Neurocomputing 611 (2025) 128681

Contents lists available at ScienceDirect

Neurocomputing

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

DM-HAP: Diffusion model for accurate hand pose prediction

Zhifeng Wang [a,*], Kaihao Zhang [b], Ramesh Sankaranarayana [a]

[a] College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia
[b] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

ARTICLE INFO

Communicated by Y. Long

Keywords:
Local and global diffusion
Hand pose prediction
Reverse diffusion process
Subtle joint displacements

ABSTRACT

Forecasting hand poses is a challenging task due to inherent uncertainties, occlusions, and inaccuracies in 3D pose estimation. Diffusion models provide a promising direction for predicting precise 3D hand poses under noisy conditions. In this work, we introduce Dual-diffusion, an innovative framework for the precise prediction of future hand poses. Our approach leverages the strengths of diffusion models by framing hand pose forecasting as a reverse diffusion process, effectively addressing the complexities of noisy hand joint movements and their subtleties. To enhance the learning of hand pose representations, we propose a unique neural architecture that simultaneously captures both local and global features. This is achieved through the deployment of Global and Local Diffusion (GLD) blocks within our network, which facilitate the exchange of information between local and global features. This diffusion-based interaction enables the integration of global hand actions and local finger actions, leading to a more powerful representation learning approach. We evaluate the effectiveness of our Dual-diffusion method on three public 3D hand pose estimation datasets (MSRA, F-PHAB, and BigHand2.2M), and our approach outperforms previous methods.

1. Introduction

Hand pose prediction is an advanced computational task that involves predicting the future positions and configurations of human hands. This technology has significant importance across various domains. The task of hand pose prediction involves forecasting future 3D hand positions, which finds application in accident prevention [1], sign language translation [2], human–robot interaction [3], and tracking [4]. Previous studies using techniques such as RNNs, LSTMs, and GRUs [5–8] have struggled with errors that accumulate over time in their predictions. While models based on convolution have helped reduce these problems, they still persist. Human motion is complex, involving both space and time, and requires advanced methods to understand how different parts of movement are connected. Graph Convolutional Networks (GCNs), as mentioned in [9], are effective at understanding these connections, but they face challenges because the relationships between joints and frames can change with different types of movements. Liu et al. [10] introduces a novel method for human motion prediction that incorporates control parameters into a composite GAN structure, featuring local GANs for specific body parts and a global GAN for overall coordination, enabling customizable and manipulable motion prediction across different activities. Barsoum et al. [11] propose HP-GAN model to use a modified Wasserstein GAN architecture with a custom loss function to probabilistically predict multiple plausible future human motion sequences from past poses, and includes a motion-quality-assessment model to evaluate the realism of these predictions, tested on major skeleton datasets for both single and multiple action types. As a result, they require meticulously designed loss functions and hyper-parameters tailored to specific loss constraints, making the application of the method quite labor-intensive. The extraction of 3D hand poses from monocular data is complicated by several factors. These include ambiguities, occlusions, and the inherent inaccuracies of pose estimation methods. To address these issues, we propose a Dual-diffusion framework for the accurate forecasting of hand movements. By using the latest observed motion as a prior condition, our method can achieve higher accuracy for long-term hand pose prediction.

Our model has two main parts: a Global and Local Diffusion (GLD) network and a method for gradually removing noise, both shown in one of our figures (Fig. 2). First, the GLD network uses past data of hand and finger movements to identify and capture important features and time-related patterns. Next, we use a step-by-step noise reduction method. Here, we start with noisy hand and finger movements and slowly clean them up to make realistic hand poses using a process in our model (shown in Fig. 2). The key part of this process is that in each step, labeled as t, we reduce the noise a bit more compared to the previous step ($t-1$). This unique feature of the reverse process allows us to use real observations as extra input, giving our model the ability to use this

* Corresponding author.
E-mail addresses: zhifeng.wang@anu.edu.au (Z. Wang), super.khzhang@gmail.com (K. Zhang), ramesh.sankaranarayana@anu.edu.au (R. Sankaranarayana).

https://doi.org/10.1016/j.neucom.2024.128681
Received 20 March 2024; Received in revised form 11 September 2024; Accepted 29 September 2024
Available online 2 October 2024

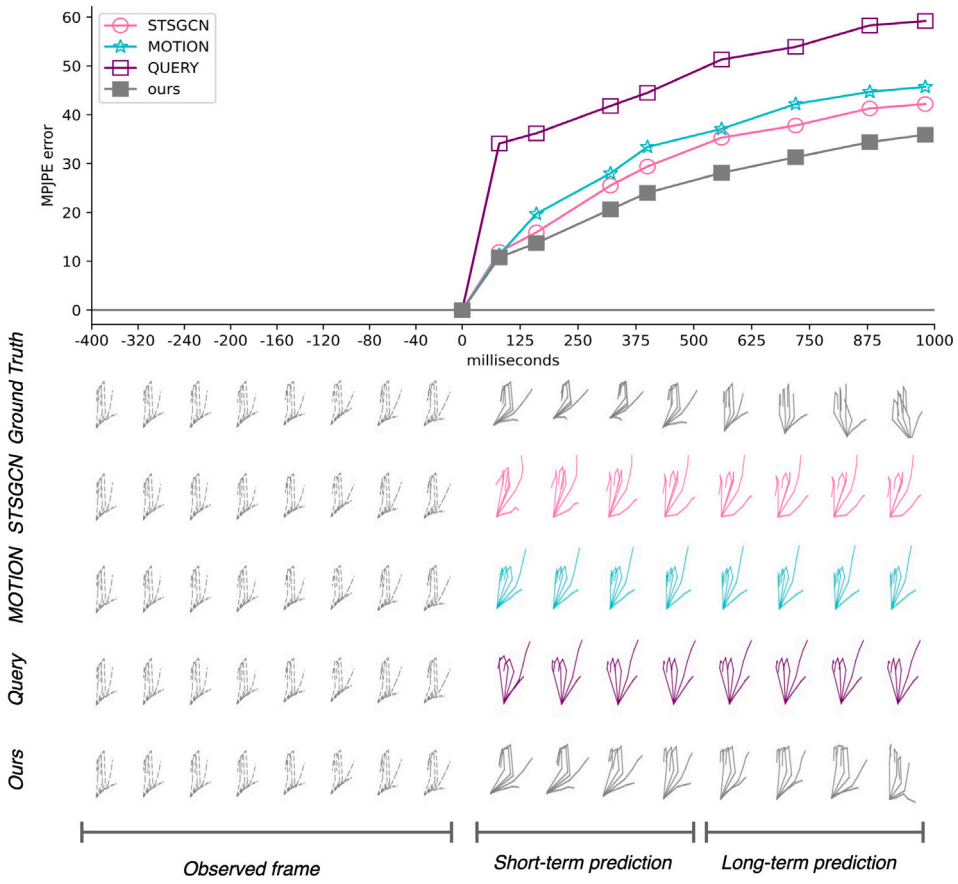

Fig. 1. Top: Mean average prediction errors corresponding to various hand pose prediction methods. Bottom: the ground truth is depicted in gray, while the short-term and long-term hand pose predictions are represented in color. Compared with previous methods, our method can accurately predict the 3D hand pose from monocular data despite challenges arising from ambiguity, occlusion, and noisy inputs.

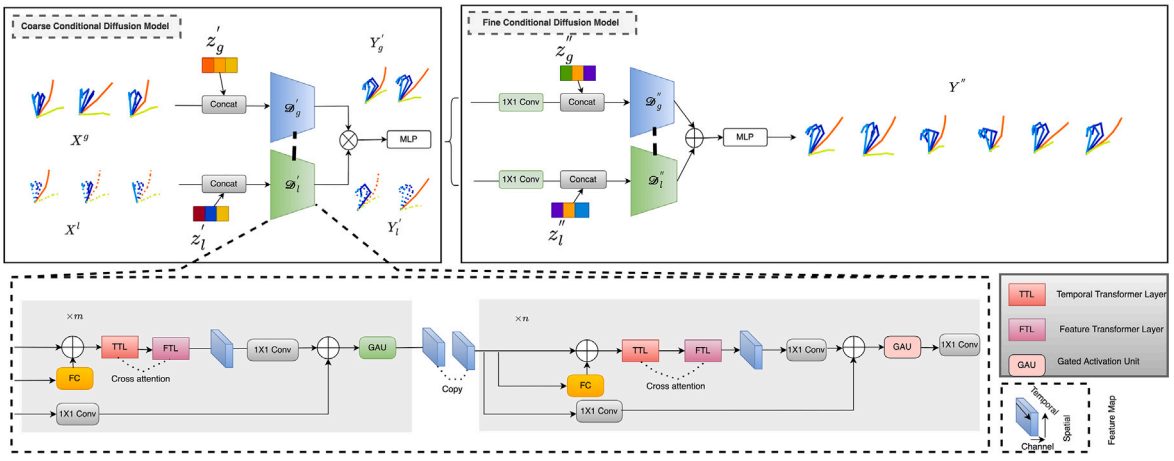

Fig. 2. The Dual-diffusion paradigm introduced in this study embodies a distinctive neural network architecture designed to concurrently acquire both local and global representations. This architectural configuration incorporates Global D_g and Local D_l Diffusion (GLD) blocks, strategically engineered to enable the seamless diffusion of information between localized and overarching features.

information to make predictions. In Fig. 1, our hand pose prediction method exhibits higher randomness and can capture a variety of hand gesture changes in short and long sequences, which other methods do not possess. Particularly in predicting long-term sequence hand poses, our method demonstrates greater accuracy.

The primary contributions of our study can be summarized as follows:

(a) We are introducing a research area focusing on predicting hand pose positions in the short and long term. We evaluated current

models using three well-known hand pose estimation datasets (MSRA [12], F-PHAB [13], and BigHand2.2M [14]) to create a detailed standard for comparison. This standard will help researchers understand and compare different techniques for predicting hand pose positions in the short and long term.

(b) Our model uniquely incorporates a two-layer attention mechanism within each residual layer of the diffusion framework, utilizing both time-based and feature-based transformers through well-designed GLD blocks. This dual transformer setup allows the model to capture critical temporal interdependencies and feature relationships within the hand pose data in both local and global diffusion processes. Such comprehensive attention to both temporal and spatial dimensions of the data enables the model to predict more accurate and dynamically consistent hand movements.

