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

Supplementary Material

A. Datasets and Evaluation Metrics

This section details the datasets and evaluation metrics used in our experiments.

1) InterHand2.6M [46]. The InterHand2.6M dataset is a large-scale dataset for 3D interacting hand pose estimation (HPE), featuring extensive variations in hand poses. Following the evaluation protocols in [32, 46], we use the union of single-hand and two-hand subsets. The training and test sets contain 1.36M and 849K RGB images, respectively, with annotations of 42 hand joints per image. We report performance using mean per-joint position error (MPJPE, mm) for single-hand (Single), two-hand (Two), and overall (All) test subsets.

2) RHP [84]. The RHP dataset is a synthetic dataset consisting of RGB images of two isolated hands. This dataset is used to evaluate how well models generalize to outdoor scenes. The training and test sets contain 41k and 2.7k images, respectively, each annotated with 24 hand joints. Endpoint error (EPE), defined as the mean Euclidean distance between the predicted and ground-truth 3D hand poses after root joint alignment, is used to measure performance.

3) NYU [60]. The NYU dataset is a depth image dataset for single-hand pose estimation. This dataset is used to evaluate models under depth-only settings. The training and test sets consist of 72k and 8.2k images, respectively. Following previous studies [32, 45, 69], we use 14 of the 36 joints in the experiments. Mean 3D distance error (Mean Error) is used to measure performance.

4) DexYCB [4]. The DexYCB dataset is a large-scale dataset that captures real-world hand-object interactions. The training and test sets consist of 582k and 163k images, respectively, each annotated with 21 hand joints. We use the set of unseen subjects. MPJPE and Area Under the Curve (AUC) are used as evaluation metrics.

5) AssemblyHands [49]. The AssemblyHands dataset is a large-scale dataset for 3D HPE from egocentric viewpoints, featuring complex hand-object interaction scenarios captured in real-world settings. The training and test sets consist of 704k and 112k images, respectively, each annotated with 21 hand joints. MPJPE and AUC are used as evaluation metrics.

B. Implementation Details

We follow the original training settings described in [4, 32, 49]. For the InterHand2.6M and RHP datasets, RGB images are resized to a resolution of 256×256. The model is trained for 42 epochs using the Adam optimizer, with a learning rate of 10^{-4} and a weight decay of 10^{-4}. For the NYU dataset, depth images are resized to 176×176, and the model is trained for 17 epochs with the same learning rate and weight decay. For DexYCB and AssemblyHands, images are resized to 128×128, and the model is trained for 20 epochs with a learning rate of 5×10^{-4}. For the ablation study of K in Table 7, we investigate the impact of different numbers of anchors. Assuming a 2D image as input (*i.e.*, $D = 2$), the anchor a_k of Sec. 3.2 is set within a uniform index range along the 2D axes, following the three variants: $i = j = 0$ ($K = 1$), $i, j \in \{-1, 0, +1\}$ ($K = 3^2$, default), and $i, j \in \{-2, -1, 0, +1, +2\}$ ($K = 5^2$).

Methods	MPJPE (mm) ↓		
	Single	Two	All
DF-Mamba	**7.94**	**10.53**	**9.32**
w/o deformable scan	8.10	10.66	9.47
w/o DSSM (CCGGGG)	8.04	10.79	9.51

Table 8. Ablation study on the InterHand2.6M dataset.

Model	FLOPs	Throughput
ResNet-50	4.1	4,977
ResNet-152	11.6	2,146
ConvNeXt-T	4.5	3,799
ConvNeXt-S	8.7	2,387
VMamba-T	4.9	1,524
VMamba-S	8.7	1,002
Swin-T	4.4	1,676
SMamba-T	4.5	4,285
DF-Mamba	4.9	5,310

Table 9. Comparison of FLOPs (G) and backbone throughput (images/sec) with a batch size of 128 and an image size of 256.

Methods	DexYCB		AssemblyHands	
	MPJPE↓	AUC↑	MPJPE↓	AUC↑
ResNet-50	19.36	84.80	19.35	85.24
ResNet-152	18.27	86.59	18.85	85.90
ConvNeXt-T	21.83	82.21	20.72	82.86
ConvNeXt-S	20.12	84.36	20.13	83.76
VMamba-T	19.84	84.45	19.64	84.89
VMamba-S	19.76	85.44	18.98	85.78
DF-Mamba	**17.80**	**87.31**	**18.78**	**86.12**

Table 10. Comparison with larger backbones.

C. Additional Results

Ablation study on InterHand2.6M. Table 8 shows the results of the ablation study on the InterHand2.6M dataset. We observe that each component consistently contributes to performance improvements, even when million-scale training data is utilized.

FLOPs and throughput. Table 9 summarizes the FLOPs and throughput for various backbones. As shown, DF-Mamba has FLOPs comparable to VMamba-T. DF-Mamba achieves higher throughput because it applies Mamba to feature maps downsampled by convolutional blocks.

Comparison with larger backbones. Table 10 presents a comparison of DF-Mamba against larger backbone variants, including ResNet, ConvNeXt, and VMamba. Although larger model variants typically achieve better performance, DF-Mamba consistently outperforms them. By adding one additional gated convolutional block into each of stages 4 and 6, we observe a 0.1 mm improvement in MPJPE for DF-Mamba on both datasets. However, adding more blocks to increase the model size comparable to other backbones does not yield significant improvements. Scaling DF-Mamba through large-scale pre-training to handle diverse 3D HPE scenarios remains future work.