Deep Learning for 3D Human Motion Prediction: State-of-the-Art and Future Trends

MATTHEW MARCHELLUS, (Student Member, IEEE), IN KYU PARK, (SENIOR MEMBER, IEEE)
Department of Electrical and Computer Engineering, Inha University, Incheon 22212, Korea (e-mail: marchellusmatthew@gmail.com, pik@inha.ac.kr)
Corresponding author: In Kyu Park (e-mail: pik@inha.ac.kr).
This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFCIT-1901-06. This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-01389, Artificial Intelligence Convergence Research Center (Inha University)).

ABSTRACT Due to the success of deep learning in wide range of computer vision and computer graphics tasks, there is an increasing number of developed methods leveraging deep neural networks to solve human motion prediction. Recent motion prediction methods focus on solving many issues to predict accurate and natural human motion in temporal domain. In this study, we present a comprehensive survey of deep-learning-based human motion prediction methods. First, we define the human motion prediction problem and the scope of this study. We then provide related background knowledge and a comprehensive list of motion prediction methods based on our proposed classification. Next, we provide a complete survey of the characteristics widely used in the literature and explain the evaluation processes. Finally, we presented a quantitative comparison of recent studies and address the remaining unsolved issues while exploring possible research directions for future research.

INDEX TERMS Future motion, human motion prediction, deep learning.

I. INTRODUCTION
The ability of machines to anticipate future human motion is beneficial because it opens up many possible applications. Human character animation in the virtual world and computer games can be further improved using a flawless human motion prediction model. More importantly, developing a prediction model that is comparable to human behavior necessitates learning how humans behave while performing certain actions, which is crucial for many computer vision tasks. This knowledge is helpful in applications such as sports, surveillance, autonomous driving cars, robotics, and smart user interfaces.

Humans can predict a set of actions that are about to transpire based on this learned experience. However, prediction is a challenging task for machines and predicting future human actions is obviously more complicated than in other applications. Even with deep learning models, for every new context or clue that might hint at a future action necessitates a proper modification of the existing model to produce plausible and substantial results.

During the last decade, attempts have been made to advance machine learning capability to predict the future. Early researchers attempted to use a recurrent neural network (RNN) to extrapolate inputs to the future [1]. It was later proven that a simple RNN is inadequate, necessitating additional modules such as long short-term memory (LSTM) [2] and gated recurrent units (GRU) [3] to improve the results.

Although many diverse studies exist in the human motion prediction domain, we limit the scope to motion prediction methods from a sequence of data using deep learning. Moreover, we further limit the scope of the survey to methods that generate human motion prediction on top of a human body representation model, such as a 3D human skeleton, parametric models, and 3D mesh.

Exhaustive surveys on a few related topics of computer vision exist, but none of them specifically discuss human motion prediction. Oprea et al. [4] provided a thorough review of deep-learning methods for video prediction. Video prediction and human motion prediction have a lot in common, but video prediction deals with images instead of human poses in each time step. Rudenko et al. [5] conducted a survey on the 2D human motion trajectory prediction problem.
This study provides an exhaustive survey of recent studies on deep learning-based human motion prediction. Based on the survey, researchers engaged in this area can easily and properly design their own advanced algorithms.

The remainder of this study is organized as follows. Section II defines the problem of human motion prediction and addresses the challenges in this field. Section III introduces the common types of deep neural network backbones used in human motion prediction. Section IV, we explain different representations of the human body and provide a list of human motion datasets that have been widely used in previous studies. Section V presents a summary of human motion prediction methods and explains their evolution over the years. In Section VI, we describe the evaluation results of the human motion prediction methods with an intensive discussion and present the limitations of the current evaluation. Finally, Section VII and Section VIII provide future research trends and concluding remarks, respectively.

II. HUMAN MOTION PREDICTION

A. PROBLEM DEFINITION

We define the problem of predicting human motion, given an input of the motion sequence, as follows: Let \( X_{0:m} \in \mathbb{R}^{j \times k} \) be the input/conditioning motion sequence \( X_{0:m} = [x_0, \ldots, x_{m-1}, x_m] \) over time step \( m \), with \( j \) joints, with each containing \( k \) dimensions. Optional input modification by concatenating a one-hot vector is possible, thereby explicitly forming the model of the action class of the input motion sequence. A prediction task involves a network \( G_\theta : X_{0:m} \rightarrow \hat{Y}_{m+1} \) parameterized by \( \theta \), where it receives \( X_{0:m} \) to predict \( \hat{Y}_{m+1} \) where \( \hat{Y}_{m+1} \in \mathbb{R}^{j \times k} \); \( \hat{Y}_{m+1} = [\hat{y}_{m+1}, \ldots, \hat{y}_{n-1}, \hat{y}_n] \) and \( \hat{y}_{m+1} \) is the next pose after \( x_m \), unless mentioned otherwise. \( G_\theta \) comprises an encoder network \( E_\theta : X_{0:m} \rightarrow \hat{X}_{0:m} \) and a decoder network \( D_\theta : \hat{X}_{0:m} \rightarrow \hat{Y}_{m+1} \) to transform the 3D body joint \( x_m \) into latent features \( \hat{x}_{m} \), and vice versa. Furthermore, \( G_\theta \) employs recurrent network \( R_\theta \) such as GRU [3] and LSTM [2] to deal with the temporal aspect of human motion prediction [6]-[13]. During training, a loss function \( L(\hat{Y}_{m+1:n}, \hat{Y}_{m+1:n}) \) evaluates the distance between \( \hat{Y}_{m+1:n} \) and \( Y_{m+1:n} \) where \( Y_{m+1:n} \) is the ground truth 3D body joint.

In other words, the problem of human motion prediction is to predict a plausible continuation of the input motion sequence. The training process involves feeding the network with numerous motion sequences of a consistent length (e.g., four-second motion sequences) and optional action class information (e.g., one-hot vector). A typical preprocessing step is to divide the motion sequence \( S \) containing multiple poses \( s \), where the first half acts as an input \( X_{0:m} := \{ s_i \}_{i=1}^{n=2m} \) and the latter half become the target \( Y_{m+1:n} := \{ s_i \}_{i=1}^{n=4m} \). The training process for a human motion prediction method is either a supervised setting (i.e., using an action label as the input) or a self-supervised setting (i.e., without an action label as the input).