(c) Our approach treats the task of predicting 3D hand positions as a process of removing noise. The main goal is to make precise predictions of 3D hand positions from single-camera images, which is challenging due to inherent ambiguity, occlusion, and input noise to imprecise 3D pose estimations.

(d) Our method outperformed previous approaches on the MSRA [12], F-PHAB [13], and BigHand2.2M [14] datasets.

2. Related work

In this part of our paper, we give a thorough explanation of the different methods previously used for motion prediction: Time Series Forecasting, Deep learning-based Motion Prediction, and Diffusion Models. These methods are well-known in motion prediction, and each has its own benefits for understanding time-related patterns and spatial features. We will provide detailed explanations of the basic ideas, structures, and main features of each approach.

2.1. Time series forecasting

In recent studies, many new methods have been developed to improve how deep learning models predict tasks over time [15]. These include techniques like Convolutional Neural Networks (CNNs) [16], Recurrent Neural Networks (RNNs) [17], and methods based on Transformers [18,19]. [20] presents a methodology for forecasting time series data utilizing Generalized Regression Neural Networks (GRNNs), focusing on leveraging their fast computation and high accuracy potential. The core of the discussion revolves around the crucial modeling decisions required for effective forecasting with GRNNs, proposing various strategies for each decision and evaluating them based on forecast accuracy and computational efficiency. SeriesNet [21] is a novel time series forecasting model designed to address the limitations of traditional forecasting methods, particularly their inability to effectively extract meaningful features from sequence data, leading to suboptimal forecasting accuracy. SCINet [22] introduces a recursive downsample-convolve-interact architecture that effectively models complex temporal dynamics by extracting valuable temporal features from downsampled sub-sequences using multiple convolutional filters. This approach has demonstrated promising results in time series analysis. Xu et al. [23] introduce a new method using a Long Short-Term Memory (LSTM) autoencoder and multitask learning to predict the levels of PM 2.5, a type of air pollutant, in different locations across a city. The method focuses on understanding the connections between pollution at various sites and includes weather information from monitoring stations to improve predictions. It uses advanced LSTM networks to capture how air pollution changes over time and space, and employs a stacked autoencoder to identify key patterns in weather that affect pollution levels. By learning from multiple pollution data series simultaneously, the model can better use information from different sites, leading to more accurate predictions. NHITS [24] aims to address the challenges associated with long-horizon forecasting in the context of neural forecasting models. Long-horizon forecasting involves predicting future values of a time series over extended time intervals, which poses difficulties due to the volatility of predictions and the computational demands. ETSFormer [25] revolutionizes the application of Transformers in time-series forecasting by integrating the principles of exponential smoothing. Triformer [26] is a novel approach for efficient and accurate long sequence multivariate time series forecasting. It addresses limitations in existing attention-based models, offering linear complexity through a novel patch attention mechanism and enhancing accuracy by employing variable-specific parameters.

2.2. Deep learning-based motion prediction

Forecasting human motion is about using past movements to predict future ones. Earlier methods like Gaussian Mixture Models work well for simple motions but struggled with complex and long-term predictions. The rise of deep learning has brought big improvements in this area, using powerful network models like Recurrent Neural Networks (RNNs) [27], Convolutional Neural Networks (CNNs) [28], and Graph Convolutional Networks (GCNs) [29,30] to make more accurate predictions of human movements. The Seq-to-seq framework [28] introduces an innovative methodology grounded in Convolutional Neural Networks (CNNs), leveraging the hierarchical structure of CNNs to adeptly capture spatial and temporal correlations. Chen et al. [31] present the Pose Guided Structured Region Ensemble Network (Pose-REN), an innovative method that improves hand pose estimation by extracting and hierarchically integrating critical regions from CNN feature maps, directed by an initially estimated pose, and using tree-structured connections to iteratively refine the pose, achieving outstanding results on public datasets. However, traditional self-attention mechanisms focus on calculating the similarities among nodes, often overlooking the graph structural data of nodes. To enhance the modeling of node structure representations in graphs with self-attention, Zhao et al. [32] develop a novel transformer architecture for 2D-to-3D pose estimation that incorporates graph convolution layers within the transformer, enabling it to capture both global information and high-order connections between joints effectively, as demonstrated by extensive testing on public datasets. Wang et al. [33] propose a new approach for predicting human movements by focusing on the speeds of body joints instead of just their positions, using a special model that learns from these speeds and a unique training method that considers body movements as a series of mechanical actions, leading to more accurate and stable predictions. Pioneering the domain, Diller et al. [34] look at predicting poses based on their meaning and goal, rather than just time. They use a probabilistic method to handle the different possible key poses a person might take. This method guesses future poses by considering how different body parts move together. Wang et al. [35] use a Multi-Range Transformers model that combines a local-range encoder for individual motions and a global-range encoder for social interactions to predict each person's motion by attending to both local and global features, enabling accurate and diverse long-term predictions, even for up to 15 individuals simultaneously. Mao et al. [36] introduce an attention-based feed-forward network that enhances human motion prediction by utilizing motion attention to capture similarities between current and historical motions for predicting future poses. Guo et al. [37] introduce SkelNet, a novel network for long-term human motion prediction that separately learns local representations for different body components using component-specific layers. Shi et al. [38] enhance traffic participants' multimodal future behavior prediction by using a small set of learnable motion query pairs to jointly optimize global intention localization and local movement refinement, improving training stability and prediction accuracy without relying on dense goal candidates. However, unlike previous local and global transformers, our method uniquely incorporates a two-layer attention mechanism within each residual layer of the diffusion framework, utilizing both time-based and feature-based transformers. This dual transformer setup allows

the model to capture critical temporal interdependencies and feature relationships within the hand pose data. Such comprehensive attention to both temporal and spatial dimensions of the data enables the model to predict more accurate and dynamically consistent hand movements. Additionally, the method employs advanced denoising diffusion probabilistic models that facilitate the reconstruction of clean hand pose trajectories from noisy inputs. By reversing the diffusion process, the method effectively recovers the original hand pose sequence from the degraded inputs, thereby significantly improving the quality of the predictions. This approach ensures robustness against input noise.

2.3. Diffusion models

The Motion Diffusion Model (MDM), as introduced by [39], presents a meticulously tailored diffusion-based generative model for the purpose of motion generation. This strategic choice permits the application of established geometric loss functions, such as foot contact loss, to the positions and velocities of the generated motion. MDM underscores its versatility as a universal methodology, accommodating diverse generation tasks and facilitating various modes of conditioning. Similarly, the MotionDiff model [40] introduces a diffusion-based probabilistic framework inspired by the diffusion process observed in nonequilibrium thermodynamics. Within this framework, human joint kinematics are likened to heated particles diffusing from their original states to a noise distribution. This diffusion process inherently "whitens" latent variables without necessitating trainable parameters, while concurrently introducing fresh noise at each diffusion step, thereby promoting enhanced motion diversity. The anticipation of future human motion is subsequently framed as a reverse diffusion process, wherein the noise distribution is transformed into realistic future motions, contingent upon the observed sequence. Furthermore, the latent diffusion model (LDM) approach [41] puts forth an inventive technique grounded in a diffusion model (DM) for the purpose of image reconstruction from human brain activity, acquired through functional magnetic resonance imaging (fMRI). LDM adopts a latent diffusion model known as Stable Diffusion, which not only mitigates the computational burden associated with DMs but also maintains their commendable generative performance.

3. Proposed method

As illustrated in Fig. 2, the Dual-diffusion model proposed in this study comprises two distinct stages, each serving a specific and well-defined purpose. The initial stage involves the prediction of short-term hand poses, whereas the subsequent stage focuses on the prediction of long-term hand poses. Hand features are inputted into the global diffusion model, while enhanced finger features are channeled into the local diffusion model. Both the global and local diffusion models are composed of a series of residual layers, with each layer containing two transformers characterized by identical input and output dimensions. The primary transformer, denoted as the temporal transformer, is tasked with modeling the temporal dynamics inherent in hand pose sequences. The resulting output from the temporal transformer is then directed to a second transformer, referred to as the spatial transformer. This transformer focuses its attention on the intricacies of hand poses within individual frames. The Dual-diffusion model, as described, follows a coherent and structured process, leveraging the interplay between global and local diffusion mechanisms to effectively forecast both short-term and long-term hand poses. The integration of multiple residual layers within each diffusion model enhances the model's depth, facilitating the recognition of intricate temporal and spatial patterns within the hand features and enriched finger features.

3.1. Problem definition

We record the hand pose of an individual, characterized by the 3D spatial coordinates of its V joints, across a temporal span of T frames. Subsequently, our objective is to predict the trajectory of the V hand joints over the ensuing K forthcoming frames. This task involves the representation of the individual joints through 3D vectors denoted as $x_{v,k}$, representing the coordinates of joint v at time k. To represent the historical sequence of hand poses, we construct the tensor $X = [x_1, x_2, \dots, x_T]$, which is composed of matrices of 3D joint coordinates $x_i \in \mathbb{R}^{3\times V}$ corresponding to frame $i = 1, 2, \dots, T$. The ultimate aspiration is to forecast the forthcoming K hand poses, designated as $Y = [x_{T+1}, x_{T+2}, \dots, x_{T+K}]$.

3.2. Denoising diffusion probabilistic models

The focus of this study centers on the acquisition of a model distribution denoted as $p_\theta(X_0)$, intended to approximate the underlying data distribution $d(X_0)$. The predicted sequence of hand poses is represented as X_t, where t ranges from 1 to K in discrete time steps. These hand pose sequences consist of latent variables that exist within the identical hand pose sample space as the initial configuration $X_0 = [x_1, x_2, \dots, x_T]$.

To achieve this objective, we employ diffusion probabilistic models, encompassing two distinct processes: the forward process and the reverse process. The forward process is characterized by the following Markov chain formulation (Eq. (1)):

$$d(X_{1:K}|X_0) := \prod_{t=1}^{K} d(X_t|X_{t-1}), \tag{1}$$

where the conditional distribution $d(X_t|X_{t-1})$ is normal distribution $N(\sqrt{1-\beta_t}X_{t-1}, \beta_t I)$. Here, β_t signifies a small positive constant representing the noise level.

