Prior-aware Dynamic Temporal Modeling Framework for Sequential 3D Hand Pose Estimation

Pengfei Ren, Jingyu Wang*, Haifeng Sun*, Qi Qi, Xingyu Liu,
Menghao Zhang, Lei Zhang, Jing Wang, Jianxin Liao
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
{rpf, wangjingyu, hfsun, qiqi8266, liuxingyu, zhangmenghao}@bupt.edu.cn;
zhangl83@chinaunicom.cn; wangjing@bupt.edu.cn; jxlbupt@gmail.com

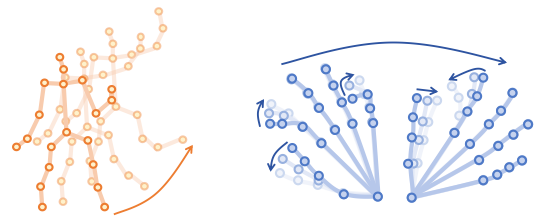
2025 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3315-8775-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCV51701.2025.00611

Abstract

3D hand pose estimation plays a critical role in various human-computer interaction tasks. Single-frame 3D hand pose estimation methods have poor temporal smoothness and are easily affected by self-occlusion, which severely impacts their practical applicability. Traditional joint-based sequential pose estimation methods primarily focus on the human body and struggle to handle the complex hand structure, high degrees of freedom in hand motion, and rapidly changing hand motion trends. To address these challenges, we propose a prior-aware dynamic temporal modeling framework for sequential 3D hand pose estimation. We introduce a flexible memory mechanism to model hand prior information, which alleviates the scale and depth ambiguity in single-frame hand pose estimation. Additionally, we propose a dynamic temporal convolution module that adjusts the receptive field size and feature aggregation weights based on the motion information at each moment, effectively capturing rapid motion trends. By decoupling dynamic temporal modeling at the joint and hand levels, our method captures both subtle short-term variations and long-term motion trends, significantly improving the smoothness and accuracy of hand pose estimation. Experiments on four public datasets demonstrate that our method achieves the state-of-the-art results in terms of hand pose estimation accuracy and temporal smoothness.

1. Introduction

3D hand pose estimation is a crucial research direction in computer vision, with widespread applications in virtual reality (VR), augmented reality (AR), and human-computer interaction (HCI). With the increasing demand for immersive and natural interactions, robust and accurate 3D hand pose estimation has garnered significant attention. However, due to challenges such as viewpoint, pose, and lighting variations, maintaining temporal smoothness in hand pose estimation based on a single frame is difficult [2, 4, 20, 40, 58, 59, 67, 76]. Hand poses and hand shapes face issues of temporal jitter and abrupt transitions, which severely impact the practical usability of hand pose estimation. Therefore, leveraging temporal information to enhance the accuracy and stability of 3D hand pose estimation is important. Among these, using temporal neural networks to directly refine the 3D poses estimated by monocular methods has received widespread attention due to its flexibility, generalization, and efficiency.



(a) Human Body Motion (b) Hand Motion

Figure 1. **Comparison of human and hand motion patterns.** Body motion typically exhibits long-term and stable motion patterns. Hand motion is more flexible, with high frequency and large amplitude changes in motion trends. It often involves the coupling of long-term, large-scale motions with short-term, subtle motions.

Current sequential pose estimation methods based on joint have primarily focused on 3D human pose estimation [17, 28, 30, 31, 53, 66, 68–70, 74]. These approaches mainly emphasize modeling structured spatial dependencies among joints and long-range temporal dependencies to capture human motion trajectories and refine the 3D joint positions. However, as shown in Fig. 1, due to the complex-

*Corresponding author

2380-7504/25/$31.00 ©2025 IEEE
DOI 10.1109/ICCV51701.2025.00611

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

ity of hand structure, high freedom of hand poses, and fast-changing trends of hand motion, existing methods face serious challenges. Specifically, these methods fail to account for the rapid and frequent changes in local and global hand motion, hard to capture subtle hand pose variations, leading to over-smoothing or even failures in hand pose refinement. Therefore, the key to improving the accuracy and temporal consistency of 3D hand pose refinement lies in jointly modeling the short-term and long-term motion of hand.

The hand poses within the same sequence inherently exhibit fixed shape priors, such as consistent bone lengths, which provide strong disambiguation cues for refining inaccurate initial hand poses, especially in cases of motion blur or occlusion in single-frame images [19, 25]. Meanwhile, hand movements exhibit characteristics of intense local short-term variations and stable global long-term trends. Therefore, it is necessary to explicitly model both local short-term dynamics and global long-term trends, while dynamically capturing motion trend features across different temporal granularities. Hand shape priors are relatively stable features across different temporal granularities, providing a robust anchor for temporal modeling. Short- and long-term dynamic temporal modeling facilitates the decoupling of hand shape from inaccurate pose sequences. Thus, hand shape priors and dynamic temporal modeling are complementary and mutually reinforcing.

Inspired by the above insights, we propose a prior-aware dynamic temporal modeling framework for sequential 3D hand pose estimation. Specifically, we implement a flexible memory mechanism for modeling hand prior information, which can alleviate the scale ambiguity and depth ambiguity problems of single-frame hand pose estimation. At the same time, we propose a dynamic temporal convolution module, which dynamically adjusts the receptive field size and feature aggregation weights based on the hand motion information, thereby accurately capturing the rapidly changing motion trends. In particular, considering the flexibility of hand movements, we perform decoupled dynamic temporal modeling at the joint level and the palm level. By combining temporal dynamic convolution with a multi-head attention mechanism, our model can capture long-term motion trends and subtle short-term variations. Our method alleviates the difficulties caused by the high degree of freedom of hand pose and complex hand movement patterns, significantly enhancing the smoothness and accuracy of temporal hand pose estimation.