B. CHALLENGES

Recent studies have successfully leveraged deep neural networks for human motion prediction; however, they are still far from being put to real-world use. Furthermore, accurately modeling human body dynamics is challenging because there are many intricacies within each human motion. The actual human motion has many variables to determine the incoming action, which comes from both inside (e.g., intention and tendencies) and outside (for example, navigating obstacles and evading dangers). Humans face no challenge in this prediction task and perform it effortlessly most of the time. However, unlike humans, machines are not innately aware of the physical constraints of humans and have difficulties understanding the context of human motion. We discuss some challenges in designing a deep neural network for human motion prediction.

1) Multimodal Nature of Human Motion

To predict future motion, a conditioning frame or sequence is necessary. Theoretically, it is possible to perform predictions with a few frames or one frame as a condition. However, the neural network model would have difficulty understanding the context, causing contextually unrelated human motion prediction under this setting. Hence, most methods use a conditioning frame (varies between methods, but at least ten frames). This becomes an even bigger problem in the evaluation process as it becomes impossible to precisely determine the accuracy of a prediction. Because of the multimodal nature of humans, there is a plethora of possible subsequent poses to which a human can conform for a given human pose. However, the most effective method for evaluating human motion is simply to calculate the distance between the predicted motion and the ground truth data, which is the continuation of the conditioning sequence.

2) Unnatural Artifacts

Within a neural network, the human body is regarded as a group of joints, in which each joint has a set of values that determine its location. The neural network predicts the future value of each joint at each time step belonging to a human pose that continues the last input pose. However, realistic human motion prediction is more than just predicting a contextually valid sequence of poses. Some defects exist in computer-generated human motion, such as foot skating and invalid body pose, resulting in artifacts that degrade the prediction quality. These artifacts are particularly challenging for neural networks because they are not aware of the physical constraints of humans and regard the human body as a set of values. Understanding details from a spatiotemporal representation of data requires specialized networks such as recurrent and generative models. Over the years, these methods have become increasingly complex. Here, we provide a brief explanation of each network type that has been utilized for human motion prediction.
TABLE 1. A list of human-motion prediction methods, negative log likelihood (NLL), mean angle error (MAE), mean squared error (MSE), mean per joint position error (MPJPE), mean average error (MAE), KL divergence (KL), exponential map (EM), joint location (JD), joint displacement (JD), lie algebra (LA), 4D-quaternions (4D – Q), ± marker location (ML), ✓ represents a network consisting of more than one component and ✓ represents a network that uses one component only.

| Proposed by            | Datasets | LSTM | GRU Transf. | TCN | GCN | E-D | C/VAE | GAN | Loss | In  | Out   | Joint Type |
|------------------------|----------|------|-------------|-----|-----|-----|-------|-----|------|-----|------|----------|
| Pragkiladaki (’15) [1] | [14]     | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | NLL, L2 | -   | -    | EM       |
| Jain (’16) [15]        | [14]     | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | L2   | 50f | 100f | ST-Graph |
| Bützepage (’17) [16]   | [14, 17] | x    | x           | x   | x   | x   | x     | x   | L1   | 40f | 100f | 3D-L     |
| Ghosh (’17) [18]       | [14, 17, 19] | x    | x           | x   | x   | x   | x     | x   | L2   | 50f | 30f  | EM       |
| Martinez (’17) [6]     | [14]     | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | L2   | 50f | 10f  | EM       |
| Zhou (’18) [20]        | [17]     | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | L2   | 10f | -    | JD       |
| Gui (’18) [7]          | [14]     | x    | ✓           | ✓   | x   | ✓   | ✓     | ✓   | Ladv(short) + Ladv(long) | 50f | 10f, 25f | EM       |
| Li (’18) [21]          | [14, 17] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | Ladv, Lproj, Lbone | 10f | 30f  | 3D-L     |
| Barsoum (’18) [8]      | [14, 22] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | Ladv, Lproj, Lbone | 10f | 30f  | 3D-L     |
| Chiu (’19) [23]        | [14, 24] | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | MAE, L2 | 50f | 25f  | EM       |
| Gopakrishnan (’19) [9] | [14]     | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | MSE  | 50f | 100f | EM       |
| Mao (’19) [25]         | [14, 17, 26] | x    | ✓           | ✓   | ✓   | x   | ✓     | ✓   | L1, MPJPE | 1s  | 1s   | EM       |
| Liu (’19) [10]         | [14, 27] | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | tenjerrynajia + L2 | 50f | 10f, 100f | LA       |
| Ruiz (’19) [28]        | [14]     | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | Ladv, Ljoint, Lbone, L2 | 50f | 50f  | 3D-L     |
| Aksan (’19) [29]       | [14, 30] | -    | -           | -   | -   | -   | -     | -   | MSE, MPJPE | 2s  | 400ms | -        |
| Zhang (’19) [31]       | [14, 24, 32] | x    | x           | x   | x   | ✓   | ✓     | ✓   | Ladv, L2D | 13f | 25f  | SMPL     |
| Alahbaphari et al. (’19) [11] | [14, 17] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | L2, MSE, KL | 16f | 60f  | 4D-Q     |
| Cui (’20) [33]         | [14, 17, 26] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | MSE, Ladv, Ljoint, Lbone | -   | 10f, 25f | 3D-L & EM |
| Mao (’20) [34]         | [14, 30, 26] | x    | x           | x   | x   | ✓   | ✓     | ✓   | MAE, MPJPE | 10f | 10f, 25f | 3D-L & EM |
| Cui (’20) [35]         | [14, 17] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | L1, L2 | 10f | 25f  | 3D-L & EM |
| Lebally (’20) [36]     | [14, 17] | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | MPJPE | 10f | 15f  | 3D-L     |
| Li (’20) [37]          | [14]     | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | MAE  | -   | -    | 3D-L     |
| Yuan (’20) [12]        | [14, 38] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | Ldivergentprior, KL | 1s  | 4s   | 3D-L     |
| Wang (’21) [39]        | [17, 19, 40] | ✓    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | MSE  | 20f | 20f  | 3D-L     |
| Cui (’21) [13]         | [14, 17, 26] | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | Ladv, Ljoint, L2 | 1s  | 3s   | 3D-L     |
| Zhang (’21) [41]       | [14]     | x    | ✓           | ✓   | ✓   | ✓   | ✓     | ✓   | Ladv, Ljoint, L2 | 1s  | 1s   | ML       |

III. NETWORK BACKBONES
Understanding details from a spatiotemporal representation of data requires specialized networks such as recurrent and generative models. Over the years, these methods have become increasingly complex. Here, we provide a brief explanation of each network type that has been utilized for human motion prediction.

