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

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

Abstract

Reconstructing 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 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 prediction tasks like 3D HPE, with improved inference speed over ResNet50. Our experiments involve extensive evaluations on five datasets that cover diverse scenarios, including single-hand and two-hands estimation, hand-only and hand-object interactions, as well as RGB and depth modalities. We demonstrate that 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 interacting with two hands [24, 28] and grasping an object [2, 7, 27], which necessitates effective and efficient inference models that reconstruct 3D hands to comprehend 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 [9, 16, 26], depth images [32], egocentric views [7, 21], etc. In parallel, developing compact models with improved inference speed has become crucial to support real-time applications, especially in AR/VR devices [3, 19]. Given these challenges, limited attention has been paid to the inductive biases introduced by backbone architectures (e.g., CNNs) and their synergy with HPE. This highlights the need for designing backbones that are both effective in capturing complex hand interactions and efficient for real-time inference.

To enable robust feature learning for complex hand interactions, it is essential to learn the relationship between local image features (e.g., for occluded joints) and global context

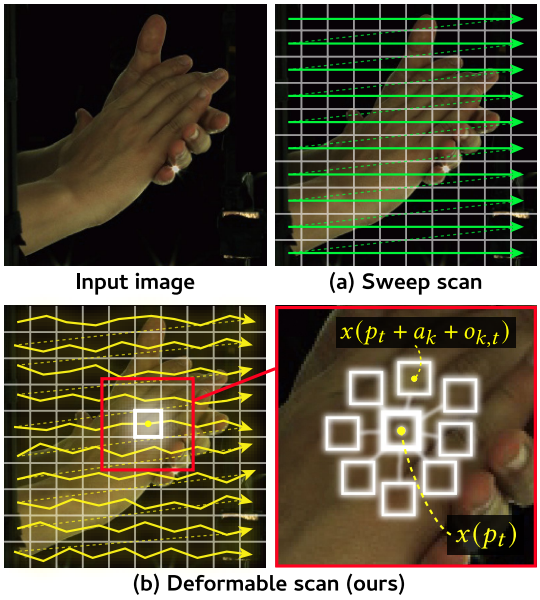

Figure 1. **Deformable scan for DF-Mamba.** (a) Conventional sweep scan uses a fixed grid pattern [22, 38] in the state space equations. (b) Our deformable scan adaptively adjusts the scanning pattern with multiple anchors $\boldsymbol{a_k}$ by predicting offset vectors $\boldsymbol{o_{k,t}}$ dependent on visual feature input.

*Equal contribution.

(e.g., cues from inter-joints, inter-hands, hand-object, or the scene). A popular backbone in HPE is CNN, particularly ResNet-50 [13] used in numerous works [6, 15, 16, 20, 21, 24, 25, 27, 29, 30, 37]. These CNN backbones rely on convolution operations with local receptive fields, resulting in a favorable balance between accuracy and inference speed. Nevertheless, these local convolutions lack an explicit capability to model global context. In contrast, existing transformer methods [12, 14, 16, 18, 31, 33, 34, 36] enable non-local feature learning; for instance, Jiang *et al*. [16] encourage self-attention across local anchor points and pyramid image features. However, these transformer methods indeed follow a hybrid approach, combining ResNet-50's feature extraction in an initial stage with a subsequent transformer that learns across the extracted features using attention." This underscores that most current HPE methods still heavily depend on CNN's inductive bias, indicating significant room for improvement in backbone architectures to achieve better feature learning for hand pose.

As an emerging foundational architecture, Mamba [10] based on state space modeling (SSM) has garnered considerable attention, which have been originally proposed for natural language processing tasks. The Mamba model excels at efficiently selecting input tokens (i.e., focusing on or ignoring particular signals), which serves to model global context cues from long sequential tokens. Several recent studies have extended it to image backbones. For example, VisionMamba [38] introduced the Vim block, which employs a 2D bidirectional scan for spatially-aware sequence modeling. VMamba [22] 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 after convolution and global context with flexible state spaces. Specifically, our DSSM blocks aggregate the convoluted 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 (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, along with a comparable model size with ResNet50. This approach efficiently leverages the complementary strengths of each block type: extracting 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 without SSM.

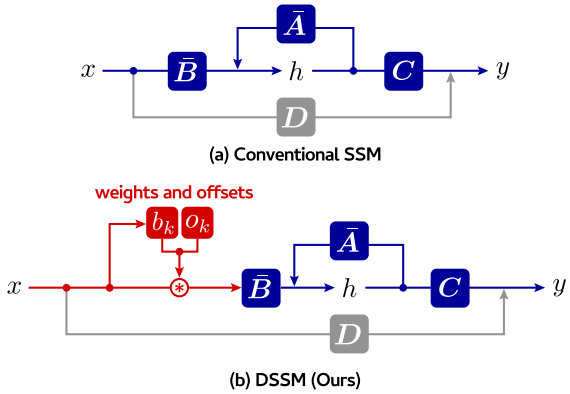

Figure 2. Computational flow of SSM and DSSM. (a) Conventional SSM utilizes four matrices $\bar{\boldsymbol{A}}, \bar{\boldsymbol{B}}, \boldsymbol{C}, \boldsymbol{D}$, to compute the output $\boldsymbol{y}$ from an input $\boldsymbol{x}$ through intermediate representation $\boldsymbol{h}$. (b) Our DSSM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

