

HandMCM: Multi-modal Point Cloud-based Correspondence State Space Model for 3D Hand Pose Estimation

Wencan Cheng, Gim Hee Lee

Department of Computer Science, National University of Singapore
 {wccheng, gimhee.lee}@nus.edu.sg

Abstract

3D hand pose estimation that involves accurate estimation of 3D human hand keypoint locations is crucial for many human-computer interaction applications such as augmented reality. However, this task poses significant challenges due to self-occlusion of the hands and occlusions caused by interactions with objects. In this paper, we propose HandMCM to address these challenges. Our HandMCM is a novel method based on the powerful state space model (Mamba). By incorporating modules for local information injection/filtering and correspondence modeling, the proposed correspondence Mamba effectively learns the highly dynamic kinematic topology of keypoints across various occlusion scenarios. Moreover, by integrating multi-modal image features, we enhance the robustness and representational capacity of the input, leading to more accurate hand pose estimation. Empirical evaluations on three benchmark datasets demonstrate that our model significantly outperforms current state-of-the-art methods, particularly in challenging scenarios involving severe occlusions. These results highlight the potential of our approach to advance the accuracy and reliability of 3D hand pose estimation in practical applications.

Introduction

3D hand pose estimation (HPE) aims to determine the 3D locations of hand keypoints from sensor-acquired visual data. As it provides a fundamental understanding of hand actions and behaviors, HPE plays a crucial role in human-computer interaction applications such as augmented/virtual reality, robotics, *etc.* However, severe self-occlusions and hand-object occlusions during hand-object interactions make HPE a challenging task.

With the emergence of deep learning technologies, early deep learning-based methods utilized conventional convolutional neural networks (CNNs) to estimate hand poses from 2D images (Tompson et al. 2014; Ge et al. 2016; Ren et al. 2019; Chen et al. 2020; Fang et al. 2020; Ren et al. 2021a; Du et al. 2019). To address the challenging non-linear mapping problem from 2D images to 3D poses, CNN architectures that leverage 3D modalities such as point clouds (Ge et al. 2018; Ge, Ren, and Yuan 2018; Cheng, Park, and Ko 2021; Cheng and Ko 2023), voxelized representations (Ge

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

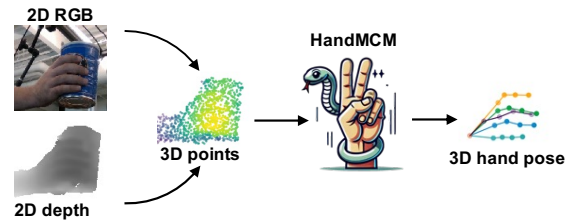


Figure 1: Illustration of our HandMCM concept. The model extracts multi-modal 3D point features from input depth images, RGB images, and corresponding point clouds. The Mamba architecture directly explores geometric and kinematic correspondences in 3D point space to accurately recover hand poses.

et al. 2017; Moon, Yong Chang, and Mu Lee 2018), or multi-modal combinations (Ren et al. 2023) have been extensively studied.

Despite the notable progress made by these approaches, accurate estimation remains challenging in scenarios involving hand-object occlusion and self-occlusion. To facilitate the reconstruction of occluded keypoints using information from visible keypoints, recent studies (Cheng and Ko 2023; Ren et al. 2023; Wang et al. 2023), such as Hamba (Dong et al. 2024), have attempted to capture the spatial kinematic relationships between keypoints through graph convolution. However, due to the wide variety of hand poses and interactions, those static graph-based methods struggle to accurately model complex occlusions and the dynamic spatial relationships between keypoints.

To address these challenges, we propose a novel architecture named *Multi-modal Correspondence Mamba* (HandMCM), which is the first to implement a state space model to capture dynamic kinematic correspondences for accurate 3D hand pose estimation under various occlusion conditions and complex poses. The proposed Correspondence Mamba mechanism demonstrates superiority in modeling dynamic kinematic correspondence compared to the straightforward combination of an auxiliary graph mechanism in Hamba (Dong et al. 2024). *This advantage stems from the intuitive motivation that spatial correspondence can be effectively processed as sequential data* (Liu et al.

2017) which is highly compatible with the Mamba model. To further enhance accuracy, HandMCM incorporates a local geometric injection and filtering mechanism that explicitly provides the Mamba module with local geometric information such as visible keypoints and occluded regions. This enables the Mamba mechanism to automatically activate kinematically corresponding features for occluded/visible keypoint estimation according to local details. Additionally, HandMCM employs a powerful multi-modal super point extraction approach that enhances the representation ability of 3D point clouds by incorporating visual cues such as occlusions and interacting objects from color and depth images.

We evaluate HandMCM on four challenging benchmarks, including the single-hand NYU dataset (Tompson et al. 2014), and the hand-object DexYCB (Chao et al. 2021) and HO3D (Hampali et al. 2020) datasets. The results show that our framework achieves state-of-the-art performance with a mean distance error of 7.06mm on the NYU dataset. Additionally, our model achieves state-of-the-art results on the hand-object DexYCB and HO3D datasets with mean distance errors of 6.67mm and 1.71cm, respectively. Our source code is publicly available at <https://github.com/cwc1260/HandMCM.git>.

Our **main contributions** are summarized as follows:

- We propose HandMCM, a novel Mamba-based architecture for accurate 3D hand pose estimation using RGBD images as multi-modal inputs.
- We propose a novel correspondence state space model designed to capture the kinematic relationships between keypoints, ensuring high reliability under occlusion.
- We conduct comprehensive experiments on large and challenging benchmarks, demonstrating the state-of-the-art performance of our proposed method.

Related Work

3D Hand Pose Estimation

Conventional methods utilizing 2D convolutional neural networks (CNNs) (Tompson et al. 2014; Ge et al. 2016; Ren et al. 2019; Chen et al. 2020; Fang et al. 2020; Ren et al. 2021a; Du et al. 2019) have become popular for 3D HPE due to their ease of implementation. However, these approaches have significant drawbacks in their inability to fully capture the complexities of 3D structures and their dependence on camera perspective.

