3D Motion Prediction in Multi-Domain Application Scenarios: Methods, Challenges and Future Directions

Chris Chen
Department of Software Engineering, Chengdu University of Information Technology, Chengdu, Sichuan Province, 610225, China
13308073885@163.com

ABSTRACT

3D motion prediction and tracking are the core technology of intelligent perception in the field of computer vision. It is widely used in the fields of complex environment perception of autonomous driving, industrial production safety monitoring, digital protection of intangible cultural heritage dance movements and so on. In recent years, the achievements in this field have emerged, but the existing research focuses on the customization and optimization of a single scene, and there are problems such as the separation of technical context, the disconnection between theoretical innovation and engineering implementation. The relevant review also has the defects of single coverage, lack of systematicness, and incomplete coverage of the latest frontier achievements. This paper takes 3D motion prediction and tracking technology in multi-domain application scenarios as the research object, systematically combs the core research results at home and abroad from 2017 to 2026, constructs a four-dimensional analysis framework, clarifies the technological evolution context, compares the applicable boundaries of different technical routes, and refines the adaptation rules of technology and scenarios, which can provide reference for relevant theoretical research and engineering application.

Keywords: Multi-Domain Application Scenarios; 3D Motion Prediction; 3D Target Tracking; Computer Vision; Deep Learning.

1. INTRODUCTION

3D motion prediction and tracking are the core research direction in the field of computer vision, and it is the key technology to realize intelligent perception and behavior understanding of dynamic targets. In recent years, with the rapid iteration of 3D sensing technology and deep learning algorithm, a large number of innovative achievements have emerged in 3D motion prediction and tracking technology in different scenarios. The technical route has been continuously optimized, and the scene adaptation ability has been continuously improved.

However, the traditional research and existing review results in this field still face multiple core challenges. The existing research is mostly scattered in a single subdivision direction such as human motion prediction [1-4] and automatic driving target tracking [5]. Not only is the technical evolution context fragmented, the cross-scene research lacks systematic integration, and there are also widespread problems such as insufficient connection between domestic scene engineering results and foreign general theoretical innovation, fragmentation of domain evaluation system, and difficulty in objectively comparing the performance of different technical methods. It not only restricts the cross-domain reuse and iterative optimization of technology, but also fails to provide clear selection guidance for multi-scene engineering landing.

In view of the above problems, this study uses the literature analysis method and the inductive summary method to sort out the domestic and foreign research results in the field of 3D motion prediction and tracking in multi-domain application scenarios from 2017 to 2025, and constructs a four-dimensional classification framework of “core technology-multimodal fusion-scene application-evaluation system”. It clarifies the complete development context of domain technology, integrates the complementary advantages of differentiated research paths at home and abroad, extracts the scene adaptation rules of different technical routes, and establishes a cross-scene evaluation index mapping system. It aims to provide comprehensive theoretical support and practical reference for the application of technologies in many fields such as automatic driving safety perception [5], industrial safety intelligent monitoring [6,7], and digital protection of intangible cultural heritage [8].

2. THEORETICAL BASIS

3D motion prediction and tracking are the core technology of dynamic target intelligent perception in the field of computer vision. Among them, 3D target tracking refers to obtaining the three-dimensional space coordinates of the target through the sensing device to achieve target identity maintenance and trajectory reconstruction between consecutive frames. 3D motion prediction refers to modeling the motion law of the target based on historical time series data, and outputting the three-dimensional pose and motion trajectory of the future frame. This article mainly involves four core scenarios: autonomous driving, industrial safety, intangible cultural heritage dance digitization, and general human posture analysis. The research focuses on related technologies that adapt to different scene requirements, and does not involve generalized full-scene coverage and non-moving target detection research.

The technical system of this study is based on three core basic modules-the first is the attitude feature representation method, which converts the three-dimensional motion information of the target into a feature vector that can be processed by the deep learning model, which is the underlying basis of technical implementation; the second is the temporal motion modeling technology, which captures the temporal dependence of target motion and is the core module of trajectory prediction and stable tracking. The third is multi-modal perception fusion technology, which integrates multi-source sensing data such as vision and point cloud to improve the robustness of the algorithm in complex environments, and is the key support for technical engineering.

