregressive way. (b): The architecture of TracjectoryCNN explained by Liu et al. [21]. The human body is transformed into trajectory space using a convolution layer. Then the data are fed to the encoder and decoder composed of convolution, relu, and dropout layers.

future poses $\{\hat{x}_t\}_{t=1}^{T}$. The loss function and a gram matrix G are defined as:

$$\mathcal{L}_{\text{gram}} = \frac{1}{T}\sum_{t=1}^{T-1} \|G(\hat{x}_t, \hat{x}_{t-1}) - G(x_t, x_{t-1})\|_2^2, \quad (3)$$

$$G(x_t, x_{t-1}) = \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} \begin{bmatrix} x_t & x_{t-1} \end{bmatrix}, \quad (4)$$

In previous studies, motion data was often represented using the Euler angle or exponential map. Such representations can infer the issues of non-uniqueness, discontinuity, and singularities within the representation space [65]. Hence, Pavllo et al. [19] introduce the QuaterNet to utilize quaternion for rotation parameterization in a two-layer Gated Recurrent Unit (GRU) networks [66]. Due to the stability of quaternion parameterization, it improves the results for short-term prediction (see Fig. 4 (a)).

In [62], the authors innovatively incorporate motion derivative information combined with joint-angle history. They emphasize the significance of motion derivatives to capture near-past motion details. These derivatives from the input x are efficiently computed using a finite backward difference approximation, given by:

$$\nabla_h^n[f](x) = \sum_{i=0}^{n} (-1)^n \binom{n}{i} f(x - ih), \quad (5)$$

where i through n denotes the indices corresponding to the order of derivatives, and h is a non-zero spacing constant. Furthermore, with growing interest in the attention mechanism, Sang et al. [67] attempt to incorporate an attention mechanism into the decoder. The attention mechanism calculates weights using the features of both the encoder and decoder. These weights are then employed to aggregate the encoder's features, which are then used as input for the decoder. The attention mechanism is further discussed in the studies by [68] and [69].

In conclusion, early works utilizing RNN-based methods have established a solid foundation for future motion prediction techniques [70], [71], [72], [73]. The early preference for RNNs in research was due to their natural fit for direct modeling that did not incorporate the structural tree of the human body. However, recent advancements have shown accuracy improvements when inducing anatomical connection information with RNNs, as further elaborated in Section III-B.

2) CNN-BASED METHODS

A defining feature of Convolutional Neural Networks (CNNs) is their exploitation of convolutional layers to extract spatial information from the input. This capability has paved the way for exploring its efficacy in human motion analysis. Li et al. [74] built upon earlier research in motion studies and proposed an encoder-decoder model immersed in a convolutional architecture. This model integrates both short-term and long-term encoders to comprehend immediate and distant temporal movements, respectively. The spatial decoder, denoted by h_d and parameterized by w_d, predicts the next pose using a combination of long-term and short-term hidden states, z_{e_l} and z_{e_s}, respectively. The recursive training process updates the predicted pose $\hat{x}_{t+k}$ using the formula:

$$\hat{x}_{t+k} = h_d\left([z_{e_l}, z_{e_s}(k)] \mid w_d\right) + \hat{x}_{t+k-1}, \quad (6)$$

The hierarchical structure of convolutional layers facilitates the capture of spatial-temporal relationships, offering advantages over the preceding RNN-based models. Following this, another method emerged that focuses on a hierarchical asymmetric structure using Velocity-Cascade Multiplicative Units (v-CMUs) [75]. This method gives priority to recent frames, offering a refined approach to human motion prediction by understanding both static and dynamic elements.

In another study [76], a novel method was introduced to transform body joint coordinates into an image sequence, placing similar joints in proximity to explore local correlations among them. Moreover, Liu et al. [21] formulated the TrajectoryCNN, a model developed to understand motion dynamics in trajectory space by leveraging intertwined spatial-temporal data, global temporal relationships, and correlations of neighboring joints (see Fig. 4 (b)). The 3D data of the j-th joint, represented as (x_j, y_j, z_j), is encoded into a latent space denoted by T. This encoded representation is then used in the decoding phase to estimate the future positions of particular joints.

$$T = \phi(x_j, y_j, z_j), \quad j = 1, 2, \ldots, N_j \tag{7}$$

$$\hat{p}_i = \psi(T), \quad i = 1, 2, \ldots, N_i \tag{8}$$

Following the innovative approaches of previous studies, another significant contribution comes from the High-Resolution Spatio-Temporal Attention Network (HR-STAN) architecture [55]. This method uniquely separates convolutions into spatial and temporal segments, enhancing the modeling of human motion. Unlike traditional techniques, HR-STAN uses dilated convolutions to capture extensive motion patterns without feature compression. Furthermore, it incorporates specialized spatial and temporal attention mechanisms, refining its prediction accuracy by focusing on distinct spatio-temporal relationships. Similarly, to attain attention to dynamic information, Tang et al. [59] presents a Temporal Fusion (TF) module that fuses information from two streams by utilizing a reinforcement Trajectory Spatial-Temporal (TST) block. This approach differentiates between immediate and extended motion predictions, ensuring continuity between predicted and provided poses.

CNNs remain relatively underexplored in human motion prediction research. While they inherently excel at extracting spatial features, they often fall short in capturing the temporal dependencies of future motion timestamps. As discussed later, although some methods have delved into geometry-aware modeling, the volume of published research on CNNs is still modest compared to other networks, such as GCNs.

3) GENERATIVE MODELS

The challenge of deciphering motion dynamics has been extensively explored using generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). These models excel at approximating data distributions and generating high-quality synthetic data that closely resembles real data. While many generative models focus on direct modeling and often overlook the prior information of human motion joints, few exceptions are noted in the work of [30], [31], and [77].

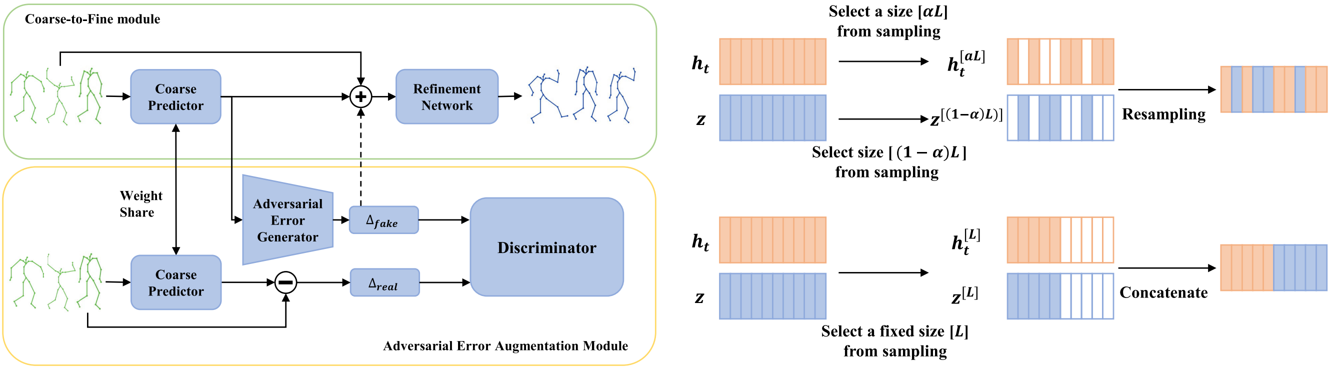
Starting with, Barsoum et al. [27] put forth the Human Prediction GAN (HP-GAN), a specialized approach for probabilistic human motion prediction. This HP-GAN model undergoes training that incorporates both a critic via the Wasserstein GAN (WGAN-GP) loss [42] and a discriminator using the GAN loss. The critic loss is instrumental in optimizing both the critic and generator networks. In contrast, the discriminator loss is reserved for training the discriminator network and serves as a yardstick to gauge the authenticity of the generated human poses. Inspired by the studies of [42] and [78], Chopin et al. [79] present a similar approach employing WGAN for the task of motion prediction.

Kundu et al. enhanced probabilistic human motion models by proposing the Bidirectional Human Motion Prediction GAN (BiHMP-GAN) [43], which includes a content loss for better motion sequence prediction. In this approach, the discriminator not only differentiates real from predicted poses but also guides future pose generation. Building on this, Hernandez et al. [56] applied the GAN framework to spatio-temporal inpainting, employing a variety of loss functions to more accurately capture the complexities of human movement.

Achieving realistic poses through adversarial training, where the generator and discriminator compete, is challenging. To address this, Chao et al. [44] introduced a refinement module to learn motion dynamics (see Fig. 5(a)). On the other hand, Lyu et al. [28] contended that while GCN-based methods are finely tuned for the skeletal kinematic chain, they have innovatively incorporated human biases into their model. They modeled each joint motion using a stochastic differential equation and employed GANs to simulate path integrals for forecasting near-future motion trajectories. This methodology represents a novel approach to motion prediction, yet it is categorized under direct modeling approaches for its partial incorporation of structural constraints within the predictive model.

In the domain of generative models, VAEs have also been thoroughly investigated for learning latent variables from human motion datasets. Noteworthy early work by Yan et al. [57] demonstrated the use of sampling to generate multimodal plausible outputs by concatenating a part of the original hidden state with the random vector. However, Aliakbarian et al. [29] observed that due to this sampling approach, the model tends to neglect randomness and relies instead on deterministic conditioning information for motion generation. To counteract this, they suggest inheriting stochasticity by combining two vectors, i.e., hidden state h_t and random vector z. The indices of the hidden state are randomly sampled $\mathcal{I} \subseteq \{1, \ldots, L\}$ and the complementary indices set $\hat{\mathcal{I}}$ is assigned to the random vector (see Fig. 5 (b)).

Whereas, Yuan and Kitani [80] argue that random sampling based on data likelihood can lead to low sample efficiency. As an alternative to random sampling, the authors proposed to generate a diverse set of samples from a pretrained generative model. This approach towards sampling is further refined by explicitly condition-dependent sampling in the work of [58].



(a)

(b)

FIGURE 5. **(a): The architecture of the Adversarial Refinement Network (ARNet) implemented by Chao et al. [44]. The architecture is configured with a Coarse-to-fine module and an adversarial error augmentation module that are both used to optimize the network. (b): The Mix-and-Match perturbation method and the perturbation by concatenation described by Aliakbarian et al. [29]. For Mix-and-Match perturbation, indices are sampled stochastically for each mini-batch. Contrastingly, the concatenation methods are sampled deterministically, with the size of halves of the original vectors.**

In the context of human motion prediction, generative models have exhibited a commendable performance in generating a variety of multimodal human motions. However, there exists an opportunity to further alleviate the impact of mode collapse to better generalize human motion dynamics.