To address these limitations, methodologies based on 3D CNNs (Ge et al. 2017; Moon, Yong Chang, and Mu Lee 2018) have been developed that use 3D voxel grids derived from depth images to represent volumetric data. Although these 3D techniques have shown progress in hand pose estimation, they require considerable memory and computational power, which can limit their practicality in real-world applications.

On the other hand, approaches based on PointNet (Ge et al. 2018; Ge, Ren, and Yuan 2018; Chen et al. 2018; Li and Lee 2019; Cheng, Park, and Ko 2021) analyze point cloud data, enabling an accurate representation of 3D structures.

PointNet is a deep learning architecture designed to process irregular and unstructured point clouds, and it has been first applied to hand pose estimation in HandPointNet (Ge et al. 2018). Subsequent improvements, such as the Point-to-Point model (Ge, Ren, and Yuan 2018) and SHPR-Net (Chen et al. 2018) have improved performance by producing point-wise probability distributions. In particular, SHPR-Net (Chen et al. 2018) integrated HandPointNet with an additional semantic segmentation network to improve performance. More recently, HandFoldingNet (Cheng, Park, and Ko 2021) introduces a novel folding approach to transform a predefined 2D hand skeleton into various hand poses, thus improving accuracy in estimation. Following this approach, HandR2N2 (Cheng and Ko 2023) proposes a folding-based recurrent framework to iteratively refine the locations of keypoints through local point region searches. However, a significant limitation of working with point clouds is the high computational cost associated with neighbor querying within dense point sets during convolution operations. As a result, many existing methods typically resort to utilizing sparse point clouds, which can hinder overall performance.

To bridge the shortcomings of images and point clouds, IPNet (Ren et al. 2023) and Keypoint-Fusion (Liu et al. 2024b) suggest the utilization of multi-modal representations, one combining 2D depth images and 3D point clouds while the other combining depth images and RGB images. As a result, the model adeptly extracts dense detail information while effectively capturing 3D spatial features, thereby facilitating accurate 3D hand pose estimation.

State Space Models

Classical state space models are originally developed in the context of linear dynamic systems in control theory and popularized by Kalman filtering (Kalman 1960), where the state variables and their transitions are described through first-order differential equations. Recently, a new series of state space models (Gu et al. 2020, 2021, 2022; Gu and Dao 2023) has emerged in natural language processing, leveraging their unique interpretability as a bridge between recurrent neural networks and convolutional neural networks to accelerate the learning of a long sequential context.

The most representative Structured State Space Sequence (S4) model (Gu et al. 2021, 2022), demonstrates its superior ability to effectively model dependencies within sequential data. Subsequently, the Mamba framework (Gu and Dao 2023) improves the S4 models by incorporating a linearly scalable projection matrix that adapts to the length of the input sequence for long-sequence models. Additionally, it includes a selection mechanism to selectively propagate or forget information along the sequence or scan path based on the current token. Although initially applied in natural language processing, Mamba-based architectures have recently been explored in computer vision tasks. With their global receptive field and dynamic weight mechanisms, these architectures show promise in image classification (Liu et al. 2024c; Zhu et al. 2024), video understanding (Li et al. 2025), and motion generation (Zhang et al. 2024). Moreover, Mamba-based architectures have also been investigated in other modalities such as 3D point clouds (Liang et al. 2024;

Liu et al. 2024a) and graph representations (Wang et al. 2024), enabling efficient 3D vision recognition and robust long-range graph prediction.

The research most closely aligned with ours involves the application of the Mamba hand reconstruction, Hamba (Dong et al. 2024). However, Hamba employs graph convolution for correspondence extraction, which fails to fully exploit the potential of the Mamba architecture. In contrast, our approach focuses on harnessing the capability of Mamba to model dynamic interaction patterns among keypoints. Furthermore, Hamba neglects the incorporation of local geometry, which, as demonstrated by our study, plays a critical role in improving the accuracy of pose estimation.

Preliminaries

State Space Model

A state space model (SSM) originated from the classical linear control theory is used to describe the dynamic relationship between a continuous-time scalar input $u(t)$ and output $y(t)$ through the following set of differential equations:

$$x'(t) = \mathbf{A}x(t) + \mathbf{B}u(t), \quad y(t) = \mathbf{C}x(t) + \mathbf{D}u(t), \quad (1)$$

where $x(t) \in \mathbb{R}^N$ represents the continuous-time state vector, while $x'(t)$ is its derivative. $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$, and $\mathbf{D} \in \mathbb{R}^{1 \times 1}$ are weighting parameters.

Discretization of SSM

The continuous SSM must be discretized to be deployed in deep learning models that generally accept discrete inputs. Let a discrete sequence of scalar inputs be u_1, \dots, u_L , the SSM can be reformulated into a discrete-time version with the following recursive expressions:

$$x_k = \bar{\mathbf{A}}x_{k-1} + \bar{\mathbf{B}}u_k, \quad y_k = \bar{\mathbf{C}}x_k + \bar{\mathbf{D}}u_k, \quad (2)$$

where $\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\mathbf{C}}, \bar{\mathbf{D}}$ represents transformations of the original parameters according to a specific discretization step.

Parallelization of SSM

The recursive relations make the discrete-time SSM resemble the structure of an RNN, where $x_k \in \mathbb{R}^N$ plays the role of a hidden state at time step k . More importantly, due to the linear structure of the recursion, it is possible to compute the entire output sequence y_1, \dots, y_L directly by applying a convolution with a fixed kernel $\bar{\mathbf{K}} \in \mathbb{R}^L$:

$$\begin{aligned} \bar{\mathbf{K}} &= (\bar{\mathbf{C}}\bar{\mathbf{B}}, \bar{\mathbf{C}}\bar{\mathbf{A}}\bar{\mathbf{B}}, \dots, \bar{\mathbf{C}}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}}), \\ y &= \bar{\mathbf{K}} * u. \end{aligned} \quad (3)$$

Once trained, the kernel $\bar{\mathbf{K}}$ serves as a comprehensive representation of the SSM, simplifying the model to a one-dimensional convolution with an extended kernel length.