Conversely, the reverse process seeks to denoise the anticipated hand pose X_t and recover the input hand pose X_0. This process is characterized by the subsequent Markov chain (Eq. (2)):

$$p_\theta(X_{0:K}) = p(X_K)\prod_{t=1}^{K} p_\theta(X_{t-1}|X_t), \quad X_T \sim N(0, I). \tag{2}$$

Within this formulation, the conditional distribution $p_\theta(X_{t-1}|X_t)$ is represented as normal distribution $N(X_{t-1}; \mu_\theta(X_t, t), \delta_\theta(X_t, t)I)$, with the parameters $\mu_\theta(X_t, t)$ and $\delta_\theta(X_t, t)$ defined as specified in Ho et al.'s Denoising Diffusion Probabilistic Models (DDPM) [42]. In light of this parameterization, the overarching process of global hand prediction is cast as an optimization problem, explicitly formulated as (Eq. (3)):

$$L_g = \min_g E_{X_0\sim d(X_0),\epsilon\sim N(0,1),t}\|\epsilon - \mathcal{D}_g(X_t, t|X_0)\|_2^2, \tag{3}$$

where $\mathcal{D}_g$ denotes the denoising function responsible for estimating the noise vector ϵ added to its noisy input X_t. Furthermore, the refined local finger prediction is characterized by the following loss function (Eq. (4)):

$$L_l = \min_l E_{X_0'\sim d(X_0'),\epsilon\sim N(0,1),t}\|\epsilon - \mathcal{D}_l(X_t', t|X_0')\|_2^2, \tag{4}$$

where $\mathcal{D}_l$ represents the denoising function dedicated to local finger prediction. The overarching loss function is then amalgamated as follows (Eq. (5)):

$$L = L_g + L_l. \tag{5}$$

The proposed methodology capitalizes on diffusion probabilistic models to effectively capture the temporal dynamics inherent in hand pose sequences and to refine local finger predictions. The formal articulation of the forward and reverse processes enables the reconstruction of realistic hand poses from predicted sequences. Through the utilization of denoising functions in the optimization process, the model achieves both elevated generative performance and diversified motion predictions.

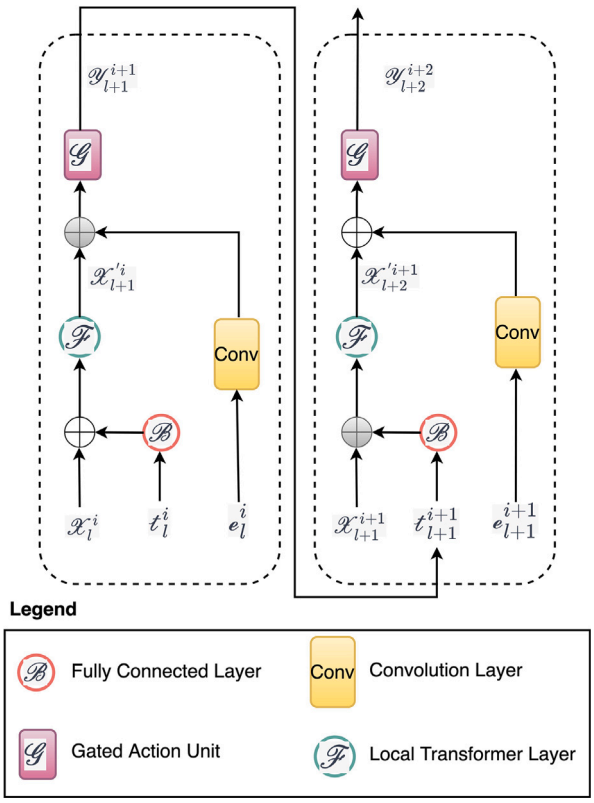

Fig. 3. GLD blocks for diffusion model.

3.3. Global and local diffusion blocks

In this section, we present the implementation details of the GLD (Global Local Diffusion) model for the prediction of hand poses. We begin by describing the essential data processing steps and by outlining the architectural framework denoted as $\mathcal{D}$.

The dataset representing hand pose information is denoted as X, M, e, wherein the hand pose space X is structured as $R^{K\times L}$, where $L = 3\times V$ signifies the 3D coordinates of the V joints. The temporal dimension is divided into two segments: N frames for observation and T frames for prediction. The mask matrix $M \in \{0,1\}^{K\times(N+T)}$ serves to indicate the presence of hand poses in the input frames (assigned the value 1) as opposed to the predicted or occluded frames (assigned the value 0). For the processing of the observed hand pose sequence X, we perform element-wise multiplication of M and X, followed by the introduction of Gaussian noise $z \sim N(0,1)$ to the unmasked region. This operation yields the modified hand pose data as $X = M \odot X + (1-M)z$. The principal objective is to forecast the hand pose sequence $Y = [X'_1, X'_2, \ldots, X'_N, X'_{N+1}, \ldots, X'_{N+T}]$, with the aim of minimizing the discrepancy $|Y - X| \odot (1-M)$, all given the initial input X.

The architectural design of $\mathcal{D}$ is underpinned by multiple residual layers, with the diffusion step denoted as T set to a value of 50. To effectively capture the temporal and feature-based relationships inherent within a specified time frame, a two-layer attention mechanism $\mathcal{F}$ is employed within each residual layer, contrasting with the conventional convolutional architecture. As depicted in Fig. 3, this involves the integration of a time-based Transformer layer and a feature-based Transformer layer, both instantiated as 1-layer Transformer encoders. To elucidate further, the time-based Transformer layer functions on tensors associated with individual features, thereby facilitating the capture of temporal interdependencies. Meanwhile, the feature-based Transformer layer focuses on tensors linked to discrete time instances, allowing for the elucidation of interdependencies related to features.

In addition to the parameters characterizing $\mathcal{D}$, auxiliary contextual information in the form of e is introduced into the model. Temporal embeddings, designated as $e_l^i = \{1 : L\}$, are integrated to facilitate the acquisition of temporal relationships. Here, $\mathcal{X}_l^i$ denotes the hand pose, and t_l^i represents the specific diffusion step. The representation of the projected hand pose can be formalized as follows (Eq. (6)):

$$\begin{aligned}
\mathcal{X}'^{i}_{l+1} &= \mathcal{F}(\mathcal{X}_l^i + \mathcal{B}(t_l^i)),\\
\mathcal{Y}^{i+1}_{l+1} &= \mathcal{G}(\mathcal{X}'^{i}_{l+1} + \mathrm{Conv}(e_l^i)),\\
\mathcal{X}'^{i+1}_{l+2} &= \mathcal{F}(\mathcal{X}^{i+1}_{l+1} + \mathcal{B}(t^{i+1}_{l+1})),\\
\mathcal{Y}^{i+2}_{l+2} &= \mathcal{G}(\mathcal{X}'^{i+1}_{l+2} + \mathrm{Conv}(e^{i+1}_{l+1})),
\end{aligned} \tag{6}$$

Here, $l = 1, 2, \ldots, L$, wherein L denotes the total count of GLD (Global Local Diffusion) blocks, and i serves as the index denoting the ith block within each residual layer. The local transformer layer $\mathcal{F}$ encompasses two distinct types of transformers: one dedicated to the comprehension of temporal dependencies, and the other dedicated to the understanding of feature dependencies.

The implementation of the GLD model adheres to the aforementioned procedural steps, encompassing data preprocessing and the utilization of attention-driven transformers within the diffusion model framework to facilitate the prediction of hand poses. The local diffusion model receives the 3D coordinates of each of the five fingers as input and incorporates a dual-layer attention mechanism within each residual layer. This mechanism integrates both a time-based and a feature-based Transformer layer, each implemented as a single-layer Transformer encoder. Meanwhile, the 3D coordinates of the entire hand are fed into the global diffusion model. Both the local and global models are optimized separately using their respective losses, L_l for local and L_g for global adjustments. Following optimization, a concatenation layer combines the features from both models. The local model specifically optimizes the positions of the individual fingers, effectively capturing subtle variations in finger movements. Additionally, each finger is consistently labeled with four 3D coordinates. This structured input not only simplifies the learning process but also ensures consistent output across each finger. The global diffusion model offers significant advantages in understanding and predicting hand movements by capturing the complex relationships between the fingers. This global diffusion is critical for generating more natural and coordinated hand movements, as it ensures that the movements of individual fingers are consistent with the overall hand structure and function. It can effectively handle complex gestures and interactions where the position and movement of one finger may influence or be influenced by others.

4. Experiments

4.1. Datasets

4.1.1. Dataset preprocessing

Normalizing hand pose data is a crucial preprocessing step to ensure that the model is not biased by the scale of the data and to help improve convergence during training. The process typically involves subtracting the mean and dividing by the standard deviation for each feature, ensuring that each feature contributes equally during model training. To address missing values, rows with empty entries are removed from the dataset. By applying these steps, the data is effectively centered and scaled, and noise or irrelevant information is minimized, resulting in a clean and standardized input for hand pose prediction tasks (see Table 1).

4.1.2. Dataset details

MSRA [12]. The MSRA dataset is an extensive repository containing more than 76,000 frames, acquired from a cohort of 9 subjects. Within this dataset, each subject's data encompasses 17 distinctive hand gestures, with an approximate count of 500 frames per gesture. Notably, each frame is accompanied by a meticulously segmented hand depth

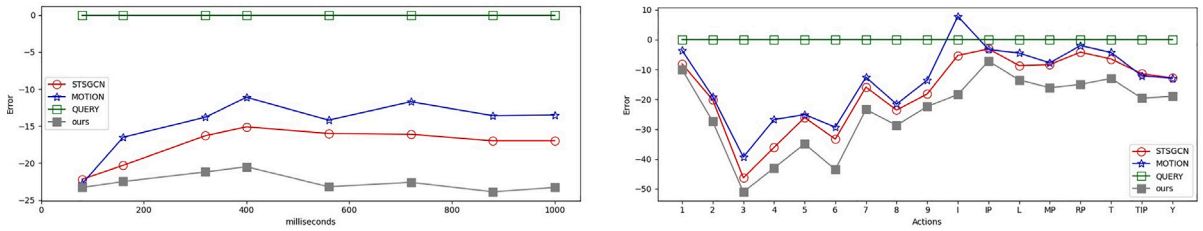

Fig. 4. Advantage analysis (MSRA).(a) Our approach demonstrates its greatest advantage at the 400 ms. (b) The standout benefit of our method is most pronounced when considering the execution of 'Action 3'.