4) TRANSFORMERS

In 2017, the transformer network was introduced using self-attention for sequence-to-sequence tasks like machine translation [45]. This self-attention mechanism enables the network to concentrate on pertinent sections of the input while filtering out irrelevant details. Given, a set of *query* Q, *key* K, and *value* V representations, the scaled-dot product attention and a Multi-Head Attention (MHU) mechanisms are defined as:

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{9}$$

$$\mathrm{MHU}(Q, K, V) = \mathrm{Concat}(head_1, \dots, head_h)W^O, \tag{10}$$

where $head_i = \mathrm{Attn}(QW_i^Q, KW_i^K, VW_i^V)$, and the respective parameter matrices are defined as $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$.

Taking that into account, the Pose Transformer (POTR) [33] architecture is introduced to showcase the capabilities of the transformer network for motion prediction tasks. They approach the prediction problem as a multi-task learning challenge, wherein the network is tasked with predicting both activity labels and future motion concurrently (see Fig. 6 (a)). The non-autoregressive nature of this model can lead to errors in long-term predictions since they rely on just one input pose. The final multi-task loss function for POTR training is defined as:

$$L_{POTR} = L_{motion} + \lambda L_{activity}, \tag{11}$$

Conversely, Aksan et al. [32] employ decoupled temporal and spatial attention blocks to extract both local and global joint dependencies in human motion. Their strategic deployment of attention mechanisms assists in modeling human motion across extended time frames effectively. Despite using a decoupled transformer to extract inter and intra-joint connections, we categorized this method under direct modeling since the input information lacks specific joint details. Extending this approach, further studies also consider decoupled spatial and temporal modeling [82], [83].

Despite several advancements, the prediction of future motion over extended periods remains a challenging task. To tackle this, Idrees et al. [81] propose an innovative approach, blending an adversarial training mechanism with a transformer encoder network. Additionally, they employ a temporal consistency loss to compel the model to learn human body movement constraints (see Fig. 6 (b)). This significantly enhances the ability of the model to mimic natural human motion for long-term predictions.

After the exceptional performance of the transformer network in language models, it has also shown promising outcomes in human motion prediction. While a basic encoder network offers commendable motion prediction, it has the potential for a deeper understanding of human motion dynamics. This could be realized by merging the transformer network with other proven techniques.

B. GEOMETRY-AWARE MODELING

1) RNN-BASED METHODS

Understanding human motion dynamics crucially depends on both local and global joint information. Therefore, studies are performed focusing on modeling this prediction problem by incorporating anatomical structural information with RNNs.

The research by Guo and Choi [39] serves as a cornerstone in this area, highlighting the importance of local dependencies. They achieved this by segmenting the input into five distinct non-overlapping sections: the left arm, right arm, left leg, right leg, and torso. Each segment is treated autonomously, and processed through specific pathways to draw out local attributes. These extracted local features are then integrated to form a comprehensive pose representation

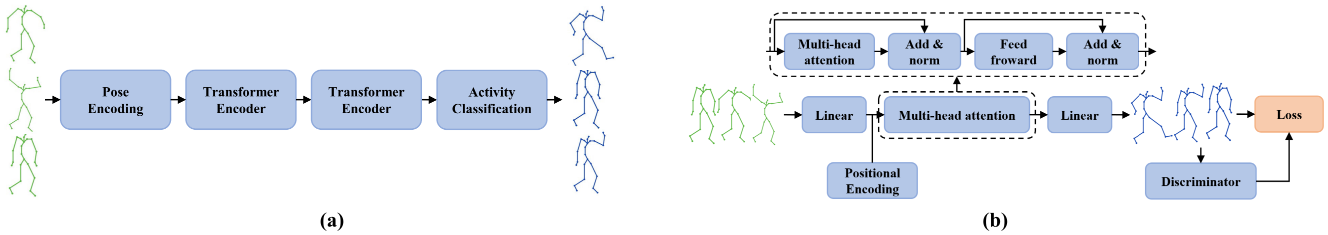


FIGURE 6. (a): The non-auto-regressive prediction model Pose Transformer (POTR) proposed by Martínez-González et al. [33]. The human poses are embedded and then processed through the encoder and decoder. The outputs of the decoder are combined with specialized class tokens to embed them as activity representations. (b): the Adversarial Motion Transformer (AdvMT) introduced by Idrees et al. [81]. Future motion is predicted using an auto-regressive approach that incorporates adversarial training techniques. The results are further enhanced through a consistency loss to restrict the model to predict more consistent and realistic motion.

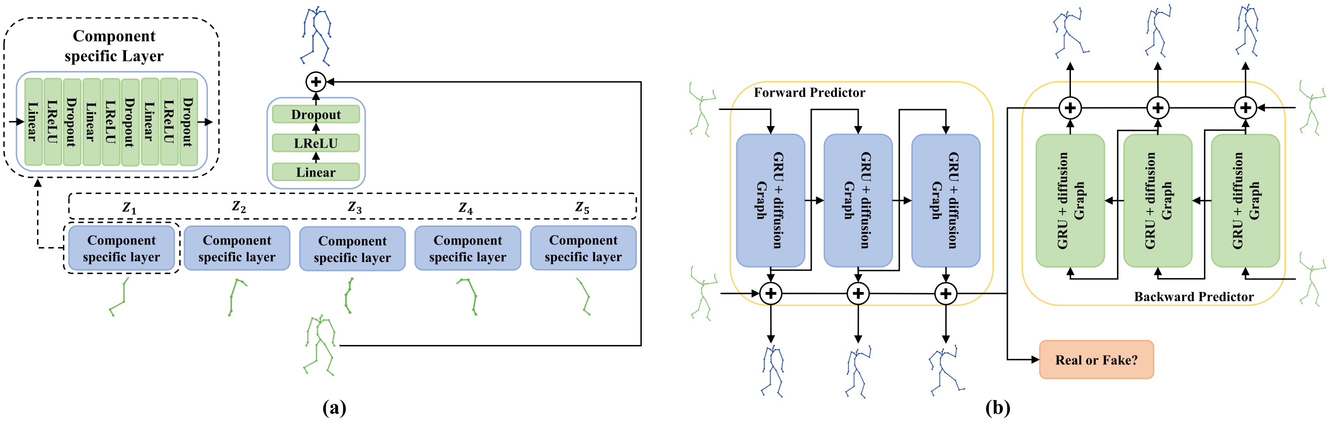


FIGURE 7. (a): The outline of SkelNet proposed by Gou and Choi [39]. The current pose is divided into five components. These non-overlapping components are fed to each branch of component-specific layers, which is a combination of linear layers, LReLUs, and dropouts. (b): The Quadruple Diffusion Convolutional Recurrent Network (Q-DCRN) mentioned by Men et al. [84]. The skeletal representations are fed to the forward and backward in chronological directions that include GRU with diffusion graphs. Then both predictions are fed to the discriminator.

(see Fig. 7 (a)). Similar to the work by [39], Aksan et al. [20] suggested a structured decomposition of the human kinematic chain into information for separate joints, such as:

$$\mathbf{x}_t = \left[x_t^{(hip)}, x_t^{(spine)}, \ldots x_t^{(lwrist)}, x_t^{(lhand)}\right], \tag{12}$$

As observed, RNNs integrating human body structure information to learn motion dynamics have been relatively unexplored. However, combining RNNs with architectures like GCNs has gained some attention recently, as seen in the works of [17] and [85].

2) CNN-BASED METHODS

CNNs inherently excel in extracting spatial features from input data, and when applied to pose prediction, the integration of structural information enables the creation of more plausible future human movements. In [22], the authors unveiled a hierarchical encoding technique, visualizing the human body in a tree-structured manner. This representation commences with individual body parts at the base layer and ascends to the entire body representation at the top layer.

Despite their advantages, hierarchical methods are computationally intensive. Addressing this, Li et al. [86] unveiled a streamlined framework featuring a Convolutional Hierarchical Module (CHM) which employs 1D convolutional layers to simultaneously infer temporal dynamics and spatial constraints from the data. Another contribution is from Men et al. [84], introducing the Quadruple Diffusion Convolutional Recurrent Network (Q-DCRN), a mechanism that transforms spatial structures into graphs and enables adaptive information diffusion through bi-directional random walks. This network, paired with a seq2seq and graph convolution, interprets temporal dependencies in the data and optimizes both retrospective and prospective human movements through the integration of a bi-directional temporal predictor and a bi-discriminator (See Fig. 7 (b)).

While CNNs excel in vision-centric research, their application to the MoCap dataset for motion prediction has been challenging. Despite efforts to incorporate tree-structured human body data through hierarchical methods, the results have not been consistently compelling.

3) GCN-BASED METHODS

Past approaches have shown that structuring joint information in a tree-like manner improves the proficiency of pose prediction models in interpreting human movement dynamics. Given that GCNs excel at extracting key features from graph-structured data, including inter-node connections and the significance of specific nodes and edges. Given $\tilde{A}$ as the adjacency matrix with added self-connections and $\tilde{D}$ as its diagonal node degree matrix, the GCN layer propagates node

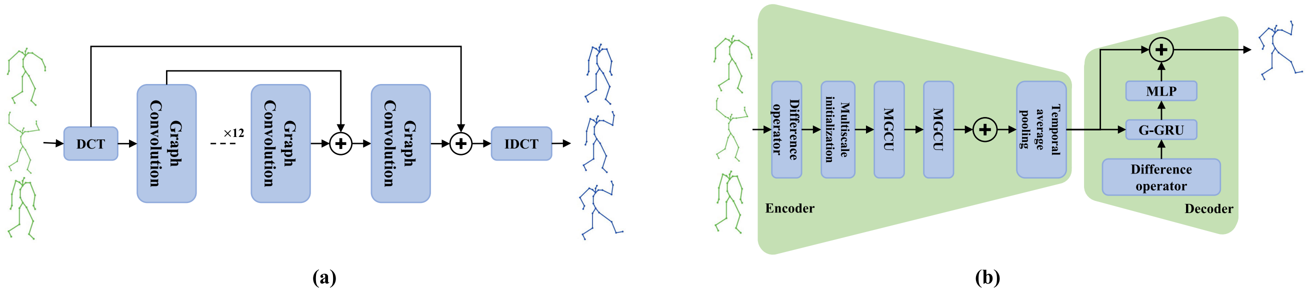


FIGURE 8. (a): The short-term graph network Quaternet by Mao et al. [37]. The temporal pose information is fed to the DCT for enhanced 3D coordinate representations. Then they are inputted into 12 residual blocks containing 2 graph convolutional layers for encoding and decoding data. (b): The Dynamic Multiscale Graph Neural Networks proposed by Li et al. [24]. The semantics of human poses are input into the encoder containing Multiscale Graph Computational Units (MGCU). Then they are passed through the decoder formed with graph-based GRU (G-GRU) to update hidden states.

features $H^{(l+1)}$ at layer $l+1$ using the following equation:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tag{13}$$

GCNs are optimally designed to capture the anatomical kinematic tree relationships. The fusion of GCNs with the tree-structured joint information led to breakthroughs in motion prediction accuracy. Therefore, GCN-based motion prediction methods are exclusively categorized under the geometry-aware methods, emphasizing their unique capability to process and understand geometric structures.