Our Method

The architecture of our HandMCM is illustrated in Figure 2. Formally, the architecture of our HandMCM accepts multi-modal inputs which includes a hand depth image $\mathbf{D}_{in} \in$

$\mathbb{R}^{H \times W}$ and an RGB image $\mathbf{R}_{in} \in \mathbb{R}^{H \times W \times 3}$. It also randomly samples a set of pixels from \mathbf{D}_{in} and projects them into 3D point coordinates $\mathbf{P} \in \mathbb{R}^{N \times 3}$ to form the additional 3D perceptivity. The architecture produces output in the form of 3D keypoint coordinates $\mathbf{J} \in \mathbb{R}^{J \times 3}$.

Initially, HandMCM samples a multi-modal super point set that integrates features from input RGB and depth images, thereby providing rich visual information including occlusion cues for subsequent processing stages. From this super point set, keypoint tokens are extracted and used as input to the Mamba block. The proposed novel correspondence Mamba block iteratively refines these keypoint tokens by leveraging the intrinsic kinematic relationships among keypoints and their local features to achieve an accurate estimation of keypoint coordinates.

Multi-Modal Super Point Encoder

Given that the input point set is irregular and unordered, our HandMCM utilizes a point set convolutional layer (Qi et al. 2017b; Liu, Qi, and Guibas 2019) as a local 3D encoder. This layer extracts a subsampled 3D super point set $\mathbf{P}' \in \mathbb{R}^{N/2 \times 3}$ alongside the corresponding 3D local geometric features $\mathbf{F}_p \in \mathbb{R}^{N/2 \times C_p}$.

Subsequently, the depth and RGB images are each processed by independent ResNet-based autoencoders, resulting in 2D local visual feature maps $\mathbf{F}_d \in \mathbb{R}^{H/2 \times W/2 \times C_d}$ and $\mathbf{F}_{rgb} \in \mathbb{R}^{H/2 \times W/2 \times C_{rgb}}$, respectively. These 2D local visual features undergo 2D-3D projection and interpolation to create 3D point features $\mathbf{F}_{d \rightarrow p} \in \mathbb{R}^{N/2 \times C_d}$ and $\mathbf{F}_{rgb \rightarrow p} \in \mathbb{R}^{N/2 \times C_{rgb}}$. Following this projection, the features are concatenated with the super points, resulting in a combined feature set $\mathbf{F} \in \mathbb{R}^{N/2 \times (C_p + C_d + C_{rgb})}$. Finally, the super points are input into the keypoint token extraction module, which generates initial keypoint tokens for the Mamba blocks.

Keypoint Token Extraction

In our HandMCM framework, the keypoint tokens $\mathbf{X} \in \mathbb{R}^{J \times C}$, where the length is the number of tokens J , and the token size is C , serve as the starting point for processing within the proposed Mamba block. To construct the initial keypoint tokens, we utilize the joint-wise embeddings initially introduced by HandFoldingNet (Cheng, Park, and Ko 2021).

The keypoint tokens initialization module processes the super points, incorporating their features \mathbf{F} through two point set convolutional layers to generate 3D global vectors. The initial keypoint tokens (embeddings) \mathbf{X}_0 are then performed sequentially using a three-layer bias-induced layer (BIL) (Cheng and Lee 2019). In this process, the global vector is replicated J times and fed into the BIL layers. This results in the generation of embeddings for the J individual keypoints. Importantly, the BIL provides keypoint-wise independent biases, which function similarly to learnable positional embeddings, which enhance the flexibility and accuracy of the keypoint representation. After the acquisition of the keypoint tokens, a linear transformation is applied to

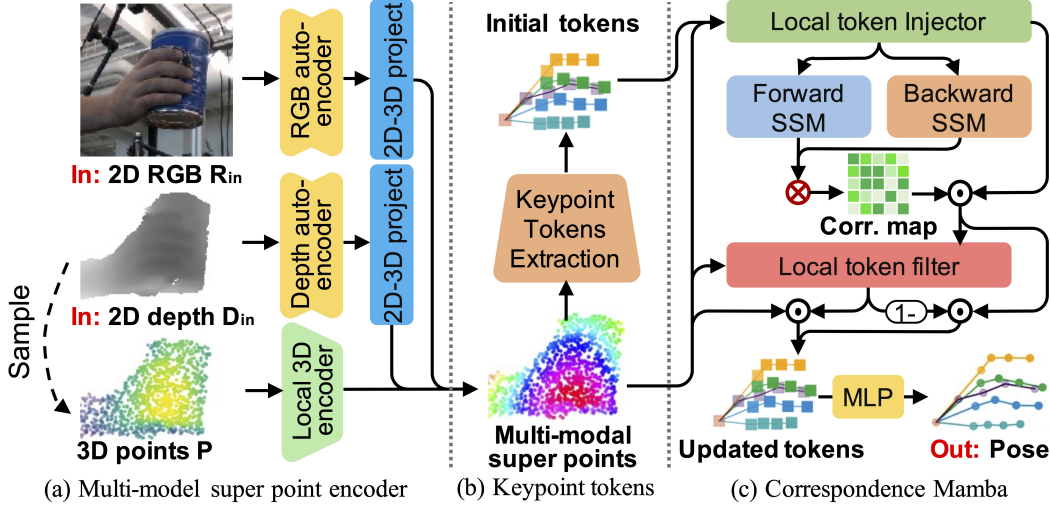


Figure 2: The HandMCM architecture. HandMCM takes as input an RGB image, a 2D depth image. The multi-modal super point encoder utilizes a PointNet-based local 3D encoder and two 2D autoencoders to extract local 3D point features and local 2D features from the RGB and depth inputs, respectively. The encoder then aggregates these projected 2D features with the 3D features, forming multi-modal super point features. Keypoint tokens are derived from the super points and fed into the novel correspondence Mamba to accurately estimate 3D keypoint coordinates.

project the high-dimensional tokens into the 3D space as the positions \mathbf{J}_0 for the local feature collection.