Table 1
In this table, we present the experimental evaluation of our proposed model and compare its performance with state-of-the-art methods on three recent, large-scale, and challenging benchmarks: MSRA [12], F-PHAB [13], and BigHand2.2M [14]. These benchmarks were chosen to rigorously test the model's performance across diverse datasets and scenarios.

Database	MSRA [12]	F-PHAB [13]	BigHand2.2M [14]
Annotation	Track + refine	Track	Automatic
No. frames	76,375	105,459	2.2M
Frame rate	25	30	60
No. joints	21	21	21
No. subjects	9	6	10
View point	3rd	3rd	full
Depth map resolution	320 × 240	1920 × 1080	640 × 480

image. The dataset also includes meticulously annotated ground-truth information for $J = 21$ joints, which includes one joint representing the wrist and four joints corresponding to each finger. To assess the efficacy of our proposed methodology, we employ a rigorous leave-one-subject-out cross-validation strategy. This strategic approach ensures that each subject's data is systematically used for both training and testing phases, while upholding a balanced rotation scheme. The results obtained from this approach are dependable, robust, and amenable to direct comparison with alternative methods.

F-PHAB [13]. The dataset utilized in this study comprises a diverse compilation of 1175 action-centric videos, spanning 45 distinct action categories. These actions are executed within three distinct scenarios and are performed by a cohort of 6 actors. Within this dataset, a meticulous annotation effort has resulted in the provision of precise hand pose annotations alongside categorical labels denoting the executed actions. Furthermore, this dataset provides comprehensive insights by delivering detailed 6-dimensional pose information, 3D coordinates, angles, and mesh models of four distinct objects.

BigHand2.2M [14]. The dataset utilized in this study provides a comprehensive representation of hand pose intricacies, incorporating 21 joints with a cumulative count of 31 degrees of freedom. These degrees of freedom include six dimensions accounting for global pose, alongside 25 angles associated with individual joint orientations. The acquisition of hand pose data is achieved through meticulous manual measurements of bone lengths for each subject. Employing a suite of six magnetic sensors, each endowed with 6 degrees of freedom encompassing both location and orientation, and leveraging a hand model, the application of inverse kinematics facilitates the inference of comprehensive hand poses. These poses are aptly represented by the spatial coordinates of the 21 pivotal joints. A noteworthy accumulation of 375,000 frames was captured, with subjects encouraged to explore the complete pose space via deliberate and random actions.

4.2. Metrics

In the context of this study, we evaluate our proposed methodology using the Mean Per Joint Position Error (MPJPE) [46] as the chosen metric. MPJPE is a widely accepted measure for quantifying the precision of 3D hand pose estimation derived from images. It computes the average Euclidean distance between the predicted joint positions and their corresponding ground-truth positions. This error metric offers a comprehensive assessment of the accuracy of hand pose estimation, encompassing all individual joint positions within the entire dataset.

The use of MPJPE provides a standardized means of evaluation, enabling a fair and equitable comparison across distinct methodologies for 3D hand pose prediction. The computation of MPJPE is expressed as follows (Eq. (7)):

$$\mathcal{L}_{3D} = \frac{1}{J \times T_f} \sum_{j=1}^{J} \sum_{t=T_h+1}^{T_h+T_f} \|\Delta\hat{X}_{t,j} - \Delta X_{t,j}\|_2, \tag{7}$$

Here, $\Delta\hat{X}_{t,j}$ and $\Delta X_{t,j}$ signify the predicted and ground-truth displacement of a specific joint j between two consecutive frames. The notation $\|.\|_2$ denotes the l_2 norm, encapsulating the Euclidean distance measure.

4.3. Implementation details

The experimental configuration for our proposed methodology includes the utilization of PyTorch and Python 3.9, deployed on the Ubuntu 22.04 operating system. During the training phase, we adopt a learning rate of 1×10^{-2}, which decays by a factor of 0.1 every 10 epochs. The training regimen extends up to a maximum of 50 epochs, with a consistent batch size of 32 across all three distinct datasets: MSRA [12], F-PHAB [13], and BigHand2.2M [14].

Experimental Setup. The assessment of hand pose prediction, as commonly practiced, is conducted using the leave-one-subject-out (LOSO) cross-validation strategy. This methodology is well-established within the field and serves as a benchmark for evaluation. By consistently implementing the LOSO cross-validation approach in all our experimental analyses, we strive to derive outcomes that are both dependable and resilient. Such results are suitable for equitable comparison with other preeminent methodologies in the realm of hand pose prediction.

Lengths of Input and Output Sequences. In alignment with the methodology put forth in the study by [45], our experimental framework entails the establishment of specific parameters. Specifically, for the MSRA [12], F-PHAB [13], and BigHand2.2M [14] datasets, we configure the input sequence length to include 10 frames, while the corresponding output sequence length is set within the range of 2 to 25 frames.

4.4. Comparison to state of the arts

Within our comprehensive assessment, a meticulous comparison is conducted between our proposed method and a selection of cutting-edge algorithms, encompassing SCINET [22], QUERY [43], STSGCN [29], PGBIG [44], MOTION [45], LTD [30], and HRIS [46]. These evaluations are performed on three notably challenging datasets: MSRA [12], F-PHAB [13], and BigHand2.2M [14]. To foster standardized

Table 2
A comparative analysis of performance is conducted among various methodologies, focusing on short-term hand pose prediction. The evaluation metric is the Mean Per Joint Position Error (MPJPE), which measures the precision of joint positions for each activity within the MSRA dataset. Notably, the most exemplary performance outcomes are denoted in boldface, underscoring their prominence in the context of predictive accuracy.

	'1'				'2'				'3'				'4'				'5'				'6'			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
SCINET [22]	15.0	15.2	23.1	24.1	16.1	16.7	24.6	26.4	19.0	18.9	28.3	29.5	20.1	21.5	32.5	34.1	19.6	21.2	31.0	33.5	17.7	18.0	27.5	28.1
QUERY [43]	28.8	25.1	31.4	27.7	22.5	29.2	30.8	35.1	56.2	58.8	54.4	66.4	48.4	49.2	51.1	56.6	44.3	43.4	51.6	54.1	40.9	48.7	48.7	59.4
STSGCN [29]	9.8	12.1	17.3	20.3	11.3	13.7	23.0	25.3	11.9	14.5	22.5	26.0	12.9	17.0	26.6	31.9	12.6	17.4	28.0	31.9	11.8	14.1	23.6	27.8
PGBIG [44]	10.6	11.3	16.9	20.0	11.6	12.1	19.9	23.0	13.0	13.5	20.8	23.9	11.6	16.2	26.9	29.6	11.6	15.4	25.5	29.3	12.7	12.9	23.6	27.4
MOTION [45]	9.0	14.3	19.7	22.2	**9.5**	17.5	22.5	27.7	10.8	21.1	23.9	31.3	12.3	23.9	31.7	38.3	14.1	26.8	32.2	37.3	**10.5**	18.9	27.1	30.5
LTD [30]	9.8	13.7	18.7	22.0	11.6	15.8	22.9	26.4	13.9	19.2	24.8	28.7	14.5	21.2	30.4	34.2	14.5	21.1	31.1	34.8	15.0	20.1	27.4	33.9
HRIS [46]	11.2	11.1	16.8	18.8	13.5	11.4	19.9	24.5	13.4	13.2	22.0	24.5	13.0	14.9	25.2	31.0	14.0	15.2	25.1	30.8	13.6	11.9	21.0	26.6
Ours	**8.3**	**9.9**	**14.6**	**17.0**	9.9	**10.8**	**15.9**	**18.8**	**10.5**	**12.4**	**18.3**	**21.0**	**10.7**	**13.3**	**20.8**	**24.4**	**10.4**	**13.7**	**20.5**	**23.7**	10.7	**11.4**	**17.0**	**19.4**
	'7'				'8'				'9'				'I'				'IP'				'L'			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
SCINET [22]	16.5	18.0	26.0	27.9	17.4	19.4	27.9	31.0	18.8	21.1	32.2	34.5	16.4	18.1	26.6	28.4	21.9	23.7	33.7	36.2	19.2	20.7	31.2	33.9
QUERY [43]	25.9	30.2	35.2	34.7	28.0	36.9	41.2	45.4	28.8	31.4	39.6	39.0	36.8	36.2	44.1	41.7	26.0	28.8	35.1	37.2	28.5	29.5	34.4	37.4
STSGCN [29]	10.5	14.4	23.2	26.6	10.8	13.6	23.6	28.2	11.2	15.9	26.1	29.9	10.7	14.6	23.3	27.2	12.6	17.6	26.7	30.3	11.1	15.1	24.5	28
PGBIG [44]	10.7	13.5	21.3	27.1	12	13.4	24.7	26.4	10.2	13.7	23.5	28.3	9.6	14.2	22.1	27.1	12	16.3	25.7	28.4	11.4	13.1	21.5	27.6
MOTION [45]	9.8	16.6	23.7	30.3	9.4	16.8	28.5	30.5	10	18.6	27.5	34.1	10.5	18.4	27.7	35.2	**11.8**	19.4	28.2	31.5	10.3	17.4	28.5	31.9
LTD [30]	12.7	18.1	25	28.6	12.9	18.2	25.9	30.1	12.6	18	26.9	30.8	12.5	17.5	24.6	28.4	13.7	19.2	27	30.2	12.4	16.8	25.4	28.8
HRIS [46]	12.6	12.3	21.6	27.6	13.3	12.3	23.6	28.7	11.9	13.2	23.8	31.1	11.2	12.8	22.9	27.4	13.4	**15.5**	26	28.2	12	13.6	23.2	28.9
Ours	**9.7**	**12.3**	**18.5**	**21.1**	**10.8**	**12.3**	**17.8**	**21.6**	**9.8**	**12.4**	**20.0**	**23.6**	**9.0**	**12.3**	**18.4**	**21.9**	12	16.2	**23.8**	**27.6**	**9.8**	**13.1**	**20.8**	**24.4**
	'MP'				'RP'				'T'				'TIP'				'Y'				Average			
Milliseconds	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
SCINET [22]	23.5	24.7	36.4	38.2	25.8	25.8	39.2	40.6	20.5	22.6	32.8	36.3	21.8	23.9	33.9	36.9	19	20.3	30.2	32.2	19.3	20.6	30.4	32.5
QUERY [43]	33.3	34.4	44.0	46.2	33.1	34.6	42.9	45.5	37.5	34.7	41.3	43.1	30.6	36	44.3	48.9	30.9	29.0	40.1	38.5	34.1	36.2	41.8	44.5
STSGCN [29]	14.0	19.0	30.0	35.2	14.3	19.9	32.1	36.0	13.0	18.2	28.5	32.7	13.2	18.3	30.1	33.5	10.6	15.6	24.8	29.3	11.9	15.9	25.5	29.4
PGBIG [44]	**12.3**	17.8	29.7	34.6	13.8	17.2	28.1	**31.8**	13.2	16.2	25.4	31.6	13.5	17.2	28.2	32.0	10.7	13.9	23.4	26.9	11.8	14.6	24.0	27.9
MOTION [45]	13.5	23.2	33.5	40.6	**13.5**	23.6	34.6	44.2	12.8	19.6	28.1	34.9	12.3	20.1	30.7	35.8	11.2	18.8	27.6	30.9	11.3	19.7	28.0	33.4
LTD [30]	15.8	22.2	31.3	35.8	16.6	22.5	31.9	35.6	15.4	20.7	30.8	34.6	15.3	21.5	31.1	34.9	12.8	17.7	26.5	29.9	13.6	19.0	27.2	31.0
HRIS [46]	14.8	17.3	29.6	33.7	15.2	**17.2**	**28.1**	33.9	14.3	15.7	26.8	30.9	14.8	16.9	28.4	34.2	11.4	**13.3**	23.9	28.6	13.2	14.0	24.0	28.8
Ours	12.9	**17.1**	**25.7**	**28.8**	13.9	18.4	28.2	33.1	**12.4**	**16.1**	**24.3**	**28.8**	**12.3**	**16.5**	**24.8**	**28.7**	**10.0**	14.1	**20.6**	**24.4**	**10.8**	**13.7**	**20.6**	**24.0**