A. RECURRENT NETWORK
A recurrent network is a class of networks with connections between nodes, representing temporal relations within the data. Most recurrent network designs apply this principle by reusing a specific value from the previous time-step calculation (e.g., Figure 1(a) and (b)). An example is the vanilla RNN hidden state [42]. LSTM [2] was then proposed to alleviate the inability of the RNN to accurately remember the past and the vanishing/exploding gradient issue with RNNs. LSTM introduces three distinct gate types that control the hidden-state value and computation flow. A similar method, called GRU [3], was proposed, which uses two gate types. Despite the lack of one gate type, a further experiment showed that the performance of GRU is comparable to that of LSTM [43].

Machine translation [44], speech recognition [45], and video captioning [46] have been accomplished using a recurrent-based network. The ability of a recurrent network to exploit past information is beneficial for these tasks because they deal with sequential data. Hence, it is natural that human motion prediction networks often use recurrent-based networks. Table 1 shows that most motion prediction methods use a subset of the recurrent network. For human motion prediction, recurrent networks are often used to process a feature in the current time step for the future or the next time step [1, 6]–[13], [15], [18], [20], [23], [37], [39], [41].

B. CONVOLUTION NETWORK
Convolution-based networks were originally designed to handle image-like data with large spatial dimensions [47]. A...
convolution-based network consists of a convolutional layer that works by thoroughly applying a filter throughout the spatial dimension in a feed-forward manner. This characteristic makes the convolutional layer and its byproduct, the convolution-based network, a suitable building block of most neural networks designed for image-related tasks. However, it has limitations when dealing with temporal data.

An extension of the convolution network is available for dealing with video-based data [48]. However, a different approach is required to model the motion sequence, which has a higher temporal dependency and less spatial information than the videos. For example, Bütepage et al. [16] experimented with different encoder types to encode a human pose and found that the best encoder considers different time scales. Furthermore, Li et al. [21] replicated the ability of the recurrent network to model the temporal context by using a convolution network with different input scales. Mao et al. [34] employed an attention mechanism to aggregate motion history into future motion prediction (i.e., Figure 1(c)).

C. GENERATIVE NETWORK

A generative network generates new plausible data from the distribution of training data. Examples of the networks built for this purpose include variational autoencoders (VAE) [49] and generative adversarial networks (GAN) [50]. Proposed as an extension of autoencoders, the VAE is capable of both reconstruction and generation. The training process forces the VAE to have a latent space in the form of a probabilistic distribution that encapsulates the training-data distribution. The GAN design consists of a generator and discriminator networks. The generator generates accurate yet fake data, whereas the discriminator spots the fake from actual data. Consequently, GAN networks can generate impressive results in a broad range of tasks, while also being difficult to train.

Within the field of human motion prediction, GANs and VAEs are used differently (see subsection sub:Stochastic Prediction). VAEs are often used in conjunction with a recurrent network for diverse predictions [11], [12], [41]. This network combination excels by keeping track of temporal information, while being able to generate diverse motions by leveraging the VAE latent space. Conversely, GAN is more versatile because the discriminator can be used to measure the prediction accuracy [7], [8], [28], [33] for both deterministic and stochastic networks or to measure the rate of convergence [8] (Figure 1(d)).

IV. HUMAN BODY REPRESENTATION AND DATASET

A. HUMAN BODY REPRESENTATIONS

The human motion prediction literature includes some terminologies regarding the 3D body, which are shared between related topics. For example, a human body/pose describes the entirety of a joint conforming to a single human body or pose in each time step. Each pose consists of several connected joints, forming a model that resembles a human skeleton. The word joint describes the location or angle of a specific joint within the human skeleton. There is no definite number of joints in a pose because it varies depending on the motion sequence dataset used.

There are a few variations in the joint representation used in many human-motion prediction networks. Typically, a joint representation within a pose is either angle- or location-
based. The angle-based/joint angle refers to an axis angle representation of the joints, or more specifically, the exponential map (Figure 2(a)). Meanwhile, the location-based joint representation defines the joint location in 3D Cartesian coordinates (Figure 2(d)). Information regarding the joint representation used in the literature is presented in Table 1. From Table 1, we can infer that the angle-based representation is the most used, followed by the location-based representation.

B. DATASET
The self-supervised nature of the deep human motion prediction method enables them to use an action sequence dataset that contains some form of 3D body annotation. Most mo-cap datasets satisfy the requirements necessary for training a deep human motion prediction network; therefore, they are widely used. This section describes widely used datasets in the literature, along with their details and standard preprocessing configurations according to each dataset.

Human3.6M [14]: is an indoor dataset comprising RGB videos and the 2D and 3D body annotations of seven actors performing a range of actions. It was captured using 15 sensors (four digital video cameras, one time-of-flight sensor, and ten motion cameras), which were combined using hardware and software synchronization. This setup allows them to provide multiple data such as pose data, time-of-flight data, scanner data, and RGB video (sampled at 50 Hz).

At the time of writing this paper, the Human3.6M dataset was the most widely used benchmark dataset for motion prediction. Most researchers divided the actions within the dataset into eight easy and seven difficult actions. The standard experimental setup for this dataset is to downsample the motion sequence two times (to 25 Hz), train on six subjects, and test on five subjects.

CMU Mocap [17]: is a motion capture database that is captured using a system consisting of 12 infrared MX-40 cameras. It contains 2235 recordings from 144 different subjects performing various complex movements. With their recording setup, they could capture 120 Hz data on some actions. Each mo-cap actor wears a black suit with markers on the body (with 41 mo-cap markers). The CMU dataset can be used as an action classification dataset for multiple tasks [16]. For human-motion prediction tasks, some researchers included tasks performed by only one actor [11], [13], [21], [25], [33], [35]–[37].

NTU RGB+D [22]: is an action-recognition dataset captured using three Microsoft Kinect V2 cameras. The dataset contains multiple annotations such as RGB videos, depth map sequences, 3D skeletal data, and infrared videos. It contains 60 action classes (40 daily actions, nine health-related actions, and 11 mutual actions) performed by 40 distinct subjects.

Holden [19] et al.: combined multiple datasets [17], [51], [52] for their training process and added data from their internal captures. The combined dataset underwent multiple postprocessing steps [19]. The resulting dataset contains approximately twice as much data as the CMU [17] sampled at 120 Hz. Methods using this dataset mostly downsample the sequences four times (to 30 Hz) to unify the sampling rate across all sequences.