3. MULTI-DOMAIN 3D MOTION PREDICTION AND TRACKING KEY TECHNOLOGIES AND RESEARCH PROGRESS

3D motion prediction and tracking technology relies on computer vision, deep learning and spatio-temporal geometric modeling. The current technology system evolves around three core routes: human motion prediction [1-4,8-15], point cloud target perception [16-20], and robust optimization of complex scenes [6-8]. This chapter sorts out the core technology system, systematically compares the research results at home and abroad, condenses the existing common technology bottlenecks in the field, and provides theoretical support for subsequent algorithm design.

3.1 Core technical route and Method system

This chapter is the core of the research. Combined with the two major technical branches of human skeleton modeling and 3D point cloud perception, a complete technical method system of pose representation optimization, spatio-temporal feature fusion and multi-modal scene adaptation is constructed, covering the whole technical link from the underlying feature coding to the upper task reasoning.

3.1.1 Human 3D motion prediction technology

Human 3D motion prediction takes the time series of skeleton joints as input, which aims to model the kinematic constraints and spatio-temporal dependence of human body and realize the accurate prediction of future posture. It is the basic technology of behavior recognition and action reproduction.

(1) Skeleton pose representation method

The pose representation is the basis for determining the prediction accuracy, which is divided into two categories: Cartesian coordinate representation and kinematic chain geometric representation. The traditional method uses the absolute coordinates of three-dimensional joints, which is simple to calculate but ignores the rigid constraints of bones, and is prone to limb length distortion [1,3]. The kinematic chain model is constructed by the frontier method, and three parametric methods of axis angle, quaternion and Stiefel manifold are derived. Among them, the Stiefel manifold, with its global continuity, avoids gradient disappearance and rotation ambiguity, and becomes the international mainstream optimization scheme [9]; domestic research combines industrial scenarios and adopts lightweight skeleton reconstruction (14-joint point streamlined model) to reduce computational complexity and adapt to edge device deployment.

(2) Spatio-temporal feature coding network

Aiming at the problem of time series dependence modeling, three mainstream architectures, cyclic network, graph convolution and Transformer, are formed. Recurrent neural network (RNN/GRU/LSTM) realizes serial time series modeling, but it has the defect of insufficient capture of long dependence [9]; the spatio-temporal graph convolutional network (ST-GCN/Info-GCN++) integrates the topological structure and temporal features of the skeleton. The domestic improved algorithm introduces attention enhancement and skeleton prediction compensation modules to solve the problem

of incomplete data recognition [6]; the hierarchical attention cycle network (AHMR) constructs a two-layer parallel architecture of local frame state-global sequence state, integrates spatio-temporal attention to adaptively allocate joint weights, and takes into account long-term timing prediction and reasoning efficiency. The test frame rate reaches 636 FPS [9]. In addition, multi-scale hypergraph convolution is proposed for non-heritage scenes in China to capture the long-distance interaction of joints and adapt to complex dance motion modeling [8].

(3) Geometric constraint loss optimization

Abandoning the geometric adaptation defects of the traditional mean square error (L2) loss, two types of optimization schemes are formed: international research proposes geodesic loss and forward kinematics loss [9], which fit the rotation manifold distance and human body level kinematics constraints respectively, and significantly reduce the angle and position prediction error; domestic research uses the Huber loss function [6] to reduce the interference of outliers in industrial scenes and improve the robustness of the model under low-quality data.