comparisons, all algorithms are provided with an input sequence length of 10 frames, equivalent to a temporal span of 400 ms. Subsequently, a dual evaluation paradigm is implemented, encompassing both short-term and long-term prediction scenarios.

In the domain of short-term prediction, the algorithms are tasked with forecasting forthcoming poses spanning 2 to 10 frames, corresponding to temporal intervals ranging from 80 to 400 ms. Meanwhile, for long-term prediction, the algorithms forecast spanning 14 to 25 frames, denoting temporal expanses of 560 to 1000 ms.

Distinct attributes characterize the specialized methodologies within this comparative cohort. SCINET and QUERY excel in time series forecasting, with SCINET harnessing temporal relationships via downsampling to achieve superior predictive accuracy. QUERY addresses computational intricacies through the utilization of sparse matrices for approximating attention matrices, yielding exceptional performance across diverse time series datasets. STSGCN leverages gating networks and adaptive adjacency matrices, rendering it adept at modeling intricate spatio-temporal dependencies inherent within human motion data. PGBIG capitalizes on a two-stage prediction paradigm and dense graph convolutional networks to extract spatiotemporal features, resulting in enhanced predictive precision. MOTION, based on multi-layer perceptrons (MLPs), exhibits efficient forecasting prowess for 3D human body pose, achieving state-of-the-art performance with an optimized parameter configuration. LTD operates within trajectory space, employing a graph convolutional network to encapsulate temporal smoothness and spatial interdependencies among body joints. HRIS introduces a novel attention-based feed-forward network, effectively harnessing the recurrent nature of human motion to capture extended motion patterns via motion attention and graph convolutional networks.

MSRA [12]. Table 2 offers a comprehensive quantitative comparison, delineating the performance of short-term predictions (time intervals below 400 ms) on the MSRA dataset. The analysis compares our proposed Dual-diffusion model against previous state-of-the-art approaches.

Meanwhile, Table 3 extends this scrutiny to long-term predictions (time intervals ranging between 400 ms and 1000 ms) within the same dataset. Across these multifaceted evaluation scenarios, our method consistently demonstrates superiority in the majority of cases. This discernible trend serves to underscore the heightened performance and efficacy of our proposed model in the realm of hand pose prediction on the MSRA dataset. The markedly enhanced accuracy and predictive quality attained by our model substantiate its potential to adeptly capture the intricate nuances inherent in hand motion. This, in turn, contributes to the generation of more precise and accurate forecasts compared to prevailing methodologies.

In Fig. 4, we present an exhaustive performance analysis and comparative study across distinct methods. Specifically, Fig. 4(a) and (b) establish QUERY as the baseline, subsequently quantifying the relative average prediction errors of STSGCN, MOTION, and our proposed approach in comparison to QUERY. The temporal evolution of these results is plotted, highlighting noteworthy trends. Notably, the performance trajectory reveals MOTION's superiority over QUERY, followed by STSGCN's superiority over MOTION, and culminating in our method's pronounced superiority over STSGCN. This heightened advantage of our method becomes most evident at the 880-ms mark.

Furthermore, in Fig. 4(b), an assessment of our method's advantage against QUERY, STSGCN, and MOTION is depicted. The findings unequivocally illustrate our method's substantial upper hand over the three benchmarked counterparts. This supremacy is especially salient for "Action 3", where our advantage is most pronounced.

In addition to the quantitative analysis, we provide qualitative comparisons in Fig. 5. These visual contrasts corroborate our method's ability to generate predictions closer to the ground truth, thereby

Table 3
The Mean Per Joint Position Error (MPJPE) measured in millimeters is employed for the purpose of evaluating the long-term prediction accuracy of 3D joint positions across the MSRA datasets. Our proposed model exhibits a substantial performance advantage over prevailing state-of-the-art methodologies across diverse time prediction horizons and action categories.

	'1'				'2'				'3'				'4'				'5'				'6'			
Milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
SCINET [22]	25.9	28.7	30.4	31.1	28.6	33.2	35.6	37.8	31.3	35.0	36.7	38.7	38.3	42.8	45.0	48.0	37.9	41.6	43.9	46.5	29.9	34.0	35.9	37.8
QUERY [43]	36.3	30.2	34.8	33.0	44.5	49.8	55.1	53.6	71.8	78.0	80.9	82.5	71.4	69.0	78.4	72.0	60.2	64.5	70.7	72.8	63.8	69.3	71.9	71.5
STSGCN [29]	22.9	24.9	26.7	27.9	30.2	31.9	34.9	35.3	30.7	32.8	34.6	35.1	36.6	39.3	42.3	46.0	38.5	39.4	44.6	44.6	34.4	35.5	38.6	41.3
PGBIG [44]	23.4	25.9	26.2	29.0	29.8	33.1	34.0	34.2	28.7	33.4	31.6	35.7	37.2	44.5	42.8	42.1	37.3	41.7	43.1	42.8	36.8	37.8	42.0	42.4
MOTION [45]	25.2	31.7	31.2	33.2	31.2	33.2	36.0	35.0	31.4	34.5	41.6	44.0	40.5	48.2	51.6	52.8	39.8	43.3	45.6	45.5	37.3	39.3	42.6	46.4
LTD [30]	23.2	25.9	26.7	27.5	30.5	34.1	31.5	34.6	31.4	34.6	36.2	42.4	39.9	42.5	46.0	46.9	40.4	43.5	46.1	47.3	39.7	44.2	44.6	48.4
HRIS [46]	24.1	25.8	29.1	31.9	31.8	33.0	36.9	36.7	32.5	36.5	36.5	38.2	36.6	40.9	43.5	44.7	38.2	41.0	44.8	45.7	34.1	37.5	45.6	42.6
Ours	**20.4**	**22.1**	**24.8**	**25.4**	**22.1**	**25.2**	**27.7**	**29.1**	**24.3**	**27.2**	**29.9**	**31.3**	**29.0**	**31.9**	**35.4**	**36.1**	**29.1**	**33.4**	**35.8**	**38.3**	**22.7**	**25.4**	**28.4**	**29.7**
	'7'				'8'				'9'				'I'				'IP'				'L'			
Milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
SCINET [22]	30.8	34.6	37.2	39.6	34.7	39.7	42.4	45.3	39.9	45.0	47.8	50.9	31.8	34.9	37.0	39.0	40.7	44.0	46.2	48.7	38.8	44.1	46.2	50.5
QUERY [43]	49.3	47.8	54.2	57.9	52.9	59.3	61.3	65.9	50.2	52.0	58.9	60.2	44.6	42.3	48.7	42.4	43.0	44.1	45.0	47.9	45.4	44.8	49.0	53.8
STSGCN [29]	33.9	36.7	38.3	40.6	32.2	35.2	37.7	40.7	35.0	38.8	40.7	42.9	34.1	36.5	43.4	40.1	35.8	38.2	41.9	39.8	33.4	36.1	40.3	41.5
PGBIG [44]	32.0	40.6	40.1	42.8	33.5	38.1	38.6	40.4	35.3	39.4	41.4	41.6	36.4	39.5	45.0	45.7	33.8	36.4	**37.2**	39.9	31.5	35.3	40.3	42.7
MOTION [45]	33.0	38.1	41.6	41.5	34.6	37.9	39.7	40.8	38.1	43.7	45.3	46.7	38.5	47.2	56.5	54.9	35.5	39.4	41.7	45.0	37.0	43.5	44.5	43.5
LTD [30]	34.6	39.7	40.2	41.1	34.5	37.8	39.5	41.3	38.0	42.2	46.0	46.4	34.3	40.4	41.2	44.9	34.5	37.8	39.5	40.4	34.4	36.6	38.0	39.9
HRIS [46]	36.0	37.7	42.7	42.1	35.0	39.4	39.5	38.2	36.9	39.5	41.2	44.3	36.5	40.8	42.3	45.4	34.2	34.8	39.6	41.5	34.9	35.7	41.6	39.4
Ours	**25.6**	**28.9**	**30.8**	**32.5**	**25.9**	**29.5**	**32.6**	**33.5**	**29.0**	**33.3**	**36.5**	**37.8**	**25.1**	**28.1**	**30.3**	**32.0**	**31.6**	**33.4**	37.8	**38.7**	**28.0**	**31.1**	**35.5**	**36.9**
	'MP'				'RP'				'T'				'TIP'				'Y'				Average			
Milliseconds	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
SCINET [22]	43.7	47.5	50.3	52.1	46.7	51.5	55.3	60.1	41.6	46.7	48.7	51.4	41.2	44.2	45.9	47.8	36.6	40.8	42.8	45.3	36.4	40.5	42.8	45.3
QUERY [43]	46.3	59.4	57.6	62.6	46.4	59.5	59.1	69.0	50.1	48.5	54.2	52.5	53.7	53.5	57.2	60.3	42.3	44.1	53.4	48.7	51.3	53.9	58.3	59.2
STSGCN [29]	43.4	44.8	49.2	51.8	44.0	48.2	54.9	54.7	40.3	44.5	47.7	47.1	40.0	42.1	45.8	47.0	34.4	37.5	40.6	40.9	35.3	37.8	41.3	42.2
PGBIG [44]	41.4	44.7	47.3	49.4	42.1	46.1	52.0	51.8	38.2	42.4	44.4	47.1	38.8	43.5	45.8	49.1	32.6	38.1	40.3	41.2	34.6	38.9	40.7	42.2
MOTION [45]	43.4	49.1	49.8	51.2	45.9	54.0	57.1	57.1	42.5	49.3	49.9	48.6	41.8	44.7	45.1	48.8	35.7	39.9	40.5	41.9	37.1	42.2	44.7	45.7
LTD [30]	42.7	47.4	51.3	50.1	42.8	48.6	51.3	53.0	41.1	46.5	49.4	52.7	40.2	43.2	45.2	45.7	34.3	37.1	40.5	41.9	36.3	40.1	42.0	43.8
HRIS [46]	43.8	48.1	47.6	49.2	44.9	46.5	53.5	54.1	40.5	43.5	48.0	50.5	39.6	42.5	44.3	46.3	32.8	34.8	38.8	39.2	36.0	38.7	42.1	42.9
Ours	**33.6**	**37.8**	**41.5**	**42.5**	**35.4**	**39.2**	**44.1**	**48.0**	**34.1**	**37.4**	**41.2**	**42.4**	**32.6**	**34.6**	**37.6**	**39.4**	**29.3**	**33.0**	**34.4**	**36.9**	**28.1**	**31.3**	**34.4**	**35.9**