To validate the effectiveness of our method, we perform experiments on four datasets, including InterHand2.6M [41], Re:InterHand [42], HanCo [77], and DexYCB [3]. These datasets have a wide range of interactive scenarios and hand pose variations. On all datasets, our method can effectively enhance the performance of single-frame methods, demonstrating the effectiveness of hand prior modeling and dynamic temporal feature fusion. Our contributions can be summarized as follows:

- We propose a prior-aware dynamic temporal modeling framework for sequential hand pose estimation, which explicitly addresses the complexity of hand motion.
- We propose a dynamic temporal convolution that adaptively models both short- and long-range temporal information, while also adopting a memory mechanism to capture hand prior information.
- Our method outperforms existing methods in multiple datasets and is less dependent on hand pose estimators.

2. Related Work

2.1. Single-frame Hand Pose Estimation

Single-frame hand pose estimation aims to estimate hand poses directly from input data at a single moment. Early single-frame 3D hand pose estimation works relied on depth data [20, 40, 48, 56, 58, 59], as it provided geometric structure information of the hand's surface, thus reducing the complexity of network learning. With the emergence of large-scale datasets and the development of deep learning, it has become possible to estimate accurate 3D hand joint positions or hand parameters from RGB images [7, 8, 15, 21, 26, 32, 33, 54, 75]. In particular, 3D hand pose estimation in complex scenarios, such as hand-object interaction and two-hand interaction [13, 18, 29, 38, 41, 52, 65], has recently become a hot research topic. However, due to self-occlusion and self-similarity, the temporal stability and smoothness of single-frame existing methods are poor, leading to significant jitter, especially in complex scenarios.

2.2. Temporal Pose Estimation

In 3D human pose estimation, a representative method is predicting 3D pose from 2D sequences. These approaches utilize fully connected networks [10, 37], graph convolution [17, 28, 69, 74], temporal convolution [6, 46], transformers [30, 31, 66, 68, 70] to perform spatio-temporal interactions between joints. Its core purpose is to restore the 3D structure of the human body through temporal information. In addition, some works [27] attempt to enhance the smoothness and local details of the 3D hand surface by introducing temporal information. Recently, some works [60, 64] have developed a plug-and-play network that can take the 3D pose estimation results from any single-frame model as input to predict more accurate and smooth 3D human poses. Our work falls into this category, but differs in that we focus on temporal 3D hand pose estimation, which involves higher degrees of freedom and more frequent variations in motion trends.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

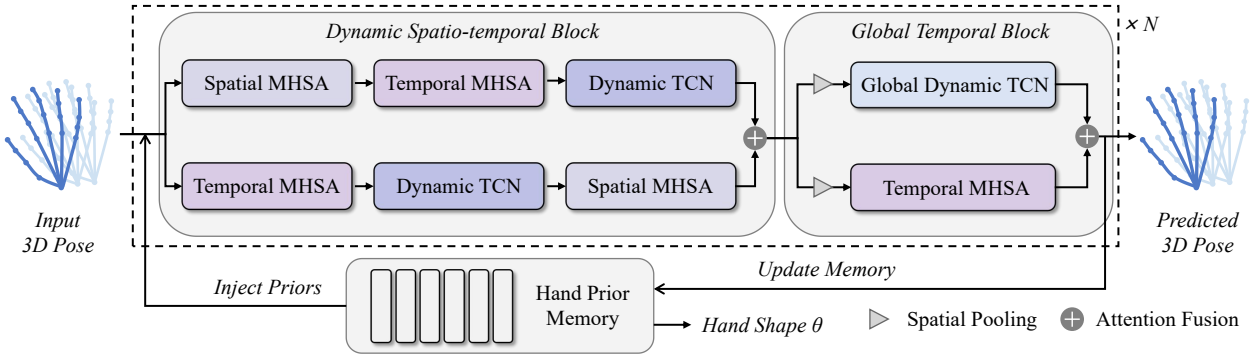


Figure 2. **Overall Architecture**. Our method comprises three primary components: the Dynamic Spatio-temporal Block, the Global Temporal Block, and the Hand Prior Memory. The Dynamic Spatio-temporal Block employs a dual-branch structure to model joint-level spatial and temporal relationships, while the Global Temporal Block extracts hand-level temporal information. We leverage hand-level temporal information to update the hand prior memory, injecting this prior knowledge into the Dynamic Spatio-temporal Block to provide robust disambiguation cues for spatio-temporal interactions. Here, N represents the number of stacked blocks.

2.3. Dynamic Convolutions

Dynamic Convolutions adaptively generate convolution kernel parameters based on input data, so as to improve the representation ability of the network while maintaining the inference efficiency. Some works [5, 63] dynamically combined multiple learnable convolution kernels, and some works [1, 22, 43, 44, 55] directly generated convolution kernel parameters through a controller. For example, Nie *et al.* [43] generated convolution kernels for human pose estimation based on human parsing features, which are able to exploit body part cues to constrain the joint locations and pose structures. Unlike the above dynamic convolution methods, which can only generate kernels with fixed receptive fields, deformable convolution [11, 61, 73] generates convolutional kernels with flexible receptive fields through learnable offsets, which combine the local inductive bias of the convolution mechanism with the strong learning capability of the attention mechanism, enabling it to better capture complex spatial dependencies and dynamic features.