3.1.2 3D point cloud target tracking and detection technology

Aiming at the characteristics of sparse, disordered and timing dislocation of point cloud of autonomous driving lidar, this technology focuses on single-frame feature coding, multi-frame timing aggregation and lightweight attention matching to realize real-time tracking and accurate detection of dynamic targets. The core method system is as follows:

(1) Single frame point cloud feature coding

With the lightweight voxelization method as the core, PointPillars algorithm encodes the three-dimensional point cloud into a two-dimensional pseudo-image feature based on the cylinder grid [10], taking into account the coding efficiency and spatial information retention, and has become a common backbone network at home and abroad. Domestic research uses PointNet++ to extract hierarchical geometric features, combined with cascaded Hoffman voting to optimize target center positioning, to solve the problem of low positioning accuracy of sparse point cloud [5,18].

(2) Multi-frame temporal feature aggregation

Aiming at the problem of single frame information loss, three methods of point splicing, feature fusion and attention aggregation are formed. The basic scheme adopts multi-frame point cloud direct splicing, which is simple in calculation but easy to blur the moving target. TransPillars, the international frontier, proposes a coarse-to-fine hierarchical aggregation strategy to complete cross-frame feature fusion at the voxel level, and optimizes the positioning accuracy through coarse features to locate the motion trajectory and fine features. The domestic TM2B algorithm constructs a spatio-temporal Transformer [5], directly models the relative motion between frames, abandons the redundant calculation of appearance matching, and realizes real-time tracking.

(3) Lightweight attention matching mechanism

The high computational complexity of the traditional Transformer global attention is optimized. The deformable attention reduces the complexity through sparse sampling. The international improved version adds query-key matching constraints to solve the problem of cross-frame feature misalignment. In China, the attention of learning relationship modeling is proposed, the interactive features are adaptively screened, and the real-time requirements of vehicle edge devices are adapted.

3.1.3 Multi-modal fusion and extreme scene robust optimization techniques

This technology is the core support for the implementation of algorithm engineering. Aiming at the four extreme scene pain points of low illumination, occlusion, data incompleteness and small sample labeling, a whole scene adaptation method system is formed:

(1) Multi-modal feature fusion

Fusion of image semantics and point cloud spatial information, domestic industrial scenes using YOLOv7 target detection+AlphaPose attitude estimation dual branch architecture [6], synchronous extraction of human action and environmental equipment information, construction of human-object interaction entry matching mechanism, to achieve accurate identification of violations; the automatic driving scene integrates the bird’s eye view (BEV) features and temporal motion information to strengthen the scene context modeling.

(2) Robust enhancement of low-quality data

In domestic coal mine scenes, skeleton prediction compensation and weak light enhancement network (Zero-DCE) are introduced to repair incomplete skeleton nodes and improve the quality of dark light images. The intangible cultural heritage scene uses multi-view data enhancement and self-supervised comparative learning to solve the problem of scarce annotated data [6,7]; international studies have used manifold parameterization to naturally constrain skeletal rigidity and avoid predicting limb distortion.

(3) End-to-end engineering deployment

Domestic research led to lightweight model cutting, skeleton joint simplification, multi-task joint optimization, and developed a real-time monitoring system with a detection speed of 81.2 frames/s [7]; international research focuses on model parallel reasoning and sparse attention acceleration to achieve a balance between high accuracy and high efficiency of the algorithm.

3.2 Domestic research progress and technical characteristics

Domestic research focuses on industrial safety, intangible cultural heritage protection, autonomous driving and other fields.

In the field of industrial safety behavior monitoring, aiming at the pain points of low illumination, skeleton loss and complex human-object interaction in coal mines, the miner behavior data set is constructed, and the improved ST-GCN and Info-GCN++ algorithms are proposed. Combining attention enhancement and skeleton compensation technology, the recognition accuracy of low-quality data is increased to 89.12% [6]; innovate the human-object interaction identification architecture, realize the integrated detection of static wear violations and dynamic behavior violations, and develop an end-to-end real-time early warning system to complete the large-scale deployment of industrial sites.

