2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3315-5511-5/26/$31.00 ©2026 IEEE | DOI: 10.1109/WACV61042.2026.00519

DF-Mamba: Deformable State Space Modeling for 3D Hand Pose Estimation in Interactions

Yifan Zhou[1*], Takehiko Ohkawa[1,3*], Guwenxiao Zhou[1], Kanoko Goto[1], Takumi Hirose[1], Yusuke Sekikawa[2], Nakamasa Inoue[1]
[1]Institute of Science Tokyo [2]Denso IT Laboratory [3]The University of Tokyo

Abstract

Modeling daily hand interactions often struggles with severe occlusions, such as when two hands overlap, which highlights the need for robust feature learning in 3D hand pose estimation (HPE). To handle such occluded hand images, it is vital to effectively learn the relationship between local image features (e.g., for occluded joints) and global context (e.g., cues from inter-joints, inter-hands, or the scene). However, most current 3D HPE methods still rely on ResNet for feature extraction, and such CNN's inductive bias may not be optimal for 3D HPE due to its limited capability to model the global context. To address this limitation, we propose an effective and efficient framework for visual feature extraction in 3D HPE using recent state space modeling (i.e., Mamba), dubbed ***Deformable Mamba (DF-Mamba)****. DF-Mamba is designed to capture global context cues beyond standard convolution through Mamba's selective state modeling and the proposed deformable state scanning. Specifically, for local features after convolution, our deformable scanning aggregates these features within an image while selectively preserving useful cues that represent the global context. This approach significantly improves the accuracy of structured 3D HPE, with comparable inference speed to ResNet-50. Our experiments involve extensive evaluations on five divergent datasets including single-hand and two-hand scenarios, hand-only and hand-object interactions, as well as RGB and depth-based estimation. DF-Mamba outperforms the latest image backbones, including VMamba and Spatial-Mamba, on all datasets and achieves state-of-the-art performance.*

1. Introduction

Daily human activities often involve complex hand interactions, such as two-hand interactions [46, 50] and object grasping [16, 49, 51], necessitating effective and efficient inference models that estimate 3D hand poses to handle such challenging scenarios. These intricate interactions with severe occlusions make it cumbersome to perform 3D hand pose estimation (HPE) from visual data, including single RGB images [32], depth images [60], egocentric views [16, 38], etc. Meanwhile, developing fast inference models has become crucial to support real-time applications, especially in AR/VR devices [6, 36]. Given these challenges, limited attention has been paid to the inductive biases introduced by backbone architectures and their synergy with 3D HPE. This highlights the need for designing backbones that are both effective in capturing complex interactions and efficient for real-time inference.

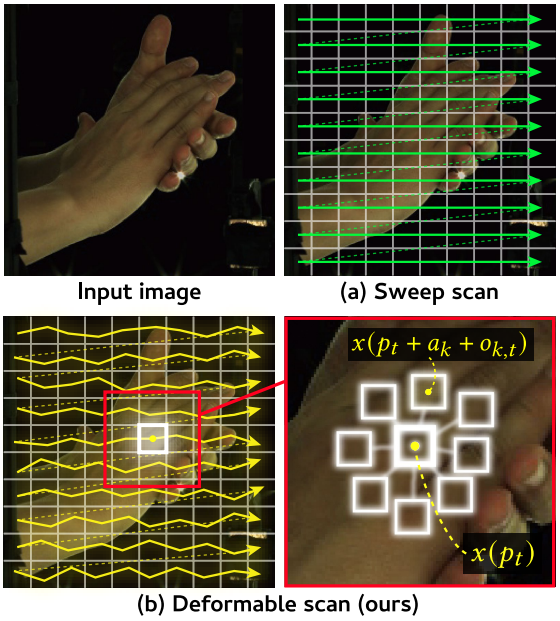

Figure 1. **Deformable scan for DF-Mamba.** (a) Conventional sweep scan uses a fixed grid pattern in the state space equations. (b) Our deformable scan adaptively adjusts the scanning pattern with multiple anchors $\boldsymbol{a_k}$ by predicting offset vectors $\boldsymbol{o_{k,t}}$ dependent on input visual features.

To enable robust feature learning for complex hand interactions, it is essential to learn the relationship between local

*Equal contribution.

2642-9381/26/$31.00 ©2026 IEEE
DOI 10.1109/WACV61042.2026.00519

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

image features (*e.g.*, for occluded joints) and global context (*e.g.*, cues from inter-joints, inter-hands, hand-object, or the scene). Convolutional neural networks (CNNs) are widely used as backbone architectures in 3D HPE, with ResNet-50 [26] being particularly popular [31, 32, 37, 38, 46, 49, 55, 73]. These CNN backbones rely on convolution operations with a local receptive field, providing a favorable balance between accuracy and inference speed. However, these local convolutions have a limited ability to capture global context. Another line of recent studies leverages the vision transformer (ViT) [12] as a feature extractor for hand-mesh reconstruction [11, 52, 80], but its higher computational complexity often becomes a bottleneck in practical scenarios. This suggests significant room for improvement in backbone architectures to achieve better feature learning of hand poses efficiently.

As an emerging foundational architecture, Mamba [20] based on state space modeling (SSM) has garnered considerable attention, which was originally proposed for natural language processing. The Mamba model excels at selectively focusing on input tokens (*i.e.*, emphasizing particular signals), thus efficiently capturing global context from long token sequences. Several recent studies have extended it to image backbones. For example, VisionMamba [81] introduced the Vim block, which employs a 2D bidirectional scan for spatially-aware sequence modeling. VMamba [40] further proposed the VSS block based on four different scanning paths. However, these scanning mechanisms employ a fixed grid as illustrated in Figure 1(a), which limits their ability to capture intricate hand pose variations when applied to 3D HPE.

Given this limitation, we introduce an effective and efficient backbone, **Deformable Mamba (DF-Mamba)**, with deformable state space modeling (DSSM) that encourages robust visual feature extraction in 3D HPE. The core idea of DF-Mamba is to perform feature extraction by dynamically modeling local features and global context with flexible state spaces. Specifically, our DSSM blocks aggregate the local features according to a deformable path and selectively store useful cues to represent the global context. The scanning path is adjusted with deformable point sampling with local anchors and learnable offsets dependent on the given input features as illustrated in Figure 1(b).

The overall architecture of DF-Mamba is a tribrid design composed of three blocks: convolution blocks, DSSM blocks, and gated convolution blocks, with a model size comparable to ResNet-50. This approach efficiently leverages the complementary strengths of each block type: extracting local features via convolution blocks at lower layers, adaptively enhancing features with DSSM blocks at higher layers after downsampling, and further refining visual representations using gated convolution blocks.