3. Method

MotionBERT [72] proposed a dual-branch spatio-temporal transformer that can efficiently perform spatio-temporal modeling and has shown strong performance in the task of 3D human pose estimation. Therefore, we adopt the dual-branch spatio-temporal transformer as the basic structure to implement the prior-aware dynamic temporal framework. On the one hand, we propose a temporal deformable convolution module to capture rapidly changing hand movements along the temporal dimension. This module dynamically generates convolutional weights and offsets based on the input information, allowing the model to effectively track sudden or subtle hand motion variations. On the other hand, we introduce a learnable token pool that acts as a hand prior memory to store critical hand information from the input sequence. The memory module is integrated into each spatio-temporal block, progressively learning sequence-specific hand shapes and providing shape priors for subsequent spatio-temporal interactions, thereby alleviating the depth ambiguities inherent in monocular hand pose estimation.

In this section, we first introduce the basic structure of MotionBERT, then introduce our proposed dynamic temporal convolution and hand prior memory respectively, and finally introduce the loss functions of our method.

3.1. Preliminary

The input 3D hand pose sequence $\mathbf{X} \in \mathbb{R}^{T\times J\times 3}$ is initially processed by a projector to obtain joint-level features $\mathbf{F} \in \mathbb{R}^{T\times J\times C}$. Specifically, the 3D hand joints sequence $\mathbf{X}$ is projected to a high-dimensional initial joint feature $\mathbf{F}^{init} \in \mathbb{R}^{T\times J\times C}$. Then, the learnable spatial positional encoding $\mathbf{P}^{\mathrm{s}}_{\mathrm{pos}} \in \mathbb{R}^{1\times J\times C}$ and temporal positional encoding $\mathbf{P}^{\mathrm{t}}_{\mathrm{pos}} \in \mathbb{R}^{T\times 1\times C}$ are added to it. Here, T denotes the sequence length, J denotes the number of hand joints, and C denotes the channel numbers of features. Then, MotionBERT uses a series of stacked spatio-temporal interaction blocks to extract high-level joint features.

The spatio-temporal interaction blocks have two core modules, namely the Spatial Multi-Head Self-Attention module (Spatial MHSA) and the Temporal Multi-Head Self-Attention module (Temporal MHSA). The spatial MHSA models the spatial relationship between joints at the same time step, and the temporal MHSA models the temporal relationship of each joint in parallel,

$$\mathrm{MHSA}(\mathbf{F}) = \mathrm{softmax}\left(\frac{(\mathbf{F}W_Q)(\mathbf{F}W_K)^T}{\sqrt{d_k}}\right)(\mathbf{F}W_V) \quad (1)$$

where W_Q,W_K,W_V are projection parameter matrices. De-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

pending on the type of self-attention module, a single feature $\mathbf{F}_i$ can be $\mathbb{R}^{J\times C}$ or $\mathbb{R}^{T\times C}$.

Considering that spatial MHSA and temporal MHSA model different aspects of information, namely intra-frame and inter-frame relationships, MotionBERT constructs two parallel branches in different orders. MotionBERT adaptively combines the results of two parallel branches through a weighted fusion mechanism. These blocks with the same structure and unshared weights are stacked multiple times to extract spatio-temporal features.

MotionBERT demonstrates strong spatio-temporal modeling capabilities of the human body, achieving competitive results across several skeleton-based body-centered tasks, including action recognition and 3D human pose estimation. However, MHSA lacks local inductive bias, resulting in limitations when modeling subtle, short-term motion changes. This drawback makes it challenging to capture the rapid and frequent variations of hand movements. Consequently, directly applying MotionBERT to sequential 3D hand pose estimation often leads to suboptimal results.

3.2. Dynamic Temporal Convolution Module

As shown in Fig. 2, to capture the rapid changes in hand movement, we propose a dynamic Temporal Convolution Network (TCN) module. Hands can convey human intentions and interact with the complex physical world, and hand pose has high degree of freedom. The global hand movement caused by the forearm and the movement of local fingers or joints may be independent of each other. For example, when waving left or right, the fingers can bend freely. In contrast, when the human body stands or squats, the movement state of the lower limbs is relatively stable. Therefore, it is important to model global and local movement state of the hand separately. We further decouple the dynamic temporal convolution module into a local joint-level dynamic temporal convolution module and a global hand-level dynamic temporal convolution module.

3.2.1. Joint-level Dynamic Temporal Convolution

Joint-level temporal module aims to capture the dynamic changes of each hand joint, as the movement of each joint may be independent of the others. Inspired by 2D deformable convolution [11, 61, 73], we first use the features of each joint to generate the weights and offsets of the 1D convolution, so that each joint at each moment has a different receptive field and response pattern. The dynamic temporal convolution can be formulated as follows:

$$\hat{\mathbf{x}}(p_0) = \sum_{g=1}^{G}\sum_{k=1}^{K} \mathbf{w}_g \mathbf{m}_{gk} \mathbf{x}_g(p_0 + p_k + \Delta p_{gk}), \quad (2)$$

where G denotes the total number of aggregation groups and K denotes the size of the kernel. For the g-th group, $\mathbf{w}_g \in \mathbb{R}^{C\times C'}$ denotes the location-irrelevant projection weights of the group, where $C' = C/G$ represents the group dimension. $\mathbf{m}_{gk} \in \mathbb{R}$ denotes the estimated modulation scalar of the k-th sampling point in the g-th group. p_k denotes the k-th location of the pre-defined grid sampling $\{..., (-2), (-1), (0), (+1), (+2), ...\}$ as in regular convolutions, Δp_{gk} is the estimated offset corresponding to the grid sampling location p_k in the g-th group. In particular, we constrain the estimated offset to be in the range $[-2, +2]$ to avoid overly large windows. Similar to [62], we do not normalize the modulation scalar using softmax, thereby enhancing the expressiveness of the convolution kernel and reducing the difficulty of network optimization.