Penn Action [24]: is an action recognition dataset comprising 2326 video sequences of people performing 15 different actions. It also provides annotations in the form of a coarse viewpoint, human body joints, and 2D bounding boxes. The study annotated 13 joints for each human subject in the form of 2D locations and visibility.

HumanEva-I [38]: is a dataset built for human motion and pose estimation in videos. The 3D body pose annotations in this dataset are acquired using a mo-cap system. The dataset also contains videos of four subjects performing six action types, captured from multiple angles (three colored videos and four grayscale videos) and sampled at 60 Hz.

Action3D [40]: is another dataset that was captured using a single Kinect camera. This dataset provides annotations attainable with a Kinect camera, such as RGB videos (sampled at 30 Hz), depth maps, and 3D skeleton joint locations. The recording consisted of ten subjects who performed actions in ten action categories.

AMASS [30]: is a collection of standardized mo-cap databases. The study proposed a new method, MoSh++ [53], for converting different mo-cap data into realistic 3D human meshes. These realistic 3D human meshes are SMPL+H [54] and SMPL+X [55] parameters. SMPL+H is a parametric 3D body model that consists of a hand pose and an articulated
body pose. SMPL-X is an extension of this work, which enables not only the hand and body but also the facial expression to be modified. The current version of this dataset provides both SMPL+H and SMPL-X annotation for all of the different mo-cap data. As a dataset that unifies other mo-cap databases, the combined dataset contains 2992.34 minutes (a little less than 5 h) of motion sequences associated with 484 subjects, with more than 11,000 motions (not unique) from all databases.

3DPW [26]: is an in-the-wild dataset with accurate 3D poses for evaluation. Additionally, this is the first dataset containing video footage captured from a mobile phone camera. Furthermore, this dataset addresses the lack of a large-recording-volume outdoor dataset. Finally, this dataset provides multiple annotations, such as 2D pose, 3D pose, camera pose (for every frame), 3D body scan Researchers have tackled motion prediction problems for many years. Early motion prediction methods were straightforward. Most of them proposed a simple architecture with common distance-based loss. However, further experiments revealed that a simple recurrent or convolutional network is insufficient for motion prediction. Furthermore, existing issues such as the multimodal nature of the future, error accumulation, and evaluation difficulty raise the need to address the problem in multiple approaches. Therefore, many motion prediction methods have been proposed. s, and 3D people models (18 variations with different clothing).

V. PREDICTION METHODS

Researchers have tackled motion prediction problems for many years. Early motion prediction methods were straightforward. Most of them proposed a simple architecture with common distance-based loss. However, further experiments revealed that a simple recurrent or convolutional network is insufficient for motion prediction. Furthermore, existing issues such as the multimodal nature of the future, error accumulation, and evaluation difficulty raise the need to address the problem in multiple approaches. Therefore, many motion prediction methods have been proposed.

We propose a taxonomy for human motion prediction methods. First, we observed that researchers often share terminologies to describe certain types of work, but no one has suggested concrete terminology beforehand. Therefore, we used similar, if not identical, words to describe motion prediction methods. Second, we split motion prediction methods into three categories: deterministic, structure-aware, and stochastic predictions. Deterministic predictions include studies that treat human joints impartially. Structure-aware prediction includes methods that use structural information of the human joint. Methods belonging to this group are designed with the joint and its neighbor in mind, usually represented in the form of an adjacency matrix. This design renders structurally aware networks more robust, enabling them to perform without a recurrent-based network module. Hence, we further split the structure-aware section into two, consisting of networks that do not use a recurrent-based network and networks that do. The stochastic prediction category includes methods that generate multiple samples from a single-motion sequence input. The final category contains a few methods that have unique characteristics, such as using 2D images as input [31] or dealing with incomplete input sequences [13], [28]. The following passage summarizes human motion prediction methods based on their classifications.

A. DETERMINISTIC PREDICTION

Fragkiadaki et al. [1] proposed the encoder recurrent decoder (ERD) network as the first deep learning-based human motion prediction method. ERD can be defined as $G_{\theta} = D_{\theta}(R_{\theta}(E_{\theta}(\cdot)))$ where $E_{\theta}$ is a multiconvolutional-layer encoder inspired by [56], $R_{\theta}$ is an LSTM layers, and $D_{\theta}$ is a multiple dense layer. The design paradigm for ERD involves evaluating $Y_t = G_{\theta}(X_{t-1})$ for every frame $t$ and using the ground truth pose $X_t$ instead of $Y_t$ during the next time step $t + 1$ in the training process. In addition to ERD, they proposed a three layered LSTM (LSTM-3LR) with dense layer for $E_{\theta}$ and $D_{\theta}$.

The quantitative results indicate that LSTM-3LR is better than ERD for human motion prediction. However, when predicting times longer than 560 ms, LSTM-3LR suffers from a mean pose convergence problem (ERD does not have this issue) [1]. Methods plagued with this issue will predict a simple pose where the spatial variations of human motion disappear. This contradicts how humans behave because normal human motion contains unique perturbations, which we consider natural motion. Additionally, early human motion prediction works [1], [15] had motion discontinuity issues between the last conditioning pose and the first predicted pose.

However, Ghosh et al. [18] argued that modeling spatiotemporal information is possible without manually designed and action-specific graphs [15]. They proposed a DAE-LSTM method to implicitly model the structural and temporal information of human motion. They employed a dropout auto-encoder (DAE), which was inspired by denoising auto-encoders, to remove the error accumulated from the recurrent network. Regarding motion prediction, they used the LSTM-3LR, as in previous studies [1], [15]. The quantitative results showed that their dropout autoencoder module successfully reduced errors for long-term prediction.

Martinez et al. [6] proposed architecture and design changes to address the issues raised in previous studies [1], [15]. The first design change addresses the frame discontinuity issue using a residual architecture that models the displacement of the joints instead of the actual value. Second, in previous studies [1], [15] ground-truth data $Y_t$ was used as the input for the network during the training process. This design leads to deteriorating prediction quality after a few seconds of prediction [20]. It is also known that using ground-truth information during training in reinforcement learning [57] and recurrent networks [58] causes issues, as the network is denied the opportunity to learn from its mistakes, mainly...
from feeding the ground-truth data as input during training. Martinez et al. [6] proposed a paradigm change by using $Y_t$ during training, mimicking test time behavior, as it prevents the deteriorating quality issues. Their architecture is defined as $G\theta = D\theta(R\theta(\cdot))$, where $R\theta$ is a GRU network with 1024 units [3] and it uses a dense layer for $D\theta$. Departing from previous network designs, their design excluded $E\theta$, as they discovered that the network produced better results without it.