In the field of intangible cultural heritage dance digitization, the technical bottleneck of clothing occlusion and small sample labeling is broken through, and a three-dimensional pose estimation algorithm combining graph convolution and Transformer is proposed, and a multi-scale hypergraph convolution is designed to capture the long-distance dependence of joints [8]. The self-supervised comparative learning is used to expand the training samples to realize the accurate prediction and visual reproduction of intangible cultural heritage actions, and the interactive digital protection system is implemented.

In the field of automatic driving point cloud perception, balancing accuracy and real-time performance, a C-DHV cascade voting tracker and TM2 B spatio-temporal motion tracking algorithm [5] are proposed to optimize point cloud layer-level feature extraction and lightweight motion modeling, adapt to vehicle-mounted edge devices, and achieve real-time target tracking with anti-occlusion and anti-interference in complex traffic scenarios.

3.3 Foreign research progress and technical characteristics

Foreign research focuses on universal modeling in the field of human motion prediction and point cloud perception.

In the field of human 3D motion prediction, AHMR hierarchical attention cycle network is proposed [9], and four kinds of pose representation methods are systematically compared to establish the optimal performance of Stiefel manifold. A local-global two-layer parallel modeling architecture is constructed to solve the long-term prediction drift problem by combining geodesic loss and forward kinematics loss. It can generate more than 50 seconds of natural human motion [9]. The generalization and reasoning efficiency of the algorithm are international leading, and it can be adapted to multi-class joint motion prediction without scene customization modification.

In the field of 3D point cloud target detection, the TransPillars multi-frame aggregation algorithm is proposed. The voxel-level feature fusion and coarse-to-fine hierarchical aggregation strategy are innovated. The deformable attention mechanism is improved to solve the problem of cross-frame matching misalignment, and the accuracy of long-distance sparse point cloud target detection is significantly improved. The algorithm is based on PointPillars lightweight backbone, and achieves SOTA performance on international benchmark datasets such as nuScenes and Waymo, taking into account high accuracy and low computational overhead [10].

3.4 Common problems in research at home and abroad

(1) Insufficient robustness in complex environments: Under extreme conditions such as occlusion, low illumination, sparse point clouds, and missing skeleton nodes, feature extraction fails, and model accuracy and stability decline sharply, which is the primary pain point in industrial and autonomous driving scenarios [5-7].

(2) Long time series prediction error accumulation: The existing algorithms are prone to limb distortion and trajectory

drift in long time series prediction, which cannot meet the long time series application requirements such as non-heritage action reproduction and long-term behavior warning [8,15].

(3) Imbalance between accuracy and real-time performance: The high-precision model has high computational complexity and is difficult to adapt to edge devices; the characteristics of the lightweight model are seriously lost, and the perception accuracy of long-distance and small targets cannot reach the standard [5,10].

(4) Weak cross-scene generalization ability: The algorithm is highly dependent on the distribution of training data. A single model is not compatible with human motion and multi-class tasks of point cloud targets, and the cost of domain migration and model reuse is extremely high.

4. EXAMPLE ANALYSIS

3D motion prediction and tracking technology takes skeleton space-time modeling and point cloud timing fusion as the core, and has achieved large-scale application in four fields: industrial safety, intangible cultural heritage digitization, automatic driving, and general human motion analysis.

4.1 Industrial safety scenarios

Aiming at the pain points of low illumination and incomplete skeleton in coal mines, domestic research adopts the improved Info-GCN++ algorithm, combined with skeleton compensation and attention optimization, to solve the problem of low-quality data recognition [6]. The fusion of YOLOv7 and AlphaPose constructs a human-object interaction detection framework to realize the early warning of miners' basic action recognition, safety helmet wearing and illegal operation [6,7]. The actual measurement shows that the accuracy of action recognition under low-quality data is 89.12%, and the accuracy of core violation recognition is over 85%, which can be integrated into the real-time monitoring system to meet the needs of industrial safety automation control [6,7].