Through the joint-level dynamic TCN, the model can accurately track the movement of each hand joint between frames, thereby capturing the subtle and short-term pose changes. Predicting kernel offset and modulation weights through joint information relies on high-quality self-motion context. Therefore, we add Joint-level Dynamic Temporal Convolution after Temporal MHSA to provide more reliable and robust joint information for offset and weight estimation. As shown in Fig. 2, the spatio-temporal interaction module in our method contains two parallel branches, each with a different spatio-temporal interaction order.

3.2.2. Hand-level Dynamic Temporal Convolution

In addition to joint-level dynamic TCN, we also introduce global dynamic TCN to capture the global hand motion generated by the forearm and wrist. Specifically, we perform spatial average pooling on all joints to generate an global feature representation of the hand at each time step. Based on the global hand features, hand-level dynamic temporal convolution predict modulation weights and offsets for 1D deformable convolutions to dynamically generate the enhanced hand motion feature. This global feature acts as a temporal anchor that complements the joint-level information. This approach enhances the temporal consistency of hand motion, which may not be directly obtained from joint-level temporal modeling.

In particular, the global hand motion also has fast-changing short-term movement and stable long-term movement. Therefore, we combine dynamic TCN and temporal attention mechanism to form two parallel branches, and use the weighted fusion mechanism as the spatio-temporal interaction module to fuse two branches feature. Furthermore, we directly add the global hand-level features to the joint-level features through residual connections. Through the fusion of joint-level and hand-level feature, the model is able to take into account both fine-grained joint information and global hand dynamics.

3.3. Hand Prior Memory

Hand poses from the same sequence typically maintain consistent hand shape and bone length over time, providing a

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

powerful disambiguation cue to refine inaccurate estimations. On the one hand, monocular RGB-based hand pose estimation suffers from depth and scale ambiguities. On the other hand, the self-occlusion and mutual occlusion problems in complex scenes will lead to unreasonable prediction of invisible hand regions. Given the prior hand shape information, these problems can be alleviated. Therefore, we use a memory mechanism to achieve long-term stable hand prior storage. The memory mechanism consists of two important parts: Memory Update and Memory Injection.

The memory update refers to updating the hand prior memory parameters using global hand information. Given a series of learnable memory tokens $\mathbf{M} \in \mathbb{R}^{M\times C}$, we use the temporal features output by the global temporal block to update the memory based on the multi-head cross-attention mechanism. Here, M denotes the number of memory. Specifically, we use the memory token as the query feature and the global temporal features as the key and value. The memory injection refers to the use of hand prior memory to assist spatio-temporal information interaction. The hand prior serves as a robust disambiguation cue in both the spatial and temporal self-attention modules in spatio-temporal block, helping to reduce network optimization complexity. Therefore, we incorporate memory tokens as additional tokens in both the temporal and spatial dimensions, denoted as $\mathbf{F}_i^s \in \mathbb{R}^{(J+M)\times C}$ and $\mathbf{F}_i^t \in \mathbb{R}^{(T+M)\times C}$. Then, we perform spatial MHSA and temporal MHSA separately for $\mathbf{F}_i^s$ and $\mathbf{F}_i^t$ in the time and space dimensions, thereby obtaining enhanced temporal and spatial features. Final, we fuse these two features using the weighted fusion mechanism. The hand prior memory mitigates the impact of low-quality single-frame estimations, enhancing the smoothness of predictions.

3.4. Pose Estimation and Supervision

Specifically, given the joint features $F^{out} \in \mathbb{R}^{T\times J\times C}$ extracted by the backbone, global features $F^g \in \mathbb{R}^{T\times C}$are obtained through joint-level average pooling. Then, a Fully Connected (FC) layer is used to regress the hand joint position $P \in \mathbb{R}^{T\times 3J}$ and the MANO pose parameters $\theta \in \mathbb{R}^{T\times 45}$, respectively. At the same time, given the memory tokens $\mathbf{M} \in \mathbb{R}^{M\times C}$, we use a FC layer to predict hand shape parameters $\beta \in \mathbb{R}^{10}$. Based on the frame-wise pose parameters θ and frame-shared shape parameters β, we can obtain the hand vertex coordinates $V \in \mathbb{R}^{T\times 3V}$ by hand model MANO [51]. Similar to previous methods [29, 50], we supervise the joints and mesh vertices with the smooth L1 loss [16, 47]. Meanwhile, we will use smooth L1 loss to supervise the pose parameters θ and shape parameters β. Furthermore, we adopt the acceleration loss [64] to enhance the smoothness of the estimated pose.