While earlier hierarchical human motion predictors grouped similar or neighboring joints to comprehend body movement, coordinating motions across distinct body parts is often neglected. Inspired by [87], Mao et al. [37] developed a GCN tailored to adaptively learn crucial joint connections without relying on joint-tree data. They argue that the joint angles or 3D joint coordinate representations employed in previous studies remain static in nature. Thus, they utilize the Discrete Cosine Transform (DCT) to capture the temporal dynamics of human motion, functioning within the trajectory space. The DCT offers a compact yet potent representation of temporal variations in human joint movement. This method is further refined by incorporating an attention layer or adopting a distinct motion generation technique, as outlined in [23], [88], and [89]. Mao et al. use specific equations to derive the DCT coefficients $\{C_{k,l}\}_{l=1}^{L}$ from trajectories $\{x_{k,l}\}_{l=1}^{L}$. Afterward, the Inverse-DCT (IDCT) applied within the learned feature space yields the future pose [23], [37], [88], [89].

$$C_{k,l} = \sqrt{\frac{2}{L}}\sum_{n=1}^{L} x_{k,n}\sqrt{\frac{1}{1+\delta_{l1}}}\cos(\beta_{n,l}), \tag{14}$$

$$x_{k,n} = \sqrt{\frac{2}{L}}\sum_{l=1}^{L} C_{k,l}\sqrt{\frac{1}{1+\delta_{l1}}}\cos(\beta_{n,l}), \tag{15}$$

where $\beta_{n,l} = \frac{\pi(2n-1)(l-1)}{2L}$ and $\delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$.

In [37], a joint connection matrix is learned without constraints (see Fig. 8 (a)). In contrast, Cui et al. [41] advocate for learning a parameterized representation of human body structure through a learnable adjacency matrix. In this approach, only the connection weights between body parts are refined during the optimization process.

Lebailly et al. [90] introduced the Temporal Inception Module (TIM) to encode human motion across multiple temporal scales, enabling the model to inspect input pose sequences through varying receptive fields. Whereas, to leverage both local and global joint connections, the Dynamic Multiscale Graph Neural Networks (DMGNN) framework [24] was crafted to leverage the structural connections of the human body across multiscale graphs. This multiscale approach effectively captures relationships among nodes. By providing insights at diverse scales, the model detects complex patterns and dependencies that might be missed by single-scale models (see Fig. 8 (b)). In general terms, Eq. (13) can be modified to generate the multi-scale GCN as follows:

$$H_{s_i}^{(l+1)} = \sigma\left(\tilde{D}_{s_i}^{-\frac{1}{2}}\tilde{A}_{s_i}\tilde{D}_{s_i}^{-\frac{1}{2}}H^{(l)}W_{s_i}^{(l)}\right), \tag{16}$$

where $W_{s_i}^{(l)}$ is the weight matrix at layer l for scale s_i, and $\tilde{A}_{s_i}$ and $\tilde{D}_{s_i}$ are the adjacency and degree matrices for scale s_i, respectively. Furthermore, the aggregated output features for multiple scales with λ as a regularization factor can be summed as:

$$H = H_{s_1} + \lambda\sum_{i=2}^{S} H_{s_i}, \tag{17}$$

Building on the approach of [24], the researchers have embraced the use of multiscale graphs for human motion learning by employing a variety of diverse methods [91], [92], [93], [94], [95], [96], [97], [98].

Improving over previous methods, Cui and Sun [38] addressed two primary challenges of error accumulation in predicted poses and handling of incomplete input sequences. Their solution, the Multi-Task Graph Convolutional Network (MT-GCN), confronts both issues by simultaneously correcting incomplete sequences and predicting human actions. Their framework architecture comprises three main components: a Shared Context Encoder (SCE) for context extraction from observed sequences, a Sequence Repairing

Module (SRM) that uses a temporal self-attention mechanism for gap-filling, and a Human Action Predictor (HAP). This structure facilitates more precise predictions, especially when fed with incomplete sequence data. In another research, Cui et al. [99] developed a distinctive method that merges the efficiency of GCNs in processing body-joint information with adversarial training to capture more realistic human motion. The proposed approach proposes the fidelity discriminator and the consistency discriminator. The former differentiates predictions from the ground truth, while the latter ensures motion consistency for long-term predictions. The model outperforms traditional recurrent models, enabling real-time applications with improved accuracy and visualization for long-term predictions.

In the work of [25], the authors tackle the over-smoothness issue observed in extended horizon predictions which often reflects unrealistic human motion. Addressing this, they present a novel approach called Skeleton Graph Scattering Network (SGCN). The key idea is to extract crucial motion information across various graph spectrum bands using graph scattering techniques. The model combines graph scattering decomposition and graph spectrum attention to learn joint-spectral representation and reflect their importance through attention scores. Further expanding on this concept, Li et al. [100] introduced Multi-Part Graph Scattering Blocks (MPGSBs) in a subsequent study, aimed at enhancing motion representation through adaptive graph scattering methods. Furthermore, Ma et al. [101] introduce a stage-wise prediction mechanism that iteratively uses the forecasted future pose as an initial input for the next stage. Each of these stages integrates both spatial and temporal GCN networks, effectively capturing the inherent spatial intricacies within a pose and the temporal dynamics across motion trajectories. Similar to this, Zhong et al. [26] employs spatio-temporal modeling to discern intra-joint connections and temporal dynamics over sequences. They further innovate by fusing spatial and temporal adjacency matrices with learnable blending coefficients to produce an adaptive adjacency matrix.

In conclusion, GCNs have emerged as a pivotal tool in human motion prediction due to their geometry-aware modeling. Due to effectively processing graph-structured data, even conventional feedforward GCNs surpass the previous RNN-based methods. The strategic combination of GCNs with advanced techniques like multiscale, spatio-temporal modeling, and adaptive adjacency matrices, promises a new frontier in achieving unparalleled motion prediction accuracy.

4) GENERATIVE MODELS

A well-known challenge with GANs pertains to mode collapsing and the issue of predicting freezing future motion. One approach to mitigate these issues involves integrating hierarchical methods that account for the local dependencies among human joints. Hence, Liu et al. [30] presented a unique strategy for blending joint geometry information into a multi-GAN model. They employed sub-GANs to learn local joint connections, and later they unified the outputs from these local GANs using a global GAN. This synergy between local and global perspectives aids in generating diverse and robust future motion predictions (see Fig. 9 (a)).

Furthermore, inspired by [16] and [102], Zhao et al. [77] introduced a Bidirectional Transformer-based Generative Adversarial Network (BiTGAN) aimed at tackling the freezing future motion dilemma. To accommodate the spatial relationships among joints, they employed a GCN predictor with a learnable adjacency matrix. In a related vein, the work of [31] advocates for the learning of disentangled human body joints information. Conditional-VAE (CVAE) is employed to concurrently learn full-body human motion and partial-body motion.

Like RNNs and CNNs, generative models have received less attention in tackling motion prediction task. The complex and variable nature of human behavior contributes to this, as it complicates the task of creating models that generalize human dynamics well. Future research that includes a blend of various architectures combined with geometry-aware modeling may enhance our grasp of human motion dynamics.

5) TRANSFORMERS

In pursuit of optimal joint representation, the transformer network has been modified with a progressive-decoding [102] and a multi-scale [34] approach. Where, Cai et al. [102] adopted a progressive-decoding technique to tap into the innate structural links among joints. Their approach initiates by predicting future motion for central joints within the kinematic chain. Subsequently, this prediction is propagated sequentially to the peripheral joints. Such a structured approach capitalizes on information about joint linkages to enhance human motion prediction accuracy (see Fig. 9 (b)).

On the other hand, Chen et al. [34] advocate for the use of the established multi-scale technique to distill hierarchical structural details of the human body. They introduced a spatio-temporal transformer network composed of distinct multi-scale modules. This design first employs a GCN to understand the spatial relationships between joint connections in multiple scales. Following this, the transformer network processes the correlations between the extracted features.

At the end, there remains ample scope to investigate more into geometry-aware motion prediction techniques with the transformer network. The critical role of structural constraints in decoding motion dynamics cannot be overstated. As demonstrated by [34], the synergy between GCNs and transformer networks offers a promising direction for future research.

6) EQUIVARIANT NEURAL NETWORKS

The ENNs represent a category of deep learning models developed to utilize symmetries within the provided data. Given a group G and two homogeneous spaces X_1 and X_2 that have corresponding G-actions, a G-ENN is a linear or nonlinear map $\psi \, : \, f(X_1) \rightarrow f'(X_2)$ that satisfies the

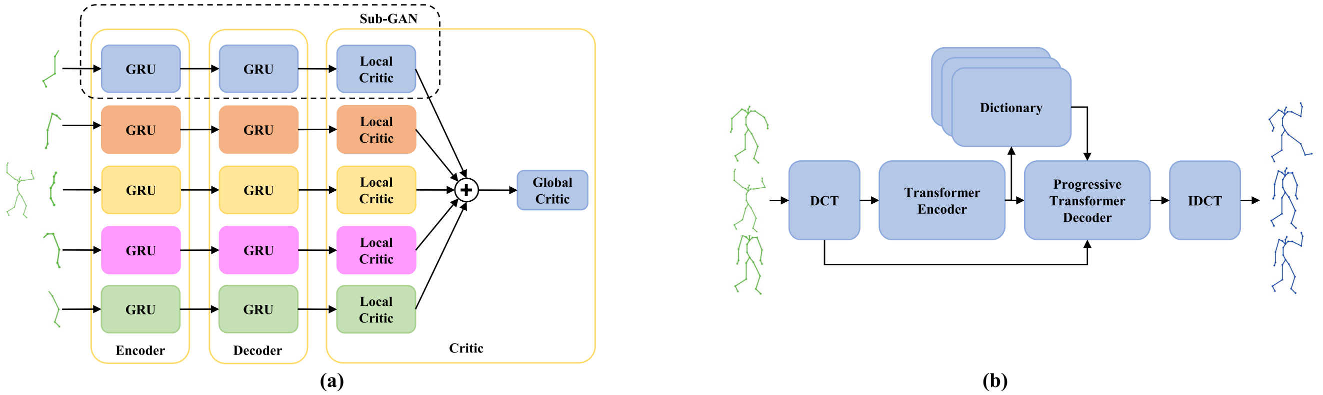


FIGURE 9. (a): The following is the Aggregated Multi-GAN (AM-GAN) assembled by Liu et al. [30]. The skeletal representations are partitioned into different sub-GANs, and their outputs are combined and fed to a global critic. (b): The progressive-decoding method was implemented by Cai et al. [102]. The human poses are encoded by the Discrete Cosine Transform (DCT) and passed to the transformer-based architecture. After they are fed to the inverse DCT (IDCT) to convert the coefficients back to human poses.