4.2 Digitalization of intangible dance movements

Aiming at the problem of serious occlusion of intangible dance and scarcity of annotated data, multi-scale hypergraph convolution combined with Transformer architecture is used to model the spatio-temporal correlation of joints, and self-supervised learning is used to optimize the training effect of small samples [8]. The technology is implemented as a lightweight digital system, which can complete three-dimensional pose reconstruction and long-term motion prediction only by relying on ordinary video, effectively avoid the problem of traditional algorithm motion distortion, realize the standardized reproduction and digital archiving of intangible cultural heritage dance, and promote the intelligent protection of cultural heritage [8].

4.3 Automatic driving

Based on the sparse characteristics of lidar point cloud, the foreign TransPillars algorithm uses coarse-to-fine hierarchical feature aggregation to improve the deformable attention to complete multi-frame point cloud fusion and improve the accuracy of long-distance target detection [10,17,18]. Domestic TM2B and C-DHV algorithms lightweight modeling inter-frame motion, taking into account vehicle real-time [5]. In the verification of nuScenes and Waymo datasets, the algorithm effectively solves the problem of occlusion and sparse point cloud positioning, and the accuracy of dynamic target tracking is significantly improved, which is suitable for the deployment of autonomous vehicle edge devices [5,10].

4.4 Human body posture analysis

The international AHMR algorithm is represented by Stiefel manifold optimization posture, constructs a hierarchical attention loop network, and combines geodesic loss optimization model training. On the H3.6M and CMUMoCap standard datasets, the short-term prediction error is as low as 0.25, which can generate more than 50 seconds of natural human action, and the inference efficiency is 636 FPS [9]. The algorithm has strong generalization and can be adapted to multi-class joint motion prediction. It is widely used in general scenarios such as human-computer interaction and medical rehabilitation, and provides basic technical support for the field [1,3,9].

4.5 Summary of this chapter

3D motion prediction technology forms a differentiated adaptation scheme in various fields. Industrial and cultural scenes focus on robust optimization in extreme environments, and autonomous driving and general scenes focus on real-time and generalization capabilities. The existing applications have completed the actual verification, and the technology has strong

landing, but there are still problems such as insufficient robustness of complex scenes and weak cross-scene reusability, which is the core optimization direction of follow-up research.

5. CONCLUSION

This paper focuses on the comprehensive review of 3D motion prediction and tracking technology. Based on the classical literature and the latest research results at home and abroad, this paper combs the core methods of this technology, the development status at home and abroad, and the practical application in many fields such as industrial safety, intangible cultural heritage digitization and automatic driving. On the whole, after years of development, 3D motion prediction and tracking technology has formed a relatively mature technical framework, and has made significant progress in human pose modeling, point cloud target perception, and temporal feature learning, which can effectively support the actual needs of various intelligent perception scenarios. This technology has been implemented in many fields, and has played an important role in coal mine safety monitoring, non-legacy action reproduction, automatic driving target perception and other scenarios, which fully proves the practical value and development potential of this technology. All kinds of application scenarios also reversely promote the optimization and upgrading of the algorithm, make the technical research more in line with the actual needs, and form a good development trend of mutual promotion between research and application.

REFERENCE

[1] Zhou, H., Zhang, Y., Guo, X., Cheng, N., Yang, H., Du, Z., & Wu, X. (2025). Modal Dynamics Features Enhancement Multi-Scale Spatial-Temporal Networks for 3D Human Motion Prediction from 2D Skeletons. Expert Systems with Applications, 130524.

[2] Kong Kangcheng, Feng Shaoquan, Li Xingxing, Zhang Zhongying, Yan Zhuohao, Yang Wuzhuo... & Hao Yushi. A 3D visual multi-target tracking algorithm considering motion constraints. Journal of Wuhan University (Information Science Edition), 1-13. https://doi.org/10.13203/j.whugis20250148.

[3] Cao, W., Zhang, J., & Zhong, J. (2025). Spectral-spatial representation progressive learning via segmented attention for 3D skeleton-based motion prediction. Applied Soft Computing, 113688.

[4] Huang, J., He, D., Cao, W., & Zhong, J. (2025). Progressively deeper attention networks for 3D human motion prediction: Progressively deeper attention networks...Multimedia Systems,31(5).