4. Experiment

4.1. Implementation Details

We train and evaluate our method on a single server with an NVIDIA 4090 GPU. The network is implemented within PyTorch. We train our network using the AdamW [36] optimizer with an initial learning rate of 1e-4 and a cosine decay learning rate schedule [35]. The whole training process takes 50 epochs with a batch size of 64. We perform data augmentation including random rotation, random scaling, random translation, and random horizontal flipping. Similar to MotionBERT, we adopt the backbone with depth $N = 5$, number of heads $h = 8$, feature size $C = 512$, kernel size $k = 8$. By default, we use sequence length $T = 27$. Please refer to the supplementary material for more details.

4.2. Datasets and Metrics

We conducted experiments on four public sequential hand datasets that contain rich hand pose and diverse interaction scenarios, including single-hand scenarios, hand-object interaction scenarios, and two-hand interaction scenarios.

Datasets. InterHand2.6M [41] provides multi-view RGB images with two-hand joint and mesh 3D annotation. It contains complex two-hand interaction poses and covers large-scale perspective changes. We use the 30 FPS version and use both the single hand data and interacting two-hand data. **Re:InterHand** [42] contains highly realistic and diverse 3D hand interaction images through relighting, closely resembling real-world appearances. With a multi-camera setup and an advanced 3D hand pose estimation network, the dataset includes precise 3D hand pose annotations. It contains approximately 739K relighted images, covering two-hand interaction data from 10 subjects. Since Re:InterHand also does not have an official train/test split, we use 520K data for training and the remaining 219K data for testing. **HanCo** [77] consists of 1,517 videos with multiple views and camera calibration. It has 860,304 frames in total, *i.e.* 107,538 time-step per view. All images are captured from the real world with 3D annotations. The dataset composited images with four types of real-world backgrounds and hands captured against a green screen. Since HanCo does not have an official train/test split, we use the first 1,200 sequences for training and the last 317 sequences for fair comparisons [71]. **DexYCB** [3] is a real hand-object dataset captured by multiple RGB-D cameras. It consists of 582,000 image frames with 10 different subjects and 20 YCB objects from 8 views. We followed previous methods and conducted experiments on the split s0.

Metrics. We report three metrics for hand pose estimation as follows. **MPJPE/MPVPE** (mean per joint/vertex position error) measures the average Euclidean distance in mm between the predicted and ground-truth joints/vertices. For a fair comparison, following previous work [18, 41, 50], we

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

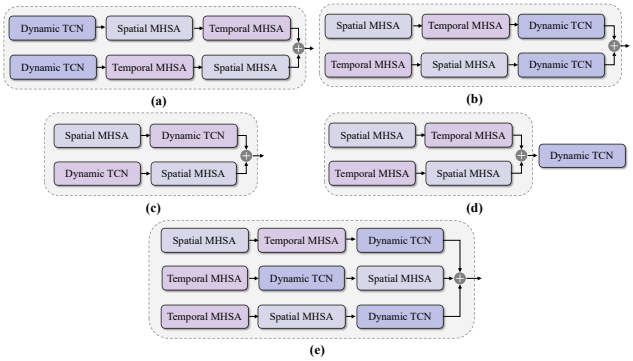


Figure 3. **Different Strategies for Dynamic Temporal Convolution**. (a) Inserting dynamic convolution into the head of each branch (b) Inserting dynamic temporal convolution into the tail of each branch (c) Directly replacing the temporal multi-head self-attention module with dynamic temporal convolution (d) Inserting dynamic temporal convolution into the back of the two-branch module (e) Using Tri-branch modules with different orders

Table 1. We report the MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) of different Insertion strategies of Joint-level dynamic temporal convolution on InterHand2.6M dataset.

Method	MPJPE	MPVPE	Accel
Dual-branch (Baseline)	8.13	8.55	5.92
Head Insertion (a)	8.03	8.29	5.58
Tail Insertion (b)	7.70	7.96	5.31
Replacement (c)	8.25	8.83	6.33
Fusion Insertion (d)	7.78	8.04	5.30
Tri-branch (e)	7.59	7.87	5.13
Dynamic Dual-branch	7.64	7.89	5.20

perform the root joint alignment and scale the estimation according to the ground-truth bone length. To measure temporal smoothness, we follow the related works [64] to adopt the acceleration error (**Accel**). The acceleration error computes an average of the difference between the predicted and ground-truth acceleration of each joint.

4.3. Ablation Study

In this section, we first explored the strategy of combining joint-level dynamic temporal convolution with spatio-temporal MHSA. Secondly, we verified the necessity of combining joint-level and hand-level dynamic temporal module. Finally, we explored the representation form and learning strategy of prior memory. We conducted ablation experiments on the InterHand2.6M dataset because this dataset is large and involves rich hand motions.

4.3.1. Joint-level Dynamic Convolution

Firstly, we examined the order of spatio-temporal attention modules and dynamic temporal convolution modules within the dual-branch architecture. As illustrated in Fig. 3a and 3b, we compared placing the dynamic temporal convolution module before and after the spatio-temporal attention module. As shown in Table 1, incorporating dynamic temporal convolution before the spatio-temporal attention module (Head Insertion) yielded limited improvement, enhancing the baseline by only 0.1 mm. However, adding dynamic temporal convolution after the spatio-temporal attention module (Tail Insertion) can significantly improve the accuracy of estimation, and the performance is close to the architecture we proposed. This indicates that generating high-quality offset and modulation weights relies heavily on contextual information. The long-range interaction provides powerful disambiguation cues for the receptive field and weight prediction of each joint. Therefore, in subsequent structural designs, we will ensure the presence of a temporal interaction module prior to the dynamic temporal convolution module whenever possible.