Correspondence Mamba

The correspondence Mamba illustrated in Figure 2 is the key contributing component of our proposed framework. It can effectively model the dynamic kinematic topology of keypoints by building a robust correspondence map by a state space model.

Bi-directional Correspondence SSM The basic architecture follows the bidirectional gated SSM (BiGS) (Wang et al. 2022). Specifically, let $\mathbf{X}_{i-1} \in \mathbb{R}^{J \times C}$ be keypoints tokens for the k -th Mamba block. The Mamba block first transforms tokens to construct the bidirectional inputs \mathbf{X}_f and \mathbf{X}_b , *i.e.*:

$$\begin{aligned} \tilde{\mathbf{X}} &= \text{LN}(\mathbf{X}_{k-1}), \quad \mathbf{V} = \text{GELU}(\mathbf{W}_v \tilde{\mathbf{X}}), \\ \mathbf{X}_f &= \text{GELU}(\mathbf{W}_f \tilde{\mathbf{X}}), \quad \mathbf{X}_b = \text{GELU}(\mathbf{W}_b \text{Reverse}(\tilde{\mathbf{X}})), \end{aligned} \quad (4)$$

where $\text{LN}(\cdot)$, $\text{GELU}(\cdot)$ and $\text{Reverse}(\cdot)$ are the layer normalization (Ba 2016), Gaussian error linear unit, and order flip operation on the token length dimension, respectively. \mathbf{W}_v , \mathbf{W}_f and \mathbf{W}_b denote the learnable linear weights.

Subsequently, the bidirectional inputs \mathbf{X}_f and \mathbf{X}_b are sent to two sequential blocks of forward and backward SSM layers to model the correspondence between keypoints:

$$\mathbf{U}_f = \mathbf{W}_{u_f} \text{SSM}(\mathbf{X}_f), \quad \mathbf{U}_b = \mathbf{W}_{u_b} \text{SSM}(\mathbf{X}_b), \quad (5)$$

where \mathbf{W}_{u_f} and \mathbf{W}_{u_b} are learnable parameter matrices for each direction. Based on the bidirectional correspondence between keypoints \mathbf{U}_f and \mathbf{U}_b , we explicitly construct a dynamic correspondence map that fuses the correlation of each pair of keypoints by an outer product (He et al. 2018):

$$\mathbf{M}_{\text{corr}} = \mathbf{W}_c (\mathbf{U}_f \otimes \text{Reverse}(\mathbf{U}_b)), \quad (6)$$

where \mathbf{W}_c is the trainable transformation weights. As shown in Figure 2 (c), each element in the correspondence map contains both forward and backward information. Finally, the correlation map is multiplied with the transformed token \mathbf{V} to obtain an updated token:

$$\mathbf{X}_k = \mathbf{M}_{\text{corr}} \mathbf{V}. \quad (7)$$

For the pose regression, a transformation weight $\mathbf{W}_r \in \mathbb{R}^{C \times 3}$ is learned to transform the updated high-dimensional tokens into the 3D keypoint coordinates,

$$\mathbf{J}_k = \mathbf{X}_k \mathbf{W}_r. \quad (8)$$

Local Token Injection & Filtering Although the suggested correspondence Mamba can capture dynamic kinematics, the performance remains unsatisfactory since the local geometric information around the keypoints is ignored. To address this issue, we propose a unique local token injection and filtering mechanism to provide the Mamba block with local information and to tailor the output appropriately.

Following previous studies (Cheng and Ko 2023; Cheng et al. 2024), we first group the local super point sets \mathcal{N}_{j_j} with corresponding features near each input keypoint $\mathbf{j}_j \in \mathbf{J}_k$ using the k -nearest neighbor algorithm. Each feature $\mathbf{f} \in \mathbf{F}$ is then concatenated with its token $\mathbf{x}_j \in \mathbf{X}$ and its local coordinate $\mathbf{p} - \mathbf{j}_j$. Subsequently, the local token is computed by the point set convolution operation (Qi et al. 2017a; Liu, Qi, and Guibas 2019), which aggregates local geometry, as follows:

$$\mathbf{x}_{loc,i} = \text{SetConv}_{(\mathbf{p}, \mathbf{f}) \in \mathcal{N}_{j_j}}([\mathbf{p} - \mathbf{j}_j, \mathbf{x}_j, \mathbf{f}]). \quad (9)$$

Once the local tokens are prepared, we inject them into the Mamba block. To achieve this, we reformulate the layer normalization $\text{LN}(\cdot)$ from Eq. (4) by multiplying the local token

with the input as follows,

$$\tilde{\mathbf{X}} = \text{LN}(\mathbf{X}_{k-1} \odot \mathbf{X}_{loc}). \quad (10)$$

For output filtering, the local geometry is first aggregated from the concatenation of local point sets and updated tokens from Eq. (7). The filter coefficients are then calculated on local features by applying the sigmoid activation:

$$\mathbf{G} = \sigma(\mathbf{X}_{loc}). \quad (11)$$

Finally, the updated tokens are filtered along with the local geometry. We can thus reformulate Eq. (8) for the final keypoint coordinate regression as the weighted sum of updated correspondence tokens and local tokens given by:

$$\mathbf{J}_k = (\mathbf{G} \odot \mathbf{X}_k + (1 - \mathbf{G}) \odot \mathbf{X}_{loc}) \mathbf{W}_r, \quad (12)$$

where \mathbf{W}_r is the same transformation weight in Eq. (8).