Martinez et al. [6] also discovered a simple zero-velocity baseline. The zero-velocity baseline works by constantly replicating the last input frame $X_m$; it outperformed prior works by a significant margin [1], [15], establishing a strong baseline for future methods. Gui et al. [7] proposed an adversarial geometry-aware encoder (AGED) network that outperformed the zero-velocity baseline. AGED uses two distinct discriminators comprising GRU cells [59]: the continuity and fidelity discriminators, which are based on the adversarial training principle [50]. The discriminators evaluated long and short sequences, respectively. Furthermore, a novel geodesic loss provides the network with a more meaningful loss by evaluating the geodesic distance between two rotations in the Riemannian manifold structure [60], in contrast to the Euclidean distance.

However, previous works faced difficulties in long-term predictions. This issue is exacerbated by the lack of a useful evaluation function, as the existing MSE-based evaluation function is only useful for evaluating short-term prediction, leaving long-term prediction results to be evaluated qualitatively. Motivated by this issue, Gopalakrishnan et al. [9] proposed a new architecture, the verso-time label noise RNN (VTLN-RNN), to tackle both short- and long-term prediction problems simultaneously alongside a more competent long-term evaluation metric, which is the normalized power spectrum similarity (NPSS). The VTLN-RNN network performs a two-level computation. The top level receives the conditioning sequence and action label and runs backward in time to generate "guide vectors." The lower level uses the "guide vectors" for motion prediction at each time step. Their novel loss, the NPSS, compares motion sequences based on their power spectrum obtained using a discrete Fourier transform. This leads to a more accurate error value as tolerable deviations (e.g., misaligned joint angles, adding extra frame, or removing frame during a motion compared to ground truth) will not be punished as severely as when using MSE.

Previous studies were limited by their reliance on action class information during the prediction process [1], [7], [15], [16], [20]. Methods with this characteristic are unrealistic because action labels are not available in real-world scenarios. Inspired by the ability of the hierarchical multiscale recurrent neural network (HM-RNN) [61] to model high-and low-level abstractions, Chiu et al. [23] proposed a similar architecture called triangular-prism RNN (TP-RNN), which does not require an action label. It contains multilevel LSTM, which consists of an increasing number of LSTM sequences as the level grows, where each sequence takes turns to process the information at lower levels. This design allows the network to encode different hierarchies in human dynamics at different time scales, thereby learning a robust motion context that spans over different time scales, allowing the network to predict without action-class information.

B. STRUCTURE-AWARE PREDICTION

It is possible to leverage spatial information of the human body for human motion prediction. Methods in this category are encouraged to learn the implicit relationship between each joint in the human body during motion (Figure 3), usually in the form of a parameterized graph.

1) Feed-Forward Based

Despite its success when dealing with periodic data, early deterministic prediction methods using LSTM networks [2] did not perform as well on aperiodic data. These characteristics, along with the need to train a separate model for each human motion, limit their generalization to novel data. Therefore, Bütepage et al. [16] proposed an autoencoder consisting of three different encoder types. One of the encoder types, the hierarchy encoder (connecting each human joint in a tree-like setup), outperformed the other encoder types in most aspects, hinting at the fact that the structural prior is beneficial for motion prediction. Quantitative results show that their network performed much better on complex actions compared to previous recurrent-based networks without incorporating a recurrent unit into their architecture.

Supporting similar arguments, Li et al. [21] proposed a convolution seq2seq network to model both temporal and spatial information from past motions. The network is divided into two parts to obtain long- and short-term hidden variables, containing long- and short-term encoders respectively. Following that, the long- and short-term hidden variables are concatenated to produce future motion. In contrast to LSTM [2] networks, their concept can preserve long-term information indefinitely because it is unaffected throughout the prediction process. Their method also allows for indefinite future predictions with the sliding window attribute of the short-term encoder.

Both Bütepage et al.’s [16] and li et al.’s [21] proposed network-applied convolutions across time on input poses. This design has a drawback in that the temporal dependencies of the network strongly depend on the convolutional filters of the network. Therefore, Mao et al. [25] proposed a
feed-forward deep network architecture using a GCN [62], [63] and automatically learned graph connectivity. Their proposed method adopts a trajectory representation based on the discrete cosine transform (DCT), inspired by [64], instead of using a pose representation like that in previous studies [16] [21]. Using DCT, they obtained a compact representation of the human motion. Their experimental results also supported another of their claim that predicting in 3D space yields better results compared to predicting in angle space. They supported their claim by further claiming that the angle representation has ambiguities because two completely different sets of angles can share the same 3D pose.

Similar to Mao et al., Lebailly et al. [36] proposed a novel human motion prediction network using GCN. The temporal inception module (TIM), which was inspired by the inception module [65], can encode the input pose sequence at different temporal scales simultaneously. The TIM module contains multiple kernels of different sizes to obtain information from different scales. Based on their experiment, a trade-off was observed because a TIM with a longer subsequence will perform better at long-term prediction and vice versa.

Cui et al. [33] suggested that a single learnable unrestricted graph is equivalent to implicitly modeling pose information. Therefore, they proposed a dynamic relationship GCN (drGCN) that uses two distinct learnable adjacency graphs. One is a connective graph that learns the weights between neighboring joints. The other is a global graph that learns the implicit connections between all the joints in the human body.

Aksan et al. [29] proposed the structured prediction layer (SPL) to decompose the human joint prediction layer into a hierarchy, following the kinematic chain of the human skeleton. SPL creates many small networks that enable the model to learn dedicated representations per joint. As a result, the gradient on the hidden layers is affected only by the gradient from each joint hierarchy connected to them instead of the entire dense layer. Furthermore, they analyzed state-of-the-art methods and found that the benchmark using the H3.6M dataset became saturated. Methods that performed well on H3.6M did not scale well on a larger and more diverse dataset. Accordingly, they evaluated their method with AMASS on top of H3.6M and found that their method was more impactful on a large dataset.

2) Recurrent Based

Jain et al. [15] proposed the structural-RNN (S-RNN) that combines an LSTM network [2] and spatiotemporal graphs. Spatiotemporal graphs have been used to model high-level spatiotemporal structures [66]–[72]. The combined network (S-RNN) is capable of both modeling long sequences and capturing the spatiotemporal structure. Jain et al. also provides a converted version of the H3.6M dataset into an exponential map, which they used in their prediction.