In our experiments, we integrate DF-Mamba into two representative 3D HPE frameworks proposed by Jiang *et al*. [32] and Zhou *et al*. [79], replacing their backbones with our method. Our evaluations are performed on five datasets: InterHand2.6M [46], RHP [84], NYU [60], DexYCB [4] and AssemblyHands [49], covering diverse scenarios, including single-hand and two-hand pose estimation, hand-only and hand-object interactions, and RGB and depth modalities. The results demonstrate that DF-Mamba outperforms the latest Mamba-based backbones, *e.g.*, VMamba [40] and SpatialMamba [68], achieving state-of-the-art performance while maintaining inference speed. Our contributions are summarized as follows.

1) We propose **DSSM**, a novel approach to modeling dynamic systems with flexible state spaces to represent the global context. DSSM incorporates the deformable scan mechanism into the state-space equations.
2) We introduce **DF-Mamba**, a novel Mamba-based backbone for 3D HPE. It adopts a tribrid design composed of convolution, DSSM, and gated convolution blocks.
3) We demonstrate the effectiveness and efficiency of DF-Mamba on five datasets. We show that DF-Mamba outperforms the latest image backbones in 3D HPE.

2. Related work

2.1. 3D Hand Pose Estimation

This task is formulated as predicting the 3D coordinates of hand joints, typically from a single image [13, 48]. This task has been studied in various interaction scenarios, such as single-hand [5, 19, 84, 85], self-contact [50, 56] and hand-object interactions [15, 16], as well as RGB-based [19, 32, 78] and depth-based [60] estimation. Most existing approaches rely on deep neural networks to predict hand poses from visual input, which necessitates the design of effective and efficient architectures.

Encoder-decoder designs are predominant for precise estimation, typically utilizing an image backbone such as ResNet [26] as the encoder and a sophisticated decoder to accurately predict joint positions. For example, Zhou *et al*. [79] propose a simple regression decoder head that estimates 2D heatmaps and depth from features extracted by ResNet-50. Recently, transformer-based decoding leveraging attention mechanisms has also gained popularity in 3D HPE [28, 32, 59, 66, 67, 73]. For instance, Jiang *et al*. [32] integrate the anchor offset-weight estimation into a transformer decoder [3, 83]. Despite these successes, ResNet [26] remains one of the most widely used backbones in 3D HPE [22, 32, 37, 39, 47, 49, 55, 73, 79]. While several recent studies have employed pre-trained vision transformer backbones [11, 52, 80] for hand-mesh reconstruction, the high computational complexity of attention modules limits their practical applicability. Furthermore, when the target data distribution significantly differs from the pre-training

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

domain (*e.g.*, two-hand interactions [46], depth images [60], or egocentric views [1, 49]), it is still challenging to perform training with ViT-based backbones. To address this, we instead propose a versatile backbone for 3D HPE based on Mamba modeling that replaces the popular ResNet-50 and functions across diverse scenarios (*e.g.*, [46, 47, 60]).

2.2. Image Backbones

CNN-based backbones. CNNs leverage simple convolution operations with local receptive fields to efficiently extract visual features. ResNet [26] remains one of the most widely adopted backbones for various vision tasks. To further enhance the adaptability and flexibility of convolutional operations, advanced techniques such as deformable convolutions [9, 82] and depthwise convolutions [7, 27] have been introduced. Recently, ConvNeXt [42] improved upon ResNet architectures by integrating design elements from vision transformers, achieving competitive performance across various tasks.

Transformer-based backbones. Since the introduction of the vision transformer (ViT) [12], various transformer-based backbones have been proposed to further enhance image backbones [8, 41, 54, 61, 65, 70]. Among them, Swin Transformer [41], which introduces a hierarchical architecture with shifted window-based attention, has shown its generality for various vision tasks due to its strong capability of extracting multi-scale pyramid features.

Mamba-based backbones. Mamba [20] is the first successful architecture based on state space modeling (SSM) to achieve comparable performance to transformers on natural language processing tasks. It particularly introduces selective SSM to adaptively remember or forget input tokens by computing matrices B and C of the state space equations [2, 33] depending on the inputs.

VisionMamba [81] is the first Mamba-based image backbone, replacing the encoder of ViT with a Mamba-based encoder and extending the original 1D SSM to a 2D bidirectional SSM. Subsequent studies have extended the scanning mechanisms and architectures [25, 29, 40, 53, 68, 71, 74]. For example, VMamba [40] enhances the 2D scan by introducing a cross-scan strategy, traversing the image along four directions. Spatial-Mamba [68] is one of the latest Mamba-based backbones. It introduces a structure-aware state fusion equation, applying a 2D convolution to the input of matrix C in the state space equations. As the fixed-grid 2D scanning employed by the previous Mamba models can limit model capability in structured prediction tasks like 3D HPE, more adaptive scanning methods are needed. Hamba [11] introduces a Mamba-based decoder for hand-mesh reconstruction built upon a ViT backbone. Other works adopt the Mamba-based modeling for human pose estimation [30, 34, 43, 77].

In contrast, we investigate a Mamba modeling for the backbone architecture of 3D HPE that can be applied to diverse scenarios, including two-hand poses [46] and egocentric views [49].

3. Deformable State-Space Modeling

We introduce Deformable State-Space Modeling (DSSM), a novel approach to modeling dynamic systems with flexible state spaces. The core idea behind DSSM is to integrate local visual cues with a deformable scan mechanism of Mamba to capture global context. Below, we begin with a preliminary review of state-space equations, and then introduce our DSSM formulation.

3.1. Preliminary

State-Space Modeling (SSM). Let $x(t) \in \mathbb{R}$ be a one dimensional continuous input, where $t \in \mathbb{R}_+$ is the time index. SSM represents a dynamic linear system [2, 33], which produces an output $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^{N\times 1}$ as follows:

$$\begin{cases} h'(t) = Ah(t) + Bx(t) \\ y(t) = Ch(t) + Dx(t) \end{cases}, \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$ is the transition matrix, $B \in \mathbb{R}^{N\times 1}$, $C \in \mathbb{R}^{1\times N}$ and $D \in \mathbb{R}^{1\times 1}$ are parameters to control the dynamic system.