Next, as shown in Fig. 3, we investigated the impact of different global structures. We evaluated several configurations: directly replacing temporal attention in the dual-branch model with dynamic temporal convolution, applying spatio-temporal attention in parallel first and then applying dynamic temporal convolution, and adopting a tri-branch structure for enhanced diversity in feature fusion. As shown in Table 1, Replacement worse than the baseline, indicating the importance of long-range temporal information interaction. Fusion Insertion is worse than ours, which shows that it is important to maintain the heterogeneity of spatio-temporal interactions between different branches, which can have an effect similar to model ensemble. The Tri-branch is slightly better than our method, but it will bring an additional 30% of computation. Considering the balance between performance and efficiency, we will use the two-branch method by default.

4.3.2. Global Temporal Block

First, as shown in Table 2, adopting hand-level dynamic TCN or global temporal attention mechanism can further improve the performance of the network, which shows the importance of explicitly modeling the global hand movement. The dual-branch parallel global temporal block can achieve the best results. Second, we explore the characteristics of hand-level dynamic TCN. As shown in Fig. 4a, we directly use the hand-level dynamic TCN to replace the joint-level dynamic TCN in spatio-temporal block. As shown in Table 2, direct replacement leads to a decrease in performance, which shows that the joint-level temporal modeling module is critical for capturing the fine-grained hand movement, and simply relying on global information may lose some local motion details. Finally, as shown in Fig. 4b, we perform global temporal modeling and spatio-temporal information interaction in parallel. This parallel structure performs worse than the serial structure, which shows that hand-level temporal modeling also re-

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

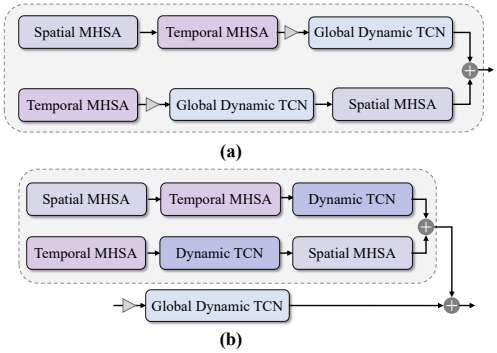


Figure 4. **Different Strategies for Global Dynamic Convolution**. (a) Replace the joint-level dynamic temporal convolution in the dual-branch block with the global temporal dynamic convolution. (b) Perform joint-level spatio-temporal interaction and global temporal interaction in parallel.

Table 2. We report the MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) of different strategies of global dynamic temporal convolution on InterHand2.6M dataset. GA represents the global temporal attention mechanism. DC represents the global dynamic temporal convolution mechanism. PS represents the parallel structure. SS represents the serialized structure.

GA	DC	PS	SS	MPJPE	MPVPE	Accel
				7.64	7.89	5.20
✓			✓	7.55	7.71	5.05
	✓		✓	7.52	7.78	5.01
✓	✓		✓	**7.42**	**7.63**	**4.93**
	✓			7.85	8.03	5.68
	✓	✓		7.48	7.72	5.03

quires high-quality joint representation.

4.3.3. Hand Prior Memory

We conducted a series of experiments on different configurations of the hand prior memory construction. First, we examined the impact of the number of memory tokens. We experimented with various quantities of memory tokens to evaluate their effect on network performance. As shown in Table 3, the appropriate number of tokens can effectively improve the performance of the model, while too many or too few tokens will lead to performance degradation. Then, we tried a simple way to update and insert the prior memory. Similar to ViT [12], we use a learnable token to store hand prior information, let it participate in all spatial attention interactions and temporal attention interactions, and predict the hand shape parameters based on the token. As shown in Table 3, this approach yielded minimal improvement in network performance. Finally, we also explored the supervision method of the prior memory, that is, block-by-block supervision and supervision only on the final updated memory. The experimental results indicate that using only the memory obtained from the final update to predict hand parameters yields better results.

Table 3. We report the MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) of different hand prior memory strategies on InterHand2.6M dataset. TN represents the memory token number, VS represents the ViT-style token interaction, and FS stands for applying supervision only on the final updated memory.

TN	VS	FS	MPJPE	MPVPE	Accel
1		✓	7.33	7.48	4.69
4		✓	7.27	7.44	4.62
8		✓	**7.21**	**7.39**	**4.54**
16		✓	7.28	7.49	4.58
1	✓		7.40	7.54	4.81
8			7.35	7.45	4.65

Table 4. Comparison with SOTA methods of MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) on InterHand2.6M.

Methods	MPJPE	MPVPE	Accel
InterWild	10.32	10.54	7.18
w/ Savitzky-Golay	10.10	10.29	6.21
w/ Gaussian1d	10.11	10.32	6.29
w/ SmoothNet	10.02	10.26	5.98
w/ MotionBERT	8.13	8.55	5.92
w/ Ours	**7.21**	**7.39**	**4.54**
DIR	10.45	10.78	8.33
w/ Ours	7.35	7.47	4.70

4.4. Comparisons with State-of-the-arts

To validate the superiority of our method, we compared it with current state-of-the-art temporal models and single-frame methods on four datasets: InterHand2.6M [41], Re:InterHand [42], HanCo [77], and DexYCB [3].