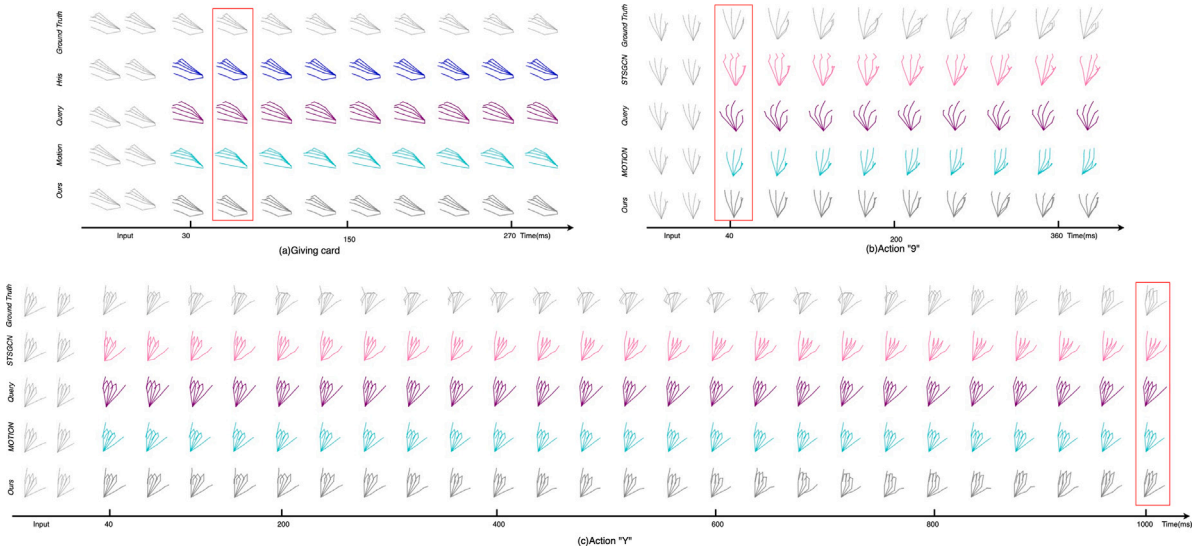

Fig. 5. Qualitative comparisons on the MSRA and F-PHAB datasets for both short-term and long-term hand pose prediction. (a) Our method produces short-term predictions that are closer to the ground truth compared to the baselines for the activity 'Giving card' on the F-PHAB dataset. (b) We showcase visual comparisons for the activity 'Action 9' on the MSRA dataset. (c) We provide visual comparisons for long-term predictions for the activity 'Action Y' on the MSRA dataset.

exceeding the performance of baselines across all three actions. This visual validation reinforces the superior precision and accuracy inherent in our proposed Dual-diffusion model for hand pose prediction.

F-PHAB [13]. In Table 4, we present a comprehensive overview of the average 3D position error outcomes. The magnitude of improvement our model exhibits in comparison to the previous state-of-the-art is indeed noteworthy, ranging from a 5% improvement for a prediction span of 3 frames to an 8% advancement for the more intricate scenario of 30 frames. Concurrently, Table 6 furnishes a detailed exposition of errors associated with diverse activities within the F-PHAB dataset. This encompassing list comprises tasks such as 'Charge cell phone', 'Clean glasses', 'Close peanut', 'Flip pages', 'Give card', 'Give coin', 'High fives', 'Light candle', and 'Open letter'—all of which have notably constituted focal points for comparisons within existing literature. Notably,

Table 4
The results pertaining to diverse models on the F-PHAB dataset, spanning distinct prediction time intervals measured in milliseconds (ms), are presented. The principal evaluation metric employed is the Mean Per Joint Position Error (MPJPE), quantified in millimeters (mm).

	Average							
Milliseconds	100	200	300	400	600	700	800	1000
SCINET [22]	18.2	22.2	25.8	28.0	34.5	37.3	40.5	44.9
QUERY [43]	38.6	42.2	45.2	48.9	53.7	56.3	59.7	62.5
PGBIG [44]	6.1	12.7	16.0	20.0	27.3	30.7	33.4	39.0
MOTION [45]	20.7	21.1	26.0	32.3	35.7	38.0	39.6	43.4
LTD [30]	13.3	18.0	23.0	26.8	33.4	36.5	38.5	43.3
HRIS [46]	9.1	18.5	21.9	22.3	27.6	31.1	33.3	37.8
Ours	**5.6**	**11.1**	**15.5**	**19.3**	**25.9**	**29.2**	**30.9**	**34.9**

Table 5
The evaluation metric used is the Mean Per Joint Position Error (MPJPE) in millimeters (mm), and lower values indicate better performance on the BigHand2.2M dataset and tested on various time horizons to assess its predictive accuracy.

	Average							
Milliseconds	100	200	300	400	600	700	800	1000
SCINET [22]	22.3	26.3	30.5	34.2	35.3	38.7	40.0	43.4
QUERY [43]	35.1	40.2	41.8	48.4	50.1	51.0	52.0	53.9
STSGCN [29]	13.6	22.1	25.7	29.8	34.3	35.7	37.1	40.8
PGBIG [44]	13.9	20.8	25.5	30.5	35.5	37.3	38.7	41.8
MOTION [45]	17.4	25.7	28.2	30.7	34.9	37.3	39.0	43.7
LTD [30]	16.0	24.8	28.1	31.6	36.0	38.6	39.9	43.1
HRIS [46]	14.7	21.8	25.5	29.9	35.6	36.8	38.9	43.2
Ours	**13.0**	**20.6**	**24.3**	**27.1**	**31.8**	**34.3**	**35.4**	**37.7**

our method consistently outperforms all baselines on average, thereby underscoring its intrinsic capacity for precision in predicting hand poses across a spectrum of distinct activities. To further underscore the efficacy of our approach, we present qualitative comparisons in Fig. 5. These visual juxtapositions serve to vividly illustrate the heightened performance achieved by our Dual-diffusion model. This is reflected in its ability to adeptly capture the intricacies of hand pose variations while concurrently aligning closely with ground truth references across diverse actions and temporal horizons.

BigHand2.2M [14]. Displayed in Table 5 are the outcomes derived from the short-term and long-term 3D prediction exercises conducted on the BigHand2.2M dataset. Notably, our model perform best in terms of average error across the entire spectrum of short-term and long-term forecasting endeavors. The attained performance improvement is notable, encompassing enhancements ranging from 4.6% at a 100-ms interval to an impressive 8.2% at the 1000-ms mark. This underscores the substantial advantages afforded by the proposed Dual-diffusion model in the realm of 3D hand pose prediction.

Qualitative evaluation. Fig. 5 presents sample predictions on the MSRA and FPHAB datasets. In Fig. 5(a), our method demonstrates accurate short-term forecasting for the 'Giving card' action, with an average accuracy of 15.5 mm at 9 frames (300 ms). The predicted hand poses visually match the ground truth and showcase a diverse range of hand shapes, indicating the effectiveness of our approach. In contrast, the MOTION method fails to capture the bending of the thumb in short-term hand pose prediction and lacks diversity in its predictions. Fig. 5(b) shows the results for 'Action 9' at 10 frames (400 ms). Our method's predicted hand poses closely align with the ground truth, whereas the poses generated by the STSGCN and QUERY methods exhibit considerable differences in the degree of bending for the middle finger, ring finger, and little finger compared to the ground truth. Furthermore, Fig. 5(c) offers a pertinent case study by presenting an assortment of predicted poses garnered from distinct methodologies, specifically in the context of the MSRA dataset. Evidently, as the prediction timeframe extends, our method consistently exhibits superior performance when juxtaposed against alternative approaches.

4.5. Ablation study

In the subsequent analysis, we aim to evaluate the effects of the various components inherent to our approach on the MSRA dataset.

Number of GLD blocks. We conduct an ablation study to assess the impact of varying the number of Global Local Diffusion (GLD) blocks, as detailed in Table 7. In our network architecture, using only 2 GLD blocks yielded favorable outcomes, showcasing a promising performance while maintaining a total parameter count of $0.066M$. Notably, our network exhibited optimal performance for short-term prediction when equipped with 7 GLD blocks, while the optimal configuration for long-term prediction required 6 GLD blocks.

