Neurocomputing 489 (2022) 345–365

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

3D human motion prediction: A survey

Kedi Lyu [a,*], Haipeng Chen [a], Zhenguang Liu [b,*], Beiqi Zhang [c], Ruili Wang [b,*]

[a] Jilin University, Changchun, Jilin, China
[b] Zhejiang Gongshang University, Hangzhou, Zhejiang, China
[c] Sichuan University, Chengdu, Sichuan, China

ARTICLE INFO

Article history:
Received 4 October 2021
Revised 8 February 2022
Accepted 19 February 2022
Available online 11 March 2022

Keywords:
Survey
Human motion prediction
3D human pose representation

ABSTRACT

3D human motion prediction, predicting future poses from a given sequence, is an issue of great significance and challenge in computer vision and machine intelligence, which can help machines in understanding human behaviors. Due to the increasing development and understanding of Deep Neural Networks (DNNs) and the availability of large-scale human motion datasets, the human motion prediction has been remarkably advanced with a surge of interest among academia and industrial community. In this context, a comprehensive survey on 3D human motion prediction is conducted for the purpose of retrospecting and analyzing relevant works from existing released literature. In addition, a pertinent taxonomy is constructed to categorize these existing approaches for 3D human motion prediction. In this survey, relevant methods are categorized into three categories: *human pose representation*, *network structure design*, and *prediction target*. We systematically review all relevant journal and conference papers in the field of human motion prediction since 2015, which are presented in detail based on proposed categorizations in this survey. Furthermore, the outline for the public benchmark datasets, evaluation criteria, and performance comparisons are respectively presented in this paper. The limitations of the state-of-the-art methods are discussed as well, hoping for paving the way for future explorations.

1. Introduction

Comprehending and predicting human motion is crucial in assisting humans and machines to interact with the outside world. Predicting future human motions is an innate ability for humans to interact with other people such as navigating through the crowds, defending against offensive players in a game, or shaking hands with others. Further, it is also important for intelligent machines to respond to human behaviors, coordinate their poses, and project paths during the interactions with humans. However, as a non-talent ability of the machine, applying this ability is a very challenging task. Further more, motion prediction has also been widely applied to autonomous driving [23,24,20], intelligent robot [39,38], human-robot collaboration [51,58,52,68,19,50,55,78], and multimedia applications [64,104], as shown in Fig. 1.

The research of motion prediction has attracted increasing attentions from academia and industry. Due to the advances in deep learning, great progress has been made in motion prediction. However, it is still challenging to predict motions accurately. For instance, humans' intentions are very complex, which act as internal stimuli to drive human to behave differently. Likewise, the surroundings of the physical world can also affect humans' motions actively or passively. In fact, various factors that affect human poses cannot always be intuitively identified or resorted by modeling the context. These indeed increase the intricacy of the issue, but provides diverse perspectives for our investigation.

1.1. Scope

There are many ways to interpret and represent human motions, such as kinematic trees [56,30], joints graph [102,57], video frames [32,82], and moving as a mass point from a starting point to a target point, which all reflect people's different comprehensions of human motion, resulting in diverse manifestations of human motion prediction. These have been inspired by various classifications of prediction methods: 1) 2D Motion trajectory prediction: it mainly predicts motion trajectories for human or moving devices [42,43,97,100,101,84,85] in a 2D plane. 2) Video prediction: it focuses on the motion prediction on the video frames [109,90,69,105]. 3) Poses sequence prediction: it predicts future 3D human motions. 3D human motion prediction is the only scope in this survey. The other two will not be included into the discussion

* Corresponding authors.
E-mail addresses: lvkd19@mails.jlu.edu.cn (K. Lyu), chenhp@jlu.edu.cn (H. Chen), liuzhenguang2008@gmail.com (Z. Liu), beiqizhang126@126.com (B. Zhang), prof.ruili.wang@gmail.com (R. Wang).

https://doi.org/10.1016/j.neucom.2022.02.045
0925-2312/

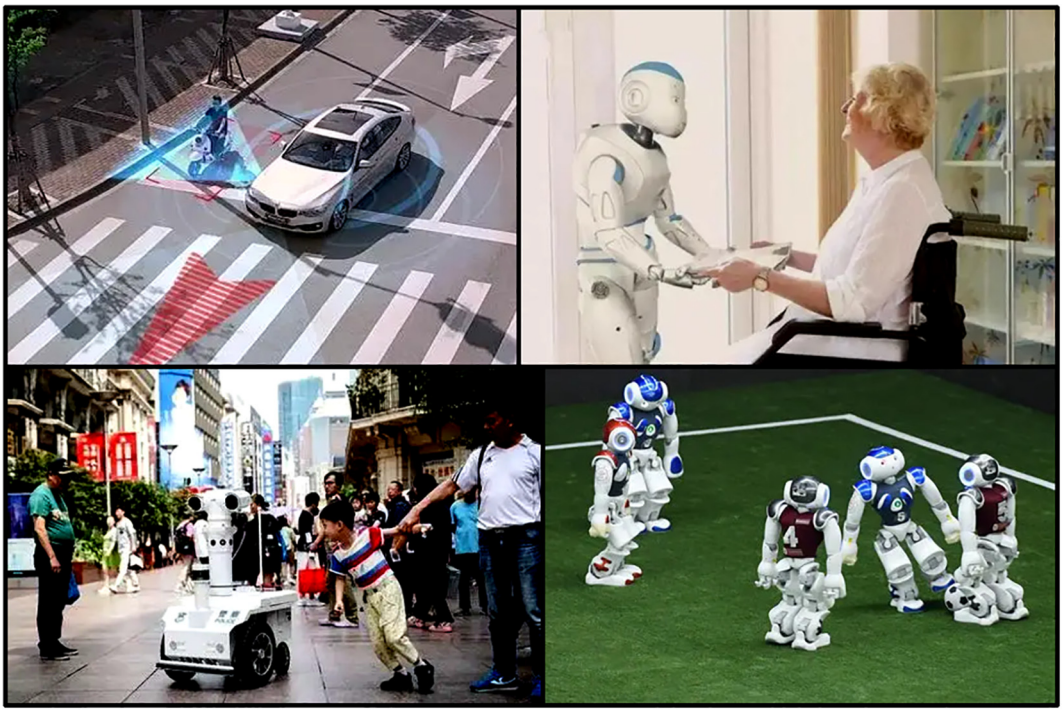


Fig. 1. Applications of 3D human motion prediction. **Top left**: Autonomous vehicles need to promptly judge the intentions and future positions of other traffic participants. **Top right**: Robots assisting humans, such as delivering documents to a human, need to predict human motion and accurately place documents in a given location. **Bottom left**: In densely populated spaces, machines should accurately predict the human motions around them to safely pass through the crowds. **Bottom right**: In the RoboCup, robots need to predict the actions of opponents to perform effective offensive, and defensive behaviors.

scope and the conventional non-deep learning methods will not be taken into consideration here. As 3D human motion prediction is an emerging research field, this survey will embrace as much literature as possible to provide a better understanding of current work and offer reference for further research in this field.

1.2. Problem formulation and challenges

In 3D human motion prediction, it is a traditional method to represent the human pose by a skeletal kinematic tree composed of joints. In anatomy, the human body contains hundreds of joints, but only the joints recorded by a motion capture system will be focused on in this study. As shown in Fig. 2, the human skeleton consists of representative joints. 3D human motion is a sequence constituted by multiple human skeletons arranged in a chronological order. Mathematically, we suppose that a pose sequence of length N is formulated as $X_{1:N} = (x_1, \ldots, x_N) \in R^{J\times N\times D}$ and each $x_i \in R^{J\times D}$ belonging to sequence X, where x_i represents a pose at the i^{th} frame; J is the number of joints; D is the dimension number of each joint; $i \in (1, N)$. The future pose sequence is $X_{1+K:N+K} = (x_{K+1}, \ldots, x_{K+N}) \in R^{J\times K\times D}$, where K denotes the frames of the future sequence. The objective of 3D human motion prediction is to generate future sequences on the basis of the already observed ones. Mathematically, motion prediction aims to construct a model F to forecast $\widehat{X}_{1+K:N+K} = F(X_{1:N})$, where $\widehat{X}_{1+K:N+K}$ is the predicted future motion sequence close to the ground truth $X_{1+K:N+K}$.

The challenge of obtaining accurate and natural sequences of future human poses lies in the complexity of human behavior and the flexibility of human body. As a result, understanding and predicting human motion is not an easy task for both humans and machines. The two key issues that need to be addressed are *Inherent kinematic problems* and *Network performance limitations*. The inherent kinematic problems of human motion prediction are triggered by its highly stochastic nature, high dimension, and non-linearity, which leads to a high degree of uncertainty for future human poses. Network performance limitations result from the inherent networks drawbacks that inevitably involve error accumulation from RNNs [108] and are stuck due to the limitation in processing standard 2D grid representations from CNNs.

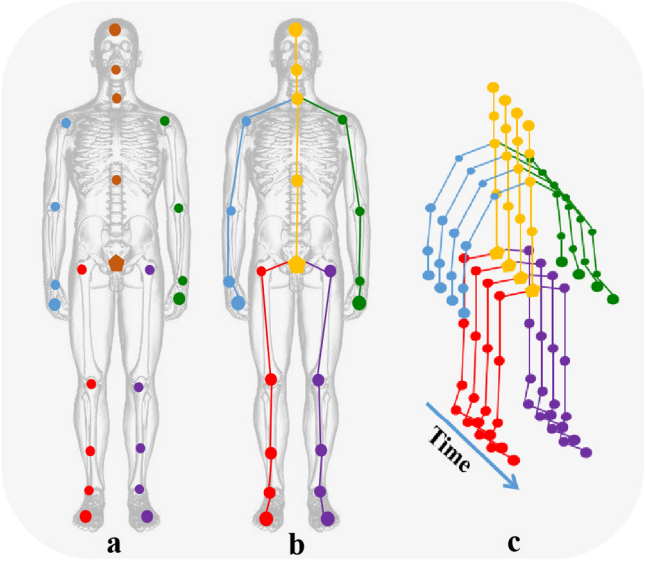


Fig. 2. As is shown in the human body structure **a** and **b**, the human body is composed of 3D joints and different colored joints which constitute different kinematic chains, such as the central torso, the left/right arms and the left/right legs. The human body contains different numbers of joints in different datasets. For **c**, it is the human motion sequence made up of 3D human bodies that changes over time.

To cope with the challenges mentioned above, different solutions have been proposed. Initially, researchers focused on modeling human motion sequence-to-sequence, which depends on a deep RNN-based architecture that specializes in sequential problems. On account of the prosperity of generative adversarial networks (GANs) [26], they are employed as a new kind of learning algorithm for human motion prediction. Further, considering the connections between joints, Graph convolution networks (GCNs) [102], a generalization of CNNs, are utilized to model the correlation between joints for the human motion prediction. Meanwhile, other CNNs [45,49] methods are also proposed.

1.3. Outline

The purpose of this paper is to survey the existing methods of 3D human motion prediction and investigate these methods by classifying them and analyzing their performance differences. Then, the public benchmark datasets and evaluation metrics in this field are also reviewed in detail. This review intends to conduce to an accessible comprehension about the similarities and differences among a variety of methods and provides more ideas for future research.

The main contributions of this paper are summarized as follows:

1) To the best of our knowledge, this paper should be hitherto the first survey on 3D human motion prediction.
2) Existing methods in this domain are classified into three categories, and the benchmark datasets and evaluation metrics available in this domain are elaborated.
3) Comparisons of existing methods are presented, and future research directions are discussed.

The remaining parts of the paper are organized as follows: The taxonomy of 3D human motion prediction is introduced in Section 2. The review and analysis of existing prediction methods are described minutely in Section 3. The experimental benchmark datasets and evaluation metrics are presented in Section 4. Then, the comparison of different human motion prediction methods is shown in Section 5. In Section 6, the most advanced methods are discussed and the possible challenges for further research are listed. Finally, Section 7 is the conclusion of this paper.

2. Taxonomy

In this section, the proposed taxonomy classifies human motion prediction methods into three categories: *Human pose representations*, *Network structure design*, and *Prediction targets*, as shown in Fig. 4. Afterwards, the explanation for the rules of taxonomy are introduced. Finally, the three categories are utilized as hints to further expound on the correlative methods, and representative papers in each category are taken as examples to promote understanding.

2.1. Classification rules

In recent years, as shown in Fig. 5, human motion prediction has attracted more and more researchers to tackle this task with various approaches, which are generally a mixture of numerous different and basic ways. Hence, certain criteria to systematically classify them is urgently needed. We use the following rules:

1) Priority is given to modeling approaches. The approaches from the category introducing the modeling approaches are classified with precedence.
2) Focus is on the major models in the composite-model approach. Considering that some methods are combined with sub-components, the major models and innovations should be the focus for classification.
3) Emphasis is given to the diversity of perceptions. Different perceptions of prediction targets and human body representations determine the approaches to problem solving, which motivates our classification.

With the highest level of abstraction, as shown in Fig. 3, the 3D human motion prediction task can be classified into three phases: *Observed sequences*, *Prediction models*, and *Prediction results*. Three main problem-solving strategies are inspired among current research results, which are summarized in this paper as: *Human pose representations*, *Network structure design*, and *Prediction targets*. Human pose representations emerge as the physical or mathematical transformation of parameterized observed human pose sequences. The network structure design aims at extracting and processing powerful features to facilitate prediction. It is the core means for the prediction models. The prediction target is not soli-

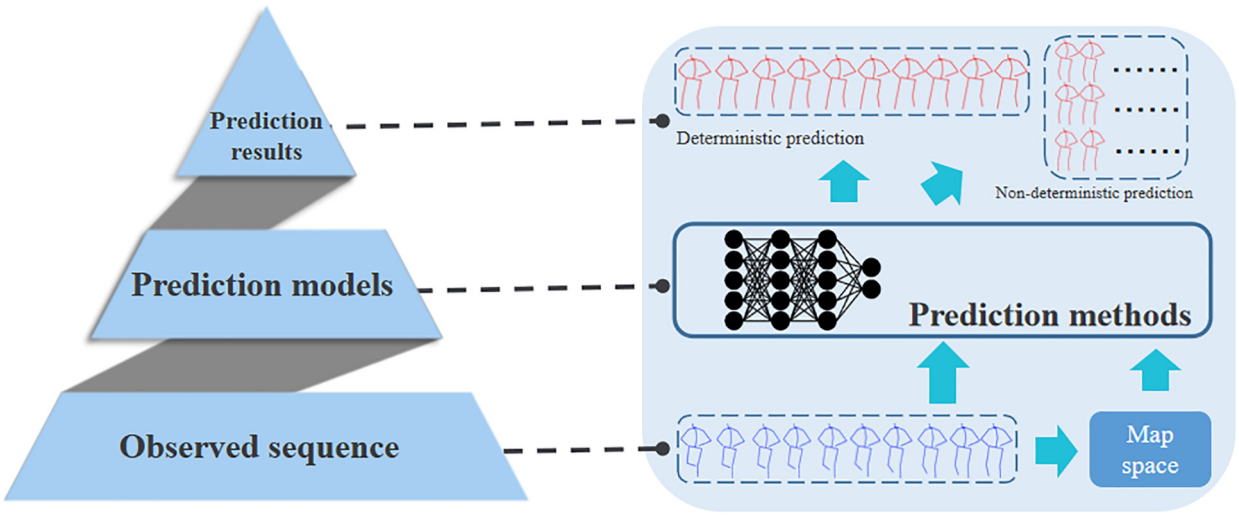


Fig. 3. Generalization of the 3D human motion prediction task. On the **left**, it is a concise summary of the task. On the **right**, it is an overview of operations of methods.