Discretization. To integrate SSM with deep neural networks, the dynamic system of Eq. (1) is discretized with a timescale Δ [18, 21]. Specifically, the discrete state-space equations are given by:

$$\begin{cases} h(t) = \bar{A}h(t-1) + \bar{B}x(t) \\ y(t) = Ch(t) + Dx(t) \end{cases}, \tag{2}$$

where the zero-order hold rule is applied as

$$\bar{A} = \exp(\Delta A),\ \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I)(\Delta B). \tag{3}$$

Figure 2(a) visualizes the computational flow of SSM. As shown, the output y is obtained by repeatedly applying $\bar{A}$ to x based on the recurrence relation in Eq. (2). In practice, this model is implemented as an SSM layer that computes $Y = X * \bar{K}$, where $X \in \mathbb{R}^{d\times s}$ is a sequence of input features, $Y \in \mathbb{R}^{d\times s}$ is the output sequence, $\bar{K} = [C\bar{B}, C\bar{A}\bar{B}, \cdots, C\bar{A}^{s-1}\bar{B}]$ is the global convolution kernel, d is the dimension, and s is the sequence length. Here, the size of each matrix is also expanded to the dimension of d, with $B \in \mathbb{R}^{N\times d}, C \in \mathbb{R}^{d\times N}$, and $D \in \mathbb{R}^{d\times d}$. The Mamba model [20] computes B, C and Δ dependent on the input X, each through a learnable linear layer.

Sweep scan. To apply SSM to image data, visual features need to be sequentialized. This process, known as the scan scheme, replaces the 1D input $x(t)$ in Eq. (2) with a 2D

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

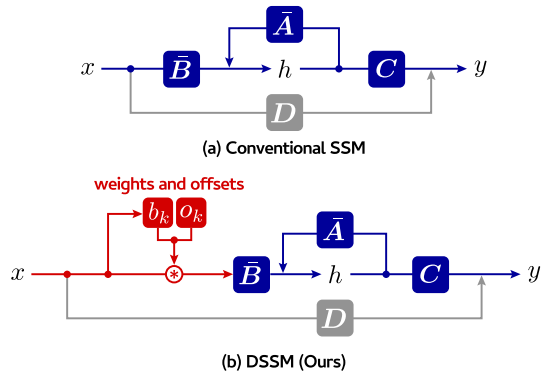

Figure 2. Computational flow of SSM and DSSM. (a) As shown in Sec. 3.1, conventional SSM utilizes four matrices $\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}, \boldsymbol{C}, \boldsymbol{D}$, to compute the output $\boldsymbol{y}$ from an input $\boldsymbol{x}$ through intermediate representation $\boldsymbol{h}$. (b) Our DSSM incorporates weights $\boldsymbol{b_k}$ and offsets $\boldsymbol{o_k}$ for deformable scan into SSM.

visual feature map $x(p_t)$, where $p_t \in \mathbb{R}^2$ indicates a 2D position. The sweep scan scheme shown in Figure 1(a) is a representative approach. Specifically, when the size of the feature map is given by $H \times W$, the sweep scan determines the 2D position as

$$p_t = \begin{pmatrix} \mathrm{mod}((t-1), W) + 1 \\ \lfloor t - 1/W \rfloor + 1 \end{pmatrix}, \tag{4}$$

with $t = 1, 2, \cdots, HW$. This scans over a fixed 2D grid.

However, in 3D HPE, the fixed grid pattern introduces inefficiency, limiting the ability to capture intricate hand variations. For example, background pixels do not provide any cues for local joint positions, while the center region of the input image exhibits a higher density of joints. These observations necessitate a flexible solution to model spatial locality bias and global context adaptively to the input.

3.2. Deformable Formulation

To overcome the SSM's limitation, we introduce data-driven, adaptive capabilities to SSM. We propose to incorporate a deformable scan, which adaptively adjusts the sampling points by introducing learnable offsets into the input sampling. This allows us to dynamically balance its receptive fields, selectively aggregating relevant local cues to accommodate intricate interactions. Specifically, we define DSSM for 1D inputs by the following dynamic system:

$$\begin{cases} h(t) = \bar{A}h(t-1) + \bar{B}x(t + \delta t) \\ y(t) = Ch(t) + Dx(t) \end{cases}, \tag{5}$$

where $\delta t \in \mathbb{R}$ is a small offset predicted from x.

DSSM improves the flexibility of SSM because the offset allows the input to be adaptively shifted. However, applying DSSM to image data is not straightforward, as the offset requires two degrees of freedom, which expands the area to be explored by the offset compared to 1D data. To address this, we effectively handle such a large exploration area by introducing spatially allocated K anchors, which sample multiple points from the input image. Specifically, we define DSSM for 2D inputs as follows:

$$\begin{cases} h(t) = \bar{A}h(t-1) + \bar{B}\sum_{k=1}^{K} b_k x(p_t + a_k + o_{k,t}) \\ y(t) = Ch(t) + Dx(p_t) \end{cases}, \tag{6}$$

where $a_k \in \mathbb{R}^2$ is a fixed anchor, $o_{k,t} \in \mathbb{R}^2$ is an offset vector, and $b_k \in \mathbb{R}$ is a weight coefficient. Both $o_{k,t}$ and b_k are predicted from x, each through a learnable linear layer. Its computational flow is shown in Figure 2(b).

Specifically, the K anchors are symmetrically defined as $a_k \in \{(i,j) : i,j \in \{-1,0,+1\}\}$, inspired by deformable convolution [9]. This provides a complete initial distribution for local relative offsets, requiring 3^D anchors for D spatial dimensions (*i.e.*, $D = 2$ and $K = 9$ for this case). From these starting points, the anchor spatial distribution becomes learnable via the offset mechanism. This balances initial coverage of the anchors with data-driven flexibility by 2D scanning, a core DSSM strength.

Theoretical support. Our DSSM is a direct generalization of the standard Mamba (SSM) block [20] and Spatial-Mamba block [68]. For instance, when the number of anchors $K = 1$, with the offset $o_{k,t} = 0$ and the scaling factor $b_k = 1$, our DSSM reduces to a normal Mamba block. While Spatial-Mamba adopts dilated operation in SSM, deformable convolutions are designed to incorporate the dilated operation as a special case [9]. Thus, DSSM acts as a general form of Spatial-Mamba block, showcasing its further enhanced flexibility over these Mamba variants.

Roles in 3D HPE. When applying DSSM to 3D HPE, the input x corresponds to local features convoluted from a single image, and their neighboring cues are probed with the guidance of K adaptive anchors. Then, these evolving local cues are aggregated into the latent state h, which can represent the global context for the given image, *e.g.*, cues from inter-joints, inter-hands, hand-object, or the scene. As such, this adaptive local-global feature learning based on DSSM helps model the dynamism of hands in 3D HPE.