Liu et al. [10] argued that LSTM and GRU face difficulties in long-term motion prediction because they often fail to capture the motion context and the long-term predicted pose degrades into a motionless state. They added that prior algorithms do not respect the physical constraints that govern human motion. Therefore, they proposed hierarchical motion recurrent (HMR), a hierarchical autoencoder structure with recurrent networks. The encoder is designed to encode the input sequence hierarchically using an update process similar to that used by a recurrent network cell state to update its value with the neighboring cell state. This encoding process runs \( n \) times, enriching the information within a cell state because it repeatedly shares information with other cell states. The decoder uses the final encoded motion context to recursively output the future motion context. During the training process, the network is guided by a loss function that penalizes the network according to the bone length (a longer bone means more loss). HMR outperforms previous works in both long- and short-term predictions, demonstrating its capability in modeling motion context.

Inspired by [19], Wang et al. [39] developed STRNN using the same principle. The network encodes the entire input one by one before splitting it into two branches (input decoding and motion prediction). To capture the spatial features better, they employed a spatial encoder that encoded the DoFs of each body part.

C. STOCHASTIC PREDICTION

One solution for predicting the multimodal nature of human motion is to predict multiple plausible motions from a single input. Barsoum et al. proposed a WGAN-GP [73]-based network called HP-GAN [8], as one of the first probabilistic prediction methods. They claimed that using MSE in a non-deterministic setup causes the model to average between two possible futures. Aware of the difficulty in training GANs, they used a loss based on skeleton physics alongside GAN loss to stabilize and improve the training process. A critic network was used to quantify the prediction quality, although it was not sufficient to determine whether the training had converged.

Maintaining a balance between accuracy and diversity is important for stochastic human motion predictions. Allakbarian et al. [11] addressed this problem by proposing mix-and-match perturbations for imposing diversity on a stochastic network. They claimed that rather than combining a noise vector with the conditioning variables in a deterministic manner, a better result can be achieved by randomly selecting and perturbing a subset of these variables.

Although much care is being taken when designing a network to model multimodal data, the problem of producing diverse sampling from networks is underexplored. Yuan et al. proposed DLow [12], a novel sampling method that produces diverse sampling results while using a pretrained network to address this issue. They proposed a mapping function that subdivides the latent space into chunks using a set of variables. These variables were obtained from a network trained using the likelihood and diversity losses. Their methods produced better quantitative results than previous methods.
state-of-the-art motion prediction methods, and produces competitive quantitative results that outperform observations while also predicting future human motion. This graph convolutional network (MT-GCN), a novel multitask imperfections in mind. Cui with most methods because they are not built with these imperfections in mind. Cui

or other objects. These types of data do not work well in certain respects. Hence, performance metrics remain a tool to assess the performance of deep human motion prediction with respect to other studies. motion prediction method.

A. PERFORMANCE METRICS
Evaluating human motion prediction methods is not a straightforward process because of the lack of a definite method for ascertaining the quality of the produced motion. The quality of the predicted motion itself comprises many aspects such as plausibility, accuracy, and diversity, for stochastic methods. Currently, no evaluation method can precisely measure the extent to which each method excels in certain respects. Hence, performance metrics remain a tool to assess the performance of deep human motion prediction with respect to other studies.

The deepest motion prediction method is compared with a distance-based metric, such as the Euclidean distance ($L_2$). The common setting for this evaluation is to calculate the distance between the predicted and ground truth 3D joint locations using the mean per joint position error (MPJPE) [14] or the Euler angle with the mean angle error (MANe). Because most methods use the exponential map representation within their network, conversion to the Euler angle is necessary for evaluation.

The methods report their quantitative evaluation results in the form of tables. A typical comparison procedure with the most frequently used dataset for comparison, Human3.6M [14], is to report the MANe result for each or some motion category over some timestamps. Table 3 and Table 2 show the evaluation scores compiled using multiple methods. Each column shows the MANe result for a specific timestamp (shown in the top column in milliseconds). Columns adjacent to each other belong to the same action category (denoted by the topmost column).

Tables 2 and Tables 3 show the MANe results from many methods; however, they are organized according to the comparison made by the respective authors of each method. Key differences between Tables 2 and 3 are the three baseline evaluation results, that is, the ERD, LSTM-3LR [1], and S-RNN [15]. Table 3 reports the results obtained from their study for these three methods. Table 3 is obtained from the study by Martinez et al. [6] because they reevaluated these three methods and produced a different score. Afterwards,
future work will compare their method to different baseline scores, which creates this diverging comparison. However, some methods compare both, which leads them to exist in both Tables 2 and Table 3.

However, simply evaluating the distance between the angles or joint locations is unreliable and often lacking. For example, angle-distance evaluation is inconsistent because of the ambiguity of angle-based representation, where two quite different angles can be attributed to the same pose. This leads to a discrepancy in the error value between the angle- and joint location distance. Zero Velocity [6] could highlight large deviations, as the motion representation of this is that it is unable to evaluate finer details of motion.

A method with diverse predictions faces a significant evaluation challenge. A diverse prediction method is likely to yield a lower $L_2$ score than a deterministic method. However, this does not mean that the produced motion is inaccurate or implausible [28]. Another study used a classifier network, trained as a discriminator to differentiate the ground truth data and generated data, to ascertain the predicted motion quality [11]. Finally, a couple of modified $L_2$ metrics are used for evaluation [12], [41], such as the average pairwise distance (APD) to measure diversity [11], average displacement error (ADE), final displacement error (FDE) to measure accuracy [78]–[80], and the multimodal versions of ADE and FDE, i.e., MMADE and MMFDE [81].

### B. LACK OF INFORMATIVE METRIC

As mentioned above, the evaluation of human motion prediction methods is challenging. Most researchers have used the L2 norm (Euclidean distance) to evaluate their prediction results. However, high L2 norm scores do not always translate into accurate predictions or good models. Hence, a new evaluation metric is required for long-term predictions where deviations from the ground truth are inevitable.

Some researchers have attempted to create new metrics for evaluating long-term predictions. Ghosh et al. [18] proposed the use of an action classifier network to evaluate their predictions. It can highlight large deviations, as the motion will not be classified correctly when it occurs. A disadvantage of this is that it is unable to evaluate finer details of motion.
VII. FUTURE TRENDS

Based on the survey study on the state-of-the-art human motion prediction, we present a few possible future trends.