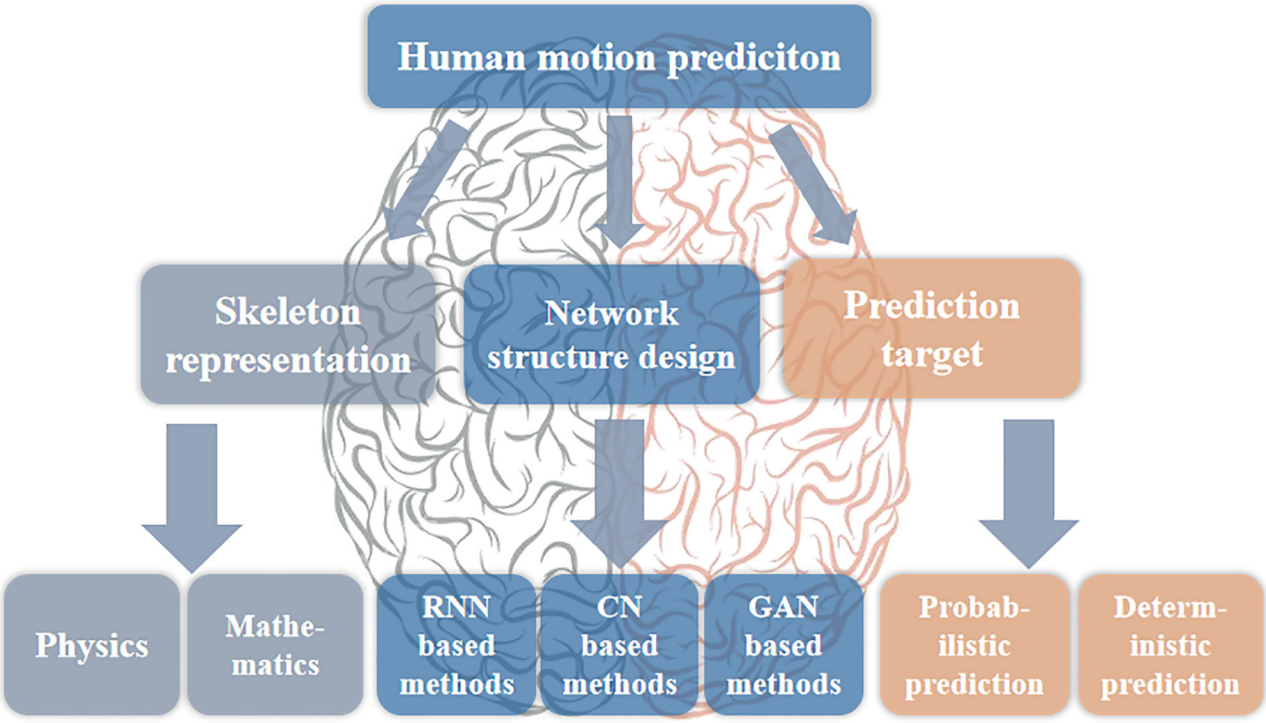


Fig. 4. Overview of the categories in our taxonomy.

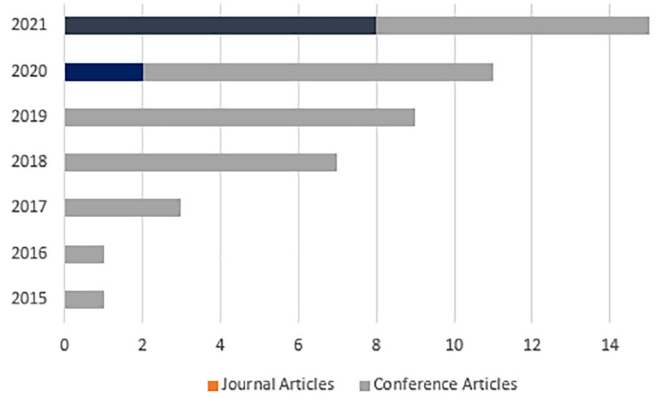


Fig. 5. Publications trends in the literature reviewed for this survey.

tary. For a given past, there will be one or more generating sequences of future frames.

2.2. Human pose representation

All the while, human pose representation plays a fundamental yet important role in the field of computer vision. An efficient human pose representation is helpful for the machine to understand the human behaviours. After induction, existing human pose representation approaches are classified into two categories: *physical representation* and *mathematical representation*.

Physical representation As a classic, skeleton-based human pose representations have been widely recognized. These benefit from their robustness to position display, kinematic representation, and real-time performance in the field of human motion prediction. Physical representations of human pose reflect the kinematic laws, basic structures, and motion forms. Employing natural 3D joint positions to denote the human pose skeleton is one of the most commonly used strategies. As a follow up of this line of work, hierarchical body parts are selected. Besides, being important motion properties, velocity and acceleration carry significant connotations of dynamics. As is shown in Table 1, representative papers in this category are: SkelNet [30], CHA [49], MGCN [110], AM-GAN [56].

Mathematical representation Mathematical representations signify the rigorous description of the abstract structures and patterns of human poses. The parameterized 3D human pose is original representation, which is derived from motion capture data.

Table 1
Human pose representation methods table.

Categories	Methods	Years
Physical representation	SkelNet [30]	2019
	DSR [96],	2019
	CHA [49]	2020
	MGCN [110]	2020
	AM-GAN [56]	2021
Mathematical representation	SRNN [36]	2016
	AGED [28]	2018
	QuaterNet [79]	2018
	LTD [67]	2019
	HMR [60]	2019
	MGCN [110]	2020
	HRI [66]	2020
	ARNet [12]	2020
	LDR [17]	2021
	LPJP [11]	2021
	TrajectoryCNN [53]	2021
	JDM [91]	2021
	MMA [68]	2021
	MPTC [57]	2021
	SGAN [62]	2021
	MT-GCN [15]	2021
	MPTC [57]	2021
	MST-GNN [47]	2021

Mathematically, these parameters are mapped to different mathematical spaces and abstracted into different distributions whose features are easy to be extracted by the network. As is shown in Table 1, representative papers in this category are: SRNN [36], AGED [28], QuaterNet [79], LTD [67], HMR [60], MGCN [110], ARNet [12], LDR [17], LPJP [11], HRI [66], JDM [91], SGRU [62]. MT-GCN [15], MPTC [57], MST-GNN [47], MMA [68].

2.3. Network structure design

As a major characteristic of deep learning approaches, automatic extracting and learning features from data is of considerable importance. Further, the quality of features is close to the network structure. Likewise, Network structure design approaches aim to extract the efficient features to enhance human motion prediction. There are mainly three kinds of methods in terms of major network structures: *RNN-based methods*, *CN-based methods*, *GAN-based methods*.

RNN-based methods are a class of structures that are mainly composed of recurrent neural networks (RNNs) [108]. Currently, LSTM [31], GRUs [13] and their variants are methods that are widely used in human motion prediction due to its preponderance in processing sequential issues. CN-based methods rely on multiple convolutional networks, such as graph convolutional networks (GCNs) [17,67,110], temporal convolutional networks (TCNs) [17], and the variants of traditional convolutional neural networks (CNNs) [45,49]. CN-based methods do well in capturing spatial dependencies. Moreover, they also perform well in capturing temporal dependencies by the rising of TCNs. GAN-based methods profit from the generative adversarial networks (GANs) [26]. These are always applied to synthetic data generation and probabilistic prediction by DNNs and a kind of novel algorithm that is utilized in network learning.

RNN-based Methods With the examples of the successful applications in voice recognition [103,92], machine translation [33,95], and sequential prediction [22,81], RNNs (LSTM, GRU, and their variants) have become a widely used framework for the human motion prediction task. It is normally deemed as sequence-to-sequence (seq2seq) prediction tasks, where RNNs are adopted in solving. However, for the human motion prediction, a major distinction from other seq2seq tasks is that the human body is a human kinematics system with high constraints. Therefore, some RNN structures particularly for human motion prediction are proposed. As shown in Table 2, these kinds of methods include: LSTM-3LR [71], ERD [21], Res-GRU [71], SRNN [36], DAE-LSTM [25], AGED [28], MHU [93], QuaterNet [79], SkelNet [30], DSR [96], HMR [60], RNN-SPL [3], VGRU [27], C-RNN [14], PVRED [94], BNN [99].

CN-based methods In a general way, RNNs seem to be a natural choice for sequential prediction, but convolutional approaches are also employed in this task [16,5,77]. Convolution networks have their inherent advantage in capturing spatial dependencies. Thereby, convolutional seq2seq methods are well adopted. Besides, Graph Convolutional Networks (GCNs) are adaptable for presenting the human skeleton as a graph, which makes GCN widely utilized in human pose representation. Additionally, widespread recognition of the Temporal Convolution Networks (TCNs), made the CN-based methods more widely accepted for solving the sequential problem. These methods include: TE [9], PAML [29], CHA [49], LDR [17], MGCN [110], LTD [67], LDR [17], MoPredNet [107], TrajectoryCNN [53], NAT [44], C-seq2seq [45], Q-DCRN [73], MT-GCN [15], LMC [110], MST-GNN [47], DA-GNN [48], JDM [91].

GAN-based Methods GANs have shown impressive performance in various tasks [75,41,35]. It not only provides a kind of novel learning algorithm but also offers a vital alternative to generative models. Empirically, two main strategies are deployed by leveraging GANs: promoting GANs framework performance and researching how to improve human motion prediction by these frameworks. These methods include: HP-GAN [7], BiHMP-GAN [40],STMI-GAN [86],GAN-poser [37],AM-GAN [56],TC-GAN [16], AGED [28], ARNet [12], SGAN [62].

Table 2
Convolution Network methods table

Categories	Methods	Years
RNN-based methods	LSTM-3LR [71]	2015
	SRNN [36]	2016
	Res-GRU [71]	2017
	DAE-LSTM [25]	2017
	AGED [28]	2018
	MHU [93]	2018
	QuaterNet [79]	2018
	SkelNet [30]	2019
	DSR [96]	2019
	HMR [60]	2019
	RNN-SPL [3]	2019
	VGRU [27],	2019
	C-RNN [14]	2020
	PVRED [94]	2021
	BNN [99]	2021
CN-based methods	TE [9]	2017
	C-seq2seq [45]	2018
	PAML [29]	2018
	CHA [49]	2019
	LTD [67]	2019
	LDR [17]	2020
	MGCN [110]	2020
	MoPredNet [107]	2020
	TrajectoryCNN [53]	2021
	NAT [44]	2021
	Q-DCRN [73]	2021
	MT-GCN [15]	2021
	LMC [110],	2021
	MST-GNN [47]	2021
	DA-GNN [48]	2021
	JDM [91]	2021
GAN-based methods	HP-GAN [7]	2018
	AGED [28]	2018
	BiHMP-GAN [40]	2019
	STMI-GAN [86]	2019
	GAN-poser [37]	2020
	ARNet [12]	2020
	AM-GAN [56]	2021
	TC-GAN [16]	2021
	SGAN [62]	2021

2.4. Prediction target

Human motion prediction mainly contains two targets. The first purpose is to generate motion predictions that are close to the Groundtruth. Simultaneously, the future is not deterministic in the real world, which leads to the second purpose that is to generate a variety of possible motion futures. In a word, these approaches are segmented into two classifications: *probabilistic prediction* and *deterministic prediction*.

Probabilistic prediction The future is not deterministic. So, the same prior poses could lead to multiple possible future motions. To generate probabilistic motion prediction, the random noise vector is aggregated to the prior pose sequence or latent variables in the prediction process. Besides, a meaningful network learning mode is gained to upgrade the noise. Conclusively, as shown in Table 3, these methods include: HP-GAN [7], Dlow [106], GAN-poser [37], Mix-and-Match [4],BiHMP-GAN [40], PVRED [94], SGRU [62].

Table 3
Prediction target methods table.

Categories	Methods	Years
Probabilistic prediction	HP-GAN [7]	2018
	BiHMP-GAN [40]	2019
	GAN-poser [37]	2020
	Mix-and-Match [4]	2020
	Dlow [106]	2020
	PVRED [94]	2021
	SGAN [62]	2021
Deterministic prediction	others	2015–2021

Deterministic prediction Most human motion prediction methods are based on deterministic models. It is usually regarded as a regression task that produces only one outcome from a prior pose sequences. A large proportion of the methods mentioned earlier are using deterministic prediction models, which are also used by some newly emerged methods with only a few applications in this field. We summarize these methods in this part.

3. Methods description

In this section the detailed analysis of different methods in different categories is carried out on the basis of the taxonomy shown in Fig. 4.

3.1. Physical representation

The physical representation of human poses follows the principles of human motion and the basic structure of the human body. These representations can be understood as a class of a prior anatomical knowledge and constraints that help the network to comprehend the input information and extract features efficiently. Generally, many existing traditional approaches [71,21] usually utilize natural 3D joint positions to display the human skeleton. These methods mainly rely on the network performance but neglect the logical mining of human poses mocap data. With the progress of human motion prediction by utilizing physical representation, gradually, researchers recognize the usefulness of physical representation. Statistically, this category can be sorted into two strategies: *human body structures* and *human motion laws*.

Human body structures As a representative for the first strategy, researchers attempt to optimize human body structure or adopt a new approach to learn the structure of the human body. Guo et al. [30] argued to learn local structure representations in different kinematic chains in different ways. Specifically, instead of taking the holistic human pose as input, it was divided into five non-overlapping parts according to the connatural kinematic chains (the central torso, the left/right arms, and the left/right legs). This method demonstrated that paying attention to the local dynamic patterns was significant for the network to understand human motion. Analogously, Liu et al. proposed a new GAN framework named Aggregated Multi-GAN (AM-GAN) [56]. AM-GAN modeled the motion of the central spine and four limbs. Then, the final whole human motion was aggregated over them. Nevertheless, it not only predicted future motion accurately but also can complete the tasks of controlled human motion prediction finely. Further, Li et al. [49] designed a convolutional network (CHA) for human motion prediction. The human pose was graded into three levels which respectively stood for joint links, five kinematic chains of human body, and the bilateral zygomorphic parts of human body. In this way, CHA can integrate multiple human body constraints into feature extraction. Mathematically, this strategy can be described as follows:

$$c_{N+T} = f_N(c_N), p_{t+T} = (c_{1+T}, c_{2+T}, \ldots, c_{N+T}) \quad (1)$$

where $N \in N^*, c$ indicates kinematic chain, p indicates single pose. In general, when N is 5, c represents the kinematic chain. If N is 2, c represents two parts of human body.

This type of strategy artificially attaches a prior knowledge related to the structure of the human body to the framework. However, it is relatively superficial compared with other kinematic knowledge. Moreover, networks may perceive only dimensional changes. Therefore, in terms of performance, such methods can improve the prediction to some extent.