4. DF-Mamba Backbone

With the proposed DSSM in Sec. 3, we finally construct the overall backbone architecture, **DF-Mamba**. DF-Mamba is a tribrid model that consists of three types of blocks: (i) convolution blocks, (ii) DSSM blocks, and (iii) gated convolution blocks. Since (i) and (iii) are existing modules, Sec. 4.1 first describes the design of (ii) DSSM blocks; see Figure 3(d). The following Sec. 4.2 further details how to combine all three blocks in a unified architecture.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

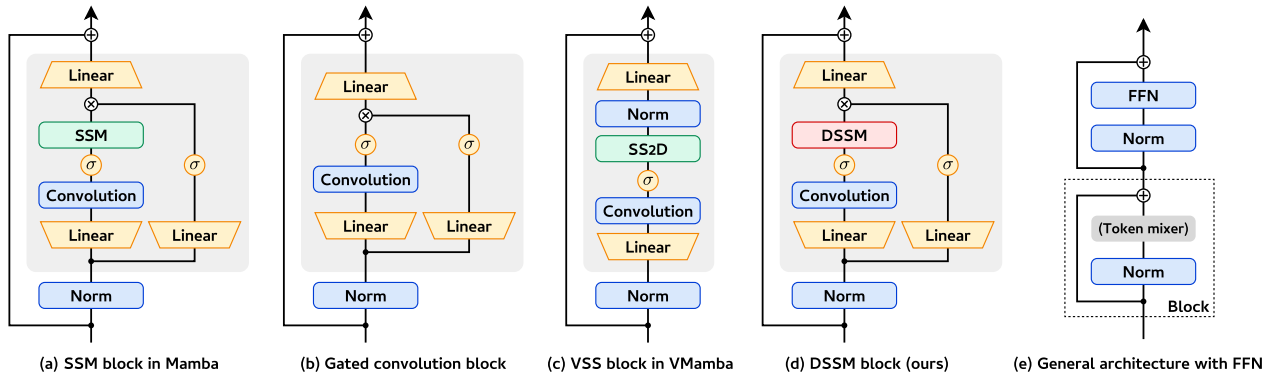

Figure 3. Block architectures. (a) Vanilla Mamba block using SSM [20] (b) Gated convolution block that omits the SSM layer from the Mamba block (*e.g.*, MambaOut [74]). (c) VSS block used in VMamba [40]. (d) Our DF-Mamba block that replaces the SSM layer with the DSSM layer in the Mamba block. (e) General architecture inspired by the transformer architecture [63]. Each subblock highlighted with a gray background acts as a token mixer.

4.1. DSSM Block

We integrate DSSM with deep neural networks by defining its blocks that can be stacked as layers. Figure 3 compares the architectures of (a) vanilla SSM block in Mamba [20], (b) gated convolution block [10] (*e.g.*, MambaOut [74]), (c) VSS block in VMamba [40] and (d) our DSSM block. All of them take a visual feature map X as input and apply layer normalization and a skip connection. The general design inspired by the original transformer [63] is shown in (e). The token mixture subblocks are highlighted with a gray box, which differ from each other in (a)–(d).

Specifically, the gated convolution block in Figure 3(b) forms the simplest structure by omitting the SSM layer. Given $X' = \mathrm{Norm}(X)$, which is the visual feature map obtained after layer normalization, the block first computes $Z_1 = \sigma(\mathrm{Conv}(\mathrm{Linear}(X')))$ and $Z_2 = \sigma(\mathrm{Linear}(X'))$, where σ is a SiLU activation function, Conv is a depth-wise 1D convolution layer, and Linear is a linear layer. An additional linear layer is then applied to the gated output $Z_1 \odot Z_2$ followed by a skip connection as $Y = \mathrm{Linear}(Z_1 \odot Z_2) + X$. The SSM block in Figure 3(a) proposed in the original Mamba model [20] inserts the SSM layer into the computation of Z_1 as $Z_1 = \mathrm{SSM}(\sigma(\mathrm{Conv}(\mathrm{Linear}(X'))))$.

Inspired by this SSM block, our DSSM block in Figure 3(d) inserts the DSSM layer as $Z_1 = \mathrm{DSSM}(\sigma(\mathrm{Conv}(\mathrm{Linear}(X'))))$. This effectively enhances visual feature map via the deformable scan mechanism. We also add the feedforward network (FFN) block to the DSSM block as shown in Figure 3(e), corresponding "DSSM+FFN block" of Figure 4. This is a common practice in transformer-based architectures [12, 41, 63] for improving performance adopted in VMamba [40].

4.2. Tribrid Architecture

Most existing transformer backbones [12, 41] and Mamba backbones [40, 68, 81] consist of a uniform stack of the same block. However, these backbones often incur significant computational complexity, particularly in the lower layers where visual feature maps have higher resolutions, making them less practical for 3D HPE.

To address this, we adopt a tribrid architecture that integrates the DSSM blocks with convolution blocks and gated convolution blocks. This approach efficiently leverages the complementary strengths of different blocks by extracting features through convolution blocks at lower layers, while adaptively enhancing visual feature maps using DSSM blocks at higher layers.

Overview. Figure 4 provides an overview of our DF-Mamba backbone, which consists of a convolution stem for fine-grained feature extraction and six stages of feature enhancement. Specifically, the six stages follow three block types: convolution, DSSM+FFN, and gated convolution. Downsampling layers are appended at the end of stages 1, 2, and 4. This architecture is designed to maintain computational complexity comparable to that of ResNet-50. Since several 3D HPE methods use pyramid features to accurately estimate joint positions [32], the figure illustrates the flow to extract pyramid features from the four output feature maps.

Convolution stem. Given an input image of size $H \times W \times C_0$, the convolution stem applies two successive convolution layers, each followed by 2D batch normalization and a ReLU activation. The first convolution outputs 32 channels, and the second outputs $C_1 = 80$ channels; both use a 3×3 kernel with a stride of 2. This results in a visual feature map of size $H/4 \times W/4 \times C_1$.

Convolution blocks. Each convolution block uses the basic block design of ResNet-18. The activation function used is GELU instead of ReLU. Both convolution layers utilize a 3×3 kernel with a stride of 1. Each of stages 1 and 2 consists of three convolution blocks.