Loss Function

Since both the keypoint token initializer and Mamba blocks generate their respective estimated coordinates, all outputs are supervised through a loss function as follows:

$$\mathcal{L} = \sum_{k=0}^K \sum_{j=1}^J L1_{smooth}(\mathbf{j}_{k,j} - \mathbf{j}_j^*), \quad (13)$$

where \mathbf{j}_j^* indicates the ground-truth coordinate of the j -th joint, and K indicates the quantity of stacked Mamba blocks. As suggested in previous regression studies (Ren et al. 2019; Cheng, Park, and Ko 2021), we employ the smooth $L1$ function because it is less sensitive to outliers than the $L2$ function. The smooth $L1$ function is defined as follows:

$$L1_{smooth}(\mathbf{x}) = \begin{cases} 0.5|\mathbf{x}|, & |\mathbf{x}| < 0.01 \\ |\mathbf{x}| - 0.005, & \text{otherwise} \end{cases}. \quad (14)$$

Experiments

Datasets and Evaluation Metrics

NYU (Tompson et al. 2014) is a dataset for *single hand* pose estimation, which provides depth images of single hands captured from three distinct views using the PrimeSense 3D sensor. It contains 72K training frames and 8K testing frames. Following prior research practices (Ge, Ren, and Yuan 2018; Cheng, Park, and Ko 2021), 14 keypoints from the complete set of 36 annotated joints and a single view are used for evaluation. This strategy aligns with previous work to maintain standardized comparisons across studies while ensuring reliable benchmarking.

DexYCB (Chao et al. 2021) is a benchmark for *hand-object* interactions, containing 582K frames annotated with 21 hand keypoints from eight camera views. It involves 10 subjects interacting with 20 standardized YCB objects and includes four evaluation protocols: S0 (seen subjects, views, and objects), S1 (unseen subjects), S2 (unseen views) and S3 (unseen objects). These protocols and standardized objects make DexYCB a robust resource for testing models in various real-world scenarios.

HO3D_v2 (Hampali et al. 2020) is another benchmark for *hand-object* interactions, including 66K training images and 11K testing images with 10 subjects and 10 objects. It provides detailed 3D pose annotations for hands, with test results evaluated through an online submission system specific to this dataset.

Evaluation Metrics. To assess the performance of HPE, we utilize a widely adopted metric: the mean keypoint error (MKE). The mean keypoint error measures the average Euclidean distance between the predicted and ground-truth keypoint positions, indicating model precision.

Implementation Details

We implement the model using PyTorch on a NVIDIA 4090 GPU and train it with AdamW optimizer (Loshchilov and Hutter 2017) ($\beta_{1} = 0.5$, $\beta_{2} = 0.999$) and a learning rate of 0.001. The input depth and RGB images are resized to 128×128 pixels, and the input point clouds are downsampled to 1,024 points from the input depth. We use a batch size of 32 for training. For data augmentation, we apply random rotations ranging from -180 to 180 degrees, 3D scaling within a range of 0.9 to 1.1, and 3D translations between -10 and 10 mm. In addition, we set the number of super points to 256. We train the model on the HO3D dataset for 24 epochs, applying learning rate decay at the 19th epoch. For the DexYCB dataset, we train for 20 epochs with decay at the 15th epoch. For the NYU dataset, training extended over 40 epochs with learning rate decay proceeding at the 30th epoch. Note that we bypass the RGB data for the training and evaluation on NYU due to its low-quality RGB data which has massive unpredictable holes and artifacts.

Comparison with State-of-the-Art Methods

NYU Dataset. We conduct a comparison of our method with various state-of-the-art techniques using the single hand NYU dataset (Oberweger and Lepetit 2017; Wan et al. 2018; Chen et al. 2020; Cheng, Kim, and Ko 2024; Du et al. 2019; Fang et al. 2020; Ren et al. 2021b,a; Ge et al. 2017; Chen et al. 2018; Ge et al. 2018; Ge, Ren, and Yuan 2018; Moon, Yong Chang, and Mu Lee 2018; Cheng, Park, and Ko 2021; Cheng and Ko 2023; Ren et al. 2023; Liu et al. 2024b; Cheng, Kim, and Ko 2024).

The results of our evaluations measured by the mean joint error on the NYU dataset are summarized in Table 1. Our proposed HandMCM achieves a mean joint error of 7.06 mm, indicating its performance at the state-of-the-art level. Furthermore, Figure 3 showcases qualitative results of HandMCM on the NYU dataset, illustrating its effectiveness in accurately estimating hand keypoints, even in scenarios involving occlusions.

DexYCB Dataset. We assess our HandMCM on the hand-object dataset DexYCB and compare its performance against several existing state-of-the-art methods (Xiong et al. 2019; Spurr et al. 2020; Lin, Wang, and Liu 2021; Tse et al. 2022; Park et al. 2022; Ren et al. 2023; Cheng, Kim, and Ko 2024; Liu et al. 2024b; Cheng et al. 2024), adhering to the official dataset split protocols. As indicated in Table 2, our method outperforms the latest state-of-the-art (SOTA) techniques across all four evaluation protocols. The results

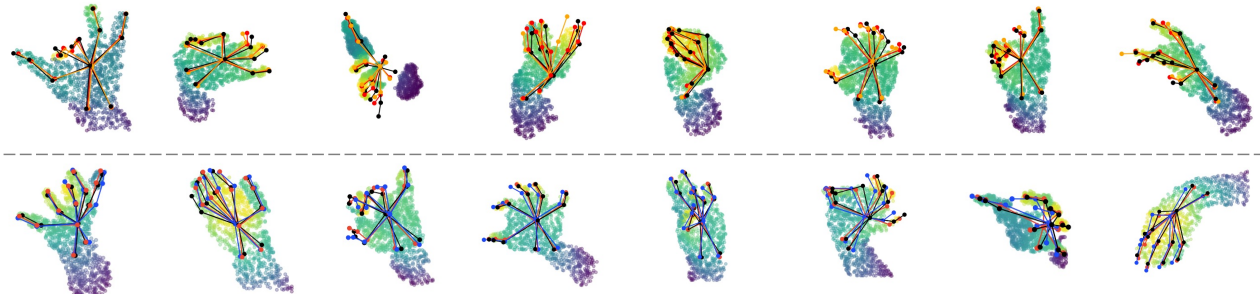


Figure 3: Qualitative results of the HandMCM model on the NYU dataset. 3D points are colored in the figure to distinctly illustrate occlusions. The ground truth joint coordinates are displayed in black, the results of the previous state-of-the-art model, HandDAGT (Cheng, Kim, and Ko 2024), are shown in orange (top), the results of Hamba (Dong et al. 2024) are shown in blue (bottom), and the estimated joint coordinates produced by our model are depicted in red. Best view in color.