Human motion laws For the second strategy, human motion is typical of mechanical motion, that is, a part of the human body is relative to another part of the body (including the original part) in terms of space and time displacement, e.g. velocity and acceleration. Following this line, Wang et al. [96] focused on that the motion of human body can be represented by velocity and acceleration instead of position. In the RNNs frameworks, Wang et al. proved that these frameworks were easier to fit the data represented by velocity and acceleration. However, this performance improvement is limited to short-term prediction (less than 400 ms) and does not effectively ameliorate the intrinsic defects of RNNs. Similarly, Li et al. [110] considered learning richer motion dynamics for human motion prediction. Therefore, the difference operators were deployed to extract multiple physical information: positions X, velocities ΔX, and accelerations $\Delta^2 X$. Summarily, the second strategy can be described as follow:

$$\Delta^n X_t = \Delta^{n-1} X_t - \Delta^{n-1} X_{t-1}, 0 < n < 3 \quad (2)$$

These methods, which follow the motion laws, are worthy of study. These data processed by motion laws, such as velocity and acceleration, comprise richer kinematic knowledge, naturally containing inter-frame information and more stable data variations. Deeper research on the usability of such methods is needed.

Summary of physical representation Physical representation plays a seminal role in prediction performance, both in terms of human structure and human motion laws. However, in both categories, the human body structure is shown to be less effective than human motion laws. The reason is that the former is explored only in terms of the shape of the human pose vector, whereas the latter provides richer information on physics that is not confined to a single frame. Therefore, attempts can be made to incorporate more physical knowledge into the field of human motion prediction.

3.2. Mathematical representation

The human pose has its immanent structure in the scientific research, which is mostly presented as the 3D skeleton representation captured by a motion capture system. The mathematical representation of human pose is a general way to strictly describe the abstraction of the human pose. It allows experts in this domain to inject their prior knowledge into the learning process of these human poses. Existing mathematical representations can be sorted into three directions, *Graph*, *Motion trajectory*, and *Mathematical encoding*.

Graph The joints and links between adjacent joints of the skeleton representation (a human skeletal kinematic tree) naturally remind the researchers of graph [67,66]. Graph $G(V, E)$ is a data format and also an encoding mode that can be used to represent social networks [83,74], communication networks [80,98], protein molecular networks [6,76], etc. The nodes E in the graph represent individuals in the network, and the linked edges V represent connections between individuals.

Inspired by this, Yan et al. [36] first constructed a spatial–temporal graph to define the human body. However, these graphs were

manually designed, thus limiting the flexibility of the algorithm. Considering this, in [17], Cui et al. designed a trainable graph embedded into the network. The adjacency matrix that represented the weights of the connected joint parts was set as the model parameters. A similar method was also used in [15,91]. Analogously, Mao et al. [67] also created a new method to learn an automatic–connected graph. Innovatively, it proposed the non-restrictive joint dependencies without the constraints of the kinematic tree or the convolutional kernel size. Nevertheless, such constructions are tantamount to treating pose sequence roughly as routine data, losing sight of significant kinematic constraints, which leads to unstable training. This strategy was also adopted in their method [66]. To overcome this drawback, they put forth a new method [68] that considered the kinematic constraints from three levels of the human body. The first level was the full-body, the second was the parts of human body, and the third was the individual joints.

Parallel to this work, considering the human kinematic structure, Li et al. proposed two multi-scale graphs to model the human body. In [46], a multi-scale graph was presented to capture multi-scale features to predict future human motions. It contained two kinds of sub-graphs, namely single-scale graphs and cross-scale graphs, which respectively connected the human body components at the same scale and cross-scale. Subsequently, they also proposed a multi-scale spatial–temporal graph (st-graph) [47] to comprehensively model human motion. The characteristics of the st-graph are trainable, multi-scaled, and decomposable. To be specific, the st-graph can be decomposed respectively into a spatial graph and a temporal graph, both are trainable. Meanwhile, Liu et al. [57] devised a semi-constrained graph to explicitly encode skeletal connections and used prior knowledge (such as limb mirror symmetry tendency and cross sides synchronization tendency). This kind of graph proves that the combination of adaptive learning and constraints is beneficial to training.

Motion trajectory The human motion trajectory is the spatial motion feature composed of the path taken by a part of the human body from the beginning position to the end. In recent years, the progress of the trajectory space for human motion prediction has aroused researchers' interest in modeling human motion in the trajectory space rather than in traditional motion space.

Based on this, Mao et al. [67,66,68] presented human motion in trajectory space for several times. Parallel to these, Su et al. [91] also adopted the same trajectory space of a joint. Mathematically, these methods can be formulated as:

$$T_j = (t_{j,1}, t_{j,2}, t_{j,3}, \ldots, t_{t,N}) \tag{3}$$

where j denotes j^{th} joint, N denotes the number of frames. As the most commonly used trajectory representation, it is favorably received by researchers. Differently, Liu et al. [57] proposed a new trajectory format, which consisted of the position displacement of the adjacent frames and motion direction. Similarly, in [53], a novel representation was invented to build the correlations between the local and global space, which were respectively representative of the same part and different parts in the human motion sequence. The motion trajectory is smoother compared with traditional space, which contributing to more stable data variation. As a result, this will bring more convenience to the network training.

Mathematical encoding It is clear that reasonable constraints are quite conducive to the network to learn the features. A mathematical expression is a formulaic constraint that is in a sense easier to program. Thereby, finding an effective mathematical way to encode the human pose is undoubtedly of great significance to human motion prediction. However, this process requires much richer prior knowledge. To inspire researchers, the existing mathematical methods are summarized and then divided into the following two categories: *algebra encoding* and *differential encoding*.

Algebra encoding Algebraic encoding deals with the problem of human motion prediction by studying the number, relationship and structure of human poses. Common types of algebraic structures are groups, loops, fields, modules, etc. Generally, each pose is described as a human body joint positions union or a 3D-joint rotations union. In most cases, human joints are represented as some algebraic form that captures dependencies more easily. In [79], they represented joint rotations with quaternions, which lie in R^4. The quadratic representation is not plagued by discontinuities and singularities. It is more stable with respect to the variation of number and is computationally more efficient, with enormous potential in studying human motion prediction. Another representation is to utilize the Lie Group [61], which is a Riemannian manifold structure. In [28,60], they both adopted this method to encode the human pose. Differently, Gui et al. [28] only made use of the Lie groups to characterize the relative 3D geometry for a single joint, but Liu et al. [60] characterized it between two successive bones. Regarding effectiveness, Liu et al.'s method was more suitable for modeling with the Lie group. After the introduction to the generality of trajectory representation above, efficient ways of encoding trajectory have emerged. The strength of trajectory representation lies in its smooth representation of the object's motion. Discrete Cosine Transform (DCT) [2] is widely used in many methods [67,66,68,12,11] for encoding trajectory. By discarding the high frequencies, it can generate a more compact representation, which can transform the 3D human joints coordinates into smoother motion state, predestining its popularity. Distinctively, the phase space trajectory representation was proposed by Su et al. [91]. They employed the joint instantaneous displacements from the joint displacement of adjacent frames to form the joint trajectory. *Differential encoding* The differential method itself is an advanced mathematical method, which sets, higher requirements on researchers' ability. Therefore, at first, researchers use differential methods only for simple problems such as [110,47]. A difference operator is utilized to compute the differences of observed poses. These differences can reflect richer dynamic information (e.g., velocities, and accelerations) for the network. They cannot play a decisive role in solving the problem. In [62], Lyu et al. proposed a novel modeling method that utilizes stochastic differential equations [88] to model human pose. The motion profile of each skeletal joint was formulated as a basic stochastic variable and modeled with the Langevin equation [1]. This approach merely required a basic and lightweight network to achieve good results.

Summary of mathematical representation The mathematical representation is a rigorous abstraction of the human pose. From a practical perspective, effective mathematical representations are quite helpful in understanding the principles, for both humans and machines. This is definitely a pretty indispensable way to meliorate predictive performance.

3.3. RNN-based methods

Human motion prediction problems are often treated as seq2-seq prediction tasks. RNNs are widely recognized for their excellent performance on such tasks, which has inspired many researchers to utilize RNN-based methods to investigate human motion prediction tasks. As is shown in Fig. 6, two basic RNNs (stack structure and stream structure) are deployed on it.

Primitively, two architectures LSTM-3LR [71] and ERD [21] were proposed. Their main frameworks were concatenated LSTM units. The difference that made ERD perform better was the nonlinear space encoder. However, these kinds of networks were still trapped in accumulated errors, discontinuity in initial frames,

which quickly produce unrealistic human motion. Therefore, the res-GRU [71] was designed to solve these issues. They modeled the velocity of joints to represent the human body. For the network, they employed the linear layer to encode the pose features and decode the hidden states. It firstly demonstrates the efficiency of the velocity modeling. The noise was combined with the input sample during training, which prompted a more robust network to alleviate drifting. However, this strategy brought about difficult training sessions. In [36], Jain et al. proposed structural RNNs (SRNNs), in which the manually designed spatial–temporal graph was later employed. Noise was also employed to SRNNs. Then, the DAE-LSTM [25] combined a LSTM-3LR with a dropout auto-encoder to model temporal and spatial structures. The above methods have laid a solid foundation for subsequent researches [79,30,96,3,99,28,94,14]. It is worth mentioning that, some methods have made innovative design in network structure.

The improved performance inspired us to focus on the kinematics during the network design. Some researchers find that not all joints are involved in human actions. In other words, some joints are motionless during the motion. Based on this, in [93], authors proposed a modified network (HMU) for human motion prediction without motionless joints for long-term prediction. Specifically, they designed a novel gate structure to filter the motionless joints. What's more, an attention module was utilized to concentrate on these kinds of motions. Then, Liu et al. [60] proposed a hierarchical recurrent network (HMR). To avoid the impact of recent frames, the input of the network was set as a single pose. To capture more long-term dependencies, in the recurrent units, adjacent joints and frames can be encoded concurrently. Further, a new RNN framework was also presented by them, which was named AHMR [59]. It not only can concurrently model the local and global context but also an attention module was employed to assist in updating the global context. Moreover, two effective loss functions were designed in this work. In this way, a longer-term and more real human motion sequence can be generated. Respectively, a hierarchically decomposed network was presented in [27]. To be more specific, they proposed a coarse-to-refine network to generate the states of each stage. In the coarse stage, the guide vectors were created as coarse states for the next stage. Then, the actual pose outputs were gained from the closed-loop predictions.

Summary of RNN-based methods The RNN-based methods, as always, show marvelous handling of timing problems. However, human motion prediction is not purely a temporal problem, as it has kinematic and human anatomical constraints. Therefore, the development of such methods has also moved from exploiting the performance of recurrent networks to incorporating effective constraints. Even so, the inherent problems of RNNs continue to plague researchers to varying degrees, which needs to be addressed on an ongoing basis.

3.4. CN-based methods

Human motion involves both spatial and temporal correlations. CN-based methods have a natural advantage in capturing spatial dependencies. However, traditional CNNs are stuck in limitations in processing standard 2D grid representations. Thanks to the GCNs, the spatial correlations can be captured efficiently. With the introduction to TCNs, the ability of CN-based methods to deal with temporal correlations becomes prominent. In summary, the CN-based methods include three sorts: *conventional convolution methods*, *GCNs methods*, and *TCN methods*.

Conventional convolution methods Traditional CNNs are widely used in human motion prediction for their inbuilt ability to capture spatial dependencies. In [9], a convolutional layer was designed to encode different time scales. It was only utilized to capture the local time scales. In QuaterNet [79], dilated convolutions was utilized in the network to capture long-term temporal dependence with the hierarchical input poses. However, these methods can only capture single and temporal information. Further, a hierarchical structure of CNN [45] was employed in the network to capture spatial and temporal dependencies. The convolutional structure was utilized in the encoder to obtain a long-term hidden state, which was fed to the decoder to generate the human poses. Then, Li et al. [49] designed a convolutional hierarchical autoencoder framework. Concretely, hierarchical topology was employed to represent the human body tree structure, and 1D convolutional layers were embedded to leverage these constraints to encode each node. A novel framework named TrajectoryCNN [53] was designed to predict future poses. It introduced a new type of trajectory space for spatio-temporal dependencies, which included a variety of local–global and spatio-temporal features, that can be captured easily by CNNs.

GCNs methods GCN actually does the same thing as CNN, which is a featured extractor, except its object is graph data. It is an ingeniously designed method for extracting features from graph data. With this method, the features extracted can be used to achieve a wide range of purposes, such as node classification, graph classification, link prediction, and incidentally, graph embedding and so on.

In [67], a human pose was encoded as a graph structure, which followed the principle of connecting every neighboring joints. Moreover, they proposed a new GCN to automatically connect the graph instead of manual defining. Then, a novel graph network was proposed as a generator in the GANs [17]. A dynamic learning graph was also used but it was different from a generatic one. Because it not only can explicitly learn the natural joint pairs but also implicitly connect geometrically separated joints. In [46], Li et al. designed a novel GCN named DMGNN which contained a dynamic multi-scale graph to represent the human skeleton struc-

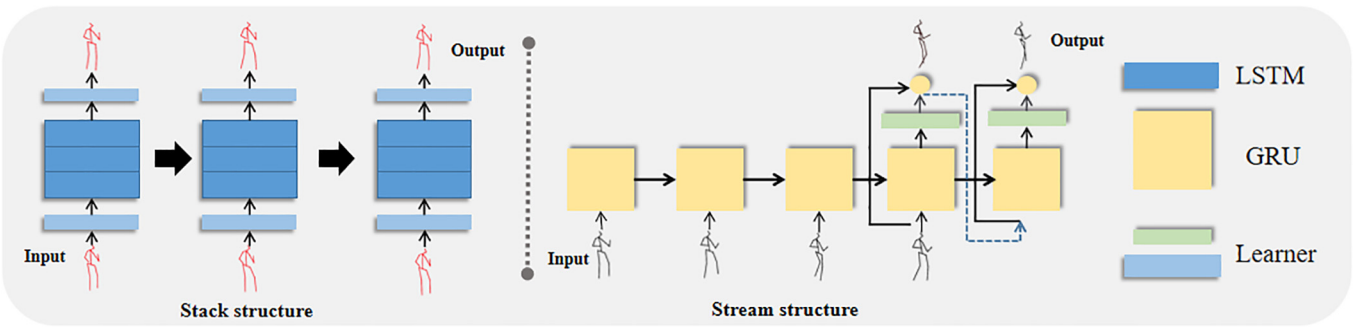


Fig. 6. The basic RNNs prediction methods. These basic networks are mainly stack structure [21] and stream structure [71].