[5] Xu Anqi. (2025). Research on 3D point cloud single target tracking method for autonomous driving road scene (Master's thesis, Hangzhou University of Electronic Science and Technology).Master https://doi.org/10.27075/d.cnki.ghzdc.2025.000360.

[6] Hou Linpeng. (2025). Research on the identification method of illegal behavior of underground miners in coal mines (master 's degree thesis, Anhui University of Science and Technology).Master https://doi.org/10.26918/d.cnki.ghngc.2025.001190.

[7] He Xinxin. (2024). recognition and application of unsafe behavior of underground miners in coal mine based on machine vision (master's degree thesis, Xi'an University of Architecture and Technology).Master https://doi.org/10.27393/d.cnki.gxazu.2024.001666.

[8] Cheng Pengyan. (2025). Research on motion prediction and action recognition method of intangible cultural heritage dance in video (master's degree thesis, North China University of Technology).Master https://doi.org/10.26926/d.cnki.gbfgu.2025.000023.

[9] Liu, Z., Wu, S., Jin, S., Ji, S., Liu, Q., Lu, S., & Cheng, L. (2022). Investigating pose representations and motion contexts modeling for 3D motion prediction. IEEE transactions on pattern analysis and machine intelligence,45(1), 681-697.ysis and Machine Intelligence, 2022.

[10] Luo, Z., Zhang, G., Zhou, C., Liu, T., Lu, S., & Pan, L. (2023). TransPillars: Coarse-to-fine aggregation for multi-frame 3D object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4230-4239).

[11] Li Zijing. (2023). Research on human action recognition and prediction algorithm based on deep learning (Master's degree thesis, Beijing Jiaotong University). Master https://doi.org/10.26944/d.cnki.gbfju.2023.000515.

[12] Ma Yatong. (2023). Research on human action recognition based on multimodal fusion (Master's degree thesis, Lanzhou Jiaotong University). Master https://doi.org/10.27205/d.cnki.gltec.2023.000319.

[13] Gu Baoxuan. (2025). Multi-modal human motion prediction based on spatio-temporal feature modeling (Master's thesis, Beijing University of Posts and Telecommunication). Master https://doi.org/10.26969/d.cnki.gbydu.2025.000384.
[14] Yang Shubin. (2025). 3D human pose estimation and motion prediction based on spatio-temporal generation model (Master's thesis, University of Electronic Science and Technology of China). Master https://doi.org/10.27005/d.cnki.gdzku.2025.001328.
[15] Du Xin. (2025). Research on 3D human motion prediction method based on representation enhancement (Master's thesis, Chongqing University of Technology). Master https://doi.org/10.27753/d.cnki.gcqgx.2025.001605.
[16] Li Xu, Zhang Yonghong, Zhu Linglong & Kan Xi. (2026). multimodal 3D object detection method based on pseudo point cloud fusion. Electronic Measurement Technology, 49 (01), 121-132.https://doi.org/10.19651/j.cnki.emt.2518691.
[17] Zhu Runwen. (2025). Research on 3D object detection algorithm of point cloud and image fusion based on sparse feature query (Master's degree thesis, Harbin Institute of Technology). Master https://doi.org/10.27061/d.cnki.ghgdu.2025.001632.
[18] Liu Zhaodi. (2024). Research on point cloud target tracking algorithm based on deep learning (master's degree thesis, North China University of Technology). Master https://doi.org/10.26926/d.cnki.gbfgu.2024.000587.
[19] Yang Haoran. (2024). 3D multi-target tracking method based on point cloud and image (Master's thesis, Qingdao University of Science and Technology). Master https://doi.org/10.27264/d.cnki.gqdhc.2024.000709.
[20] Zhang Jingwen. (2023). 3D single target tracking based on point cloud data (Master's thesis, Harbin Institute of Technology). Master https://doi.org/10.27061/d.cnki.ghgdu.2023.006845.