Network architecture. Table 8 details an ablation study for different network configurations of our model DM-HAP on the MSRA dataset, with each row showing the mean per joint position error (MPJPE) in millimeters for varying configurations across multiple forecast lengths ranging from 80 ms to 1000 ms. The table compares the effects of different diffusion and conditioning methods: global diffusion (g.only), local diffusion (l), coarse conditional diffusion (co), and fine conditional diffusion (fi). For instance, the configuration labeled 'g.only, co+fi (Ours)' starts at 9.9 mm at 80 ms, indicating a substantial performance enhancement over the comparative methods 'LTD [30]' and 'PGBIG [44]' which start at 13.6 mm and 12.7 mm respectively. The inclusion of the fine conditional model (fi) alongside global diffusion notably improves long-term prediction accuracy, demonstrating a reduced error at 1000 ms with 37.8 mm compared to the coarse-only model, which shows 40.8 mm. Similarly, the addition of local diffusion in 'g+l, co.only (Ours)' helps in significantly reducing the error to 8.9 mm at 80 ms by optimizing the positions of the five fingers separately. This model effectively captures intricate variations in finger positioning. Moreover, each finger is labeled with four 3D coordinates in a consistent order. This ordered information is encoded into the input, making it simpler for the model to learn and maintain consistent output for each finger. Moreover, the full configuration 'g+l, co+fi (Ours)' shows consistent performance improvements across all lengths, confirming the synergistic effect of combining both local and global diffusions with fine and coarse conditionings, culminating in an MPJPE of 35.9 mm at 1000 ms, the best among the configurations for long-term predictions. This detailed analysis illustrates the substantial impact of each component on the network's overall efficacy, particularly highlighting the value of fine conditioning in enhancing long-term predictions and local diffusion in boosting short-term accuracy. Fine conditioning enhances long-term predictions because the input for the fine conditioning module includes both the ground truth and the initial output from the previous coarse diffusion network. This initial output improves the performance of the fine conditioning module. In short-term hand pose prediction, the changes in fingers are minimal. Labeling the fingers with four 3D coordinates in a consistent order benefits local diffusion, ensuring accurate short-term prediction.

Number of feature embedding. Fig. 6 provides a comparative analysis of the mean 3D error in millimeters across all actions encompassed by the MSRA dataset, comparing distinct feature embedding configurations within the feature transformer layer. It is observed that the dual-diffusion mechanism yields the most favorable average error when instantiated with a parameter value of $N = 6$. Notably, increasing the feature dimensions does not yield discernible empirical enhancements in performance.

Number of temporal embedding. Table 9 presents the outcomes of our ablation study, in which we systematically vary the number of temporal embeddings employed within the temporal transformer layer. Notably, our investigation reveals that the use of 64-dimensional temporal embeddings yields the most optimal results.

Table 6
The outcomes pertaining to both short-term and long-term prediction errors on the F-PHAB dataset are reported across all distinct actions. Remarkably, our proposed methodology exhibits sustained superiority over current state-of-the-art approaches across a majority of temporal intervals.

	Charge cell phone				Clean glasses				Close peanut butter				Flip pages				Give card			
Milliseconds	200	400	800	1000	200	400	800	1000	200	400	800	1000	200	400	800	1000	200	400	800	1000
SCINET [22]	11.9	12.4	20.7	25.4	17.8	16.5	21.6	23.5	18.8	25.2	43.3	48.1	9.8	11.9	20.7	24.7	18.0	24.7	46.4	59.4
QUERY [43]	30.5	26.3	31.8	35.8	46.1	58.6	41.6	51.1	29.5	41.2	42.2	46.1	35.7	36.6	33.3	38.2	47.6	51.6	68.5	81.9
PGBIG [44]	8.4	11.2	20.3	24.0	12.6	**15.3**	19.6	22.3	14.1	20.9	35.6	39.2	7.5	11.4	19.5	22.9	10.8	20.7	43.8	54.1
MOTION [45]	18.0	24.8	24.5	28.0	19.5	24.9	25.1	27.7	25.2	32.9	43.4	47.2	19.5	28.1	26.9	29.4	22.5	34.5	43.1	51.4
LTD [30]	11.6	16.1	22.3	27.1	14.6	20.0	23.2	25.7	16.7	25.9	37.5	42.7	13.9	18.2	24.5	27.9	15.9	27.4	40.8	49.6
HRIS [46]	13.8	12.4	16.4	23.0	17.1	16.8	19.3	21.6	16.5	21.3	32.7	40.0	12.1	11.4	18.8	21.2	17.3	25.9	43.3	50.8
Ours	**5.5**	**9.3**	**14.8**	**18.4**	**10.0**	15.8	**18.6**	**20.9**	**11.2**	**19.4**	**31.0**	**36.9**	**4.6**	**10.1**	**15.2**	**18.1**	**9.1**	**14.6**	**30.0**	**31.3**
	Give coin				High fives				Light candle				Open letter				Average			
Milliseconds	200	400	800	1000	200	400	800	1000	200	400	800	1000	200	400	800	1000	200	400	800	1000
SCINET [22]	19.6	26.0	41.3	46.6	49.4	73.3	105.0	100.1	9.8	12.3	21.9	27.1	17.6	21.3	36.2	44.1	22.2	28.0	40.5	44.9
QUERY [43]	35.6	41.2	46.6	54.0	52.7	83.7	87.9	97.3	34.7	50.8	42.5	48.9	45.3	58.6	65.4	72.3	42.2	48.9	59.7	62.5
PGBIG [44]	12.8	22.7	36.9	42.4	26.8	54.6	76.3	83.9	8.1	11.3	27.9	30.7	11.0	16.8	32.0	38.3	12.7	20.0	33.4	39.0
MOTION [45]	20.3	32.3	40.0	45.5	44.0	76.0	90.0	94.4	16.9	25.8	31.5	34.5	16.8	25.3	35.4	40.6	21.1	32.3	39.6	43.4
LTD [30]	19.2	28.4	40.5	46.0	33.4	60.5	83.4	87.7	13.5	19.4	30.8	38.4	15.7	23.8	36.0	42.4	18.0	26.8	38.5	43.3
HRIS [46]	18.0	24.8	37.5	43.0	29.0	52.9	76.9	85.3	11.3	12.0	25.4	25.9	18.5	20.6	31.6	41.2	18.5	22.3	33.3	37.8
Ours	**11.1**	**19.9**	**31.0**	**35.9**	**19.6**	**51.4**	**74.7**	**82.8**	**3.8**	**8.5**	**19.8**	**24.7**	**8.4**	**14.8**	**25.4**	**31.1**	**11.1**	**19.3**	**30.9**	**34.9**

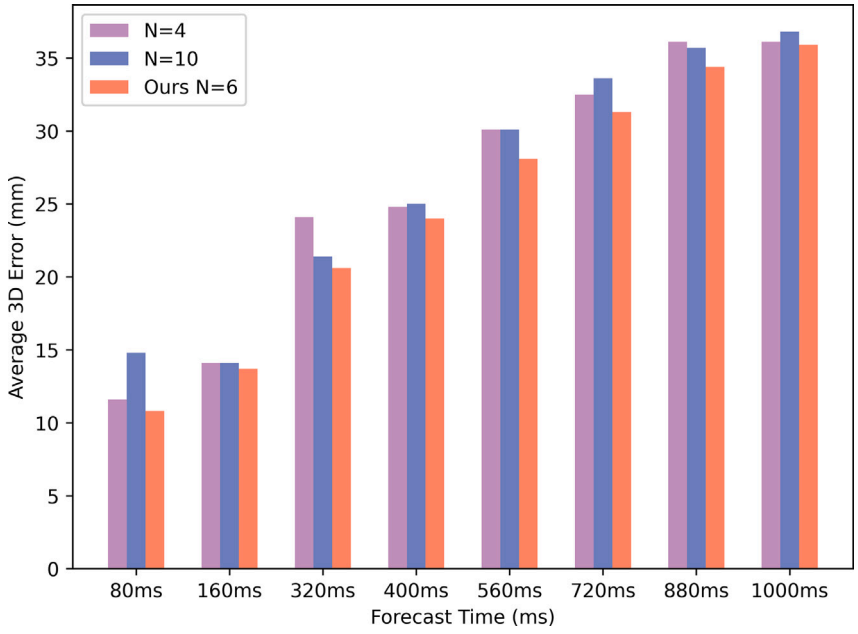

Fig. 6. Comparison of average 3D error in mm over all actions of the MSRA dataset at different feature embedding for feature transformer.

Table 7
An ablation study to assess the impacts of number of GLD blocks. Merely 2 GLD blocks yield a favorable result.

No. Blocks	# Param.(M)	MPJPE (mm)							
		80	160	320	400	560	720	880	1000
1	0.036	30.7	36.3	51.4	54.5	55.1	56.6	62.6	70.7
2	0.066	11.9	15.2	24.3	26.4	30.8	35.4	37.4	38.2
4	0.12	13.8	16.0	22.0	24.3	28.7	33.4	35.4	37.0
6 (Ours)	0.17	10.8	13.7	**20.6**	**24.0**	**28.1**	**31.3**	**34.4**	**35.9**
7	0.21	**8.8**	**12.7**	22.0	26.4	31.4	36.1	38.1	36.1
8	0.23	9.1	12.9	22.9	26.6	32.4	35.9	38.4	40.0

Table 8
An ablation study to evaluate the impact of different components of our network on the MSRA dataset. g.only denotes global diffusion. l denotes local diffusion. co denotes coarse conditional diffusion model. fi denotes fine conditional diffusion model.

Ablation	MPJPE (mm)							
	80	160	320	400	560	720	880	1000
LTD [30]	13.6	19.0	27.2	31.0	36.3	40.1	42.0	43.8
PGBIG [44]	12.7	12.9	23.6	27.4	34.6	38.9	40.7	42.2
g.only, co+fi (Ours)	9.9	13.7	21.2	23.9	28.4	32.6	35.1	37.8
g.only, co.only (Ours)	9.0	12.9	24.3	27.2	32.9	36.1	39.3	40.8
g+l, co.only (Ours)	**8.9**	**12.3**	20.7	24.5	30.3	34.3	37.6	39.5
g+l, co+fi (Ours)	10.8	13.7	**20.6**	**24.0**	**28.1**	**31.3**	**34.4**	**35.9**

5. Conclusion

This paper has addressed the intricate challenge of accurate hand pose prediction through the introduction of a pioneering framework known as Dual-diffusion. The proposed framework not only demonstrates promise in generating high-quality 3D hand poses from noisy data but also establishes a novel perspective by formulating hand pose forecasting as a reverse diffusion process. More specifically, it addresses

Table 9
Comparison of average 3D error in mm over all actions of the MSRA dataset at different time embedding for temporal transformer.