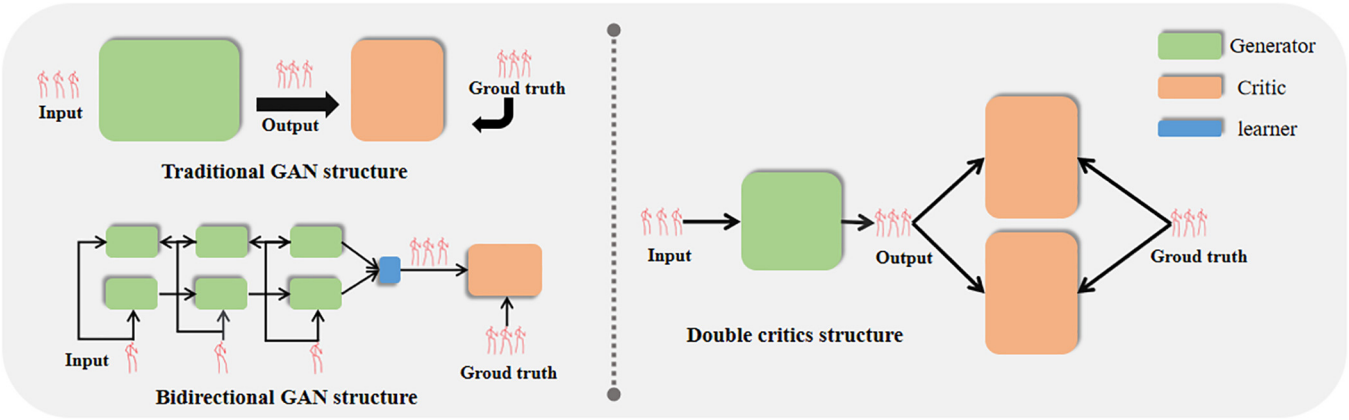


Fig. 7. The basic GANs prediction structure. Structurally, there are two kinds of frameworks, traditional GAN structure [7], bidirectional GAN structure [40], and double critics structure [28].

ture. The multi-scale graph can comprehensively model the internal relations of a human body. It also can be used during dynamic across network layers learning. To generate future poses, a proposed graph-based gate recurrent unit was employed for this task. In order to generate high-fidelity human motion predictions from incomplete observations, [15] also proposed a novel multi-task graph convolutional network (MT-GCN), which included a shared context encoder (SCE). In addition to graph structure, a temporal self-attention mechanism was selected to the most related information from the whole sequence to repair the corrupted pose. This kind of multi-scaled method was also present in [110] to capture the correlation of body components.

TCN methods In recent years, TCNs display their outstanding performance in dealing with sequential tasks. TCNs are mainly composed of Causal Convolution and Dilated Convolution. Causal convolution makes judgments based on historical and current information, rather than future information, which is in line with the characteristics of time sequence. Atrous convolution is not restricted by the size of the convolution kernel, which enables CNN to process longer time series with certain parameters. Owing to these, TCNs are also extensively utilized in human motion prediction. In [107], a novel approach named Motion Prediction Network (MoPredNet) for few-short human motion prediction was proposed. MoPredNet can be adapted for predicting new motion dynamics using limited data, and it delicately captured long-term dependency in motion dynamics. Thereinto, a deformable Spatio-Temporal Convolution Network (DSTCN) was proposed to adaptively model sub-motion dynamics and spatial correlation in the entire motion sequence to capture long-term dependency for motion generation. Further, Cui et al. presented the residual TCN (ResTCN) in the TCGAN [16]. It was utilized as the generator for motion prediction. ResTCN was employed to model sequential prediction tasks due to its few parameters and high efficiency.

Summary of CN-based methods In summary, all three methods effectively employ convolutional modules to handle the task of human motion prediction. Traditional methods focus on the ability to use convolutional networks for spatially dependent capture. GCNs benefit from the effective a priori knowledge-graph that can efficiently represent the relationships between each human body part. TCNs have become popularized when dealing with its outstanding performance for sequential problems, even surpassing RNNs.

3.5. GAN-based methods

On account of the extensive use of generative adversarial networks (GANs) [26], it provides many new directions for human motion generation by the probability density function and network learning algorithms. In Fig. 7, two classic GANs structures are shown.

As the first application of GANs in human motion prediction, Barsoum et al. [7] proposed a novel sequence-to-sequence model for probabilistic human motion prediction, trained with a modified version of improved Wasserstein generative adversarial networks (WGAN-GP). Different from previous works, HPGAN [7] represented the input of the network as probability distribution during the training. The probability density function can assist to predict multiple future pose sequences by inputting the same prior sequence with a different vector z. This is also a common practice to produce diverse futures. However, they failed to assess expressiveness of such a generative approach against deterministic counterparts. After this work, a new probabilistic generative approach called Bidirectional Human motion prediction GAN, or BiHMP-GAN [40] was designed. Similarly, BiHMP-GAN was also able to generate probabilistic future motions with the same input incorporating a random vector r from predefined distribution. Likewise, a bidirectional GAN framework was designed to avoid the mode-collapse. In this method, the discriminator was also trained to regress the vector r. Further, BiHMP-GAN made the comparison with other deterministic approaches. Also as a bidirectional GAN framework, GAN-poser [37] was employed to avert mode collapse and to further regularize the training. Similar to the previous method, it also utilized a random factor to generate multiple future motions. In spite of being in a probabilistic framework, the modified discriminator architecture allowed considering the randomness of the prediction. In addition, a new evaluation was provided in this work for assessing the prediction results. Obviously, these bidirectional GAN frameworks performed better than general GANs.

Parallel to these works, inspired by the adversarial training mechanism, AGED [28] presented a novel GAN framework with two global recurrent discriminators. One discriminator was utilized to promote the fidelity of the generation sequence, the other discriminator was trained on the joint-level to guarantee the continuity of generated future sequences. STMI-GAN [86] also adopted the adversarial learning for spatio-temporal tensor of 3D skeleton coordinates over long periods of time. Dexterously, Adversarial Refinement Network (ARNet) [12] designed a novel adversarial error augmentation. Different from the normal adversarial learning, the discriminator was utilized as a middle module to produce the prediction errors, which were transferred to the refinement module. Interestingly, unlike previous works, Lyu et al. [62] utilized the GANs to simulate path integrals for solving the stochastic differential equations and predict future motion profiles.

However, it should not be overlooked that it was challengeable to train GANs. Since it was difficult to reach the Nash equilibrium between the generator and discriminator, Cui et al. [16] presented a new GAN with spectral normalization to avoid mode collapse. There's another strategy called AMGAN [56], which was designed from a composite GAN structure, consisting of local GANs for different low-dimensional body parts and a global GAN for the high-dimensional whole body. This method proved that dimension reduction can efficaciously improve the training efficiency of GANs.

In summary, the strategy of utilizing GANs mainly can be divided into two categories. (*i*) Being used as a learning algorithm to help with network learning. (*ii*) Utilizing data distribution to generate multiple prediction targets. GANs, as a network with noticeable advantages and disadvantages, can also give rise to certain challenges for researchers' work.

3.6. Probabilistic motion prediction

The inherent stochastic nature of motions always besets the machine to understand and predict human motion. What's more, for human future motions, the longer the time period is, the more uncertain the result will be. Even the subject of motion cannot be sure about his/her movement over a long period of time. Therefore, probabilistic motion prediction is indispensable for the research of human motion prediction. According to the implementation means and network framework of existing methods, these are mainly divided into two categories: *GAN-based methods* and *VAE-based methods*. As shown in Fig. 8, it shows the basic probabilistic motion prediction structures.

GAN-based methods As a common strategy for the GAN approaches, they generate probabilistic future motions by merging the random noise into the observed pose sequence. The HPGAN [7] initially proposed to utilize the GAN framework to fulfill the probabilistic human motion prediction task. It generated multiple future pose sequences only from a sing observed pose. In this strategy, a random vector r was combined with the prior pose sequence. Each random vector can assist to produce a possible future motion. This idea was also expressed in [62,37]. Nevertheless, the issue of discontinuity of motion cannot be overcame by this method. To hit the target, BiHMP-GAN [40] utilized a single random vector to generate the entire sequence. This vector was employed to alter the initialization of the decoder and concatenated with a pose embedding at each iteration of the RNNs. By relying on concatenation as a method to fuse the condition and the random vector, these two methods contained parameters that were specific to the random vector, and thus brought the model flexibility to ignore this information.

VAE-based methods The second category of probabilistic motion prediction is achieved by an additional CVAE. It is similar to GAN in that it changes the output by fusing the noise. Differently, the random noise will turn into the intermediate variable before the passing of decoder. In [4], directly, the human pose was utilized as a conditioning variable to combine the random noise. However, it generated low-grade human motion due to the incoherence of random noise at each step. However, in [10], the RNN decoder hidden state taking place of the pose was employed as a conditioning variate. This method effectively promoted performance but neglects the flexibility of the random vector. Yuan et al. skillfully designed a new sampling approach named DLow [106]. Different from the normal random sampling, a set of learnable mapping functions was employed to produce a set of correlating latent variates only from an input of single random noise.

In summary, probabilistic human motion prediction relies on a combination of random noise and conditional variables. Definitely, the network structure that can work on this is equally important. The existing underlying networks are GANs and VAEs. The combined progress of these two paths will unquestionably be a major boost.

3.7. Deterministic motion prediction

Deterministic motion prediction is the most widely applied human motion prediction means. Generally, human motion prediction is always regarded as a regression task. RNNs are acknowledged to be skilled in handling such tasks. Therefore, many researchers attempted to utilize RNNs to cope with human motion prediction. Typically, these approaches [21,71,36,25,79,30,96,99,28,94,14,93,60,27] effectually predict future human motions in the recurrent framework. Similarly, the feed-forward models [16] are also deployed for this task. Recognizing the positive impact of kinematics on prediction, many convolutional methods [49,45,9,53], especially the graph neural networks [17,67,46,47,86,28,12], are utilized.

With the popularity of attention networks, there are many proposed solid methods. In [107], a novel approach named Motion Prediction Network (MoPredNet) for few-short human motion prediction was presented. Specifically, MoPredNet dynamically selected the most informative poses in the streaming motion data as masked poses. In addition, MoPredNet enhanced its encoding capability of motion dynamics by adaptively learning spatio-temporal structure from the observed poses and masked poses. In [11], a transformer-based method was proposed to solve the human motion prediction issue. This framework can concurrently capture temporal and spatial dependencies of human motion. With the employing of the global attention mechanism, the network can acquire more effective long-term dependencies.

4. Datasets and performance evaluation

It goes without saying that the design of experiments is a critical part of academic research. In this section, the datasets and the evaluation criteria of human motion prediction are expatiated.

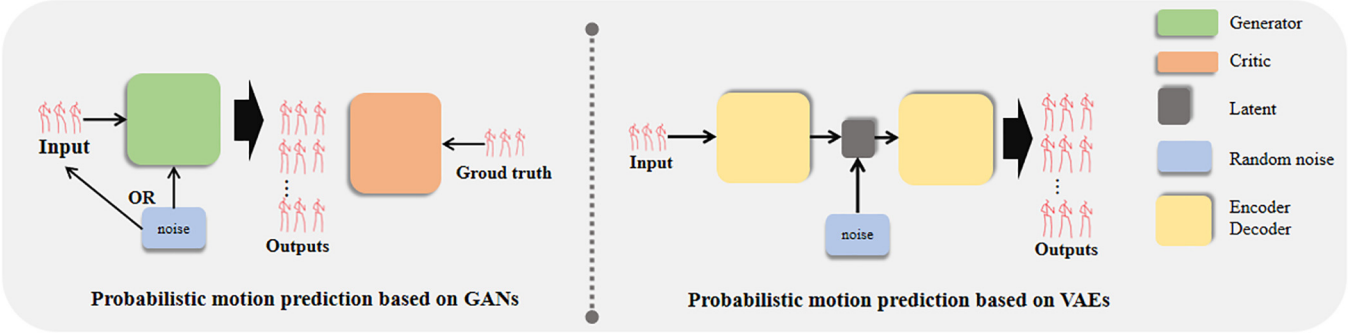


Fig. 8. The basic probabilistic motion prediction structure. Structurally, there are two kinds of frameworks, GANs structure [7,40], and VAEs structure [106].

4.1. Datasets

Datasets are a pivotal part of algorithm research in different fields. Generally, they play a principal role that facilitates the learning of networks and measures the performance as a common ground. Apart from this, the improved quality of datasets has also driven the field to be more valuable and challenging consequences. In particular, deep learning has drawn much more attention recently and it is helpful in part of the huge amount of data. Hence, more and more datasets are established to tackle the problems. For human motion prediction, seven standard datasets are applied to assist learning, as shown in Table 4. They are Human 3.6 M (H36M) [34], CMU motion capture (CMU Mocap), 3D Poses in the Wild (3DPW) [70], The Archive of Motion Capture as Surface Shapes (AMASS) [63], G3D [8], NTU RGB + D [89], Filtered NTU RGB + D (FNTU) [54], Whole-Body Human Motion (WBHM) [65]. The above-mentioned datasets will be explicated respectively in the following part.

Human 3.6 M(H36M) H36M public dataset records human motion data, which includes 3D human poses and their corresponding images with 5 females and 6 males. It totally contains 3.6 Million data recorded by a vicon motion capture system with 4 different viewpoints. These poses involve 15 different scenarios of complex actions, which include directions, discussion, eating, greeting, phoning, posing, purchases, sitting, smoking, taking photo, waiting, walking, walking dog, and walking together. In addition, each scenario includes many types of asymmetries, e.g. walking with a hand in a pocket, walking with a bag on the shoulder. The pose parametrizations comprise joint positions and joint angle skeleton representations and a full skeleton consists of 32 skeletal joints. Experimentally, researchers always divided these poses into 7 subjects (S1, S5, S6, S7, S8, S9, S11) and removed duplicate points of the human pose and 25 points are retained. A down-sampling is applied to set 25 frames per second (FPS). Datasets have been publicly available athttps://vision.imar.ro/human3.6m.

CMU motion capture (CMU) In 2003, Carnegie Mellon University released a public dataset CMU recorded by 12 infrared cameras. There are 41 markers taped on the human body. 144 different subjects are embodied into this dataset, such as basketball, basketball-signal, directing-traffic, jumping, running, soccer, walking, wash-window, and so on. 38 joints make up the parameterized human pose. In general, these samples are split into training set and test set in the experiment. The sequences are down sampled to meet the frame rate of 25fps. This dataset has been publicly available at http://mocap.cs.cmu.edu/.

3D Poses in the Wild (3DPW) 3DPW dataset is mainly presented for wild scenes. It is a kind of large-scale publicly available dataset, which contains 60 video sequences, more than 51, 000 indoor or outdoor poses. This dataset is recorded by a hand-held smartphone camera or IMU. Generally, two actors equipped the IMU to act different actions, such as shopping, doing sports, hugging, discussing, capturing selfies, riding bus, playing guitar, relaxing. There were totally 7 actors with 18 clothing styles. Each pose consists of 17 joints. The frame rate is 30fps. Datasets have been publicly available athttp://virtualhumans.mpi-inf.mpg.de/3DPW.

G3D It is a gaming dataset collected with Microsoft Kinect device and Windows SDK. It provides real-time action recognition in gaming containing synchronized video, depth, and skeleton data. In total, there are 210 samples and 10 subjects performing 20 gaming actions. These actions include punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing backhand, tennis serve, throw bowling ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap, and clap. Each pose is made up of 20 defined joints. The frame rate is 30fps. This dataset has been publicly available at http://dipersec.king.ac.uk/G3D/.

The Archive of Motion Capture as Surface Shapes (AMASS) AMASS dataset is a large and public dataset for parameterized human motions. It contains 15 different optical markers to record the human body. It entirely includes 344 subjects and 11265 motions within 42 h. Habitually, 52 joints are utilized to represent a human pose. Datasets have been publicly available at https://amass.is.tue.mpg.de/.

NTU RGB-D As a large-scale action recognition dataset, the NTU RGB-D records a lot of the corresponding RGB videos, depth map sequences, 3D skeletal data, and infrared videos by three Kinects from different viewpoints. 60 action classes, 56,880 actions, 40 different subjects, and 4 Million frames are recorded by the Microsoft Kinect API. Each action is made up of 25 major joints. This dataset has been publicly available athttp://rose1.ntu.edu.sg/datasets/actionrecognition.asp.