In our experiments, we integrate DF-Mamba into two representative 3D HPE frameworks proposed by Jiang *et al*. [16] and Zhou *et al*. [37]. Since both frameworks originally utilize a ResNet50 backbone for feature extraction, we replace the backbone with our DF-Mamba. We then perform extensive evaluations on five public datasets: InterHand2.6M [24], RHP [39], NYU [32], DexYCB [2] and AssemblyHands [27]. These datasets cover diverse interaction scenarios, including single-hand and two-hands pose estimation, hand-only and hand-object interactions, as well as RGB and depth modalities. We demonstrate that DF-Mamba outperforms the latest image backbones, including VMamba [22] and SpatialMamba [35], on all datasets and achieves state-of-the-art performance. We also find that DF-Mamba maintains computational complexity comparable to or even lower than that of ResNet50. These results suggest that DF-Mamba is a more effective and efficient backbone than ResNet-50 in 3D HPE.

In summary, our contributions are three-fold:

1) We propose **DSSM**, a novel approach to modeling dynamic systems with flexible state spaces to represent the global context.
2) We introduce **DF-Mamba**, a novel Mamba-based backbone for 3D HPE. It adopts a tribrid design composed of six stages using three types of blocks: convolution blocks, DSSM blocks and gated convolution blocks.
3) DF-Mamba outperforms the latest image backbones in 3D HPE scenarios and consistently improves performance on five different datasets with faster inference.

2. Method

We introduce **Deformable State-Space Modeling (DSSM)**, a novel approach to modeling dynamic systems

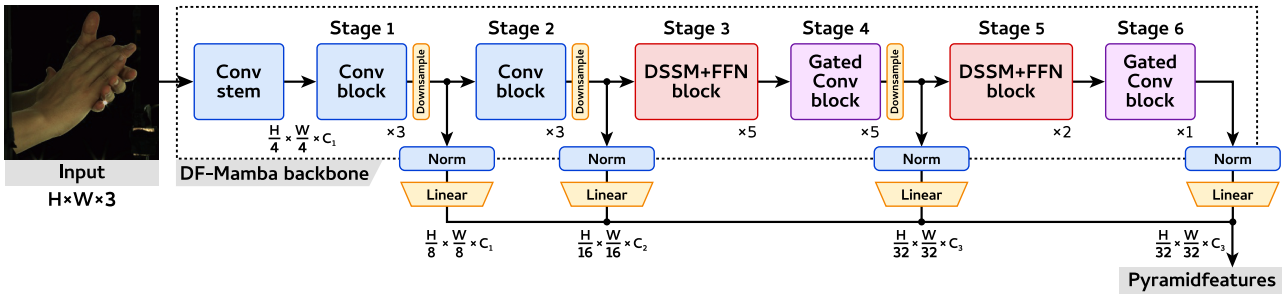

Figure 3. DF-Mamba backbone architecture. By combining three types of blocks, DF-Mamba improves the accuracy of 3D HPE while maintaining computational complexity comparable to or even lower than that of ResNet50.

with flexible state spaces. The core idea behind DSSM is to integrate local visual cues after convolution with a deformable scan mechanism that serves to capture global context. The proposed deformable scanning in DSSM aggregates these local features within an image while selectively preserving useful cues as the global features. Below, we begin with a preliminary review of state-space equations, and then introduce our DSSM.

Preliminary. The Mamba model [10] follows state-space modeling (SSM), where the input $x(t)$ and output $y(t)$ are linked by a dynamic linear system with a latent state $h(t)$ [1, 17]; see Figure 2(a). The discrete state-space equations [8, 11] are given by:

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

Gu *et al*. [10] introduce selective SSM that adaptively remembers or forgets inputs by computing $\bar{B}$ and C 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. (1) with a 2D 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.

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 has a higher density of joints. These observations necessitate a flexible solution to model spatial locality bias and global context adaptively to the input.

2.1. Deformable State-Space Modeling

To overcome the SSM's limitation, we introduce data-driven, adaptive capabilities to the SSM blocks. Our DSSM incorporates 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:

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

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. To accommodate this, we effectively handle such a large freedom by introducing spatially allocated K anchors, which sample multiple points from the input image. Specifically, we define DSSM for 2D inputs as follows:

$$\begin{cases} h(t) = \bar{A}h(t-1) + \bar{B}\sum_{k=1}^{K} b_k x(p_t + a_k + o_{k,t}) \\ y(t) = Ch(t) + Dx(p_t) \end{cases}, \tag{3}$$

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.