Temporal embedding	MPJPE (mm)							
	80	160	320	400	560	720	880	1000
16	9.4	14.0	22.0	24.5	30.7	33.7	36.8	39.1
32	12.4	14.4	21.0	24.7	29.6	32.9	38.5	38.3
48	12.3	14.3	21.9	26.8	29.7	33.3	35.6	38.1
64 (Ours)	**10.8**	**13.7**	**20.6**	**24.0**	**28.1**	**31.3**	**34.4**	**35.9**
72	14.4	19.2	21.6	24.4	29.9	32.4	36.1	37.3

the intricate task of capturing subtle joint motions, thereby enhancing hand pose predictions. This is achieved by introducing a parallel neural network architecture that amalgamates local and global representations through the incorporation of Global and Local Diffusion (GLD) blocks, ultimately leading to improved prediction accuracy. Experimental results show that the proposed method achieves state-of-the-art performance in hand pose prediction on three datasets.

CRediT authorship contribution statement

Zhifeng Wang: Writing – original draft, Visualization, Methodology, Investigation, Data curation. **Kaihao Zhang:** Writing – review & editing, Methodology, Formal analysis. **Ramesh Sankaranarayana:** Writing – review & editing, Supervision, Methodology, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zhifeng Wang reports financial support was provided by Australian National University. Zhifeng Wang reports a relationship with Australian National University that includes: non-financial support and travel reimbursement. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by Australian Government Research Training Program Scholarship.

References

[1] A. Brunetti, D. Buongiorno, G.F. Trotta, V. Bevilacqua, Computer vision and deep learning techniques for pedestrian detection and tracking: A survey, Neurocomputing 300 (2018) 17–33.
[2] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, H. Li, Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation, Adv. Neural Inf. Process. Syst. 33 (2020) 12034–12045.
[3] X. Yu, Y. Li, S. Zhang, C. Xue, Y. Wang, Estimation of human impedance and motion intention for constrained human–robot interaction, Neurocomputing 390 (2020) 268–279.
[4] Z. Sun, J. Chen, M. Mukherjee, C. Liang, W. Ruan, Z. Pan, Online multiple object tracking based on fusing global and partial features, Neurocomputing 470 (2022) 190–203.
[5] M. Woźniak, M. Wieczorek, J. Siłka, D. Połap, Body pose prediction based on motion sensor data and recurrent neural network, IEEE Trans. Ind. Inform. 17 (3) (2020) 2101–2111.
[6] X. Du, R. Vasudevan, M. Johnson-Roberson, Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction, IEEE Robot. Autom. Lett. 4 (2) (2019) 1501–1508.
[7] Y. Yang, Z. Ren, H. Li, C. Zhou, X. Wang, G. Hua, Learning dynamics via graph neural networks for human pose estimation and tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8074–8084.
[8] W. Dong, Z. Zhang, C. Song, T. Tan, Identifying the key frames: An attention-aware sampling method for action recognition, Pattern Recognit. 130 (2022) 108797.
[9] C. Zhong, L. Hu, Z. Zhang, Y. Ye, S. hong Xia, Spatio-temporal gating-adjacency GCN for human motion prediction. 2022 IEEE, in: CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 6437–6446.
[10] Z. Liu, K. Lyu, S. Wu, H. Chen, Y. Hao, S. Ji, Aggregated multi-gans for controlled 3d human motion prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 2225–2232.
[11] E. Barsoum, J. Kender, Z. Liu, Hp-gan: Probabilistic 3d human motion prediction via gan, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1418–1427.
[12] X. Sun, Y. Wei, S. Liang, X. Tang, J. Sun, Cascaded hand pose regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 824–832.
[13] G. Garcia-Hernando, S. Yuan, S. Baek, T.-K. Kim, First-person hand action benchmark with rgb-d videos and 3d hand pose annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 409–419.
[14] S. Yuan, Q. Ye, B. Stenger, S. Jain, T.-K. Kim, Bighand2. 2 m benchmark: Hand pose dataset and state of the art analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4866–4874.
[15] B. Lim, S. Zohren, Time-series forecasting with deep learning: a survey, Phil. Trans. R. Soc. A 379 (2194) (2021) 20200209.
[16] A. Dalca, M. Rakic, J. Guttag, M. Sabuncu, Learning conditional deformable templates with convolutional networks, Adv. Neural Inf. Process. Syst. 32 (2019).
[17] H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time series forecasting: Current status and future directions, Int. J. Forecast. 37 (1) (2021) 388–427.
[18] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 11106–11115.
[19] N.F. Sopelsa Neto, S.F. Stefenon, L.H. Meyer, R.G. Ovejero, V.R.Q. Leithardt, Fault prediction based on leakage current in contaminated insulators using enhanced time series forecasting models, Sensors 22 (16) (2022) 6121.
[20] F. Martínez, F. Charte, M.P. Frías, A.M. Martínez-Rodríguez, Strategies for time series forecasting with generalized regression neural networks, Neurocomputing 491 (2022) 509–521.
[21] Z. Shen, Y. Zhang, J. Lu, J. Xu, G. Xiao, A novel time series forecasting model with deep learning, Neurocomputing 396 (2020) 302–313.
[22] M. Liu, A. Zeng, M. Chen, Z. Xu, Q. Lai, L. Ma, Q. Xu, Scinet: Time series modeling and forecasting with sample convolution and interaction, Adv. Neural Inf. Process. Syst. 35 (2022) 5816–5828.
[23] X. Xu, M. Yoneda, Multitask air-quality prediction based on LSTM-autoencoder model, IEEE Trans. Cybern. 51 (5) (2019) 2577–2586.
[24] C. Challu, K.G. Olivares, B.N. Oreshkin, F.G. Ramirez, M.M. Canseco, A. Dubrawski, NHITS: Neural hierarchical interpolation for time series forecasting, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 6989–6997.
[25] G. Woo, C. Liu, D. Sahoo, A. Kumar, S. Hoi, Etsformer: Exponential smoothing transformers for time-series forecasting, 2022, arXiv preprint arXiv:2202.01381.
[26] R.-G. Cirstea, C. Guo, B. Yang, T. Kieu, X. Dong, S. Pan, Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting–full version, 2022, arXiv preprint arXiv:2204.13767.
[27] J. Martinez, M.J. Black, J. Romero, On human motion prediction using recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2891–2900.
[28] C. Li, Z. Zhang, W.S. Lee, G.H. Lee, Convolutional sequence to sequence model for human dynamics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5226–5234.
[29] T. Sofianos, A. Sampieri, L. Franco, F. Galasso, Space-time-separable graph convolutional network for pose forecasting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11209–11218.
[30] W. Mao, M. Liu, M. Salzmann, H. Li, Learning trajectory dependencies for human motion prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9489–9497.
[31] X. Chen, G. Wang, H. Guo, C. Zhang, Pose guided structured region ensemble network for cascaded hand pose estimation, Neurocomputing 395 (2020) 138–149.
[32] W. Zhao, W. Wang, Y. Tian, Graformer: Graph-oriented transformer for 3d pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20438–20447.
[33] H. Wang, L. Wang, J. Feng, D. Zhou, Velocity-to-velocity human motion forecasting, Pattern Recognit. 124 (2022) 108424.
[34] C. Diller, T. Funkhouser, A. Dai, Forecasting characteristic 3D poses of human actions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15914–15923.
[35] J. Wang, H. Xu, M. Narasimhan, X. Wang, Multi-person 3d motion prediction with multi-range transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 6036–6049.

[36] W. Mao, M. Liu, M. Salzmann, H. Li, Multi-level motion attention for human motion prediction, Int. J. Comput. Vis. 129 (9) (2021) 2513–2535.
[37] X. Guo, J. Choi, Human motion prediction via learning local structure representations and temporal dependencies, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 2580–2587.
[38] S. Shi, L. Jiang, D. Dai, B. Schiele, Motion transformer with global intention localization and local movement refinement, Adv. Neural Inf. Process. Syst. 35 (2022) 6531–6543.
[39] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, A.H. Bermano, Human motion diffusion model, 2022, arXiv preprint arXiv:2209.14916.
[40] D. Wei, H. Sun, B. Li, J. Lu, W. Li, X. Sun, S. Hu, Human joint kinematics diffusion-refinement for stochastic motion prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 6110–6118.
[41] Y. Takagi, S. Nishimoto, High-resolution image reconstruction with latent diffusion models from human brain activity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14453–14463.
[42] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[43] J. Klimek, J. Klimek, W. Kraskiewicz, M. Topolewski, Long-term series forecasting with query selector–efficient model of sparse attention, 2021, arXiv preprint arXiv:2107.08687.
[44] T. Ma, Y. Nie, C. Long, Q. Zhang, G. Li, Progressively generating better initial guesses towards next stages for high-quality human motion prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6437–6446.
[45] A. Bouazizi, A. Holzbock, U. Kressel, K. Dietmayer, V. Belagiannis, Motionmixer: Mlp-based 3d human body pose forecasting, 2022, arXiv preprint arXiv:2207.00499.
[46] W. Mao, M. Liu, M. Salzmann, 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 16, Springer, 2020, pp. 474–489.

Zhifeng Wang is presently pursuing a Ph.D. degree in College of Engineering and Computer Science, The Australian National University (ANU), Canberra, ACT, Australia. He graduated with an MS degree in computer vision and machine learning from the Australian National University in 2021. He His study focuses on computer vision and machine learning.

Kaihao Zhang received the Ph.D. degree from the College of Engineering and Computer Science, The Australian National University (ANU), Canberra, ACT, Australia. He has more than 40 referred publications in international conferences and journals, including CVPR, ICCV, ECCV, NeurIPS, AAAI, ACMMM, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), and IEEE TRANSACTIONS ON MULTIMEDIA (TMM). His research interests focuses on computer vision and deep learning

Ramesh Sankaranarayana received the ME degree from the Indian Institute of Science, India, and the Ph.D. degree from the University of Alberta, Canada. He is currently the associate director (Educational partnerships), Research School of Computer Science, Australian National University, Canberra, Australia. He research interests include information retrieval, human-centered computing and software engineering. He is a member of the ACM.