Filtered NTU RGB + D (FNTU) FNTU is closely connected with NTU. It is specialized for human motion task due to the noisy of the human skeletal data. These noisy skeletal sequences are not suitable for pose prediction. FNTU filters the mutual actions and selects the relative forward skeleton of the human body. This dataset is composed of 18,102 samples, among which 12,001 samples are selected for training and the rest for testing. This dataset has been publicly available at https://drive.google.com/drive/folders/ 1bqNyIk2O 0NIf5Hv%202sMfwsuPjwbpZK-n5.

Whole-Body Human Motion (WBHM) WBHM is a large-scale publicly available whole-body dataset containing 3D raw data of multiple individuals and objects. It is constituted by captured raw motion data as well as the corresponding post-processed motion. This database serves as a key element for a broad variety of research questions related e.g. to human motion analysis, imitation learning, action recognition and motion generation in robotics. The motion database is comprised of motion data of a total run length of 7.68 h. 43 different subjects (31 males and 12 females) and 41 different objects have been included into the database. 289 motion capture experiments, defined as a combination of subject and motion type, have been conducted in our motion capture lab. The human pose consists of 56 joints. Datasets have been avail-

Table 4
Datasets table. We collected all the datasets used in relevant papers and explained some of their attributes.

Datasets	Year	Location	Number of Joints	FPS	Sensors
H36M [34]	2014	indoor	32	25	10 Vicon T40
CMU [18]	2003	indoor	38	25	12 infrared cameras
3DPW [70]	2018	outdoor	17	30	A hand-held smartphone camera
G3D [8]	2012	indoor	20	30	The Microsoft Kinect device
AMASS [63]	2019	indoor	52	25	OptiTrack mocap system
NTURGB-D [89]	2016	indoor, outdoor	25	25	Kinect V2
FNTU [54]	2019	indoor	25	25	Kinect V2
WBHM [65]	2015	indoor	56	25	Vicon MX

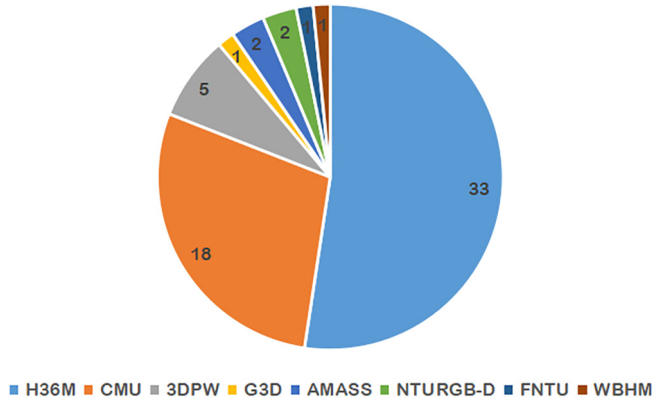


Fig. 9. The datasets are used in this domain.

able athttps://motion-database.humanoids.kit.edu/anthropometrictable/.

As shown in Fig. 9, H36M dataset is the most widely used in human motion prediction. It contains a large amount of data and a wide variety of actions. Since it was used by researchers early on, H36M has been used extensively for comparison. However, other datasets are still rarely used by researchers. This shows the broad development space in this field.

4.2. Performance evaluation

It is difficult to predict future human motion, because of the high dimension and stochastic nature of human motion. In this case, a momentous challenge for human motion prediction is how to evaluate the prediction performance, which can measure the similarity between the predicted and the actual motion, and further compare with other methods. Different classes of tasks require different evaluations. To sum up, there are mainly two metrics in our tasks: *Geometric accuracy metrics* and *Probabilistic accuracy metrics*.

4.2.1. Geometric accuracy metrics

In most circumstances, Geometric accuracy metrics are used in academic researches or application domains. To conclude, accurate metrics can be classified into four kinds of methods: Mean Angle Error (MAE), Mean Per Joint Position Error (MPJPE), Average Displacement Error (ADE), Final Displacement Error (FDE).

Mean Angle Error (MAE) MAE is an immensely popular metric for human motion prediction. When given the Euler angles $\hat{a}_{n,k}$, MAE is utilized as metrics. In terms of form, the MAE can be described as follow:

$$MAE = (N_a K)^{-1} \sum_{n=1}^{N_a} \sum_{k=1}^{K} |\hat{a}_{k,n} - a_{k,n}| \tag{4}$$

where $\hat{a}_{k,n}$ denotes the predicted k^{th} angle in frame n and $x_{k,n}$ is the ground truth.

Mean Per Joint Position Error (MPJPE) 3D coordinates joints are the original representation for human body. MPJPE is also a widely accepted metric. Formally, the MPJPE can be described as follow:

$$MPJPE = \frac{1}{N_j K} \sum_{n=1}^{N_j} \sum_{k=1}^{K} ||\hat{j}_{k,n} - j_{k,n}||^2 \tag{5}$$

where $\hat{j}_{k,n} \in R^3$ is the predicted j^{th} 3D joint position in frame n and $j_{k,n}$ is the ground truth.

Normalized power spectrum similarity (NPSS)

NPSS is proposed to evaluate the long-term predictive ability of motion synthesis models, complementing the popular mean-squared error (MSE) measurement of Euler joint angles over time. NPSS is meant to complement MSE by addressing some of its drawbacks for application as a quantitative evaluation metric for long-term synthesis. Formally, the NPSS can be described as follow:

$$NPSS = \frac{\sum_i \sum_j p_{i,j} * end_{i,j}}{\sum_i \sum_j p_{i,j}} min_{x \in X} ||\hat{x} - x|| \tag{6}$$

where $p_{i,j}$ is the total power of i^{th} feature in j^{th} sequence.

Average Displacement Error (ADE) ADE is utilized to average L_2 distance over all time steps between the ground truth motion $\hat{x}$ and the closest sample. The formula is described as follow:

$$ADE = \frac{1}{T} min_{x \in X} ||\hat{x} - x|| \tag{7}$$

Final Displacement Error (FDE) FDE utilizes L_2 distance between the final ground truth pose x^T and the closest sample's final pose, which is computed as min $x \in X$. The formula is described as follow:

$$FDE = ||\hat{x}^T - x^T|| \tag{8}$$

4.2.2. Probabilistic accuracy metrics

For probabilistic accuracy metrics, [7] utilizes a discriminator network, whose sole purpose is to learn the probability that whether a given sequence is a valid human motion. In [106], Average Pairwise Distance (APD) is used to measure sample diversity. APD average L2 distance between all pairs of motion samples to measure diversity within samples, which is computed as:

$$APD = \frac{1}{(K-1)K} \sum_{i=1}^{K} \sum_{j \neq k}^{K} ||x_i - x_j|| \tag{9}$$

where K is the number of joint and x is the motion sample.

5. Comparison of different methods

In this section, the comparison of different methods. Human motion prediction is conventionally classified into short-term prediction and long-term prediction. Short-term prediction is less than 400 ms and long-term prediction is between 400 ms and 1,000 ms. On this basis, the comparison of performance is drawnamong all the methods for human motion prediction in these years. These models are: LSTM-3LR [21], ERD [21], Res-GRU [71], SRNN [36], DAE-LSTM [25], TE [9], HP-GAN [7], C-seq2seq [45], AGED [28], PAML [29], MHU [93], QuaterNet [79], BiHMP-GAN [40], SkelNet [30], CHA [49], DSR [96], VGRU [27], HMR [60], STMI-GAN [86], LTD [67], RNN-SPL [3], GAN-poser [37], At-seq2seq [87], ARNet [12], Mix-and-Match [4], C-RNN [14], MGCN [110], LDR [17], Dlow [106], HRI [66], LPJP [11], MoPredNet [107], MMA [68], AM-GAN [56], NAT [44], Q-DCRN [79], TCGAN [16], TrajectoryCNN [53], DA-GAN [48], MT-GCN [15], POTR [72], LMC [110], PVRNN [94], MST-GNN [47], SGAN [62], JDM [91].

Short-term and long-term human motion prediction In order to provide a performance comparison for different human motion prediction algorithms, we select the most widely used benchmark dataset: H36M. For human motion prediction, it is always classified into short-term prediction (less than 400 ms) and long-term prediction (400–1,000 ms). For demonstration purposes, we show representative short-term prediction (320 ms) and long-term prediction (1,000 ms) separately for comparison. Then,

Table 5
Performance evaluation (in MAE) on H36M dataset for both short-term and long-term human motion prediction.

Types	Directions		Discussion		Eating		Greeting		Phoning		Photo		Posing		Purchases									
Terms	short		long		short		long		short		long		short		long		short		long		short		long	
LSTM-3LR [21]	–		–		2.25		2.93		1.35		3.42		–		–		–		–		–		–	
ERD [21]	–		–		2.68		2.92		1.66		2.41		–		–		–		–		–		–	
SRNN [36]	–		–		1.83		–		1.35		–		–		–		–		–		–		–	
Res-GRU [71]	1.27		1.59		1.45		1.86		0.92		1.34		2.19		2.03		0.88		1.89		1.64		2.56	
DAE-LSTM [25]	–		–		1.53		1.73		1.86		2.01		–		–		–		–		–		–	
TE [9]	–		–		0.18		0.22		0.17		0.26		–		–		–		–		–		–	
HP-GAN [7]	–		–		0.91		1.77		0.70		1.20		–		–		–		–		–		–	
C-seq2seq [45]	0.80		1.45		0.94		1.86		0.58		1.24		1.21		1.72		1.51		1.81		–		–	
AGED [28]	0.63		–		0.76		1.30		0.51		0.93		1.30		–		0.50		–		0.81		–	
PAML [29]	–		–		–		–		–		–		–		–		–		–		–		–	
MHU [93]	–		–		0.93		1.88		–		–		1.27		1.87		–		0.84		1.35		1.22	
QuaterNet [79]	–		–		0.85		–		0.58		–		–		–		–		–		–		–	
BiHMP-GAN [40]	–		–		0.91		1.77		0.54		1.20		–		–		–		–		–		–	
SkelNet [30]	–		–		0.90		–		0.55		–		–		–		–		–		–		–	
CHA [49]	0.79		–		0.85		–		0.53		–		1.28		–		1.51		–		0.80		-	
DSR [96]	–		–		0.91		–		0.57		–		–		–		–		–		–		–	
VGRU [27]	–		–		0.95		–		0.64		–		–		–		–		–		–		–	
HMR [60]	–		–		0.83		1.72		–		–		1.25		1.73		–		–		–		–	
LTD [67]	0.71		–		0.77		–		0.50		–		0.58		–		1.35		0.58		1.01		–	
RNN-SPL [3]	–		–		0.95		–		0.55		–		–		–		–		–		–		–	
GAN-poser [37]	–		–		1.11		–		0.97		–		–		–		–		–		–		–	
AT-seq2seq [87]	–		–		0.56		–		0.49		–		–		–		–		–		–		–	
ARNet [12]	0.65		–		0.81		–		0.49		–		0.90		–		1.28		–		0.55		–	
Mix-and-Match [4]	–		–		0.83		1.30		0.41		0.91		–		–		–		–		–		–	
MGCN [110]	0.65		–		0.92		–		0.49		–		0.94		–		1.29		–		0.58		–	
LDR [17]	0.59		0.95		0.72		0.84		0.49		0.97		0.87		1.33		0.63		1.33		0.45		1.20	
HRI [66]	0.69		1.27		0.87		1.63		0.60		1.10		0.95		1.57		1.31		1.68		0.58		1.08	
LPJP [11]	–		–		0.62		–		0.50		–		0.94		–		–		–		–		–	
MoPredNet [107]	–		–		0.90		1.17		0.45		0.73		–		–		–		–		–		–	
MMA [68]	0.63		1.30		0.77		1.71		0.47		1.09		0.92		1.56		1.31		1.67		0.57		1.57	
AM-GAN [56]	0.75		1.41		0.81		1.58		0.49		1.15		1.24		1.70		14.8		1.79		0.83		1.40	
NAT [44]	0.58		1.15		0.72		1.22		0.48		0.87		0.79		1.24		1.15		1.40		0.54		1.04	
Q-DCRN [73]	0.62		–		0.98		–		0.56		–		1.11		–		1.39		–		0.64		–	
TCGAN [16]	0.59		–		0.77		–		0.46		–		0.98		–		0.47		–		0.73		–	
yCNN [53]	–		–		–		–		–		–		–		–		–		–		–		–	
DA-GAN [48]	0.56		–		0.77		–		0.49		–		0.87		–		0.63		–		0.51		–	
MT-GCN [15]	–		–		–		–		–		–		–		–		–		–		–		–	
POTR [72]	0.79		–		0.85		–		0.53		–		1.17		–		1.50		–		0.71		–	
LMC [110]	–		–		–		–		–		–		–		–		–		–		–		–	
PVRED [94]	0.66		–		0.63		–		0.54		–		1.00		–		1.26		–		0.66		–	
MST-GNN [47]	0.60		–		0.83		–		0.47		–		0.95		–		1.25		–		0.57		–	
SGAN [62]	–		–		–		–		–		–		–		–		–		–		–		–	
JDM [91]	–		–		–		–		–		–		–		–		–		–		–		–	

we chose the two commonly used evaluation indicators as evaluation standards: Mean Angle Error (MAE), Mean Per Joint Position Error (MPJPE). We compare all actions from the H36M.

As is shown in Table 5 and Table 6, the performance is evaluated using MAE on H36M. All the actions in the datasets are listed. The results are derived from published papers. In the Table 7 and Table 8, the performance is evaluated using MPJPE on H36M. We also list all the actions in the datasets. The results are also derived from published papers.

From the distribution of the results, it is not difficult to note that MAE has been the dominant evaluation. MPJPE has been popular in the last year due to the widespread use of convolution methods. From the vertical quantitative comparison in each table,it is found that the methods in recent years possess tremendous progress.The methods for evaluating performance using MPJPE have become active in recent years. This indicates that researchers are demanding a more comprehensive look on algorithm performance. The success rate of the prediction method is also rising year by year. The success of these methods, especially the recent approaches [47,91], proves that reasonable prior knowledge and kinematic constraints facilitate human motion prediction.

From the horizontal quantitative comparison in each table, it can be seen that there remain many difficulties to conquer in terms of long-term prediction. In contrast, the error rate for long-term predictions is at least twice that of short-term predictions. Improving the accuracy of long-term prediction is a terribly challenging task. For actions, it is still difficult to predict actions that are not cyclical and involve a large scale of movements, e.g. purchases, discussions, etc. In summary, the prediction of long-term, non-periodic and large movements remains a huge challenge. Innovative network structures coupled with valid prior knowledge are more helpful in accomplishing the task.

6. Discussion

In the past few years, striking progress has been made in the development of advanced prediction technology in terms of method diversity, performance, and correlation application scenarios. More issues worthy of our discussion have also emerged. In this section, two questions for the 3D human motion prediction are discussed.

Table 6
Performance evaluation (in MAE) on H36M dataset for both short-term and long-term human motion prediction.