A. EXTENDED MOTION DATASET

A human motion prediction method gains the ability to predict by learning from a lot of motion sequences. Currently, the most used dataset consists of common daily human motions such as walking, eating, and discussion as in Human3.6M dataset [14]. However, a few action types cannot represent diverse human motion in real life. For example, human motion prediction to predict a particular sport or dancing action requires a large scale motion sequence dataset with such action types, which is infeasible in many cases.

In this context, it would be a promising approach to automatically generate large scale datasets with diverse action types using existing computer vision techniques and public domain videos such as YouTube, Dailymotion, and Vimeo. In addition, developing a data augmentation technique for a human motion dataset might be a useful technique to employ.

B. SEMANTIC EVALUATION METRICS

We observe that most methods evaluate using distance-based metrics. Controlled deterministic prediction methods are evaluated using MAnE or MPJPE, while diverse prediction methods are evaluated using a set of modified distance metrics. However, simply calculating the distance between predicted and ground-truth joints is not an ideal way to measure the suitability of the predicted human motion. Note that distance-based metrics ignore the semantic nature of human motion which becomes more important unless we predict very short sequences. Therefore, further research should be conducted to account for more informative aspects or semantic correctness of the predicted motion.

C. PREDICTING CHALLENGING MOTION TYPES

It is observed in Table 2 that the discussion and smoking actions are more difficult to predict than others. This shows that existing methods are still struggling to perform well on these actions because they are more aperiodic compared to walking action. In addition, most methods are developed to outperform for all action types and there is no method that aims to specifically handle various aperiodic motions.

In the future research, general and complex motions should be the target to be predicted. To this end, more rigorous analysis of human motion is necessary and the designed predictor is desired to be more customized to particular motion domains (e.g., soccer, dancing, etc).

D. FULL 3D HUMAN MESH PREDICTION

Existing methods commonly utilize the skeleton-type joints for representing the human body. However, as mentioned by Zhang et al. [41], the skeleton-type joint structure leaves out many important details because it is invisible to the naked eye. For example, the bone length may vary and the predicted joints may not conform to a valid human body. This can be alleviated by predicting the whole 3D human mesh while enforcing the physical limitation of the dynamic human body.

E. CONDITIONING ON A DYNAMIC VARIABLE

Throughout this study, we survey human motion prediction methods that generate determined input motion sequences. We believe that adding a dynamic input variable to control human motion prediction would be beneficial for many applications. For example, in video game applications, it would be challenging but promising to add a dynamic directional input or a dynamic action type variable to simulate a variety of real life scenarios by changing the variable dynamically.

F. APPLICATIONS OF HUMAN MOTION PREDICTION

We observe that most human motion prediction methods are designed to predict the continuation from common daily actions. Conversely, accurately predicting human motion can be beneficial for many applications, including sports broadcasting, video game, surveillance, and safety.

Accurately predicting a future motion of a sports player could reveal a potential outcome of a sport match. Variations of similar prediction in sports games might produce a huge commercial impact on the sport broadcasting industry. Human motion prediction methods can also be extended for surveillance and security purposes by detecting and predicting abnormal actions with potential hazard. In an autonomous driving car system, a camera in the vehicle predicts the future motion of multiple pedestrians and the possibility of a car accident can be reduced dramatically.

VIII. CONCLUSION

In this report, we summarize the recent advances in human motion prediction. We observed that this topic has garnered increasing attention every year and has seen a significant improvement from earlier works. However, human motion prediction remains a challenging problem, and most studies have struggled to produce accurate results in long-term prediction despite excelling in short-term prediction. Dataset availability is also a limitation, as new motion sequence datasets cannot be constructed with ease, and motion prediction methods can only learn the distribution of action available in the dataset. These issues, coupled with the error accumulation and the tendency to converge to the mean pose in long-term prediction, means that we are far from applying motion prediction methods in real-world scenarios. However, we observed an increase in large motion sequence datasets, which could facilitate future research on human motion prediction. Furthermore, the deep learning field has experienced fast and steady growth; therefore, it is reasonable to expect many developments in the future.

ACKNOWLEDGMENT

We would like to thank to Jonathan Samuel Lumentut, Joshua Santoso, Prof. Ju Yong Chang, and Prof. Taehyun Kim for the helpful discussion and constructive comments.
REFERENCES

[1] K. Fragiadakis, S. Levine, P. Felsen, and J. Malik, “Recurrent network models for human dynamics,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., December 2015, pp. 4346–4354.

[2] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[3] K. Cho, B. van Merrienboer, Çaglar Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder–decoder for statistical machine translation,” in Proc. Empir. Meths Natur. Lang. Process., October 2014, pp. 1724–1734.

[4] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orta-Escalon, I. A. Garcia-Rodriguez, and A. A. Argyros, “A review on deep learning techniques for video prediction,” IEEE Trans. Pattern Anal. Mach. Intell., Vol. PP, 2020.

[5] A. Ruddenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, “Human motion trajectory prediction: a survey,” J. Robotics Res., vol. 39, no. 8, pp. 895–935, 2020.

[6] J. Martinec, M. J. Black, and J. Romero, “On human motion prediction using recurrent neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., July 2017, pp. 4674–4683.

[7] 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. Springer, vol. 11208, September 2018, pp. 823–842.

[8] 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. Workshop, June 2018, pp. 1418–1427.

[9] A. Gopalakrishnan, A. Mali, D. Kifer, C. L. Giles, and A. Ororbia, “A neural temporal model for human motion prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2019, pp. 1208–12117.

[10] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Lu, R. Zimmermann, and L. Cheng, “Towards natural and accurate future motion prediction of humans and animals,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2019, pp. 9966–10004.

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

[12] Y. Yuan and K. Kitani, “Dow: Diversifying latent flows for diverse human motion prediction,” in Proc. Eur. Conf. Comput. Vis. Springer, vol. 12354, August 2020, pp. 346–364.

[13] Q. Cui and H. Sun, “Towards accurate 3d human motion prediction from incomplete observations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2021, pp. 4801–4810.

[14] 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, 2014.

[15] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-rnn: Deep learning on spatio-temporal graphs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2019, pp. 6518–6526.

[16] T. Lebailly, S. Kirciroglu, M. Salzmann, P. Fua, and W. Wang, “Motion prediction using temporal inception module,” in Proc. Asian Conf. Comput. Vis., vol. 12623, November 2020, pp. 651–665.