The computational flow of DSSM is shown in Figure 2(b). The K anchors are symmetrically defined as $a_k \in \{(i,j) : i,j \in \{-1,0,+1\}\}$, inspired by deformable convolution operations [4]. 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, allowing dynamic adaptation of sampling locations based on input context. Technically, the weight coefficients and offset vectors are predicted from the input visual features. This balances initial coverage of the anchors with data-driven flexibility by 2D scanning, a core DSSM strength.

2.2. DF-Mamba Backbone

With the proposed DSSM, we construct the overall backbone architecture, **DF-Mamba**, as shown in Figure 3. To

Table 1. Performance comparison of backbone architectures across various datasets. Best results are highlighted in bold. Numbers in parentheses indicate performance differences compared to ResNet50, with green denoting improvement and red denoting degradation.

Backbones	Venue	FPS ↑	Size ↓	InterHand2.6M			RHP	NYU	DexYCB		AssemblyHands	
				Single ↓	Two ↓	All ↓	EPE ↓	Mean Err. ↓	MPJPE ↓	AUC ↑	MPJPE ↓	AUC ↑
ResNet50 [13]	CVPR16	109.2	**42M**	8.10	10.96	9.63	17.75	8.43	19.36	84.80	19.35	85.24
ViT-S [5]	ICLR21	75.4	**42M**	–	–	–	–	–	24.63 (+4.73)	78.76 (-6.04)	23.11 (+3.76)	79.29 (-5.95)
Swin-T [23]	ICCV21	103.4	45M	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 [22]	NeurIPS24	100.8	46M	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)
SpatialMamba-T [35]	ICLR25	92.8	43M	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 (Ours)	–	**112.2**	**42M**	**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)

balance its effectiveness and efficiency, DF-Mamba consists of a convolution stem for fine-grained feature extraction and six stages of feature enhancement, including a tribrid design from three types of blocks: (i) convolution blocks, (ii) DSSM blocks, and (iii) 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.

Specifically, the gated convolution block 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 proposed in the original Mamba model [10] 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 inserts the DSSM layer as $Z_1 = \mathrm{DSSM}(\sigma(\mathrm{Conv}(\mathrm{Linear}(X'))))$. This effectively enhances visual feature map through the deformable scan.

3. Experiments

We conduct extensive experiments to validate the ability of DF-Mamba over state-of-the-art backbone architectures. Specifically, we perform evaluations on five datasets that cover diverse scenarios, including InterHand2.6M [24] for two-hands pose, RHP [39] for single-hand pose, NYU [32] for depth-based estimation, DexYCB [2] for object interaction, and AssemblyHands [27] for egocentric perception. We integrate DF-Mamba into two HPE frameworks proposed by Jiang *et al*. [16] (for InterHand2.6M, RHP, and NYU) and Zhou *et al*. [37] (for the rest). We choose baselines whose model size is closest to that of ResNet-50.

Comparison to SOTA backbones. Table 1 compares DF-Mamba with state-of-the-art backbones on the five datasets. We observe that Swin-T and VMamba-T improve performance on InterHand2.6M and RHP, while vision transformer baselines (ViT-S and Swin-T) exhibit suboptimal results, especially in DexYCB and AssemblyHands, because CNN's locality bias is better suited for heatmap regression [37]. In contrast, our DF-Mamba achieves the best and consistently improves over ResNet50 in all the scenarios, namely two interacting hands [24], synthetic data with diverse backgrounds [39], depth images [32], hand-object scenes [2], and egocentric views [27].

Not only do we find performance gains, but DF-Mamba also improves inference speed compared to ResNet50, while preserving a smaller model size against Swin-T and the other Mamba baselines. We observe that feature extraction at lower layers of ViT-S, Swin-T, VMamba-T, and SpatialMamba-T increases computational time (lower FPS), as applying attention or state-space modeling to high-resolution feature maps is expensive. We also find a major bottleneck of ViT-based backbones in inference speed, limiting the potential for real-time applications [3, 19]. Overall, our tribrid architecture successfully addresses these limitations by efficiently extracting features at lower layers similar to ResNet50, and effectively enhancing these features via DSSM at higher layers. These results highlight the superiority of DF-Mamba over those SOTA backbones.

4. 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- and two-hand interactions, hand-object interactions, and both RGB and depth modalities, we demonstrate that DF-Mamba consistently outperforms the existing backbone architectures while reducing computational cost.

References

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

[2] 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. 1, 2, 4

[3] 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, 4

[4] 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

[5] 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. 4

[6] 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

[7] 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

[8] 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

[9] 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. 1

[10] 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

[11] 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

[12] 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

[13] 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, 4

[14] 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

[15] 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

[16] 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, 4

[17] 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

[18] 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. 2

[19] 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, 4

[20] 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

[21] 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

[22] 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. 1, 2, 4

[23] 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. 4

[24] 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, 4

[25] 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

[26] 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. 1

[27] 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, 4

[28] 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

[29] 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. 2

[30] 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

[31] 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

[32] 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, 4

[33] 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

[34] 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

[35] 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, 4

[36] 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. 2

[37] 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, 4

[38] 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. 1, 2

[39] 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, 4