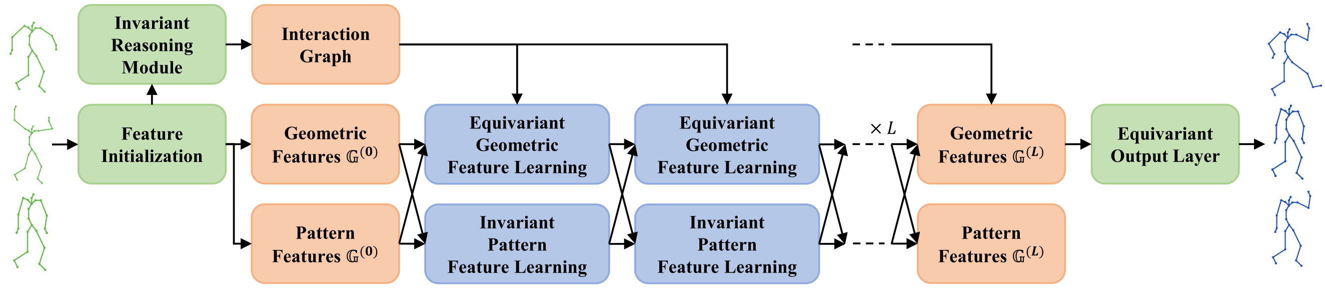

Invariant Reasoning Module
Interaction Graph
Feature Initialization
Geometric Features $\mathbb{G}^{(0)}$
Pattern Features $\mathbb{G}^{(0)}$
Equivariant Geometric Feature Learning
Invariant Pattern Feature Learning
× L
Geometric Features $\mathbb{G}^{(L)}$
Pattern Features $\mathbb{G}^{(L)}$
Equivariant Output Layer

FIGURE 10. Eqmotion by Xu et al. [35], is the motion prediction model that introduces motion equivariance. The initial feature layer adeptly extracts both geometric and pattern features. Additionally, the interacting graphs are derived from the invariant reasoning module, capturing invariant relationships under transformations. These extracted features and graphs are then iteratively fed to the equivariant geometric feature learning layers and invariant pattern feature learning layers. Finally, these features are integrated into an equivariant output layer for the model to attain the final prediction.

following:

$$\psi[T_g f(x)] = T'_g \psi[f(x)], \tag{18}$$

where T_g and T'_g denote the G-function transformations on the respective spaces. These networks excel at learning and analyzing structured data such as images [103], 3D structures [104], and point clouds [105], particularly those with inherent symmetries. The term 'symmetry' denotes the invariance of a system under specific group transformations. In the case of image analysis, typical symmetries include scaling, translation, rotation, and reflection. Conventional neural networks lack an explicit design to account for these symmetries, leading them to overlook inherent symmetrical structures. In contrast, ENNs are architecturally devised to respect and acknowledge the symmetries evident within the data.

ENNs have exhibited promising results in diverse fields of study. For example, in computer vision [103], equivariant networks effectively handle rotation and translation symmetries in images. In molecular chemistry [106], equivariant networks capture the rotational and translational symmetries of molecules, enabling more accurate predictions of molecular properties. In robotics, [107], [108], [109] demonstrated remarkable data efficiency in visual manipulation, while [110] highlighted better stability in manipulation control. Equivariance has also been employed in various fields such as 3D object segmentation [111], [112], shape reconstruction [113], and reinforcement learning [114], [115].

Several studies have integrated equivariance in the domain of human pose prediction. For instance, Yeh et al. [50] proposed the Chirality Nets. Their model utilizes chirality equivariance for pose regression tasks through odd and even symmetric parameter sharing. Due to their parameter sharing, the model demonstrates enhanced data efficiency and computational reduction. Similarly, Xu et al. [51] present equivariance in the teacher-student learning framework. In their approach, the teacher network is 3D rotation invariant, while the student network is 3D rotation equivariant. The equivariant student network utilizes a graph convolution layer with a cycle-consistent loss for 3D rotations, adding flexibility and eventually boosting the prediction accuracy. The rotation equivariant loss is defined utilizing the knowledge distillation loss, which is as follows:

$$\mathcal{L}_{REC} = \frac{1}{N} \|\mathcal{F}_s(\mathcal{P}(\boldsymbol{R}\hat{\boldsymbol{Y}}^s)) - \boldsymbol{R}\hat{\boldsymbol{Y}}^s)\|_F^2, \tag{19}$$

where $R \in SO(3)$ represents a random rotation matrix, $\mathcal{F}_s$ denotes the student network, and $\hat{\boldsymbol{Y}}^s$ is the 3D pose estimated by the student network. The $\mathcal{P}(\cdot)$ projects the 3D rotation to generate a new 2D pose. As a result, the student

TABLE 1. MAE evaluation results for different methods on the H36M dataset [116]. The best results are not highlighted as the number of joints used for evaluation differs among the methods.

		Walking			Eating			Smoking			Discussion			Directions			Sitting Down			Average		
	Time (milliseconds)	160	400	1000	160	400	1000	160	400	1000	160	400	1000	160	400	1000	160	400	1000	160	400	1000
RNN	ERD [36]	1.18	1.78	-	1.45	1.80	-	1.95	2.42	-	2.47	2.76	-	-	-	-	-	-	-	-	-	-
	LSTM-3LR [36]	1.00	1.47	-	1.09	1.46	-	1.65	2.16	-	2.12	2.23	-	-	-	-	-	-	-	-	-	-
	Res. Sup. [18]	0.49	0.81	-	0.39	0.76	-	0.61	1.15	-	0.68	1.09	-	0.47	0.84	-	0.81	1.62	-	0.67	1.15	-
	AGED [64]	0.36	0.67	0.91	0.28	0.64	0.93	0.43	0.84	1.21	0.56	0.83	1.30	0.39	0.69	-	0.62	1.10	-	0.54	0.97	-
	MHU [61]	0.53	0.77	1.06	-	-	-	-	-	-	0.66	1.00	1.88	-	-	-	-	-	-	0.68	1.13	1.80
	QuaterNet [19]	0.34	0.62	-	0.35	0.70	-	0.47	0.90	-	0.60	0.93	-	-	-	-	-	-	-	-	-	-
	SkelNet [39]	-	-	-	0.39	0.71	-	0.47	0.89	-	0.62	1.00	-	-	-	-	-	-	-	-	-	-
CNN	H-TE* [22]	0.18	-	0.31	0.20	-	0.37	0.26	-	0.41	0.17	-	0.24	-	-	-	-	-	-	-	-	-
	CEM [74]	0.54	0.73	0.92	0.36	0.71	1.24	0.49	0.92	1.62	0.67	1.01	1.86	0.60	0.91	1.45	0.78	1.31	2.06	0.68	1.13	1.77
	CHA [86]	0.45	0.74	0.92	0.34	0.66	1.21	0.48	0.93	1.66	0.62	0.91	1.72	0.62	0.88	1.40	0.79	1.30	2.05	0.66	1.11	1.74
	Q-DCRN [84]	0.36	0.60	0.69	0.32	0.67	1.18	0.43	0.84	1.58	0.69	1.04	1.56	0.45	0.70	1.31	0.73	1.15	1.95	0.57	1.02	1.60
GCN	LTD-10-10 [37]	0.31	0.56	0.77	0.29	0.62	1.10	0.41	0.80	1.58	0.51	0.85	1.75	0.45	0.79	1.35	0.61	1.00	1.67	0.52	0.94	1.62
	HisRepeat [23]	0.30	0.51	0.64	0.29	0.60	1.10	0.42	0.80	1.58	0.52	0.87	1.63	0.43	0.69	1.27	0.63	1.04	1.70	0.52	0.94	1.57
	LDR [41]	0.29	0.57	0.71	0.27	0.64	0.97	0.38	0.82	1.08	0.45	0.81	0.84	0.43	0.59	0.68	0.62	0.93	1.42	0.45	0.86	1.20
	DMGNN [24]	0.31	0.58	0.75	0.30	0.59	1.14	0.39	0.77	1.52	0.65	0.99	1.45	0.44	0.71	-	0.65	1.05	-	0.52	0.95	1.21
	TCGAN [99]	0.34	0.61	0.83	0.27	0.59	0.91	0.41	0.77	1.17	0.53	0.81	1.23	0.36	0.67	1.39	0.57	1.06	1.82	0.51	0.89	1.45
	SGCN [25]	0.30	0.53	-	0.27	0.58	-	0.39	0.78	-	0.50	0.84	-	0.39	0.68	-	-	-	-	0.48	0.91	-
Gen.	BiHMP-GAN [43]	0.52	0.67	0.85	0.33	0.70	1.20	0.50	0.86	1.11	0.65	0.96	1.77	-	-	-	-	-	-	-	-	-
	ARNet [44]	0.31	0.55	0.69	0.28	0.61	1.07	0.42	0.81	1.51	0.51	0.89	1.68	0.43	0.75	-	0.29	0.97	-	0.33	0.67	1.24
	Mix-and-Match [29]	0.48	0.58	0.68	0.34	0.50	0.91	0.42	0.77	1.25	0.60	0.89	1.30	-	-	-	-	-	-	-	-	-
	AM-GAN [30]	0.51	0.66	0.84	0.31	0.66	1.15	0.46	0.88	1.10	0.55	0.92	1.58	0.57	0.89	1.41	0.72	1.20	1.85	0.67	1.43	1.75
Trans.	ProgDec [102]	0.30	0.55	-	0.29	0.61	-	0.40	0.78	-	0.39	0.69	-	-	-	-	-	-	-	0.49	0.94	-
	POTR-GCN [33]	0.40	0.73	-	0.29	0.68	-	0.39	0.82	-	0.56	0.96	-	0.45	0.91	-	0.63	1.12	-	0.56	1.01	-
	S-Transformer [82]	0.41	0.66	-	0.32	0.71	-	0.46	0.93	-	0.64	1.08	-	-	-	-	-	-	-	-	-	-
	STTG-net [83]	0.33	0.57	-	0.30	0.62	-	0.61	1.15	-	0.47	0.78	-	0.43	0.76	-	0.45	0.94	-	0.50	0.89	-
	MSTP-net [34]	0.43	0.54	0.68	0.29	0.61	1.08	0.40	0.78	1.51	0.54	0.83	1.51	0.42	0.69	1.17	0.62	0.96	1.60	0.53	0.93	1.51

* Trained with single action separately.

network remains adaptive to the camera view and enhances the training.

In addition, Xu et al. [35] introduced a novel framework named EqMotions, which enables seq2seq motion equivariance under Euclidean geometric transformation. EqMotions extracts both equivariant geometric features and invariant pattern features through a multi-layer approach. These features are channeled through an equivariant output layer to produce the final predictions. These layers exhibit equivariant such as

$$\mathbb{G}^{(l+1)}\boldsymbol{R}+\boldsymbol{t} = \mathbb{F}^{(l)}_{EGFL}(\mathbb{G}^{(l)}\boldsymbol{R}+\boldsymbol{t}, \{c_{ij}\}), \tag{20}$$

$$\hat{\mathbb{Y}}\boldsymbol{R}+\boldsymbol{t} = \mathbb{F}^{(l)}_{EOL}(\mathbb{G}^{(l)}\boldsymbol{R}+\boldsymbol{t}). \tag{21}$$

where the rotation matrix is $R \in SO(3)$ and the translation matrix is $t \in (R)^n$ The geometric feature is defined as $\mathbb{G}^L = [G_1^{(L)}, \ldots, G_M^{(L)}] \in \mathbb{R}^{M\times C\times n}$ where the M is the agent in a multi-agent system, C is the geometric coordinates, and n is the dimension. The $\mathbb{F}^{(l)}_{EGFL}$ and $\mathbb{F}^{(l)}_{EOL}$ represents the equivariant geometric feature learning layers and the equivariant output layer. The c_{ij} denotes the set of all the interaction categorical vectors (see Fig. 10).