Method	MKE (mm)	Input
DeepPrior++ (Oberweger et al. 2017)	12.24	Depth
Pose-Ren (Chen et al. 2020)	11.81	Depth
SHPR-Net (Chen et al. 2018)	10.78	Point
HandPointNet (Ge et al. 2018)	10.54	Point
DenseReg (Wan et al. 2018)	10.20	Depth
CrossInfoNet (Du et al. 2019)	10.08	Depth
Point-to-Point (Ge, Ren, and Yuan 2018)	9.10	Point
HandFolding (Cheng et al. 2021)	8.58	Point
V2V (Moon et al. 2018)	8.42	Voxel
Hamba [†] (Dong et al. 2024)	8.38	Depth
JGR-P2O (Fang et al. 2020)	8.29	Depth
HandGCNFormer (Wang et al. 2023)	7.43	Point
SSRN (Ren et al. 2021b)	7.37	Depth
HandDiff (Cheng et al. 2024)	7.38	Depth
PHG (Ren et al. 2021a)	7.39	Depth
HandR2N2 (Cheng and Ko 2023)	7.27	Point
IPNet (Ren et al. 2023)	7.17	Depth
HandDAGT (Cheng, Kim, and Ko 2024)	7.12	Depth
HandMCM (Ours)	7.06	Depth

Table 1: Comparison of the proposed method with previous state-of-the-art methods on the NYU datasets. Note, Voxels & Points are transformed and sampled from Depth.

highlight the exceptional capability of our HandMCM in managing occlusions, as it significantly exceeds other models by 11.5% on the DexYCB dataset. Furthermore, the qualitative results presented in Figure 4 demonstrate that HandMCM effectively estimates poses during hand-object interactions, even in the presence of various occlusions.

HO3D Dataset. Additionally, we benchmark the method on the HO3D dataset against leading methods (Li et al. 2021; Yang et al. 2022; Park et al. 2022; Malik et al. 2021; Ren et al. 2023; Cheng, Kim, and Ko 2024; Liu et al. 2024b), with the state-of-the-art results shown in Table 3. Our proposed HandMCM achieves the mean keypoint error of 1.71 cm. It also surpasses Hamba (Dong et al. 2024). These findings demonstrate that our model is highly robust and accurate in scenarios with dynamic occlusions.

Runtime cost. Our model achieves an average runtime of 15.4 ms per frame on the input RGBD modality with 0.3ms of point subsampling, when tested on an NVIDIA RTX 4090

Method	Mean Keypoint Error (mm)					Input
	S0	S1	S2	S3	AVG	
A2J	23.93	25.57	27.65	24.92	25.52	Depth
Spurr et al.	17.34	22.26	25.49	18.44	18.44	RGB
METRO	15.24	-	-	-	-	RGB
Tse et al.	16.05	21.22	27.01	17.93	20.55	RGB
HandOccNet	14.04	-	-	-	-	RGB
IPNet	8.03	9.01	8.60	7.80	8.36	Depth
HandDiff	7.66	8.73	8.40	7.53	8.07	Depth
HandDAGT	7.72	8.68	8.22	7.52	8.03	Depth
K-Fusion	6.94	8.64	7.56	7.02	7.54	RGBD
HandMCM (Ours)	6.27	7.86	6.20	6.30	6.67	RGBD

Table 2: Comparison of the proposed method with previous state-of-the-art methods on the DexYCB datasets.

Method	MKE (cm)	Input
Hamba (Dong et al. 2024)	54.56	RGB
Hybrik (Li et al. 2021)	2.89	RGB
ArtiBoost (Yang et al. 2022)	2.53	RGB
HandOccNet (Park et al. 2022)	2.49	RGB
HandVoxNet++ (Malik et al. 2021)	2.46	Voxel
IPNet (Ren et al. 2023)	1.81	Depth
HandDAGT (Cheng, Kim, and Ko 2024)	1.81	Depth
K-Fusion (Liu et al. 2024b)	1.79	RGBD
HandMCM (Ours)	1.71	RGBD

Table 3: Comparison of the proposed method with previous state-of-the-art methods on the HO3D datasets.

GPU. It runs faster than the recent SOTA K-Fusion, which requires 17.8 ms on the same GPU.

Ablation Study

To assess the effectiveness of each component in our proposed method, we perform additional ablation studies on different configurations of HandMCM. These experiments are mainly conducted using the NYU dataset.

Analysis of Different Proposed Components. We conduct an ablation study to evaluate the impact of each proposed component in the key correspondence Mamba block, specifically focusing on the SSM type, using local feature injection, and local feature filtering. The results are summarized