On the InterHand2.6M dataset, we compared our method with the two SOTA methods DIR [49] and InterWild [39]. Specifically, since the data reported in the papers for these two methods were trained on the 5 FPS version, we re-trained their models using the official code on the 30 FPS version. As shown in Table 4, for the SOTA two-hand reconstruction method, our approach reduces the MPJPE by 30.1% and accelerate error by 36.8%. Meanwhile, compared with SOTA human pose smoothing methods Savitzky-Golay [64], Gaussian1d [64] and SmoothNet [64] and refinement methods [72] , our method brings very significant improvements. In particular, we observed that SmoothNet [64] for human pose temporal smoothing brought very small improvements, which shows that smoothing alone is far from enough for hand motion refinement, which may ignore some small movements of fingers. Similarly, we re-trained DIR and InterWild on the Re:InterHand dataset. As shown in Table 5, our method

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

Table 5. Comparison with SOTA methods of the MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) on Re:InterHand.

Methods	MPJPE	MPVPE	Accel
DIR	7.32	7.57	6.88
DIR w/ Ours	5.05	5.38	4.02
InterWild	7.55	7.69	6.14
InterWild w/ Ours	**4.86**	**5.03**	**3.84**

Table 6. The comparison with SOTA methods of the MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) on HanCo.

Methods	MPJPE	MPVPE	Accel
EpipolarPose	10.5	-	-
MobRecon	9.9	-	-
HaMuCo	11.1	-	-
HAMER	10.1	10.0	7.4
HAMER w/ Ours	**7.3**	**7.4**	**5.6**

significantly improves both the accuracy and smoothness of the estimations. Specifically, for InterWild, our approach reduces MPJPE, MPVPE, and Accle by 35.6%, 33.5%, and 44.2%, respectively. The improvement in temporal smoothness is even more pronounced compared to the InterHand2.6M dataset, which may be attributed to the more accurate annotations in the Re:InterHand dataset.

On the HanCO dataset, we compared our method with single-frame methods including MobRecon [4], EpipolarPose [23], and HaMuCo [71]. We used a strong hand pose estimation model, HAMER [45], as the single-frame pose estimator to generate the initial 3D hand pose. As shown in Table 6, our approach improves hand pose estimation accuracy and temporal smoothness through spatio-temporal interaction. On the DexYCB dataset, we compared our method with temporal approaches including VIBE [24] , TCMR [9], S2HAND(V) [57], MeshGraphormer [34], and Deformer [14]. As shown in Table 7, our approach significantly outperformed existing temporal methods, demonstrating strong robustness in handling complex hand movements and interactions with objects. Specifically, because the training set of HAMER includes the data from HanCo and DexYCB, we utilized official code and conducted model training using a training set that excludes these data.

In summary, our proposed method exhibited superior performance across multiple datasets and different input data, especially in scenarios with high spatio-temporal variation. These results substantiate the effectiveness of the dynamic temporal convolution and hand prior memory in handling complex, dynamic hand movements. Furthermore, we present some qualitative results on multiple datasets in the supplementary materials and supplementary videos, which further illustrate the advantages of our method in temporal smoothness and the robustness of our method to low-quality monocular hand pose estimation.

Table 7. The comparison with SOTA methods of the MPJPE (mm), MPVPE (mm) and Accel ($mm/frame^2$) on DexYCB.

Methods	MPJPE	MPVPE	Accel
VIBE	16.95	-	36.4
TCMR	16.03	-	34.3
S2HAND(V)	19.67	-	41.6
Deformer	13.64	-	31.7
HAMER	10.18	9.90	12.97
HAMER w/ Ours	**8.42**	**8.43**	**7.51**

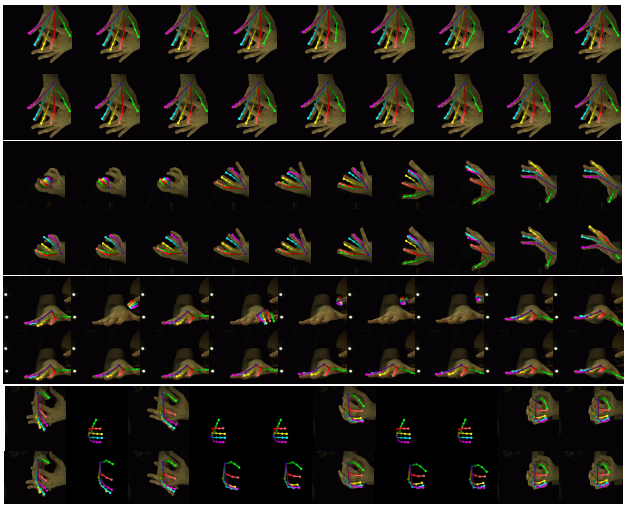


Figure 5. For each sequences, the top is the initial pose and the bottom is the refined pose (we do not use the input image).

4.5. Qualitative Analysis

We provide visualizations of four typical cases in Fig. 5, including joint errors, global hand shape errors, global pose errors caused by two-hand-interaction, and image loss. Our method can effectively refine these low-quality initial pose.