Types	Sitting	SittingDown	Smoking	Waiting	WalkDog	Walking	WalkTogether							
Terms	short	long	short	long	short	long	short	long	short	long	short	long	short	long
LSTM-3LR [21]	–	–	–	–	2.04	3.42	–	–	–	–	1.29	2.20	–	–
ERD [21]	–	–	–	–	2.35	3.82	–	–	–	–	1.59	2.38	–	–
SRNN [36]	–	–	–	–	1.94	3.23	–	–	–	–	1.16	2.13	–	–
Res-GRU [71]	1.36	2.14	1.17	2.72	1.31	1.83	1.55	2.34	0.95	1.86	1.31	1.14	1.25	1.42
DAE-LSTM [25]	–	–	–	–	1.15	1.77	–	–	–	–	1.39	1.39	–	–
TE [9]	–	–	–	–	0.19	0.27	–	–	–	–	0.20	0.24	–	–
HP-GAN [7]	–	–	–	–	0.91	1.11	–	–	–	–	–	0.63	0.85	–	–
C-seq2seq [45]	1.02	1.67	1.16	2.06	0.96	1.62	1.09	2.50	1.32	1.92	0.68	0.92	0.71	1.28
AGED [28]	1.05	–	0.98	–	0.82	1.21	0.55	–	1.15	–	0.55	0.91	0.56	–
PAML [29]	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MHU [93]	–	–	–	–	–	–	–	–	1.21	1.90	0.69	1.06	–	–
QuaterNet [79]	–	–	–	–	0.93	–	–	–	–	–	0.56	–	–	–
BiHMP-GAN [40]	–	–	–	–	0.91	1.11	–	–	–	–	0.67	0.85	–	–
SkelNet [30]	–	–	–	–	0.91	–	–	–	–	–	–	–	–	–
CHA [49]	1.03	–	1.15	–	0.89	–	1.12	–	0.75	–	1.16	–	0.75	–
DSR [96]	–	–	–	–	0.90	–	–	–	–	–	0.69	–	–	–
VGRU [27]	–	–	–	–	0.85	–	–	–	–	–	0.64	–	–	–
HMR [60]	–	–	–	–	–	–	–	–	1.20	1.84	–	–	–	–
LTD [67]	−0.80	–	0.90	–	0.86	–	0.91	–	1.12	–	0.49	–	0.52	–
RNN-SPL [3]	–	–	–	–	0.96	–	–	–	–	–	0.67	–	–	–
GAN-poser [37]	–	–	–	–	0.87	–	–	–	–	–	0.84	–	–	–
AT-seq2seq [87]	–	–	–	–	0.49	–	–	–	–	–	0.51	–	–	–
ARNet [12]	0.80	–	0.87	–	0.86	–	0.90	–	1.11	–	0.49	–	0.53	–
Mix-and-Match [4]	–	–	–	–	0.79	1.25	–	–	–	–	0.56	0.68	–	–
MGCN [110]	0.76	–	0.93	–	0.81	–	0.88	–	1.16	–	0.49	–	0.50	–
LDR [17]	0.69	1.38	0.87	1.42	0.79	1.08	0.84	1.21	0.93	1.38	0.46	0.71	0.49	1.38
HRI [66]	0.83	1.55	0.92	1.70	0.86	2.30	0.92	1.82	1.05	0.64	0.46	1.16	1.07	2.22
LPJP [11]	–	–	–	–	0.85	–	–	–	–	–	0.51	–	–	–
MoPredNet [107]	–	–	–	–	0.53	1.88	–	–	–	–	0.43	0.83	–	–
MMA [68]	0.81	1.54	0.91	1.68	0.86	1.57	0.90	2.27	1.05	1.81	0.48	0.63	0.50	1.18
AM-GAN [56]	0.98	1.60	1.08	1.85	0.88	1.10	1.10	2.15	1.18	1.80	0.62	0.84	0.71	1.19
NAT [44]	0.84	1.46	0.94	1.58	0.81	1.26	0.79	1.58	0.86	1.44	0.45	0.50	0.51	1.07
Q-DCRN [73]	0.88	–	1.03	–	0.87	–	0.99	–	1.10	–	0.56	–	0.57	–
TCGAN [16]	0.79	–	0.74	–	0.67	–	0.74	–	1.01	–	0.52	–	0.51	–
TrajectoryCNN [53]	–	–	–	–	–	–	–	–	–	–	–	–	–	–
DA-GAN [48]	0.73	–	0.86	–	0.75	–	0.85	–	1.02	–	0.44	–	0.46	–
MT-GCN [15]	–	–	–	–	–	–	–	–	–	–	–	–	–	–
POTR [72]	0.92	–	1.00	–	0.84	–	1.14	–	1.21	–	0.62	–	0.63	–
LMC [110]	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PVRED [94]	0.84	–	1.03	–	0.81	–	0.93	–	1.13	–	0.54	–	0.59	–
MST-GNN [47]	0.75	–	0.88	–	0.78	– 0.88	–	1.11	–	0.49	–	0.51	–	
SGAN [62]	–	–	–	–	–	–	–	–	–	–	–	–	–	–
JDM [91]	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 7
Performance evaluation (in MPJPE) on H36M dataset for both short-term and long-term human motion prediction.

Types	Directions		Discussion		Eating		Greeting		Phoning		Photo		Posing		Purchases							
Terms	short		long		short		long		short		long		short		long		short		long		short	
LSTM-3LR [21]	46.6	135.1	85.5	135.1	78.1	121.1	73.1	177.6	68.8	133.3	71.1	155.6	130.1	176.5	88.0	142.9						
ERD [21]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
SRNN [36]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
Res-GRU [71]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
DAE-LSTM [25]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
TE [9]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
HP-GAN [7]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
C-seq2seq [45]	44.5	78.3	35.2	67.4	37.2	53.8	66.3	129.7	29.6	116.4	33.1	85.8	58.3	113.7	62.8	112.6						
AGED [28]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
PAML [29]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
MHU [93]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
QuaterNet [79]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
BiHMP-GAN [40]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
SkelNet [30]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
CHA [49]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
DSR [96]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
VGRU [27]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
HMR [60]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
LTD [67]	48.2	89.1	39.6	78.5	25.3	44.3	74.2	148.4	37.9	94.3	38.2	125.7	66.2	143.5	64.4	127.2						
RNN-SPL [3]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
GAN-poser [37]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
AT-seq2seq [87]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
ARNet [12]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
Mix-and-Match [4]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
MGCN [110]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
LDR [17]	47.2	106.5	46.3	144.6	24.8	43.1	47.2	127.3	87.9	143.2	25.4	45.8	31.2	112.6								
HRI [66]	44.5	106.5	52.1	119.8	28.7	75.7	63.8	105.0	39.0	115.9	40.7	178.2	58.5	138.8	60.4	134.2						
LPJP [11]	–	–	48.0	–	37.4	–	64.5	–	–	–	–	–	–	–	–	–						
MoPredNet [107]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
MMA [68]	50.6	105.7	37.7	117.5	45.2	73.7	68.2	130.7	37.7	104.6	38.0	115.2	62.2	172.9	58.4	115.0						
AM-GAN [56]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
NAT [44]	30.7	–	41.5	–	33.8	–	46.0	–	35.5	–	35.0	–	50.7	–	50.1	–						
Q-DCRN [73]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
TCGAN [16]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
TrajectoryCNN [53]	50.2	104.2	41.3	103.0	37.0	71.5	67.3	84.3	37.0	113.5	36.2	86.6	62.9	210.9	64.3	115.5						
DA-GAN [48]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
MT-GCN [15]	85.3	147.9	70.2	131.2	53.0	89.2	77.8	151.2	44.2	139.9	49.6	137.1	70.3	176.4	74.2	149.3						
POTR [72]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
LMC [110]	–	–	48.9	–	37.7	–	68.4	–	–	–	–	–	–	–	–	–						
PVRED [94]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
MST-GNN [47]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–						
SGAN [62]	42.5	85.8	31.1	59.8	24.0	41.5	40.0	92.1	28.9	80.6	25.2	85.5	29.9	80.6	51.2	102.2						
JDM [91]	37.6	72.3	–	–	21.0	52.5	46.8	83.1	24.3	102.7	23.8	82.4	49.2	123.8	52.3	112.2						

Table 8
Performance evaluation (in MPJPE) on H36M dataset for both short-term and long-term human motion prediction.

Types	Sitting		SittingDown		Smoking		Waiting		WalkDog		Walking		WalkTogether		Average	
Terms	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long
LSTM-3LR [21]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ERD [21]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SRNN [36]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Res-GRU [71]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
DAE-LSTM [25]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TE [9]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
HP-GAN [7]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
C-seq2seq [45]	47.2	106.5	46.3	144.6	24.8	43.1	47.2	127.3	87.9	143.2	25.4	45.8	31.2	79.2	57.4	124.1
AGED [28]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PAML [29]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MHU [93]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
QuaterNet [79]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
BiHMP-GAN [40]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SkelNet [30]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
CHA [49]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
DSR [96]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
VGRU [27]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
HMR [60]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
LTD [67]	24.6	119.8	56.4	163.9	25.3	44.3	57.5	157.2	102.2	185.4	29.2	50.9	35.3	102.4	57.4	154.1
RNN-SPL [3]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
GAN-poser [37]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
AT-seq2seq [87]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ARNet [12]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Mix-and-Match [4]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MGCN [110]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
LDR [17]	44.5	78.3	35.2	67.4	37.2	53.8	66.3	129.7	33.1	85.8	29.6	116.4	58.3	133.7	50.1	97.6
HRI [66]	44.3	115.9	59.1	143.6	29.9	69.5	43.4	108.2	73.3	146.9	34.2	58.1	35.1	64.9	46.9	116.2
LPJP [11]	–	–	–	–	24.1	–	–	–	–	29.1	–	–	–	–	–	–
MoPredNet [107]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MMA [68]	53.8	115.0	54.6	141.8	25.4	68.7	55.7	105.1	100.3	141.4	25.5	57.1	33.2	63.2	49.6	114.1
AM-GAN [56]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
NAT [44]	44.0	–	34.8	–	23.8	–	50.5	–	55.3	–	23.3	–	28.0	–	34.2	–
Q-DCRN [73]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TCGAN [16]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TrajectoryCNN [53]	49.4	116.3	55.1	123.8	23.7	58.7	53.4	165.9	98.1	181.3	30.0	46.4	33.9	77.3	50.1	114.1
DA-GAN [48]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MT-GCN [15]	56.7	131.8	60.2	158.5	43.9	85.5	57.7	130.9	92.5	176.3	45.6	78.3	43.1	80.2	61.0	133.9
POTR [72]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
LMC [110]	–	–	–	–	22.6	–	–	–	–	–	27.1	–	–	–	–	–
PVRED [94]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
MST-GNN [47]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SGAN [62]	40.2	99.8	38.2	113.2	40.0	92.1	69.3	131.5	22.6	39.8	–	–	28.8	75.2	42.1	102.1
JDM [91]	36.9	93.6	32.7	101.9	–	–	37.5	95.6	62.3	126.1	–	–	–	–	35.7	97.2

Q1: Are evaluation criteria effective to measure performance? This will be discussed in Section 6.1 by reviewing the existing benchmarking practices including metrics, experiments, and datasets.

Q2: How should relevant research be further carried out in the future? This will be discussed in Section 6.3 by outlining open challenges and potential future research directions.

6.1. Benchmarking

Evaluating the performance of motion prediction algorithms requires the selection of appropriate experimental scenarios and accuracy metrics, as well as the robustness of research methods and a large number of diverse datasets.

Existing datasets, summarized in Section 4.1, are varied and contain a wealth of information. However, datasets are missed when regarding the interaction between humans and the environment. The environment is a highly decisive factor to human motion. Further, all of the motion scenes are solitary, which brings some limitations for researching multiple persons' motions. Furthermore, existing datasets tend to be more deterministic prediction than probabilistic prediction.

As is shown in Section 4.2, the measurement of human motion prediction has revealed a problem that needs to be improved. *For the accuracy metrics*, there is a lack of specific measurements for long-term prediction. Only a measurement strategy to this problem is presented in the method. Now that we've been trying to divide forecasts into long-term and short-term, it is a bit inappropriate to merely use the same metric. *For the probabilistic metrics*, it reflects the stochastic nature of human motion. Although more and more colleagues have been engaging in research in this area, there are no effective metrics for probabilistic prediction. Moreover, current methods rely more on the comparison of visual results.

6.2. Future directions

To solve complex problems with simple methods is the pursuit of all research. Vague understandings of the problems intensify the intricacy of problems. Nowadays, more and more advanced techniques in different fields are being applied to human motion prediction. Furthermore, some methods carry a novel understanding of human motion, which promotes human motion prediction. With these trends in mind, we want to propose several potential and interesting directions for future research.

Prediction methods are driven by the intrinsic logic from researchers' understanding of human motion. As is shown in Fig. 3, Our proposed directions are inseparable from the pyramid representation of human motion prediction (*Human pose representation*, *Network structure design*, and *Prediction target*).

For example, the most efficient RNNs structure is utilized to predict future poses, because human motion prediction is a sequential seq2seq task. Firstly, the form of data (sequential data) is differentiated. Then, a pertinent model will be selected to deal with the data. Finally, the output will in return test our method. At the same time, it becomes known to us that human motion is different from the general seq2seq task because it is subject to kinematic constraints. Therefore, many kinematic methods were developed, such as Graph modeling, kinematics tree modeling to represent the human pose. Further, the new raised mathematical thinking that can map the human body into other spaces will become another solid approach. Furthermore, more can be learnt from other related disciplines, which benefits our understanding of the underlying logic of the problem. Such as, quantum physics, mechanics, anatomy, kinematics, etc. Owning to these logics, the human pose representations will become more diverse and multi-scaled. Correspondingly, matched network is naturally sought or designed to learn about them. Likewise, our prediction target also influences the network. In summary, the above perspectives of thinking inspires a three-phase idea for the future directions: *Data phase*, *Network phase*, and *Prediction result phase*.

Data phase In the data phase, there are mostly two directions: *richer datasets* and *effective pose representation*. *For the former*, the current datasets are only captured single pose data. However, these datasets disregard the external stimuli (physical environment, passers-by) in the real world. This may cause limitations to the application of the methods using these datasets, such as the influence of interaction between human and environment (or others) on motion prediction. Moreover, existing datasets are all targeted at the deterministic prediction, which are inconsistent with objective facts. Therefore, a chief direction is to enrich the datasets, In which three means can be considered: (*i*) Adding interactive physical environment elements, such as desk, chair, and door. (*ii*) Capturing multiple poses simultaneously. (*iii*) Establishing probabilistic motion prediction datasets. *For latter*, Our mindsets should not be confined to apparent ways, like kinematics and anatomy, but should be expanded to different fields, such as advanced mathematics, quantum mechanics, and so on.

Network phase The network is vital for any deep learning work. The objective is to extract and process powerful features for the task. Yet, the human motion prediction task possesses a prior kinematic constraint. Most of the presented methods for human motion prediction are designed for specific tasks or scenarios. These make the network less robust. Practically speaking, the applicability of the network will thus be hindered. Therefore, a robust network needs to be proposed, so that it can adapt to different actions or situations. Meanwhile, interpretability is always associated with deep learning. The role of each particular layer cannot be precisely determined. Therefore, there is a lack of systematic guidance when designing networks. Besides, current methods mainly focus on the model structure, failing to notice the importance of network training. An effective training strategy can improve network performance. To sum up, there are three ways for improving the network performance: (*i*) A robust network. (*ii*) An interpretable network. (*iii*) An efficient network training method.