[17] M. Li, S. Chen, Y. Zhao, Y. Zhang, and Q. Tian, “Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2020, pp. 211–220.

[18] L. Sigal, A. O. Balan, and M. J. Black, “Humanane: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,” Int. J. Comput. Vis., vol. 87, no. 1–2, pp. 4–27, 2010.

[19] X. Liu, X. Shen, J. Liu, and N. Magnenat-Thalmann, “Learning progressive joint propagation for human motion prediction,” in Proc. Eur. Conf. Comput. Vis. Springer, vol. 12352, August 2020, pp. 226–242.

[20] M. Marchellus: Deep Learning for 3D Human Motion Prediction: State-of-the-Art and Future Trends

[21] W. Zhang, M. Zhu, and K. G. Derpanis, “From actemes to action: A strongly-supervised representation for detailed action understanding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., December 2013, pp. 2228–2235.

[22] S. Mao, M. Liu, M. Salzmann, and H. Li, “Learning trajectory dependencies for human motion prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., October 2019, pp. 9488–9496.

[23] C. Xu, L. N. Govindarajan, Y. Zhang, and L. Cheng, “Lie-v: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups,” Int. J. Comput. Vis., vol. 123, no. 3, pp. 454–478, 2017.

[24] A. H. Ruiz, J. Gall, and F. Moreno-Noguer, “Human motion prediction via spatio-temporal inpainting,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., October 2019, pp. 7113–7122.

[25] E. Aksan, M. Kaufmann, and O. Hilliges, “Structured prediction helps 3d human motion modeling,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., October 2019, pp. 1712–1721.

[26] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, “AMASS: Archive of motion capture as surface shapes,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., October 2019, pp. 5442–5451.

[27] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik, “Predicting 3d human dynamics from video,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., October 2019, pp. 7113–7122.

[28] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, “Learning 3d human dynamics from video,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2019, pp. 5607–5616.

[29] Q. Cui, H. Sun, and F. Yang, “Learning dynamic relationships for 3d human motion prediction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2020, pp. 6518–6526.

[30] W. Mao, M. Liu, and M. Salzmann, “History repeats itself: Human motion prediction via motion attention,” in Proc. Eur. Conf. Comput. Vis. Springer, vol. 12359, August 2020, pp. 474–489.

[31] Y. Cai, L. Huang, Y. Wang, T. Cham, C. Jai, Y. Tian, L. Xiu, Y. Yang, Y. Zhu, X. Shen, D. Liu, J. Liu, and N. Magnenat-Thalmann, “Learning progressive joint propagation for human motion prediction,” in Proc. Eur. Conf. Comput. Vis. Springer, vol. 12352, August 2020, pp. 226–242.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” CoRR, vol. abs/1412.3555, 2014.

[33] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in Proc. Adv. Neural Inf. Process. Syst., December 2014, pp. 3104–3112.

[34] H. Sak, A. W. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling,” in Proc. Int. Speech Commun. Assoc., September 2014, pp. 338–342.

[35] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, “Video captioning with attention-based lstm and semantic consistency,” IEEE Trans. Multim., vol. 19, no. 9, pp. 2045–2055, 2017.
[47] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, “Object recognition with gradient-based learning,” in Shape, Contour and Group, in Comput. Vis., vol. 1681, 1999, p. 319.

[48] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 4489–4497, 2015.

[49] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” Proc. Int. Conf. on Learning Rep., April 2014.

[50] J. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., December 2014, pp. 2672–2680.

[51] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, “Documentation mocap database hdm05,” Universität Bonn, Tech. Rep. CG-2007-2, June 2007.

[52] F. Offi, R. A. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Berkeley mhid: A comprehensive multimodal human action database,” IEEE Workshop on Appl. of Comput. Vis., pp. 53–60, January 2013.

[53] M. M. Loper, N. Mahmood, and M. J. Black, “MoSh: Motion and shape capture from sparse markers,” ACM Trans. Graph., vol. 33, no. 6, pp. 220:1–220:13, November 2014.

[54] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: A skinned multi-person linear model,” ACM Trans. Graph., vol. 34, no. 6, pp. 248:1–248:16, 2015.

[55] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, “Expressive body state capture: 3d hands, face, and body from a single image,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2019, pp. 10975–10985.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Commun. of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[57] S. Ross, G. J. Gordon, and J. A. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning,” in Proc. Artif. Intell. and Stat., vol. 15, April 2011, pp. 627–635.

[58] S. Bengio, O. Vinyals, N. Jaitly, and N. M. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in Proc. Adv. Neural Inf. Process. Syst., December 2015, pp. 1171–1179.

[59] K. Cho, B. V. Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” in Proc. Syntax, Semantics, and Struct. in Stats. Trans., October 2014.

[60] W. Rossmann, Lie Groups: An Introduction Through Linear Group. Oxford University Press, 2006.

[61] J. Chung, S. Ahn, and Y. Bengio, “Hierarchical multiscale recurrent neural networks,” Proc. Int. Conf. on Learning Rep., April 2017.

[62] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. Int. Conf. on Learning Rep., April 2017.

[63] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio’, and Y. Bengio, “Graph attention networks,” in Proc. Int. Conf. on Learning Rep., April 2018.

[64] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, “Nonrigid structure from motion in trajectory space,” in Proc. Adv. Neural Inf. Process. Syst., December 2008, pp. 41–48.

[65] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 1–9, June 2015.

[66] Y. Li and R. Nevatia, “Key object driven multi-category object recognition, localization and tracking using spatio-temporal context,” in Proc. Eur. Conf. Comput. Vis. Springer, vol. 5305, October 2008, pp. 409–422.

[67] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, “Track to the future: Spatio-temporal video segmentation with long-range motion cues,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2011, pp. 3369–3376.

[68] W. Brendel and S. Todorovic, “Learning spatiotemporal graphs of human activities,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., November 2011, pp. 778–785.

[69] B. Douillard, D. Fox, and F. Ramos, “A spatio-temporal probabilistic model for multi-sensor multi-class object recognition,” in Proc. Int. Symp. Robotics Res., vol. 66, November 2007, pp. 123–134.

[70] H. S. Koppula and A. Saxena, “Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation,” in Proc. Int. Conf. Mach. Learn., vol. 28, June 2013, pp. 792–800.

[71] X. Zhang, P. Jiang, and F. Wang, “Overtaking vehicle detection using a spatio-temporal CRF,” in Proc. Intel. Vehicle Symp., June 2014, pp. 338–343.