Notably, Xu et al. [35] demonstrate the data efficiency and generalization capability of equivariance in their model by experimenting with various tasks, including human skeleton motion prediction. Their model outperforms baseline models across multiple tasks while requiring fewer data and training parameters. This work effectively highlights the potential of ENNs and sets a promising direction for future research in human motion prediction.

IV. HUMAN MOTION DATASET

The performance of a specific deep learning approach is intrinsically tied to the quality and volume of the dataset utilized. Obtaining a human motion dataset is both time-consuming and resource-intensive, as it requires multiple individuals to record their movements repeatedly for each action category. While it is feasible to project 3D human motion data synthetically in a simulated environment, such projections often fall short of mimicking realistic, dynamic human movements. In the context of motion reconstruction, [117] endeavored to estimate 3D motion from a singular RGB image input. This paper also lists several motion datasets used in studies of pose reconstruction, motion prediction, and human behavior analysis.

A. HUMAN3.6M [116]

The Human3.6M dataset is extensively used as a benchmark for both human motion prediction and motion reconstruction tasks. Encompassing more than 3.6 million frames, this motion capture data was recorded using a marker-based system and includes natural full-body 3D human poses at the posture level. Additionally, it provides camera images from a static camera for the duration of each action. The dataset features motions from 11 actors: 5 females and 6 males, with each actor performing 15 different daily actions. These actions are categorized into upper-body movements (e.g., greeting, posing, taking photo), moving actions (e.g., walking, walking dog, walking together), and stationary actions (e.g., eating, smoking, sitting). Human motion within the dataset is represented using 32 body joints. The dataset can be accessed from the official website, though a signup is required.

B. 3D POSES IN THE WILD (3DPW) [118]

The 3DPW dataset is a distinctive collection of 60 video sequences, totaling 51,000 frames or approximately 1700 seconds of video. Alongside the video data, the dataset provides invaluable IMU data, 3D scans, and 3D models of participants wearing 18 different clothing variations. This comprehensive set of data offers researchers a deeper insight into human pose in relation to attire and environment. The dataset showcases 7 actors performing a variety of activities in challenging outdoor scenarios. Their actions range from walking and jogging to other dynamic movements, reflecting the complexities of real-world human motion. Given the richness of the dataset, with its mix of 2D video sequences and corresponding 3D pose information, 3DPW dataset stands out as an exceptional resource for those aiming to push the boundaries of human pose estimation techniques.

C. MULTIMODAL HUMAN ACTION DATABASE (MHAD) [119]

The MHAD provides access to mocap data in addition to other modalities such as multi-view video, depth, acceleration, and audio. The dataset contains motion data for 11 actions, including jumping in place, waving hand, clapping hands, throwing a ball, sit down, stand up, etc. This set of activities is performed by 12 actors (including 7 male and 5 female). Each action was performed 5 times, resulting in approximately 660 action sequences.

D. HumanEVA [14]

The HumanEva dataset consists of synchronized video and motion capture data for several human subjects performing a variety of activities, such as walking, running, and jumping. The authors released two versions of datasets, namely HumanEva-I and HumanEva-II. HumanEva-I includes 6 actions performed by 4 subjects, whereas HumanEva-II contains only one action from 2 actors. The dataset was created with the intent to evaluate the performance of various algorithms on the challenging task of 3D human pose estimation. The dataset is available for download on its respective website, although accessing it requires registration and additional steps.

E. NTU RGB+D [13]

The NTU RGB+D dataset is a large-scale dataset for human activity recognition, containing over 56,000 video clips of human actions, captured by Kinect v2 sensors. It contains 40 daily actions (eating, reading, sitting, etc.), 9 medical condition actions (sneezing, chest pain, falling down, etc.), and 11 actions for two-person interaction (pushing, hugging, shaking hands, etc.). Each action is performed twice and is annotated with 3D skeletal data, which provides information about the positions and orientations of various body parts. The pose information is provided in the configuration of 25 body joints.

F. KIT WHOLE-BODY HUMAN MOTION DATABASE [12]

The Whole-Body Human Motion Database comprises motion annotations sourced from various markers placed on the bodies of people and objects. For each frame, pose data is derived from 56 markers situated on the participant's body. The dataset encompasses 715 motion-capture experiments, executed by 53 distinct participants (37 male and 16 female). It offers manipulation motions involving objects, such as drinking, pouring, whisking, and throwing, which aid in evaluating context-aware human motion learning techniques.

G. AMASS [11]

The Archive of Motion Capture as Surface Shapes (AMASS) is, to date, the largest publicly available database of human motions. A significant contribution of the AMASS dataset is its ability to unify other public human motion datasets into a standardized representation. They have consolidated previously available motion capture datasets into an expansive collection under a common body representation. As they obtain new datasets, these are continually added. Currently, there are 25 aggregated datasets in the collection. The AMASS dataset is freely available for download after account registration from their official repository.

V. EVALUATION

Human motion prediction is undoubtedly a challenging task due to the complex dynamics and inherent uncertainty of human actions. Therefore, it is crucial to select performance metrics carefully to guarantee a fair comparison among different proposed prediction methodologies. Primarily, the evaluation of prediction algorithms has focused on two metrics, Mean Angle Error (MAE) and Mean Per Joint Position Error (MPJPE).

The MAE is predominantly used to evaluate prediction methods that utilize angle representation to decode motion dependencies. It calculates the mean error between the ground truth and the predicted per-joint Euler angle. On the other hand, MPJPE is the go-to metric for algorithms that represent future motion prediction in terms of 3D joint positions. As with MAE, MPJPE evaluates the error between the ground truth and the predicted future poses on a per-joint basis.

However, in their study, the authors of [80] pointed out that both MAE and MPJPE might not entirely capture the diversity of predicted motions. They advocate for metrics like Average Displacement Error (ADE), Final Displacement Error (FDE), and Average Pairwise Distance (APD).

1) Average Displacement Error (ADE) measures the accuracy between the whole ground truth motion and the closest sample by computing L^2 distance over it.
2) Final Displacement Error (FDE) computes the reconstruction loss for the final future predicted pose.

TABLE 2. MPJPE evaluation results (in millimeters) for different methods on the H36M dataset [116]. The best results are not highlighted as the number of joints used for evaluation differs among the methods.

		Walking			Eating			Smoking			Discussion			Directions			Sitting Down			Average		
	Time (milliseconds)	160	400	1000	160	400	1000	160	400	1000	160	400	1000	160	400	1000	160	400	1000	160	400	1000
GCN	LTD [37]	15.7	33.4	51.3	18.9	47.2	68.9	14.9	28.7	60.5	22.1	44.1	103.5	24.4	58.4	-	27.6	67.6	-	25.0	61.3	-
	HisRepeat [23]	19.5	39.8	58.1	14.0	36.2	75.7	14.9	36.4	69.5	23.4	65.4	119.8	18.4	56.5	106.5	30.7	72.0	143.6	22.6	58.3	112.1
	LDR [41]	14.9	29.9	45.8	15.9	41.7	53.8	13.4	24.9	43.1	20.3	41.2	67.4	23.7	50.9	78.3	21.4	59.3	144.6	18.1	31.2	79.2
	MGCN [93]	15.0	31.3	43.7	18.4	44.5	68.7	12.9	27.0	55.2	-	-	72.9	22.7	60.1	-	-	-	-	23.2	60.1	-
	MSR-GCN [91]	22.6	45.2	63.0	17.1	40.4	77.1	16.3	38.1	71.6	26.8	69.7	117.6	19.7	53.8	100.6	31.6	76.8	-	25.6	62.9	86.0
	AM-GCN [97]	-	-	-	-	-	-	13.9	30.1	62.5	21.1	48.3	94.3	16.7	43.5	97.0	26.4	61.5	116.4	22.1	57.9	102.6
	MT-GCN [38]	18.8	41.7	60.4	17.2	42.3	74.9	15.7	39.7	70.8	22.7	64.6	115.7	23.2	56.7	108.5	29.6	70.8	140.2	21.3	44.2	69.0
	SGSN [25]	15.0	31.5	43.6	17.4	43.7	68.2	13.8	28.2	58.8	19.1	40.4	96.5	21.6	56.3	101.0	24.7	60.2	126.8	22.9	56.9	100.8
	SPGSN [100]	19.4	41.5	53.6	14.9	37.9	73.4	13.8	34.6	68.6	23.8	67.1	-	17.1	50.3	100.5	27.7	70.7	-	22.3	58.3	109.6
	STS-GCN [120]	16.9	32.9	51.8	11.3	25.4	52.4	11.6	25.8	50.0	16.8	40.2	78.8	13.5	34.7	71.0	23.7	47.9	94.3	17.1	38.3	75.6
	GAGCN [26]	16.1	32.4	51.1	11.5	25.2	51.4	11.8	24.3	48.7	17.1	38.9	76.9	12.8	34.5	69.9	24.8	47.4	84.1	16.9	38.5	72.9
	CGHMP [98]	18.6	37.0	53.7	14.4	36.2	72.6	14.1	34.8	68.6	24.3	68.8	120.6	17.1	50.6	100.7	28.5	73.1	-	22.8	59.3	-
Others	TrajectoryCNN [21]	14.9	35.4	46.4	18.4	44.8	71.5	12.8	27.8	58.7	20.0	47.8	103.0	22.3	61.7	104.2	28.8	62.9	123.8	23.2	59.7	110.6
	TF-CNN [59]	14.8	34.4	-	17.1	42.6	-	12.7	27.6	-	18.7	46.8	-	22.9	65.1	-	27.4	60.7	-	22.6	58.4	-
	HR-STAN [55]	17.0	51.9	-	12.1	38.5	-	10.4	33.0	-	18.1	61.9	-	13.9	49.1	- 17.5	58.1	-	17.0	56.4	-	
	MA-WGAN [79]	21.1	42.4	68.2	15.9	35.0	65.3	14.3	30.4	63.4	22.7	54.6	102.2	20.9	47.0	83.5	28.2	64.5	125.2	22.5	50.8	96.4
	BiTGAN [77]	-	-	60.5	-	-	73.0	-	-	70.0	-	-	116.4	-	-	106.3	-	-	141.3	-	-	111.1
	ProgDec [102]	14.5	34.5	41.2	18.1	45.3	67.9	13.2	27.5	58.3	-	-	103.1	22.7	58.4	-	-	-	-	23.8	60.2	-
	AdvMT [81]	23.9	39.9	55.0	18.3	36.1	59.3	22.8	45.5	77.7	36.0	66.7	101.0	-	-	103.5	-	-	142.2	-	-	106.6
	siMLPe [47]	-	39.6	55.7	-	36.1	74.5	-	36.3	69.3	-	64.3	116.3	-	55.8	106.7	-	70.8	142.4	-	57.3	109.4
	G-G Mixer [48]	-	38.6	55.6	-	34.1	74.3	-	36.3	68.8	-	63.4	115.0	-	55.7	106.8	-	69.2	138.6	-	56.7	108.6
	EqMotion [35]	17.5	39.2	52.8	13.6	36.5	73.0	11.3	29.3	63.4	18.9	53.9	105.6	15.8	50.1	-	26.5	70.7	-	20.1	55.0	106.9

3) Average Pairwise Distance (APD) calculates the diversity within the predicted future motion. L^2 distance is calculated between all motion samples.