DSSM+FFN blocks. The DSSM+FFN block architecture follows the Figure 3(e)'s style, where the DSSM block in

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

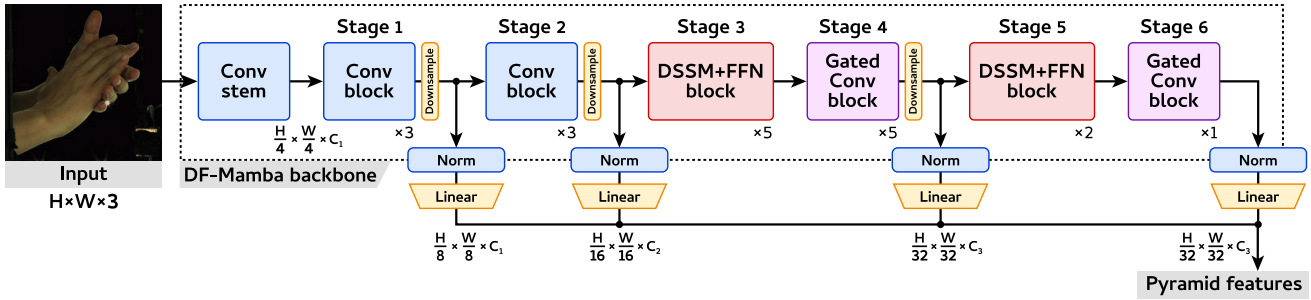

Figure 4. DF-Mamba backbone architecture. By combining three types of blocks, DF-Mamba improves the accuracy of 3D HPE while maintaining computational complexity comparable to or even lower than that of ResNet-50.

Figure 3(d) is utilized as the token mixer. We construct five and two DSSM+FFN blocks for stages 3 and 5, respectively. We adopt DSSM for these stages because applying deformable scanning for such downsampled high-level features mitigates spatial information loss and performs the state scanning more efficiently.

Gated convolution blocks. As in [74], the gated convolution block without SSM is computationally efficient and remains effective as a token mixer. Consequently, we construct five and one gated convolution blocks for stages 4 and 6, respectively, which are shown in Figure 3(b), where the FFN is omitted to reduce computational complexity. Applying gated convolution blocks after DSSM+FFN blocks enhances visual features while maintaining efficiency.

Downsampling. Downsampling layers are appended at the end of stages 1, 2, and 4. Each consists of a 2D convolutional layer with a 3×3 kernel and a stride of 2.

Pyramid features. Visual feature maps are extracted from stages 1, 2, 4, and 6. Layer normalization and a linear layer are applied to each feature map, and the resulting features are concatenated to represent pyramid features.

5. Experiments

5.1. Experimental Settings

Datasets and metrics. We conduct extensive experiments using DF-Mamba on five divergent datasets: InterHand2.6M [46] with two-hands poses, RHP [84] with single-hand poses, NYU [60] with depth images, DexYCB [4] with hand-object interactions, and AssemblyHands [49] with egocentric views. We report joint position errors following the evaluation protocols specified for each dataset, and also provide AUC scores for DexYCB and AssemblyHands. More details are provided in Appendix.

Baselines. We compare DF-Mamba with six backbones: ResNet-50 [26], ConvNeXt-T [42], ViT-S [12], Swin-T [41], VMamba-T [40], and Spatial-Mamba-T [68]. To perform 3D HPE, we integrate each backbone into two representative frameworks: one proposed by Jiang et al. [32] for InterHand2.6M, RHP, and NYU, and another by Zhou et al. [79] for DexYCB and AssemblyHands. These frameworks include decoders and are designed for use with ResNet-50; therefore, the other backbones are selected to have comparable model sizes.

Implementation details. We follow the original training settings described in [4, 32, 49]. All models are trained on the same data split from scratch. Detailed hyperparameters are provided in Appendix. All experiments are conducted using a single NVIDIA H100 GPU with FP32 precision. The inference speed (FPS) is measured with a batch size of 1 on InterHand2.6M.

5.2. Experimental Results

Table 1 compares DF-Mamba with other backbones across the five datasets. DF-Mamba consistently achieves superior performance, outperforming alternative backbones in terms of joint position errors and AUC scores on all datasets. Although several conventional backbones (*e.g.*, Swin-T, VMamba-T) improve performance on InterHand2.6M and RHP, their effectiveness diminishes in more challenging scenarios, such as depth-based estimation (NYU) and egocentric estimation (AssemblyHands). In these scenarios, the inductive biases inherent in convolutional networks often facilitate efficient learning. Consequently, Transformer-based and Mamba-based architectures face greater challenges. DF-Mamba successfully addressed this limitation through its tribrid architecture with deformable scan.

In terms of computational complexity, DF-Mamba maintains a model size and FPS comparable to those of ResNet. In Table 1, ViT-S has lower FPS, indicating that applying Transformer blocks to high-resolution feature maps is computationally expensive. Swin-T and VMamba-T address this limitation through a hierarchical architecture, and DF-Mamba further enhances efficiency by utilizing convolutional and gated convolutional blocks.

Below, we compare DF-Mamba with state-of-the-art 3D HPE methods on each dataset and discuss its performance.

InterHand2.6M. Table 2 shows results on InterHand2.6M. DF-Mamba achieves the best performance in intricate two-hands interactions, with MPJPE scores of 7.94 mm (Sin-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

Backbones	Size ↓	FPS ↑	InterHand2.6M (MPJPE)			RHP	NYU	DexYCB		AssemblyHands	
			Single ↓	Two ↓	All ↓	EPE ↓	ME↓	MPJPE ↓	AUC ↑	MPJPE ↓	AUC ↑
ResNet-50 [26]	**42M**	109.2	8.10	10.96	9.63	17.75	8.43	19.36	84.80	19.35	85.24
ConvNeXt-T [42]	45M	101.1	8.26 (+0.16)	11.04 (+0.08)	9.75 (+0.12)	17.96 (+0.21)	8.85 (+0.42)	21.83 (+2.47)	82.21 (-2.59)	20.72 (+1.37)	82.86 (-2.38)
ViT-S [12]	**42M**	75.4	–	–	–	–	–	24.63 (+4.73)	78.76 (-6.04)	23.11 (+3.76)	79.29 (-5.95)
Swin-T [41]	45M	103.4	8.15 (+0.05)	10.84 (-0.12)	9.59 (-0.04)	17.65 (-0.10)	8.48 (+0.05)	23.52 (+4.16)	80.59 (-4.21)	19.88 (+0.53)	84.03 (-1.21)
VMamba-T [40]	46M	100.8	8.06 (-0.04)	10.97 (+0.01)	9.61 (-0.02)	17.22 (-0.53)	8.62 (+0.19)	19.84 (+0.48)	84.45 (-0.35)	19.64 (+0.29)	84.89 (-0.35)
SMamba-T [68]	43M	92.8	8.44 (+0.34)	10.93 (-0.03)	9.77 (+0.14)	17.96 (+0.21)	8.78 (+0.35)	22.73 (+3.37)	80.37 (-4.43)	21.44 (+2.09)	81.88 (-3.36)
DF-Mamba	**42M**	**112.2**	**7.94** (-0.16)	**10.53** (-0.43)	**9.32** (-0.31)	**17.16** (-0.59)	**7.96** (-0.47)	**17.80** (-1.56)	**87.31** (+2.51)	**18.78** (-0.57)	**86.12** (+0.88)