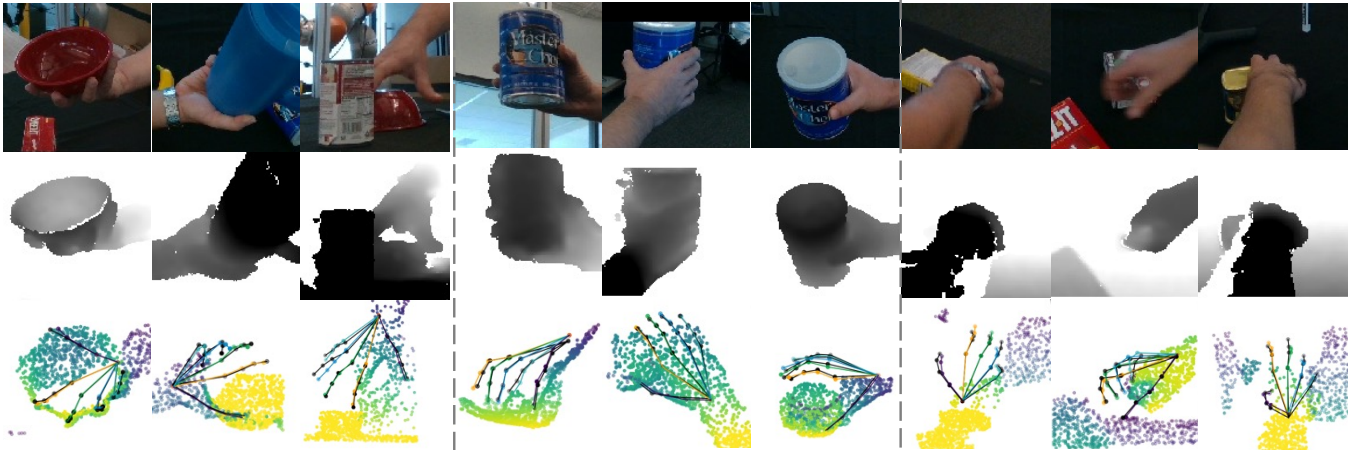


Figure 4: Qualitative results of HandMCM on the DexYCB datasets illustrating different object occlusions (left), view-angle-dependent occlusions (middle) and self-occlusions (right). Input 3D points and output keypoint coordinates are depicted within the same frame (third rows) and rotated at top-down view angles in order to clearly *view occlusions from above* as shown in the figure. Ground truth is shown in black and the estimated joint coordinates of our model are shown in colors. Best view in color.

ID	Configurations			Mean Keypoint Error (mm)
	SSM Type	Local Inject	Local Filter	
1)	×	×	×	8.47
2)	Transformer	×	×	8.35
3)	Standard	×	×	8.51
4)	Graph-guided	×	×	8.38
5)	Correspondence	×	×	8.15
6)	Correspondence	×	✓	7.24
7)	Correspondence	✓	×	7.22
8)	Standard	✓	✓	7.27
9)	Correspondence	✓	✓	7.06

Table 4: Ablations of different configurations of the transformer. All ablation models are trained and tested on NYU.

# of Stacks	Mean Keypoint Error (mm)
1	8.83
2	7.12
3	7.06
4	7.10

Table 5: Ablations of different depth of Mamba.

in Table 4. The specific configurations for the ablation studies are as follows: 1) baseline model without the correspondence Mamba block, 2) Transformer instead of Mamba, 3) standard SSM (S), 4) graph-guided SSM proposed in Hamba (Dong et al. 2024), 5) novel SSM with the proposed correspondence modeling, 6) correspondence SSM added with filtered local features, 7) correspondence SSM with only local token injection, 8) local token injection and filtering both activated on standard SSM, and 9) all proposed components. The baseline model resulted in a mean joint error of 8.47 mm. The ablation using the standard SSM 3) demonstrates an increased error of 8.51 mm, whereas the proposed novel SSM 5) integrated with correspondence modeling plausibly reduces the error by 0.3 mm. This reduction highlights the

critical role and effectiveness of the novel SSM. Meanwhile, the proposed correspondence block outperforms the graph-guided SSM 4) proposed in Hamba (Dong et al. 2024), revealing the superior correspondence modeling ability of SSM over explicitly applying graphs. Furthermore, incorporating local tokens 6) & 7) drastically reduced the error to 7.2 mm. Combining local token injection and filtering 9) with Mamba further reduces the error to 7.06 mm, indicating the essential contribution of effective local geometry information.

Analysis of Mamba Depth. Theoretically, increasing the depth of a network can improve performance by allowing the model to capture more complex features. To investigate this, we progressively expanded the depth of the Mamba block with the aim of identifying its optimal configuration. As shown in Table 5, we systematically adjusted the stack depth to assess its effect on model performance. The results indicate that increasing the depth of the Mamba stack generally enhances performance.

However, the performance gains become negative beyond three stacks at specifically a 3-stack Mamba. We thus conclude that a 3-stack configuration is optimal for our model as further depth increases in the transformer stack yield a negative effect.

Conclusion

This paper introduces HandMCM, an innovative Mamba-based architecture designed for accurate 3D hand pose reconstruction. Experimental results demonstrate that our proposed approach substantially outperforms previous state-of-the-art methods across three challenging datasets. Furthermore, extensive experiments validate the effectiveness of each component proposed in our framework. Nevertheless, despite its effectiveness, our HandMCM still has a limitation in handling scenarios involving bi-manual interacting hands. Future research can focus on extending the model to address this limitation and further enhance its capabilities.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008).