The importance of robust datasets and comprehensive evaluation metrics becomes increasingly clear as the complexity of human motion prediction tasks is recognized. As elaborated, the Human3.6M dataset [116] stands out as a frequently utilized resource for these tasks, and many motion prediction methods have been benchmarked against it. According to the literature, future motion prediction is quantified in two distinct scenarios. The first is short-term prediction, focusing on future human poses up to 400ms. The second scenario is long-term prediction, which includes poses that exceed this 400ms threshold.

This paper has summarized the outcomes of various motion prediction algorithms discussed earlier. We have primarily used the MAE and MPJPE as our evaluation metrics, reflecting their common usage among these methods. The data presented in Tables 1 and 2 are sourced directly from the corresponding published papers. Since the methods are trained and evaluated with different numbers of joints, we refrained from highlighting the best results to maintain a fair and uniform comparison across all approaches. We have included results for both short-term predictions (at 160ms and 400ms) and long-term predictions (at 1000ms). First, we include an analysis of four primary actions from the Human3.6M dataset, namely walking, eating, smoking, and discussion. To deepen our comparison, we also examine more challenging actions such as directions and sitting down, with the latter presenting considerable challenges due to occlusion and a wide range of movements. The findings suggest that GCN-based methods tend to perform better in terms of prediction accuracy, benefiting from their ability to leverage the geometrical structure of the human body.

VI. FUTURE RESEARCH DIRECTION

A. META-LEARNING

The meta-learning approach shows promise in advancing human motion prediction, especially in situations with limited data and varied environments [121]. Meta-learning models undergo training across a multitude of tasks during meta-training, which in turn equips them with the capacity to rapidly adjust to novel situations using only a small amount of task-specific data. Such a capability is indispensable in situations where immediate predictions are paramount, and the acquisition of new data is infeasible.

Additionally, meta-learning enhances model generalization by integrating knowledge across a wide variety of tasks [122]. This accrued knowledge encapsulates recurring human motion patterns prevalent in various contexts, ensuring precise predictions even when confronted with subtle motion deviations. Analogous to how humans draw from prior experiences to swiftly adapt to new circumstances, these models can benefit from their past learnings, capitalizing on their meta-knowledge to render predictions that are both efficient and of high quality.

In the context of human motion prediction, the potential of meta-learning is evident. These models, with their adaptability and proficiency in transfer learning, suggest a promising avenue for future research. By transferring insights from one specific context to diverse scenarios, they can provide well-informed predictions even as conditions change. This versatility positions meta-learning as a valuable direction for further advancements in human motion prediction, particularly in environments that are constantly evolving.

B. CAUSAL LEARNING

Causal models enable counterfactual reasoning, which involves asking questions about how outcomes might differ if certain conditions were altered [123]. For instance, researchers can inquire about the effect of changing the starting position on their future trajectory. The models can simulate interventions virtually. If there is a need to understand how adding an obstacle affects the walking trajectory, the model can simulate the changed scenario. By adjusting input variables, the model could project how the altered conditions would impact the predicted motion.

Beyond predicting motion based on observed patterns, causal models quantify the effects of interventions. They provide numerical estimates of how much motion changes due to specific alterations, offering a deeper understanding of the implications of interventions. By predicting outcomes of interventions, causal models assist in making informed decisions, especially when optimizing motions for desired outcomes while considering various constraints.

Causal models let us explore hypothetical scenarios, simulate interventions, and understand the ramifications of changes. This could aid in making informed decisions, optimizing motions, and adapting motion predictions for various real-world situations.

C. PROBABLISTIC MODELING

Probabilistic modeling has greatly enhanced human motion prediction by tackling the inherent uncertainties in complex movements. These models quantify uncertainty, offering not only predictions but also the associated confidence levels. For instance, the confidence-aware motion prediction approach integrates a Bayesian belief model to maintain and update its confidence in the human motion predictions [124]. This method continuously updates its confidence based on real-time observations, making it robust to unanticipated human behaviors.

Probabilistic models enhance decision-making by providing insights into likely outcomes and can handle multimodal data, capturing diverse behaviors by modeling multiple possible trajectories. For example, the Bayesian Neural Network (BNN) approach addresses the problem of human motion prediction by extending deterministic models to probabilistic ones [125]. It replaces deterministic weight parameters with distributions over these parameters, enabling the model to generate a range of possible future motions given an observed sequence. This approach effectively captures both uncertainty about the model itself and inherent noise in the data, offering a more comprehensive and reliable prediction framework.

However, these contemporary methods face specific challenges in the context of human motion prediction. Real-time updates and inference with Bayesian models can be computationally demanding, limiting their practical use. Additionally, the accuracy of these methods relies heavily on the quality of the training data. The models may produce unreliable predictions when the data is limited or biased. Furthermore, although probabilistic models can generate multiple potential outcomes, distinguishing between these outcomes and identifying the most accurate prediction remains challenging, especially in scenarios with highly variable human behaviors.

Despite these challenges, probabilistic models can adapt to changing patterns and environments, making them suitable for dynamic situations. Their proficiency in uncertainty estimation is particularly beneficial in fields like human-robot interaction, where safety and reliability are paramount. Overall, further exploration and refinement of probabilistic modeling approaches, such as improving computational efficiency and ensuring high-quality training data, can lead to significant advancements in the accuracy and reliability of human motion prediction.

D. EQUIVARIANT SYMMETRY LEARNING

Efficient generalization in machine learning is often challenging and typically requires large datasets. However, the principle of equivariant symmetry offers a practical approach to this issue. Models built on this principle are adept at utilizing inherent symmetries in the data, significantly boosting data efficiency and generalizability.

Additionally, equivariance increases the models' ability to detect complex features within the data. By identifying symmetries in the data, these models achieve a deeper understanding and improved performance. For instance, utilizing models like the Equiformer [126] or Tensor Field Networks [127], we can extract type-l features (e.g., type-0, type-1, etc.). A type-0 features, like colors, remain invariant to rotations, whereas a type-1 features are equivariant to rotations, such as to 3D vectors. Exploiting these features from the data enhances the models' robustness compared to when a conventional model is employed.

Therefore, equivariant symmetries present significant opportunities for data efficiency and robust generalization. Their ability to utilize symmetries and related information decreases the amount of data required for effective training and facilitates the extraction of more meaningful features. Applying this concept in human motion prediction yields considerable benefits, guaranteeing reliable results and data efficiency, particularly in situations where data collection is expensive.

VII. CONCLUSION AND DISCUSSION

In this survey paper, we have presented an overview of the latest developments in human motion prediction, including their limitations and potential future research paths. This topic has attracted significant attention in recent years, leading to considerable improvements in the performance of motion prediction models compared to earlier efforts. However, human motion prediction remains a daunting challenge. While many studies have shown success in short-term prediction, long-term prediction accuracy still presents difficulties.

In this field, the two primary modeling strategies that have gained prominence are direct modeling and geometry-aware methods. Direct modeling, while straightforward and often effective for short-term predictions, tends to encounter difficulties when dealing with long-term predictions. On the other hand, geometry-aware methods, which incorporate the inherent structural and spatial information of human joints, have shown promise in addressing some of the challenges faced by direct modeling. By leveraging the spatial relationships and constraints of human anatomy, geometry-aware models often produce more realistic and precise motion predictions over extended durations.

Given the current landscape, it is prudent to advocate for a more profound exploration of geometry-aware methods. Their innate ability to capture and utilize the spatial connections of human body joints makes them potentially more robust and reliable for diverse scenarios. However, this does not diminish the value of direct modeling, which, when combined with other techniques, can still offer valuable insights and solutions. Accordingly, it is essential to explore the potential of equivariant learning in human motion prediction. ENNs could further enhance the robustness of the predictions by exploiting the geometric features, which increases the accuracy and reduces the volume of data.

Our focus should also pivot towards engineering real-time human motion prediction systems that can provide instantaneous feedback and adapt to dynamic environments, enabling applications in real-time human-robot interaction, virtual reality, and interactive games [128], [129]. The necessity for these systems is driven by their potential to enhance user experiences and improve safety in interactive applications. However, developing these real-time human motion prediction systems is not straightforward. It requires high computational efficiency for real-time processing, as well as the ability to handle diverse and dynamic human movements while maintaining robust performance under varying conditions. Addressing these aspects would not only yield more dependable forecasts but also empower systems to make informed decisions rooted in prediction confidence.

Nevertheless, the journey of human motion prediction is interspersed with challenges. While geometry-aware methods currently hold an edge, the dynamic nature of research could lead to breakthroughs in direct modeling or even hybrid approaches. Robustly addressing environment interactions, occlusions, and varied scenes also present persistent challenges. We remain optimistic that advancements in human motion prediction will catalyze transformative impacts across sectors such as robotics, virtual reality, sports analytics, and healthcare.

REFERENCES

[1] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon, "A survey of robots in healthcare," *Technologies*, vol. 9, no. 1, p. 8, Jan. 2021.

[2] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *Int. J. Robot. Res.*, vol. 39, no. 8, pp. 895–935, Jul. 2020.

[3] H. Liu and L. Wang, "Human motion prediction for human–robot collaboration," *J. Manuf. Syst.*, vol. 44, pp. 287–294, Jul. 2017.

[4] R. Jahanmahin, S. Masoud, J. Rickli, and A. Djuric, "Human–robot interactions in manufacturing: A survey of human behavior modeling," *Robot. Comput.-Integr. Manuf.*, vol. 78, Dec. 2022, Art. no. 102404.

[5] L. Antonyshyn, J. Silveira, S. Givigi, and J. Marshall, "Multiple mobile robot task and motion planning: A survey," *ACM Comput. Surv.*, vol. 55, no. 10, pp. 1–35, Feb. 2023.

[6] H. Guo, F. Wu, Y. Qin, R. Li, K. Li, and K. Li, "Recent trends in task and motion planning for robotics: A survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–36, Jul. 2023.