Table 1. Comparison of backbone architectures across five datasets. Following the evaluation protocols for each dataset, Mean Per Joint Position Error (MPJPE), End Point Error (EPE), Mean Error (ME), and Area Under the Curve (AUC) scores are reported. For InterHand2.6M, MPJPE is reported for single-hand (Single), two-hand (Two), and overall (All) test sets. Numbers in parentheses indicate performance differences compared to ResNet-50, with green denoting improvement and red denoting degradation. Best results are highlighted in bold.

Methods	MPJPE			FPS ↑	Size ↓
	Single↓	Two↓	All↓		
Model-based					
Zhang *et al.* [76]	–	13.48	–	71.6	143M
Meng *et al.* [44]	8.51	13.12	10.97	65.0	55M
Li *et al.* [35]	–	8.79	–	76.0	39M
Model-free					
Moon *et al.* [46]	12.16	16.02	14.22	**330.5**	47M
Fan *et al.* [14]	11.32	15.57	–	87.1	41M
Hampali *et al.* [23]	10.99	14.34	12.78	82.7	48M
Jiang *et al.* [32]	8.10	10.96	9.63	109.2	42M
DF-Mamba (Ours)	**7.94**	**10.53**	**9.32**	112.2	42M

Table 2. Comparison with SOTA methods on InterHand2.6M.

gle hand), 10.53 mm (Two hands), and 9.32 mm (All). Compared to the InterHand architecture proposed by Moon *et al.* [46], the computational cost remains approximately three times higher. While our work focused on the backbone architecture, further optimization of the decoder is required to achieve a better trade-off between performance and efficiency in future work.

RHP. Table 3 presents a comparison on RHP, which contains a limited number of training images and rendered diverse backgrounds. DF-Mamba achieves the best performance without relying on ground-truth scale and handedness information provided in the dataset (*i.e.*, the GT-free setting). These results indicate that DF-Mamba effectively learns from limited data and generalizes to diverse scenes.

NYU. Under the depth-only setting, we observe from Table 1 that only DF-Mamba improves performance, achieving a mean error of 7.96 mm. The other baselines produce suboptimal performance, probably because they are primarily designed for training with RGB images. This result suggests that our DSSM provides additional flexibility in learning from depth images. Table 4 also shows that DF-Mamba outperforms strong baselines specialized in depth-image processing (*e.g.*, [17])[1], indicating that DF-Mamba delivers strong performance on depth-based 3D HPE.

DexYCB. Table 5 shows the results on DexYCB, which focuses on hand-object interactions captured from fixed cameras. DF-Mamba presents robust performance, achieving a significant improvement with an MPJPE of 17.80 mm. It also outperforms the method proposed by Spurr *et al.* [58], which uses the HRNet32 backbone[64].

AssemblyHands. Table 6 shows a comparison with state-of-the-art methods on AssemblyHands, a more challenging scenario involving real-world hand-object interactions captured from egocentric views. Our DF-Mamba achieves the lowest MPJPE among the compared methods, with an MPJPE of 18.78 mm. This demonstrates our superiority in dynamic, egocentric scenarios involving object interactions.

5.3. Ablation Study and Analysis

Necessity of deformable scan. Table 7 presents results without using the deformable scan, reducing the model to the vanilla SSM of Mamba. We observe a significant drop in performance, highlighting the importance of incorporating the deformable scan.

Architectures. To justify the proposed tribrid architecture design, we conduct experiments with alternative hybrid and monomodal architectures. Our architecture can be represented as "CCDGDG" across stages 1 to 6, where C, D, and G denote Conv, DDSM+FNN, and Gated-Conv blocks, respectively. We further construct hybrid architectures combining two block types (*i.e.*, CCGGGG and CCDDDD), as well as monomodal architectures consisting exclusively of a single block type. Experimental results in Table 7 demonstrate that our tribrid architecture significantly outperforms both hybrid and monomodal variants, highlighting the effectiveness of integrating complementary block structures.

Hyperparameter K. We evaluate the effect of the hyperparameter K by varying its value. The design of K is detailed in Appendix. The results in Table 7 indicate that $K = 3^2$ achieves the best performance. Even with $K = 1$, the

[1] "MA" in Table 4 indicates modality-agnostic (MA) models, which are not specialized for depth images.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

Methods	GT free	EPE↓
Yang *et al*. [72]		19.95
Spurr *et al*. [57]		19.73
Moon *et al*. [46]	✓	20.89
Jiang *et al*. [32]	✓	17.75
Ours	✓	**17.16**

Table 3. Comparison on RHP.

Methods	MA	ME↓
Moon *et al*. [45]		9.22
Xiong *et al*. [69]		8.61
Fang *et al*. [17]		8.29
Jiang *et al*. [32]	✓	8.43
Ours	✓	**7.96**

Table 4. Comparison on NYU.

Methods	MPJPE↓
Xiong *et al*. [69]	25.57
Spurr *et al*. [58]	22.26
Tse *et al*. [62]	21.22
Zhou *et al*. [79]	19.36
Ours	**17.80**

Table 5. Comparison on DexYCB.

Methods	MPJPE↓
Han *et al*. [24]	32.91
Ohkawa *et al*. [49]	21.92
Zhou *et al*. [79]	19.17
Ours	**18.78**

Table 6. Comparison on AssemblyHands.