References

- Ba, J. L. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; et al. 2021. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9044–9053.
- Chen, X.; Wang, G.; Guo, H.; and Zhang, C. 2020. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, 395: 138–149.
- Chen, X.; Wang, G.; Zhang, C.; Kim, T.-K.; and Ji, X. 2018. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access*, 6: 43425–43439.
- Cheng, W.; Kim, E.; and Ko, J. H. 2024. HandDAGT: A Denoising Adaptive Graph Transformer for 3D Hand Pose Estimation. In *ECCV*.
- Cheng, W.; and Ko, J. H. 2023. HandR2N2: Iterative 3D Hand Pose Estimation Using a Residual Recurrent Neural Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20904–20913.
- Cheng, W.; and Lee, S. 2019. Point Auto-Encoder and Its Application to 2D-3D Transformation. In *International Symposium on Visual Computing*, 66–78. Springer.
- Cheng, W.; Park, J. H.; and Ko, J. H. 2021. HandFoldingNet: A 3D Hand Pose Estimation Network Using Multiscale-Feature Guided Folding of a 2D Hand Skeleton. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11260–11269.
- Cheng, W.; Tang, H.; Van Gool, L.; and Ko, J. H. 2024. HandDiff: 3D Hand Pose Estimation with Diffusion on Image-Point Cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2274–2284.
- Dong, H.; Chharia, A.; Gou, W.; Carrasco, F. V.; and De la Torre, F. 2024. Hamba: Single-view 3D Hand Reconstruction with Graph-guided Bi-Scanning Mamba. *arXiv preprint arXiv:2407.09646*.
- Du, K.; Lin, X.; Sun, Y.; and Ma, X. 2019. Crossinfonet: Multi-task information sharing based hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9896–9905.
- Fang, L.; Liu, X.; Liu, L.; Xu, H.; and Kang, W. 2020. JGR-P2O: Joint Graph Reasoning based Pixel-to-Offset Prediction Network for 3D Hand Pose Estimation from a Single Depth Image. In *European Conference on Computer Vision*, 120–137. Springer.
- Ge, L.; Cai, Y.; Weng, J.; and Yuan, J. 2018. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8417–8426.
- Ge, L.; Liang, H.; Yuan, J.; and Thalmann, D. 2016. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3593–3601.
- Ge, L.; Liang, H.; Yuan, J.; and Thalmann, D. 2017. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1991–2000.
- Ge, L.; Ren, Z.; and Yuan, J. 2018. Point-to-point regression pointnet for 3d hand pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 475–491.
- Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; and Ré, C. 2020. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487.
- Gu, A.; Goel, K.; Gupta, A.; and Ré, C. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35: 35971–35983.
- Gu, A.; Johnson, I.; Goel, K.; Saab, K.; Dao, T.; Rudra, A.; and Ré, C. 2021. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34: 572–585.
- Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3196–3206.
- He, X.; Du, X.; Wang, X.; Tian, F.; Tang, J.; and Chua, T.-S. 2018. Outer product-based neural collaborative filtering. *arXiv preprint arXiv:1808.03912*.
- Kalman, R. E. 1960. A new approach to linear filtering and prediction problems.
- Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3383–3393.
- Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2025. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 237–255. Springer.
- Li, S.; and Lee, D. 2019. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11927–11936.
- Liang, D.; Zhou, X.; Xu, W.; Zhu, X.; Zou, Z.; Ye, X.; Tan, X.; and Bai, X. 2024. PointMamba: A Simple State Space Model for Point Cloud Analysis. In *Advances in Neural Information Processing Systems*.

- Lin, K.; Wang, L.; and Liu, Z. 2021. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1954–1963.
- Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; and Pande, V. 2017. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10): 1103–1113.
- Liu, J.; Yu, R.; Wang, Y.; Zheng, Y.; Deng, T.; Ye, W.; and Wang, H. 2024a. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*.
- Liu, X.; Qi, C. R.; and Guibas, L. J. 2019. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 529–537.
- Liu, X.; Ren, P.; Gao, Y.; Wang, J.; Sun, H.; Qi, Q.; Zhuang, Z.; and Liao, J. 2024b. Keypoint Fusion for RGB-D Based 3D Hand Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3756–3764.
- Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024c. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Malik, J.; Shimada, S.; Elhayek, A.; Ali, S. A.; Theobalt, C.; Golyanik, V.; and Stricker, D. 2021. Handvoxnet++: 3d hand shape and pose estimation using voxel-based neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 8962–8974.
- Moon, G.; Yong Chang, J.; and Mu Lee, K. 2018. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5079–5088.
- Oberweger, M.; and Lepetit, V. 2017. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *Proceedings of the IEEE international conference on computer vision workshops*, 585–594.
- Park, J.; Oh, Y.; Moon, G.; Choi, H.; and Lee, K. M. 2022. Handocnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1496–1505.
- Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.
- Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, 5099–5108.
- Ren, P.; Chen, Y.; Hao, J.; Sun, H.; Qi, Q.; Wang, J.; and Liao, J. 2023. Two heads are better than one: image-point cloud network for depth-based 3D hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2163–2171.
- Ren, P.; Sun, H.; Hao, J.; Qi, Q.; Wang, J.; and Liao, J. 2021a. Pose-Guided Hierarchical Graph Reasoning for 3-D Hand Pose Estimation From a Single Depth Image. *IEEE Transactions on Cybernetics*.
- Ren, P.; Sun, H.; Huang, W.; Hao, J.; Cheng, D.; Qi, Q.; Wang, J.; and Liao, J. 2021b. Spatial-aware stacked regression network for real-time 3D hand pose estimation. *Neuro-computing*, 437: 42–57.
- Ren, P.; Sun, H.; Qi, Q.; Wang, J.; and Huang, W. 2019. SRN: Stacked Regression Network for Real-time 3D Hand Pose Estimation. In *BMVC*, 112.
- Spurr, A.; Iqbal, U.; Molchanov, P.; Hilliges, O.; and Kautz, J. 2020. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, 211–228. Springer.
- Tompson, J.; Stein, M.; Lecun, Y.; and Perlin, K. 2014. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5): 1–10.
- Tse, T. H. E.; Kim, K. I.; Leonardis, A.; and Chang, H. J. 2022. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1664–1674.
- Wan, C.; Probst, T.; Van Gool, L.; and Yao, A. 2018. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5147–5156.
- Wang, C.; Tsepa, O.; Ma, J.; and Wang, B. 2024. Graphmamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*.
- Wang, J.; Yan, J. N.; Gu, A.; and Rush, A. M. 2022. Pretraining without attention. *arXiv preprint arXiv:2212.10544*.
- Wang, Y.; Chen, L.; Li, J.; and Zhang, X. 2023. HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5675–5684.
- Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, J. T.; and Yuan, J. 2019. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 793–802.
- Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2750–2760.
- Zhang, Z.; Liu, A.; Reid, I.; Hartley, R.; Zhuang, B.; and Tang, H. 2024. Motion Mamba: Efficient and Long Sequence Motion Generation. In *European Conference on Computer Vision*, 265–282. Springer.
- Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, 62429–62442.