[7] M. Marchellus and I. K. Park, "Deep learning for 3D human motion prediction: State-of-the-Art and future trends," *IEEE Access*, vol. 10, pp. 35919–35931, 2022.

[8] M. Kim, S. Lee, J. Lim, J. Choi, and S. G. Kang, "Unexpected collision avoidance driving strategy using deep reinforcement learning," *IEEE Access*, vol. 8, pp. 17243–17252, 2020.

[9] J. Park, H. Yong, S. Ha, J. Lee, and J. Choi, "Customer-specific robotic attendant for VR simulators," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 1901–1910, Oct. 2020.

[10] H. Lee, W. Byun, H. Lee, Y. Kang, and J. Choi, "Integration and evaluation of an immersive virtual platform," *IEEE Access*, vol. 11, pp. 1335–1347, 2023.

[11] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5441–5450.

[12] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The KIT whole-body human motion database," in *Proc. Int. Conf. Adv. Robot. (ICAR)*, Jul. 2015, pp. 329–336.

[13] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[14] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, 2010.

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

[16] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K., Aug. 2020, pp. 387–404.

[17] E. Corona, A. Pumarola, G. Alenya, and F. Moreno-Noguer, "Context-aware human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6990–6999.

[18] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA. Washington, DC, USA: IEEE Computer Society, Jul. 2017, pp. 4674–4683.

[19] D. Pavllo, D. Grangier, and M. Auli, "QuaterNet: A quaternion-based recurrent model for human motion," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 675–688.

[20] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3D human motion modelling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7144–7153.

[21] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "TrajectoryCNN: A new spatio-temporal feature learning network for human motion prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2133–2146, Jun. 2021.

[22] J. Bütepage, M. J. Black, D. Kragic, and H. Kjellström, "Deep representation learning for human motion prediction and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6158–6166.

[23] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Proc. 16th Eur. Conf. Comput. Vision (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 474–489.

[24] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 211–220.

[25] M. Li, S. Chen, Z. Liu, Z. Zhang, L. Xie, Q. Tian, and Y. Zhang, "Skeleton graph scattering networks for 3D skeleton-based human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 854–864.

[26] C. Zhong, L. Hu, Z. Zhang, Y. Ye, and S. Xia, "Spatio-temporal gating-adjacency GCN for human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6437–6446.

[27] E. Barsoum, J. Kender, and Z. Liu, "HP-GAN: Probabilistic 3D human motion prediction via GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1499–149909.

[28] K. Lyu, Z. Liu, S. Wu, H. Chen, X. Zhang, and Y. Yin, "Learning human motion prediction via stochastic differential equations," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4976–4984.

[29] S. Aliakbarian, F. Sadat 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. (CVPR)*, Jun. 2020, pp. 5222–5231.

[30] Z. Liu, K. Lyu, S. Wu, H. Chen, Y. Hao, and S. Ji, "Aggregated multi-GANs for controlled 3D human motion prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2225–2232.

[31] C. Gu, J. Yu, and C. Zhang, "Learning disentangled representations for controllable human motion prediction," *Pattern Recognit.*, vol. 146, Feb. 2024, Art. no. 109998.

[32] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A spatio-temporal transformer for 3D human motion prediction," in *Proc. Int. Conf. 3D Vis. (3DV)*, Los Alamitos, CA, USA. Washington, DC, USA: IEEE Computer Society, Dec. 2021, pp. 565–574.

[33] A. Martínez-González, M. Villamizar, and J.-M. Odobez, "Pose transformers (POTR): Human motion prediction with non-autoregressive transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2276–2284.

[34] L. Chen, R. Liu, W. Zhang, Y. Hou, Q. Zhang, and D. Zhou, "MSTP-Net: Multiscale spatio-temporal parallel networks for human motion prediction," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3318–3331, Feb. 2023.

[35] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang, "EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1410–1420.

[36] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4346–4354.

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

[38] Q. Cui and H. Sun, "Towards accurate 3D human motion prediction from incomplete observations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4799–4808.

[39] X. Guo and J. Choi, "Human motion prediction via learning local structure representations and temporal dependencies," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 2580–2587.

[40] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[41] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3D human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6518–6526.

[42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5769–5779.

[43] J. N. Kundu, M. Gor, and R. V. Babu, "BiHMP-GAN: Bidirectional 3D human motion prediction GAN," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8553–8560.

[44] X. Chao, Y. Bin, W. Chu, X. Cao, Y. Ge, C. Wang, J. Li, F. Huang, and H. Leung, "Adversarial refinement network for human motion prediction," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 454–469.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[46] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[47] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to MLP: A simple baseline for human motion prediction," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4798–4808.

[48] X. Wang, Q. Cui, C. Chen, S. Zhao, and M. Liu, "Graph-guided MLP-mixer for skeleton-based human motion prediction," in *Proc. ACM Multimedia Asia*, Dec. 2023, pp. 1–7.

[49] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999.

[50] R. Yeh, Y.-T. Hu, and A. Schwing, "Chirality nets for human pose regression," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Red Hook, NY, USA: Curran Associates, 2019.

[51] C. Xu, S. Chen, M. Li, and Y. Zhang, "Invariant teacher and equivariant student for unsupervised 3D human pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, May 2021, pp. 3013–3021.

[52] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 3104–3112.

[53] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[54] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668.

[55] O. Medjaouri and K. Desai, "HR-STAN: High-resolution spatio-temporal attention network for 3D human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2539–2548.

[56] A. Hernandez, J. Gall, and F. Moreno, "Human motion prediction via spatio-temporal inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7133–7142.

[57] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 265–281.

[58] S. Aliakbarian, F. Saleh, L. Petersson, S. Gould, and M. Salzmann, "Contextually plausible and diverse 3D human motion prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11313–11322.

[59] J. Tang, J. Zhang, and J. Yin, "Temporal consistency two-stream CNN for human motion prediction," *Neurocomputing*, vol. 468, pp. 245–256, Jan. 2022.

[60] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 7444–7452.

[61] Y. Tang, L. Ma, W. Liu, and W.-S. Zheng, "Long-term human motion prediction by modeling motion context and enhancing motion dynamics," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 935–941.

[62] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G. Ororbia, "A neural temporal model for human motion prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12108–12117.

[63] H. Flashner and J. L. McNitt-Gray, "3D kinematics: Using quaternions for modeling orientation and rotations in biomechanics," in *Biomechanical Principles and Applications in Sports*. Cham, Switzerland: Springer, 2019, pp. 155–233.

[64] L.-Y. Gui, Y.-X. Wang, X. Liang, and J. M. F. Moura, "Adversarial geometry-aware human motion prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 786–803.

[65] F. S. Grassia, "Practical parameterization of rotations using the exponential map," *J. Graph. Tools*, vol. 3, no. 3, pp. 29–48, Jan. 1998.

[66] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar. Baltimore, MD, USA: Association for Computational Linguistics, 2014, pp. 1724–1734.

[67] H.-F. Sang, Z.-Z. Chen, and D.-K. He, "Human motion prediction based on attention mechanism," *Multimedia Tools Appl.*, vol. 79, nos. 9–10, pp. 5529–5544, Mar. 2020.

[68] A. F. Al-aqel and M. Ali Khan, "Attention mechanism for human motion prediction," in *Proc. 3rd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Mar. 2020, pp. 1–6.

[69] R. Zhang, X. Shu, R. Yan, J. Zhang, and Y. Song, "Skip-attention encoder–decoder framework for human motion prediction," *Multimedia Syst.*, vol. 28, pp. 413–422, Jun. 2022.

[70] H. Wang, J. Dong, B. Cheng, and J. Feng, "PVRED: A position-velocity recurrent encoder–decoder for human motion prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 6096–6106, 2021.

[71] C. Guo, R. Liu, C. Che, D. Zhou, Q. Zhang, and X. Wei, ''Fusion learning-based recurrent neural network for human motion prediction,'' *Intell. Service Robot.*, vol. 15, no. 3, pp. 245–257, Jul. 2022.

[72] Y. Yu, N. Tian, X. Hao, T. Ma, and C. Yang, ''Human motion prediction with gated recurrent unit model of multi-dimensional input,'' *Appl. Intell.*, vol. 52, no. 6, pp. 6769–6781, Apr. 2022.

[73] 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. (CVPR)*, Jun. 2022, pp. 8151–8161.

[74] 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.*, Salt Lake City, UT, USA. Washington, DC, USA: IEEE Computer Society, Jun. 2018, pp. 5226–5234.

[75] J. Tang, J. Liu, and J. Yin, ''A hierarchical static-dynamic encoder–decoder structure for 3D human motion prediction with residual CNNs,'' *Math. Problems Eng.*, vol. 2020, Aug. 2020, Art. no. 7064910.

[76] X. Liu, J. Yin, H. Liu, and Y. Yin, ''PISEP^2: Pseudo-image sequence evolution-based 3D pose prediction,'' *Vis. Comput.*, vol. 38, no. 7, pp. 2603–2616, 2022.

[77] M. Zhao, H. Tang, P. Xie, S. Dai, N. Sebe, and W. Wang, ''Bidirectional transformer GAN for long-term human motion prediction,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 5, pp. 1–19, Apr. 2023.

[78] Z. Huang, J. Wu, and L. Van Gool, ''Manifold-valued image generation with Wasserstein generative adversarial nets,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3886–3893.

[79] B. Chopin, N. Otberdout, M. Daoudi, and A. Bartolo, ''Human motion prediction using manifold-aware Wasserstein GAN,'' in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[80] Y. Yuan and K. Kitani, ''DLow: Diversifying latent flows for diverse human motion prediction,'' in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Berlin, Germany: Springer-Verlag, Aug. 2020, pp. 346–364.

[81] S. Idrees, J. Choi, and S. Sohn, ''AdvMT: Adversarial motion transformer for long-term human motion prediction,'' 2024, *arXiv:2401.05018*.

[82] C. Zhong, L. Hu, and S. Xia, ''Spatial–temporal modeling for prediction of stylized human motion,'' *Neurocomputing*, vol. 511, pp. 34–42, Oct. 2022.

[83] L. Chen, R. Liu, X. Yang, D. Zhou, Q. Zhang, and X. Wei, ''STTG-net: A spatio-temporal network for human motion prediction based on transformer and graph convolution network,'' *Vis. Comput. for Ind., Biomed., Art*, vol. 5, no. 1, p. 19, Dec. 2022.

[84] Q. Men, E. S. L. Ho, H. P. H. Shum, and H. Leung, ''A quadruple diffusion convolutional recurrent network for human motion prediction,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3417–3432, Sep. 2021.

[85] K. Tonchev, A. Manolova, R. Petkova, and V. Poulkov, ''Human skeleton motion prediction using graph convolution optimized GRU network,'' in *Proc. XXX Int. Sci. Conf. Electron. (ET)*, Sep. 2021, pp. 1–5.