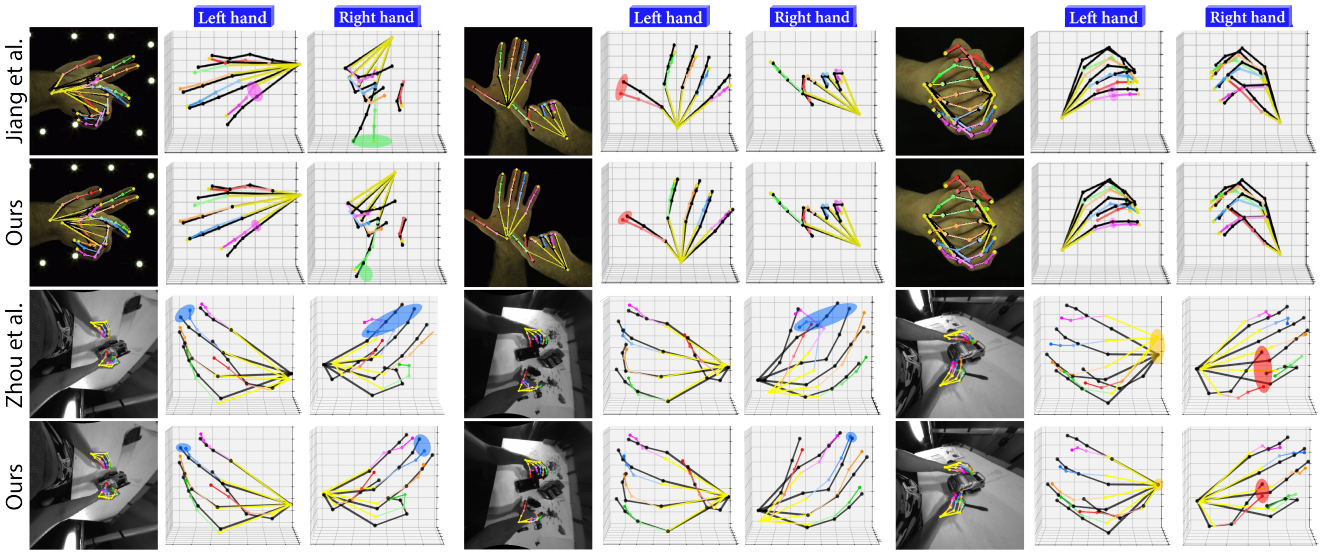

Figure 5. Qualitative examples. Predicted hand joints are color-coded by finger, with ground truth shown in black. The 3D visualizations provide rotated views, where several joint errors are highlighted using colored ellipses. The top two rows show examples from InterHand2.6M, and the bottom two rows from AssemblyHands.

Methods	DexYCB		AssemblyHands	
	MPJPE↓	AUC↑	MPJPE↓	AUC↑
DF-Mamba (Tribrid)	**17.80**	**87.31**	**18.78**	**86.12**
w/o deformable scan	18.90	85.59	19.38	84.93
w/ CCCCCC (Monomodal)	19.36	84.80	19.35	85.24
w/ GGGGGG (Monomodal)	18.70	85.98	19.13	85.67
w/ DDDDDD (Monomodal)	21.61	85.57	19.82	84.44
w/ CCGGGG (Hybrid)	18.67	86.12	18.97	85.77
w/ CCDDDD (Hybrid)	19.01	85.25	19.08	85.61
w/ CCDGDG (Tribrid)	**17.80**	**87.31**	**18.78**	**86.12**
$K = 1$	18.62	86.09	19.11	85.77
$K = 3^2$ (default)	**17.80**	**87.31**	**18.78**	**86.12**
$K = 5^2$	18.11	86.68	18.83	86.06

Table 7. Ablation and hyperparameter studies.

model outperforms alternative backbones, while $K = 5^2$ yields the second-best result.

Qualitative examples. Figure 5 presents qualitative examples from InterHand2.6M and AssemblyHands. DF-Mamba demonstrates more accurate joint estimation, as evidenced by reduced positional errors highlighted using colored ellipses. In egocentric scenarios, our approach accurately estimates joint positions, even when they are occluded or not directly visible from an egocentric viewpoint, highlighting the enhanced robustness in real-world settings.

6. Conclusion

We propose DF-Mamba, a novel backbone architecture for 3D hand pose estimation that combines efficient convolutional feature extraction in the lower layers with a deformable state-space representation in the higher layers. Through extensive experiments on five datasets covering single-hand, two-hand and hand-object interactions, as well as both RGB and depth modalities, we demonstrate that DF-Mamba consistently outperforms the existing backbone architectures while achieving faster inference. Promising future studies include 1) extending DF-Mamba to broader tasks like 3D human pose estimation, 2) pre-training it using vast external data, 3) applying it to recent hand-mesh reconstruction decoders (*e.g*., [11, 75]), and 4) its foundational model training that generalizes across different scenarios without data-specific fine-tuning.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

Acknowledgment. This work was supported by DENSO IT LAB Recognition, Control, and Learning Algorithm Collaborative Research Chair (Science Tokyo). This work was also supported by JSPS KAKENHI Grant Numbers 23H00490 and 25K03135. We thank Nie Lin and Ruicong Liu for their helpful support regarding implementation.

References

[1] Tatsuro Banno, Takehiko Ohkawa, Ruicong Liu, Ryosuke Furuta, and Yoichi Sato. Assemblyhands-x: Modeling 3d hand-body coordination for understanding bimanual human activities. 2025. 3

[2] William L. Brogan. *Modern Control Theory*. Quantum Publishers, 1974. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 2

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6

[5] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Hui Tang, Yufan Xue, Xiaohui Xie, Yen-Yu Lin, and Wei Fan. Generating realistic training images based on tonality-alignment generative adversarial networks for hand pose estimation. *arXiv preprint arXiv:1811.09916*, 2018. 2

[6] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20544–20554, 2022. 1

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017. 3

[8] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 3

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 764–773, 2017. 3, 4

[10] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *Proc. International Conference on Machine Learning (ICML)*, pages 933–941, 2017. 5

[11] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 37:2127–2160, 2024. 2, 3, 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 5, 6, 7

[13] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108 (1-2):52–73, 2007. 2

[14] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *2021 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2021. 7

[15] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[16] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proc. European Conference on Computer Vision (ECCV)*, pages 428–448. Springer, 2024. 1, 2

[17] Linpu Fang, Xingyan Liu, Li Liu, Hang Xu, and Wenxiong Kang. Jgr-p2o: Joint graph reasoning based pixel-to-offset prediction network for 3d hand pose estimation from a single depth image. In *Proc. European Conference on Computer Vision (ECCV)*, pages 120–137, 2020. 7, 8

[18] Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *Proc. International Conference on Learning Representations (ICLR)*, 2023. 3

[19] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10833–10842, 2019. 2

[20] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proc. Conference on Language Modeling (COLM)*, 2024. 2, 3, 4, 5

[21] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1474–1487, 2020. 3

[22] Shaoxiang Guo, Qing Cai, Lin Qi, and Junyu Dong. Cliphand3d: Exploiting 3d hand pose estimation via context-aware prompting. In *Proc. ACM International Conference on Multimedia (ACMMM)*, 2023. 2