Prediction result phase In the prediction result phase, there are mainly two directions: *Prediction target* and *Evaluation criteria*. For the *former*, human motion prediction is a kind of motion generation, whether probabilistic prediction or deterministic prediction. Beyond this, multiple targets can be produced along with this idea. (*i*) Action transformation. One action can smoothly transfer to another one. (*ii*) Longer-term motion prediction. For existing methods, their long-term prediction is from 400 ms to 1,000 ms. Most of the methods tend to degrade into motionless states or drift away to non–human-like motions for longer-term prediction. Thereby, it is of extraordinary research value. And probabilistic prediction or deterministic prediction is surely worth studying as well. For the *latter*, the existing evaluation criteria mainly evaluate the deterministic prediction. For other means (e.g., probabilistic prediction and action transfer), the researchers only depend on visual results. However, This can only result in the making of an intuitive judgment on the results with a large gap, which is rather unfavorable to the research. Therefore, it is imperative to propose a new evaluation criterion for more diverse researches.

7. Conclusion

A survey of the 3D human motion prediction methods is presented in this paper. The literatures across multiple domains are involved in the survey, and a taxonomy of human motion prediction approaches is presented. The existing 3D human motion pre-

diction tasks are divided into three general categories: *Human pose representation*, *Network structure design*, and *Prediction target*. Accordingly, the existing methods are combed in detail. Moreover, the metrics of evaluating these methods are summarized and extensive experimental results are illustrated. In the end, we retrospect and discussed the existing methods and proposed three potential future research directions. We hope this survey can contribute to the progress of the field.

CRediT authorship contribution statement

Kedi Lyu: Conceptualization, Writing - review & editing. **Haipeng Chen:** Conceptualization, Writing - review & editing. **Zhenguang Liu:** Conceptualization, Writing - review & editing. **Beiqi Zhang:** Conceptualization, Writing - review & editing. **Ruili Wang:** Conceptualization, Writing - review & editing.

Acknowledgement

This research is partly supported by the National Key Research and Development Program of China, Grant No.: 2018YFB0804202, 2018YFB0804203; Regional Joint Fund of NSFC, Grant No.: U19A2057; National Natural Science Foundation of China, Grant No.: 61876070; Jilin Province Science and Technology Development Plan Project, Grant No.: 20190303134SF; the Natural Science Foundation of Zhejiang Province, China (LY18F020008).

References

[1] L. Accardi, A. Boukas, Control of quantum langevin equations, Open Syst. Inf. Dyn. 10 (2003) 89–104, https://doi.org/10.1023/A:1022927426053.
[2] Akhter, I., Sheikh, Y., Khan, S., Kanade, T., 2008. Nonrigid structure from motion in trajectory space, in: Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8–11, 2008, pp. 41–48.https://proceedings.neurips.cc/paper/2008/hash/dc82d632c9fcecb0778afbc7924494a6-Abstract.html.
[3] Aksan, E., Kaufmann, M., Hilliges, O., 2019. Structured prediction helps 3d human motion modelling, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 – November 2, 2019, pp. 7143–7152. DOI: 10.1109/ICCV.2019.00724.
[4] M.S. Aliakbarian, F.S. Saleh, M. Salzmann, L. Petersson, S. Gould, A stochastic conditioning scheme for diverse human motion prediction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020 June 13–19, 2020, Seattle, WA, USA, 2020, pp. 5222–5231, https://doi.org/10.1109/CVPR42600.2020.00527, URL:https://openaccess.thecvf.com/content_CVPR_2020/html/Aliakbarian_A_Stochastic_Conditioning_Scheme_for_Diverse_Human_Motion_Prediction_CVPR_2020_paper.html.
[5] Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. CoRR abs/1803.01271.http://arxiv.org/abs/1803.01271.
[6] F. Baldassarre, D.M. Hurtado, A. Elofsson, H. Azizpour, Graphqa: protein model quality assessment using graph convolutional networks, Bioinform. 37 (2021) 360–366, https://doi.org/10.1093/bioinformatics/btaa714.
[7] Barsoum, E., Kender, J., Liu, Z., 2018. HP-GAN: probabilistic 3d human motion prediction via GAN, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18–22, 2018, pp. 1418–1427.http://openaccess.thecvf.com/content_cvpr_2018_workshops/w29/html/Barsoum_HP-GAN_Probabilistic_3D_CVPR_2018_paper.html, DOI: 10.1109/CVPRW.2018.00191.
[8] V. Bloom, D. Makris, V. Argyriou, G3D: A gaming action dataset and real time action recognition evaluation framework, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops June 16–21, 2012, Providence, RI, USA, 2012, pp. 7–12, https://doi.org/10.1109/CVPRW.2012.6239175.
[9] J. Bütepage, M.J. Black, D. Kragic, H. Kjellström, Deep representation learning for human motion prediction and classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 July 21–26, 2017, Honolulu, HI, USA, 2017, pp. 1591–1599, https://doi.org/10.1109/CVPR.2017.173.
[10] J. Bütepage, H. Kjellström, D. Kragic, Anticipating many futures: Online human motion prediction and generation for human-robot interaction, in: 2018 IEEE International Conference on Robotics and Automation, ICRA 2018 May 21–25, 2018, Brisbane, Australia, 2018, pp. 1–9, https://doi.org/10.1109/ICRA.2018.8460651.
[11] Cai, Y., Huang, L., Wang, Y., Cham, T., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., Shen, X., Liu, D., Liu, J., Magnenat-Thalmann, N., 2020. Learning progressive joint propagation for human motion prediction, in: Computer Vision - ECCV 2020–16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII, pp. 226–242. DOI: 10.1007/978-3-030-58571-6_14.
[12] Chao, X., Bin, Y., Chu, W., Cao, X., Ge, Y., Wang, C., Li, J., Huang, F., Leung, H., 2020. Adversarial refinement network for human motion prediction, in: Computer Vision - ACCV 2020–15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part II, pp. 454–469. DOI: 10.1007/978-3-030-69532-3_28.
[13] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1724–1734. DOI: 10.3115/v1/d14-1179.
[14] Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F., 2020. Context-aware human motion prediction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, pp. 6990–6999.https://openaccess.thecvf.com/content_CVPR_2020/html/Corona_Context-Aware_Human_Motion_Prediction_CVPR_2020_paper.html, DOI: 10.1109/CVPR42600.2020.00702.
[15] Cui, Q., Sun, H., 2021. Towards accurate 3d human motion prediction from incomplete observations, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021, pp. 4801–4810.https://openaccess.thecvf.com/content/CVPR2021/html/Cui_Towards_Accurate_3D_Human_Motion_Prediction_From_Incomplete_Observations_CVPR_2021_paper.html.
[16] Q. Cui, H. Sun, Y. Kong, X. Zhang, Y. Li, Efficient human motion prediction using temporal convolutional generative adversarial network, Inf. Sci. 545 (2021) 427–447, https://doi.org/10.1016/j.ins.2020.08.123.
[17] Cui, Q., Sun, H., Yang, F., 2020. Learning dynamic relationships for 3d human motion prediction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020, pp. 6518–6526.https://openaccess.thecvf.com/content_CVPR_2020/html/Cui_Learning_Dynamic_Relationships_for_3D_Human_Motion_Prediction_CVPR_2020_paper.html, DOI: 10.1109/CVPR42600.2020.00655.
[18] graphics lab motion capture database, C., 2003.http://mocap.cs.cmu.edu.
[19] Z. Ding, C. Yang, Z. Wang, X. Yin, F. Jiang, Online adaptive prediction of human motion intention based on semg, Sensors 21 (2021) 2882, https://doi.org/10.3390/s21082882.
[20] N. Djuric, V. Radosavljevic, H. Cui, T. Nguyen, F.C. Chou, T.H. Lin, N. Singh, J. Schneider, Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2095–2104.
[21] K. Fragkiadaki, S. Levine, P. Felsen, J. Malik, Recurrent network models for human dynamics, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, 2015, pp. 4346–4354, https://doi.org/10.1109/ICCV.2015.494.
[22] R. Fujii, J. Vongkulbhisal, R. Hachiuma, H. Saito, A two-block rnn-based trajectory prediction from incomplete trajectory, IEEE Access 9 (2021) 56140–56151, https://doi.org/10.1109/ACCESS.2021.3072135.
[23] Z. Gao, L. Guo, W. Guan, A. Liu, T. Ren, S. Chen, A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2, IEEE Trans. Image Process. 30 (2021) 767–782, https://doi.org/10.1109/TIP.2020.3038372.
[24] Ge, S., Zhao, S., Gao, X., Li, J., 2019. Fewer-shots and lower-resolutions: Towards ultrafast face recognition in the wild, in: Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21–25, 2019, pp. 229–237. DOI: 10.1145/3343031.3351082.
[25] Ghosh, P., Song, J., Aksan, E., Hilliges, O., 2017. Learning human motion models for long-term predictions, 458–466. DOI: 10.1109/3DV.2017.00059.
[26] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp. 2672–2680.https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.
[27] Gopalakrishnan, A., Mali, A.A., Kifer, D., Giles, C.L., II, A.G.O., 2019. A neural temporal model for human motion prediction, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, pp. 12116–12125.http://openaccess.thecvf.com/content_CVPR_2019/html/Gopalakrishnan_A_Neural_Temporal_Model_for_Human_Motion_Prediction_CVPR_2019_paper.html, DOI: 10.1109/CVPR.2019.01239.
[28] Gui, L., Wang, Y., Liang, X., Moura, J.M.F., 2018a. Adversarial geometry-aware human motion prediction, in: Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IV, pp. 823–842. DOI: 10.1007/978-3-030-01225-0_48.
[29] Gui, L., Wang, Y., Ramanan, D., Moura, J.M.F., 2018b. Few-shot human motion prediction via meta-learning, in: Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VIII, pp. 441–459. DOI: 10.1007/978-3-030-01237-3_27.
[30] Guo, X., Choi, J., 2019. Human motion prediction via learning local structure representations and temporal dependencies, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019,

Honolulu, Hawaii, USA, January 27 – February 1, 2019, pp. 2580–2587. DOI: 10.1609/aaai.v33i01.33012580.
[31] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.
[32] Y. Hu, Reliability analysis of multi-objective spatio-temporal segmentation of human motion in video sequences, Int. J. Distributed Syst. Technol. 12 (2021) 16–29, https://doi.org/10.4018/IJDST.2021010102.
[33] Iida, S., Kimura, R., Cui, H., Hung, P., Utsuro, T., Nagata, M., 2019. A multi-hop attention for RNN based neural machine translation, in: Proceedings of The 8th Workshop on Patent and Scientific Literature Translatio@ MTSummit 2019, Dublin, Ireland, August 20, 2019, pp. 24–31.https://aclanthology.org/W19-7203/.
[34] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1325–1339, https://doi.org/10.1109/TPAMI.2013.248.
[35] Jabbar, A., Li, X., Omar, B., 2022. A survey on generative adversarial networks: Variants, applications, and training. ACM Comput. Surv. 54, 157:1–157:49., 10.1145/3463475.
[36] Jain, A., Zamir, A.R., Savarese, S., Saxena, A., 2016. Structural-rnn: Deep learning on spatio-temporal graphs, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp. 5308–5317, DOI: 10.1109/CVPR.2016.573.
[37] D.K. Jain, M. Zareapoor, R. Jain, A. Kathuria, S. Bachhety, Gan-poser: an improvised bidirectional GAN model for human motion prediction, Neural Comput. Appl. 32 (2020) 14579–14591, https://doi.org/10.1007/s00521-020-04941-4.
[38] H.S. Koppula, A. Saxena, Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation, in: Proceedings of the 30th International Conference on Machine Learning, ICML 2013 USA, 16–21 June 2013, Atlanta, GA, 2013, pp. 792–800.
[39] H.S. Koppula, A. Saxena, Anticipating human activities using object affordances for reactive robotic response, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 14–29, https://doi.org/10.1109/TPAMI.2015.2430335.
[40] Kundu, J.N., Gor, M., Babu, R.V., 2019. Bihmp-gan: Bidirectional 3d human motion prediction GAN, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 – February 1, 2019, pp. 8553–8560, DOI: 10.1609/aaai.v33i01.33018553.
[41] W. Lan, X. Wu, Q. Chen, W. Peng, J. Wang, Y.P. Chen, GANLDA: graph attention network for lncrna-disease associations prediction, Neurocomputing 469 (2022) 384–393, https://doi.org/10.1016/j.neucom.2020.09.094.
[42] P.A. Lasota, T. Fong, J.A. Shah, A survey of methods for safe human-robot interaction, Found. Trends Robotics 5 (2017) 261–349, https://doi.org/10.1561/2300000052.
[43] Lee, N., Kitani, K.M., 2016. Predicting wide receiver trajectories in american football, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7–10, 2016, pp. 1–9. DOI: 10.1109/WACV.2016.7477732.
[44] B. Li, J. Tian, Z. Zhang, H. Feng, X. Li, Multitask non-autoregressive model for human motion prediction, IEEE Trans. Image Process. 30 (2021) 2562–2574, https://doi.org/10.1109/TIP.2020.3038362.
[45] Li, C., Zhang, Z., Lee, W.S., Lee, G.H., 2018. Convolutional sequence to sequence model for human dynamics, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, pp. 5226–5234.http://openaccess.thecvf.com/content_cvpr_2018/html/Li_Convolutional_Sequence_to_CVPR_2018_paper.html, DOI: 10.1109/CVPR.2018.00548.
[46] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020 June 13–19, 2020, Seattle, WA, USA, 2020, pp. 211–220, https://doi.org/10.1109/CVPR42600.2020.00029, URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Dynamic_Multiscale_Graph_Neural_Networks_for_3D_Skeleton_Based_Human_CVPR_2020_paper.html.
[47] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, Q. Tian, Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction, IEEE Trans. Image Process. 30 (2021) 7760–7775, https://doi.org/10.1109/TIP.2021.3108708.
[48] Li, Q., Chalvatzaki, G., Peters, J., Wang, Y., 2021c. Directed acyclic graph neural network for human motion prediction, in: IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 – June 5, 2021, pp. 3197–3204. DOI: 10.1109/ICRA48506.2021.9561540.
[49] Y. Li, Z. Wang, X. Yang, M. Wang, S.I. Poiana, E. Chaudhry, J. Zhang, Efficient convolutional hierarchical autoencoder for human motion prediction, Vis. Comput. 35 (2019) 1143–1156, https://doi.org/10.1007/s00371-019-01692-9.
[50] H. Liu, L. Wang, Human motion prediction for human-robot collaboration, J. Manuf. Syst. 44 (2017) 287–294.
[51] Q. Liu, Z. Liu, B. Xiong, W. Xu, Y. Liu, Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function, Adv. Eng. Informatics 49 (2021), 101360, https://doi.org/10.1016/j.aei.2021.101360.
[52] R. Liu, C. Liu, Human motion prediction using adaptable recurrent neural networks and inverse kinematics, IEEE Control. Syst. Lett. 5 (2021) 1651–1656, https://doi.org/10.1109/LCSYS.2020.3042609.
[53] X. Liu, J. Yin, J. Li, P. Ding, J. Liu, H. Liu, Trajectorycnn: A new spatio-temporal feature learning network for human motion prediction, IEEE Trans. Circuits Syst. Video Technol. 31 (2021) 2133–2146, https://doi.org/10.1109/TCSVT.2020.3021409.
[54] Liu, X., Yin, J., Liu, H., Yin, Y., 2019a. Pisep2: Pseudo image sequence evolution based 3d pose prediction. CoRR abs/1909.01818.http://arxiv.org/abs/1909.01818.
[55] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, J. Chen, Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing, Procedia CIRP 83 (2019) 272–278.
[56] Liu, Z., Lyu, K., Wu, S., Chen, H., Hao, Y., Ji, S., 2021c. Aggregated multi-gans for controlled 3d human motion prediction, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021, pp. 2225–2232.https://ojs.aaai.org/index.php/AAAI/article/view/16321.
[57] Z. Liu, P. Su, S. Wu, X. Shen, H. Chen, Y. Hao, M. Wang, Motion prediction using trajectory cues, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13299–13308.
[58] Z. Liu, X. Wang, Y. Cai, W. Xu, Q. Liu, Z. Zhou, D.T. Pham, Dynamic risk assessment and active response strategy for industrial human-robot collaboration, Comput. Ind. Eng. 141 (2020), 106302, https://doi.org/10.1016/j.cie.2020.106302.
[59] Z. Liu, S. Wu, S. Jin, S. Ji, Q. Liu, S. Lu, L. Cheng, Investigating pose representations and motion contexts modeling for 3d motion prediction, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 1–16 (2022), https://doi.org/10.1109/TPAMI.2021.3139918.
[60] Liu, Z., Wu, S., Jin, S., Liu, Q., Lu, S., Zimmermann, R., Cheng, L., 2019c. Towards natural and accurate future motion prediction of humans and animals, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, pp. 10004–10012.http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Towards_Natural_and_Accurate_Future_Motion_Prediction_of_Humans_and_CVPR_2019_paper.html, DOI: 10.1109/CVPR.2019.01024.
[61] Lu, M., Li, F., 2020. Survey on lie group machine learning. Big Data Min. Anal. 3, 235–258. 10.26599/BDMA.2020.9020011.
[62] Lyu, K., Liu, Z., Wu, S., Chen, H., Zhang, X., Yin, Y., 2021. Learning human motion prediction via stochastic differential equations, in: MM '21: ACM Multimedia Conference, Virtual Event, China, October 20–24, 2021, pp. 4976–4984. DOI: 10.1145/3474085.3475630.
[63] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J., 2019. AMASS: archive of motion capture as surface shapes, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 – November 2, 2019, pp. 5441–5450. DOI: 10.1109/ICCV.2019.00554.
[64] D. Majoe, L. Widmer, J. Gutknecht, Enhanced motion interaction for multimedia applications, in: Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia, 2009, pp. 13–19.
[65] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, T. Asfour, The KIT whole-body human motion database, in: International Conference on Advanced Robotics, ICAR 2015 July 27–31, 2015, Istanbul, Turkey, 2015, pp. 329–336, https://doi.org/10.1109/ICAR.2015.7251476.
[66] Mao, W., Liu, M., Salzmann, M., 2020. History repeats itself: Human motion prediction via motion attention, in: Computer Vision – ECCV 2020–16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV, pp. 474–489. DOI: 10.1007/978-3-030-58568-6_28.
[67] Mao, W., Liu, M., Salzmann, M., Li, H., 2019. Learning trajectory dependencies for human motion prediction, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 – November 2, 2019, pp. 9488–9496. DOI: 10.1109/ICCV.2019.00958.
[68] W. Mao, M. Liu, M. Salzmann, H. Li, Multi-level motion attention for human motion prediction, Int. J. Comput. Vis. 129 (2021) 2513–2535, https://doi.org/10.1007/s11263-021-01483-7.
[69] von Marcard, T., 2019. Human motion capture with sparse inertial sensors and video. Ph.D. thesis. University of Hanover, Hannover, Germany. https://d-nb.info/1206259906.
[70] von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G., 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera, in: Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part X, pp. 614–631. DOI: 10.1007/978-3-030-01249-6_37.
[71] J. Martinez, M.J. Black, J. Romero, On human motion prediction using recurrent neural networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 July 21–26, 2017, Honolulu, HI, USA, 2017, pp. 4674–4683, https://doi.org/10.1109/CVPR.2017.497.
[72] Martínez-González, Á., Villamizar, M., Odobez, J., 2021. Pose transformers (POTR): human motion prediction with non-autoregressive transformers, in: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11–17, 2021, pp. 2276–2284. DOI: 10.1109/ICCVW54120.2021.00257.