[86] Y. Li, Z. Wang, X. Yang, M. Wang, S. I. Poiana, E. Chaudhry, and J. Zhang, ''Efficient convolutional hierarchical autoencoder for human motion prediction,'' *Vis. Comput.*, vol. 35, nos. 6–8, pp. 1143–1156, Jun. 2019.

[87] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, ''Graph attention networks,'' *Stat*, vol. 1050, p. 20, Oct. 2017.

[88] W. Mao, M. Liu, M. Salzmann, and H. Li, ''Multi-level motion attention for human motion prediction,'' *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2513–2535, Sep. 2021.

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

[90] T. Lebailly, S. Kiciroglu, M. Salzmann, P. Fua, and W. Wang, ''Motion prediction using temporal inception module,'' in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 651–665.

[91] 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. (ICCV)*, Oct. 2021, pp. 11447–11456.

[92] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, ''Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3316–3333, Jun. 2022.

[93] H. Zhou, C. Guo, H. Zhang, and Y. Wang, ''Learning multiscale correlations for human motion prediction,'' in *Proc. IEEE Int. Conf. Develop. Learn. (ICDL)*, Aug. 2021, pp. 1–7.

[94] 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.* Switzerland: Springer, 2022, pp. 251–269.

[95] W. Cao, S. Li, and J. Zhong, ''QMEDNet: A quaternion-based multi-order differential encoder–decoder model for 3D human motion prediction,'' *Neural Netw.*, vol. 154, pp. 141–151, Oct. 2022.

[96] Q. Li, Y. Wang, and F. Lv, ''Semantic correlation attention-based multi-order multiscale feature fusion network for human motion prediction,'' *IEEE Trans. Cybern.*, vol. 54, no. 2, pp. 825–838, Feb. 2024.

[97] P. Su, X. Shen, Z. Shi, and W. Liu, ''Adaptive multi-order graph neural networks for human motion prediction,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[98] J. Li, H. Pan, L. Wu, C. Huang, X. Luo, and Y. Xu, ''Class-guided human motion prediction via multi-spatial–temporal supervision,'' *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9463–9479, May 2023.

[99] Q. Cui, H. Sun, Y. Kong, X. Zhang, and Y. Li, ''Efficient human motion prediction using temporal convolutional generative adversarial network,'' *Inf. Sci.*, vol. 545, pp. 427–447, Feb. 2021.

[100] M. Li, S. Chen, Z. Zhang, L. Xie, Q. Tian, and Y. Zhang, ''Skeleton-parted graph scattering networks for 3D human motion prediction,'' in *Proc. Eur. Conf. Comput. Vis.* Switzerland: Springer, 2022, pp. 18–36.

[101] 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. (CVPR)*, Jun. 2022, pp. 6427–6436.

[102] Y. Cai, L. Huang, Y. Wang, T.-J. Cham, J. Cai, J. Yuan, J. Liu, X. Yang, Y. Zhu, X. Shen, D. Liu, J. Liu, and N. M. Thalmann, ''Learning progressive joint propagation for human motion prediction,'' in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 226–242.

[103] T. S. Cohen and M. Welling, ''Steerable CNNs,'' 2016, *arXiv:1612.08498*.

[104] M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. S. Cohen, ''3D steerable CNNs: Learning rotationally equivariant features in volumetric data,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 10402–10413.

[105] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. Guibas, ''Vector neurons: A general framework for SO(3)-equivariant networks,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12180–12189.

[106] J. Guan, W. Wei Qian, X. Peng, Y. Su, J. Peng, and J. Ma, ''3D equivariant diffusion for target-aware molecule generation and affinity prediction,'' 2023, *arXiv:2303.03543*.

[107] H. Ryu, J. Kim, H. An, J. Chang, J. Seo, T. Kim, Y. Kim, C. Hwang, J. Choi, and R. Horowitz, ''Diffusion-EDFs: Bi-equivariant denoising generative modeling on SE(3) for visual robotic manipulation,'' 2023, *arXiv:2309.02685*.

[108] H. Ryu, H.-I. Lee, J.-H. Lee, and J. Choi, ''Equivariant descriptor fields: SE(3)-equivariant energy-based models for end-to-end visual robotic manipulation learning,'' in *Proc. 11th Int. Conf. Learn. Represent.*, 2023.

[109] J. Kim, H. Ryu, J. Choi, J. Seo, N. P. S. Prakash, R. Li, and R. Horowitz, ''Robotic manipulation learning with equivariant descriptor fields: Generative modeling, bi-equivariance, steerability, and locality,'' in *Proc. RSS Workshop Symmetries Robot Learn.*, 2023.

[110] J. Seo, N. P. S. Prakash, A. Rose, J. Choi, and R. Horowitz, ''Geometric impedance control on SE(3) for robotic manipulators,'' *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 276–283, 2023.

[111] J. Lei, C. Deng, K. Schmeckpeper, L. Guibas, and K. Daniilidis, ''EFEM: Equivariant neural field expectation maximization for 3D object segmentation without scene supervision,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4902–4912.

[112] C. Deng, J. Lei, B. Shen, K. Daniilidis, and L. Guibas, ''Banana: Banach fixed-point network for pointcloud segmentation with inter-part equivariance,'' 2023, *arXiv:2305.16314*.

[113] E. Chatzipantazis, S. Pertigkiozoglou, E. Dobriban, and K. Daniilidis, ''SE(3)-equivariant attention networks for shape reconstruction in function space,'' 2022, *arXiv:2204.02394*.

[114] D. Wang, R. Walters, X. Zhu, and R. Platt, ''Equivariant *q* learning in spatial action spaces,'' in *Proc. Conf. Robot Learn.*, 2022, pp. 1713–1723.

[115] A. K. Mondal, P. Nair, and K. Siddiqi, ''Group equivariant deep reinforcement learning,'' 2020, *arXiv:2007.03437*.

[116] 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.

[117] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A joint model for 2D and 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3634–3641.

[118] T. von Marcard, R. Henschel, M. 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. (ECCV)*, Sep. 2018, pp. 601–617.

[119] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 53–60.

[120] 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. (ICCV)*, Oct. 2021, pp. 11189–11198.

[121] L.-Y. Gui, Y.-X. Wang, D. Ramanan, and J. M. F. Moura, "Few-shot human motion prediction via meta-learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 432–450.

[122] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," 2020, *arXiv:2004.05439*.

[123] J. Berrevoets, K. Kacprzyk, Z. Qian, and M. van der Schaar, "Causal deep learning," 2023, *arXiv:2303.02186*.

[124] D. Fridovich-Keil, A. Bajcsy, J. F. Fisac, S. L. Herbert, S. Wang, A. D. Dragan, and C. J. Tomlin, "Confidence-aware motion prediction for real-time collision avoidance," *Int. J. Robot. Res.*, vol. 39, nos. 2–3, pp. 250–265, 2020.

[125] J. Xu, X. Chen, X. Lan, and N. Zheng, "Probabilistic human motion prediction via a Bayesian neural network," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 3190–3196.

[126] Y.-L. Liao and T. Smidt, "Equiformer: Equivariant graph attention transformer for 3D atomistic graphs," 2022, *arXiv:2206.11990*.

[127] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, "Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds," 2018, *arXiv:1802.08219*.

[128] T. B. Tuli and M. Manns, "Real-time motion tracking for humans and robots in a collaborative assembly task," *Proceedings*, vol. 42, no. 1, p. 48, 2020.

[129] E. Coronado, S. Itadera, and I. G. Ramirez-Alpizar, "Integrating virtual, mixed, and augmented reality to human–robot interaction applications using game engines: A brief review of accessible software tools and frameworks," *Appl. Sci.*, vol. 13, no. 3, p. 1292, Jan. 2023.

SARMAD IDREES (Graduate Student Member, IEEE) received the B.E. degree in electronics engineering from the NED University of Engineering and Technology, Karachi, Pakistan, in 2017. He is currently pursuing the integrated M.S. and Ph.D. degree in mechanical engineering with Yonsei University, Seoul, South Korea. He is a member of the Machine Learning and Control System Laboratory. His research interests include computer vision, human–robot collaboration, and human motion prediction.

JIWOO KIM is currently pursuing the B.S. degree in electrical and electronic engineering with Yonsei University, Seoul, Republic of Korea. He is an Undergraduate Research Intern with the Machine Learning and Control System Laboratory. His research interests include robotics, machine learning, deep reinforcement learning, lie group equivariance, and symmetry.

JONGEUN CHOI (Member, IEEE) received the B.S. degree in mechanical design and production engineering from Yonsei University, Seoul, Republic of Korea, in 1998, and the M.S. and Ph.D. degrees in mechanical engineering from the University of California at Berkeley, in 2002 and 2006, respectively.

He was a Visiting Scholar with UC Berkeley, in 2023. From 2020 to 2022, he was the Graduate Program Chair of the School of Mechanical Engineering, Yonsei University. From 2019 to 2023, he was the Chairperson of the Department of Vehicle Convergence Engineering, Yonsei University, funded by Hyundai Motor Company. Since 2020, he has been with the Department of Artificial Intelligence, Yonsei University. He is currently a Professor with the School of Mechanical Engineering, Yonsei University. Prior to joining Yonsei University, he worked for ten years as an Associate Professor, from 2012 to 2016, and an Assistant Professor, from 2006 to 2012, with the Department of Mechanical Engineering and the Department of Electrical and Computer Engineering, Michigan State University. His current research interests include machine learning, systems and control, deep reinforcement learning, and Bayesian methods with applications to robotics, autonomous driving, human and robot interaction, and AI in healthcare. He is a member of ASME. He received the Best Paper Award at the RSS 2023 Workshop on Symmetries in Robot Learning, in 2023, and the 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), in 2015. His papers were finalists for the Best Student Paper Award at the 24th American Control Conference (ACC), in 2005, and the Dynamic System and Control Conference (DSCC), in 2011 and 2012. He was a recipient of the NSF CAREER Award, in 2009. He has been a Senior Editor of *International Journal of Control, Automation, and Systems* (IJCAS) since 2023. He served as a Guest Editor with a two-year term (2021 and 2022) for the IEEE/ASME TMECH/AIM Focused Section on Emerging Topics. He co-organized and co-edited Special Issues on Stochastic Models, Control, and Algorithms in Robotics in *Journal of Dynamic Systems, Measurement and Control*, from 2014 to 2015. He served as an Associate Editor for IEEE Robotics and Automation Letters, in 2018; *Journal of Dynamic Systems, Measurement and Control*, from 2014 to 2019; and *International Journal of Precision Engineering and Manufacturing*, from 2017 to 2018. He served as a Senior Editor for *Ubiquitous Robots*, in 2020, and an Associate Editor for the 2021 IEEE International Conference on Robotics and Automation (ICRA).

SEOKMAN SOHN received the B.S. degree in mechanical engineering from Yonsei University, in 1993, and the M.S. degree from Korea Advanced Institute of Science and Technology, in 1995. He had investigated on the diagnosis of rotating machinery. He is currently a Leader of the Worker Safety Technology Team, Korea Electric Power Research Institute, KEPCO. His research interest includes worker safety technology development.

• • •