[23] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identifi-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

cation in challenging hands and object interactions for accurate 3d pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11090–11100, 2022. 7

[24] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 8

[25] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 3, 6, 7

[27] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[28] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation. In *Proc. ACM International Conference on Multimedia (ACMMM)*, pages 3136–3145, 2020. 2

[29] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. LocalMamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 3

[30] Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. Posemamba: Monocular 3d human pose estimation with bidirectional global-local spatio-temporal state space model. In *Proc. AAAI Conference on Artificial Intelligence*, pages 3842–3850, 2025. 3

[31] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proc. European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2

[32] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8846–8856, 2023. 1, 2, 5, 6, 7, 8

[33] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(1):35–45, 1960. 3

[34] Bo Lang and Mooi Choo Chuah. Event-guided fusion-mamba for context-aware 3d human pose estimation. In *Proc. IEEE Conference on Winter Conference on Applications of Computer Vision (WACV)*, pages 950–960, 2025. 3

[35] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2761–2770, 2022. 7

[36] Guan Ming Lim, Prayook Jatesiktat, and Wei Tech Ang. Mobilehand: Real-time 3d hand shape and pose estimation from color image. In *Proc. International Conference on Neural Information Processing (ICONIP)*, pages 450–459, 2020. 1

[37] Nie Lin, Takehiko Ohkawa, Mingfang Zhang, Yifei Huang, Minjie Cai, Ming Li, Ryosuke Furuta, and Yoichi Sato. SiMHand: Pre-training for 3d hand pose estimation with contrastive learning on large-scale hand images in the wild. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. 2

[38] Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, and Yoichi Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 677–686, 2024. 1, 2

[39] Ruicong Liu, Takehiko Ohkawa, Tze Ho Elden Tse, Mingfang Zhang, Angela Yao, and Yoichi Sato. Leveraging rgb images for pre-training of event-based hand pose estimation. 2025. 2

[40] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 5, 6, 7

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 3, 5, 6, 7

[42] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6, 7

[43] Ye Lu, Jie Wang, Jianjun Gao, Rui Gong, Chen Cai, and Kim-Hui Yap. A structure-aware and motion-adaptive framework for 3d human pose estimation with mamba. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[44] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. *arXiv preprint arXiv:2207.11061*, 2022. 7

[45] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5079–5088, 2018. 8

[46] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand 2.6M: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Proc. European Conference on Computer Vision (ECCV)*, pages 548–564, 2020. 1, 2, 3, 6, 7, 8

[47] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Ryosuke Furuta, Kris M Kitani, and Yoichi Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Proc. European Conference on Computer Vision (ECCV)*, pages 68–87. Springer, 2022. 2, 3

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

[48] Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Efficient annotation and learning for 3d hand pose estimation: A survey. *International Journal of Computer Vision*, 131(12): 3193–3206, 2023. 2

[49] Takehiko Ohkawa, Kun He, Fadime Sener, Tomáš Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 1, 2, 3, 6, 8

[50] Takehiko Ohkawa, Jihyun Lee, Shunsuke Saito, Jason Saragih, Fabian Prado, Yichen Xu, Shoou-I Yu, Ryosuke Furuta, Yoichi Sato, and Takaaki Shiratori. Generative modeling of shape-dependent self-contact human poses. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2

[51] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[52] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 2

[53] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024. 3

[54] Yaopeng Peng, Milan Sonka, and Danny Z. Chen. Group vision transformer. In *Proc. ACM International Conference on Multimedia (ACMMM)*, 2024. 3

[55] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2

[56] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8014–8025, 2023. 2

[57] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 89–98, 2018. 8

[58] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *Proc. European Conference on Computer Vision (ECCV)*, pages 211–228, 2020. 7, 8

[59] Naru Suzuki, Takehiko Ohkawa, Tatsuro Banno, Jihyun Lee, Ryosuke Furuta, and Yoichi Sato. Affordance-guided diffusion prior for 3d hand reconstruction. 2025. 2

[60] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014. 1, 2, 3, 6

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. International Conference on Machine Learning (ICML)*, pages 10347–10357, 2021. 3

[62] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1674, 2022. 8

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 5

[64] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(10): 3349–3364, 2020. 7

[65] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, 2021. 3

[66] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[67] Yilin Wen, Hao Pan, Takehiko Ohkawa, Lei Yang, Jia Pan, Yoichi Sato, Taku Komura, and Wenping Wang. Generative hierarchical temporal transformer for hand pose and action modeling. *Proc. European Conference on Computer Vision Workshops (ECCVW)*, 2024. 2

[68] Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. Spatial-mamba: Effective visual state space models via structure-aware state fusion. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 4, 5, 6, 7

[69] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 793–802, 2019. 8

[70] Guanyu Xu, Jiawei Hao, Li Shen, Han Hu, Yong Luo, Hui Lin, and Jialie Shen. Lgvit: Dynamic early exiting for accelerating vision transformer. In *Proc. ACM International Conference on Multimedia (ACMMM)*, pages 1–12, 2023. 3

[71] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J. Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. In *Proc. British Machine Vision Conference (BMVC)*, 2024. 3

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.

[72] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9877–9886, 2019. 8

[73] Huan Yao, Changxing Ding, Xuanda Xu, and Zhifeng Lin. Decoupling heterogeneous features for robust 3d interacting hand poses estimation. In *Proc. ACM International Conference on Multimedia (ACMMM)*, 2024. 2

[74] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 5, 6

[75] Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12955–12964, 2023. 8

[76] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11354–11363, 2021. 7

[77] Xuan Zhang, Qian Bao, Qixuan Cui, Wankou Yang, and Qing Liao. Pose magic: Efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network. In *Proc. AAAI Conference on Artificial Intelligence*, pages 10248–10256, 2025. 3

[78] Yumeng Zhang, Li Chen, Yufeng Liu, Junhai Yong, and Wen Zheng. Adaptive wasserstein hourglass for weakly supervised hand pose estimation from monocular rgb. In *Proc. ACM International Conference on Multimedia (ACMMM)*, pages 1944–1952, 2020. 2

[79] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5346–5355, 2020. 2, 6, 8

[80] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1367–1376, 2024. 2

[81] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proc. International Conference on Machine Learning (ICML)*, 2024. 2, 3, 5

[82] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9308–9316, 2019. 3

[83] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 2

[84] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4903–4911, 2017. 2, 6

[85] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 2

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 15:52:38 UTC from IEEE Xplore. Restrictions apply.