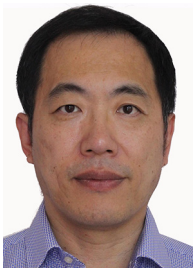
[73] Q. Men, E.S.L. Ho, H.P.H. Shum, H. Leung, A quadruple diffusion convolutional recurrent network for human motion prediction, IEEE Trans. Circuits Syst. Video Technol. 31 (2021) 3417–3432, https://doi.org/10.1109/TCSVT.2020.3038145.
[74] S. Min, Z. Gao, J. Peng, L. Wang, K. Qin, B. Fang, STGSN – A spatial-temporal graph neural network framework for time-evolving social networks, Knowl. Based Syst. 214 (2021), https://doi.org/10.1016/j.knosys.2021.106746 106746.
[75] M.G.S. Murshed, C. Murphy, D. Hou, N. Khan, G. Ananthanarayanan, F. Hussain, Machine learning at the network edge: A survey, ACM Comput. Surv. 54 (2022) 170:1–170:37, https://doi.org/10.1145/3469029.
[76] E. Nasiri, K. Berahmand, M. Rostami, M. Dabiri, A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding, Comput. Biol. Medicine 137 (2021), 104772, https://doi.org/10.1016/j.compbiomed.2021.104772.
[77] N. Nikhil, B. Tran Morris, Convolutional neural network for trajectory prediction, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.
[78] O.S. Oguz, V. Gabler, G. Huber, Z. Zhou, D. Wollherr, Hybrid human motion prediction for action selection within human-robot collaboration, International Symposium on Experimental Robotics, Springer (2016) 289–298.
[79] Pavllo, D., Grangier, D., Auli, M., 2018. Quaternet: A quaternion-based recurrent model for human motion, in: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, p. 299.http://bmvc2018.org/contents/papers/0675.pdf.
[80] Pujol-Perich, D., Suárez-Varela, J., Galmés, M.F., Wu, B., Xiao, S., Cheng, X., Cabellos-Aparicio, A., Barlet-Ros, P., 2021. IGNNITION: fast prototyping of graph neural networks for communication networks, in: SIGCOMM '21: ACM SIGCOMM 2021 Conference, Virtual Event, August 23–27, 2021, Poster and Demo Sessions, pp. 71–73. DOI: 10.1145/3472716.3472853.
[81] L. Qu, J. Lyu, W. Li, D. Ma, H. Fan, Features injected recurrent neural networks for short-term traffic speed prediction, Neurocomputing 451 (2021) 290–304, https://doi.org/10.1016/j.neucom.2021.03.054.
[82] T. Ren, W. Li, Z. Jiang, X. Li, Y. Huang, J. Peng, Video-based human motion capture data retrieval via motionset network, IEEE Access 8 (2020) 186212–186221, https://doi.org/10.1109/ACCESS.2020.3030258.
[83] Rizi, F.S., 2021. Graph Representation Learning for Social Networks. Ph.D. thesis. University of Passau, Germany. https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/921.
[84] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, Learning social etiquette: Human trajectory understanding in crowded scenes, European conference on computer vision, Springer (2016) 549–565.
[85] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, Learning social etiquette: Human trajectory understanding in crowded scenes, European conference on computer vision, Springer (2016) 549–565.
[86] Ruiz, A.H., Gall, J., Moreno, F., 2019. Human motion prediction via spatio-temporal inpainting, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 – November 2, 2019, pp. 7133–7142. DOI: 10.1109/ICCV.2019.00723.
[87] H. Sang, Z. Chen, D. He, Human motion prediction based on attention mechanism, Multim. Tools Appl. 79 (2020) 5529–5544, https://doi.org/10.1007/s11042-019-08269-7.
[88] van Schuppen, J.H., 1990. Review of 'stochastic integration and differential equations - A new approach', (protter, p.; 1990). IEEE Trans. Inf. Theory 36, 1188.
[89] Shahroudy, A., Liu, J., Ng, T., Wang, G., 2016. NTU RGB+D: A large scale dataset for 3d human activity analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp. 1010–1019. DOI: 10.1109/CVPR.2016.115.
[90] Shi, M., Aberman, K., Aristidou, A., Komura, T., Lischinski, D., Cohen-Or, D., Chen, B., 2020. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Trans. Graph. 40, 1:1–1:15. 10.1145/3407659.
[91] Su, P., Liu, Z., Wu, S., Zhu, L., Yin, Y., Shen, X., 2021. Motion prediction via joint dependency modeling in phase space, in: MM '21: ACM Multimedia Conference, Virtual Event, China, October 20–24, 2021, pp. 713–721. DOI: 10.1145/3474085.3475237.
[92] Tang, R., Yang, G., Wei, H., Mao, Y., Türe, F., Lin, J., 2018a. Streaming voice query recognition using causal convolutional recurrent neural networks. CoRR abs/1812.07754.http://arxiv.org/abs/1812.07754.
[93] Tang, Y., Ma, L., Liu, W., Zheng, W., 2018b. Long-term human motion prediction by modeling motion context and enhancing motion dynamics, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13–19, 2018, Stockholm, Sweden, pp. 935–941. 10.24963/ijcai.2018/130.
[94] H. Wang, J. Dong, B. Cheng, J. Feng, PVRED: A position-velocity recurrent encoder-decoder for human motion prediction, IEEE Trans. Image Process. 30 (2021) 6096–6106, https://doi.org/10.1109/TIP.2021.3089380.
[95] Wang, R., Panju, M., Gohari, M., 2017. Classification-based RNN machine translation using grus. CoRR abs/1703.07841.http://arxiv.org/abs/1703.07841.
[96] Y. Wang, X. Wang, P. Jiang, F. Wang, RNN -based human motion prediction via differential sequence representation, in: 6th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2019 December 19–21, 2019, Singapore, 2019, pp. 138–143, https://doi.org/10.1109/CCIS48116.2019.9073734.
[97] S. Xiao, Z. Wang, J. Folkesson, Unsupervised robot learning to predict person motion, in: IEEE International Conference on Robotics and Automation, ICRA 2015 USA, 26–30 May, 2015, Seattle, WA, 2015, pp. 691–696, https://doi.org/10.1109/ICRA.2015.7139254.
[98] W. Xiao, A. Dey, L.H. Son, A study on regular picture fuzzy graph with applications in communication networks, J. Intell. Fuzzy Syst. 39 (2020) 3633–3645, https://doi.org/10.3233/JIFS-191913.
[99] Xu, J., Chen, X., Lan, X., Zheng, N., 2021. Probabilistic human motion prediction via A bayesian neural network, in: IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 – June 5, 2021, pp. 3190–3196. DOI: 10.1109/ICRA48506.2021.9561665.
[100] Xu, Y., Piao, Z., Gao, S., 2018. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, pp. 5275–5284.http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_Encoding_Crowd_Interaction_CVPR_2018_paper.html, DOI: 10.1109/CVPR.2018.00553.
[101] H. Xue, D.Q. Huynh, M. Reynolds, A location-velocity-temporal attention LSTM model for pedestrian trajectory prediction, IEEE Access 8 (2020) 44576–44589, https://doi.org/10.1109/ACCESS.2020.2977747.
[102] Yan, S., Xiong, Y., Lin, D., 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 7444–7452.https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135.
[103] J. Yao, J. Zhang, J. Li, L. Zhuo, Anchor voiceprint recognition in live streaming via rawnet-sa and gated recurrent unit, EURASIP J. Audio Speech Music. Process. 2021 (2021) 45, https://doi.org/10.1186/s13636-021-00234-3.
[104] M. Yeasin, E. Polat, R. Sharma, A multiobject tracking framework for interactive multimedia applications, IEEE Trans. Multimedia 6 (2004) 398–405.
[105] A. Yiannakides, A. Aristidou, Y. Chrysanthou, Real-time 3d human pose and motion reconstruction from monocular RGB videos, Comput. Animat. Virtual Worlds 30 (2019), https://doi.org/10.1002/cav.1887.
[106] Yuan, Y., Kitani, K., 2020. Dlow: Diversifying latent flows for diverse human motion prediction, in: Computer Vision – ECCV 2020–16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX, pp. 346–364. DOI: 10.1007/978-3-030-58545-7_20.
[107] Zang, C., Pei, M., Kong, Y., 2020. Few-shot human motion prediction via learning novel motion dynamics, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp. 846–852. 10.24963/ijcai.2020/118.
[108] Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. CoRR abs/1409.2329.http://arxiv.org/abs/1409.2329.
[109] Zhang, H., Wang, J., Liu, H., 2021. Video-based reconstruction of smooth 3d human body motion, in: Pattern Recognition and Computer Vision - 4th Chinese Conference, PRCV 2021, Beijing, China, October 29 – November 1, 2021, Proceedings, Part II, pp. 42–53. doi: 10.1007/978-3-030-88007-1_4, DOI: 10.1007/978-3-030-88007-1_4.
[110] Zhou, H., Guo, C., Zhang, H., Wang, Y., 2021. Learning multiscale correlations for human motion prediction, in: IEEE International Conference on Development and Learning, ICDL 2021, Beijing, China, August 23–26, 2021, pp. 1–7. DOI: 10.1109/ICDL49984.2021.9515609.

Kedi Lyu received the B.E. and ME. degrees from the Jilin University, Changchun, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with Jilin University, Changchun, China. His research interests include computer vision, machine learning, and deep learning.

Haipeng Chen is a professor of Jilin University, Changchun, China. He received his Ph.D. and B.E. degrees from Jilin University, Changchun, China. His research interests include computer vision, machine learning.

Beiqi Zhang received his Bachelor's degree from Sichuan University in 2020. His current research interest is multimedia data analysis. Recently he has been researching on human motion and pose estimation at Zhejiang Gongshang University.

Zhenguang Liu is a professor of Zhejiang Gongshang University, Hangzhou, China. He had been a research fellow in National University of Singapore and A*STAR (Agency for Science, Technology and Research, Singapore) for several years. He respectively received his Ph. D. and B.E. degrees from Zhejiang University and Shandong University, China. His research interests include multimedia data analysis and smart contract security. Various parts of his work have been published in top-tier venues including TIP, CVPR, ICCV, TKDE, AAAI, ACM MM, INFOCOM, IJCAI, WWW, TMC, WWW. Professor Liu has served as technical program committee member for top-tier conferences such as ACM MM, CVPR, AAAI, IJCAI, and ICCV, session chair of ICGIP, local chair of KSEM, and reviewer for top-tier journals IEEE TVCG, IEEE TPDS, IEEE TMM, ACM TOMM.



Ruili Wang is a Professor in the School of Computer and Information Engineering at Zhejiang Gongshang University. His research areas include image and video processing, speech and natural language processing, machine learning, and deep learning, data mining. He has supervised 26 PhD students to completion, and has published over 190 papers. He is an Associate Editor (or Editor Board member) for the journals of IEEE Transactions on Emerging Topics in Computational Intelligence, Neurocomputing (Elsevier), Applied Soft Computing (Elsevier), Knowledge and Information Systems (Springer), Health Information Science and Systems (Springer), and Complex and Intelligent Systems (Springer).