5. Conclusion

In this paper, we proposed a prior-aware dynamic temporal modeling framework for sequential 3D hand pose estimation. On one hand, we introduce a flexible memory mechanism to capture hand priors. Injecting hand prior information into the spatio-temporal interaction block provides strong disambiguation cues, reducing the interference of low-quality input poses and alleviating ambiguity issues. On the other hand, we propose a dynamic temporal convolution module, which can effectively capture subtle short-term pose variations and stable long-term motion trends. Experimental results on four public datasets demonstrate that our method outperforms existing state-of-the-art approaches in both accuracy and temporal smoothness.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

6. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants (62406039, 62321001, 62471055, U23B2001, 62171057, 62201072, 62071067), the High-Quality Development Project of the MIIT (2440STCZB2584), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), the Fundamental Research Funds for the Central Universities (2024PTB-004), the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grants (2023TQ0039, 2024M750257, GZC20230320).

References

[1] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in neural information processing systems*, pages 523–531, 2016. 3

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 1

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, pages 9044–9053, 2021. 2, 5, 7

[4] 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 *CVPR*, pages 20544–20554, 2022. 1, 8

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 3

[6] Yu Cheng, Bo Wang, Bo Yang, and Robby T. Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1157–1165, 2021. 2

[7] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, pages 342–359. Springer, 2022. 2

[8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787. Springer, 2020. 2

[9] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 8

[10] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2271, 2019. 2

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3, 4

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 7

[13] 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 *3DV*, pages 1–10. IEEE, 2021. 2

[14] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23600–23611, 2023. 8

[15] 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 *CVPR*, pages 10833–10842, 2019. 2

[16] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 5

[17] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 1, 2

[18] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, pages 11090–11100, 2022. 2, 5

[19] Shangchen Han, Po-Chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang. Umetrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022*, 2022. 2

[20] Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. Awr: Adaptive weighting regression for 3d hand pose estimation. In *AAAI*, pages 11061–11068, 2020. 1, 2

[21] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 2

[22] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in neural information processing systems*, pages 667–675, 2016. 3

[23] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 8

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 8

[25] Deying Kong, Linguang Zhang, Liangjian Chen, Haoyu Ma, Xiangyi Yan, Shanlin Sun, Xingwei Liu, Kun Han, and Xiaohui Xie. Identity-aware hand mesh estimation and personalization from rgb images. *arXiv preprint arXiv:2209.10840*, 2022. 2

[26] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 2

[27] Jihyun Lee, Junbong Jang, Donghwan Kim, Minhyuk Sung, and Tae-Kyun Kim. Fourierhandflow: Neural 4d hand representation using fourier query flow. *Advances in Neural Information Processing Systems*, 36:29239–29251, 2023. 2

[28] Han Li, Bowen Shi, Wenrui Dai, Hongwei Zheng, Botao Wang, Yu Sun, Min Guo, Chenglin Li, Junni Zou, and Hongkai Xiong. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1296–1304, 2023. 1, 2

[29] 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 *CVPR*, pages 2761–2770, 2022. 2, 5

[30] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293, 2022. 1, 2

[31] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 1, 2

[32] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 2

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 2

[34] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 8

[35] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[37] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2

[38] 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. In *ECCV*, 2022. 2

[39] Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17028–17037, 2023. 7

[40] Gyeongsik Moon, 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 *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 1, 2

[41] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, pages 548–564. Springer, 2020. 2, 5, 7

[42] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mize Mallorie, Philippe Bree, Tomas Simon, Bo Peng, Shubham Garg, Kevyn McPhail, and Takaaki Shiratori. A dataset of relighted 3D interacting hands. In *NeurIPS Track on Datasets and Benchmarks*, 2023. 2, 5, 7

[43] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2108, 2018. 3

[44] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6942–6950, 2019. 3

[45] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 8

[46] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 2

[47] Pengfei Ren, Haifeng Sun, Qi Qi, Jingyu Wang, and Weiting Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, page 112, 2019. 5

[48] Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. Mining multi-view information: A strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20555–20565, 2022. 2

[49] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8014–8025, 2023. 7

[50] Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5

[51] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. 5

[52] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, pages 432–441. IEEE, 2021. 2

[53] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *European Conference on Computer Vision*, pages 461–478. Springer, 2022. 1

[54] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98, 2018. 2

[55] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 3

[56] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33(5):169:1–169:10, 2014. 2

[57] Zhigang Tu, Zhisheng Huang, Yujin Chen, Di Kang, Linchao Bao, Bisheng Yang, and Junsong Yuan. Consistent 3d hand reconstruction in video via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8

[58] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2018. 1, 2

[59] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10853–10862, 2019. 1, 2

[60] Tao Wang, Lei Jin, Zheng Wang, Jianshu Li, Liang Li, Fang Zhao, Yu Cheng, Li Yuan, Li Zhou, Junliang Xing, et al. Synsp: Synergy of smoothness and precision in pose sequences refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1824–1833, 2024. 2

[61] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 3, 4

[62] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. *arXiv preprint arXiv:2401.06197*, 2024. 4

[63] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, pages 1307–1318, 2019. 3

[64] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022. 2, 5, 6, 7

[65] 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 *CVPR*, pages 11354–11363, 2021. 2

[66] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022. 1, 2

[67] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 1

[68] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023. 1, 2

[69] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20438–20447, 2022. 2

[70] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11656–11665, 2021. 1, 2

[71] Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, and Jingyu Wang. Hamuco: Hand pose estimation via multi-view collaborative self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5, 8

[72] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 7

[73] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[74] Yiran Zhu, Xing Xu, Fumin Shen, Yanli Ji, Lianli Gao, and Heng Tao Shen. Posegtac: Graph transformer encoder-decoder with atrous convolution for 3d human pose estimation. In *IJCAI*, pages 1359–1365, 2021. 1, 2

[75] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. 2

[76] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.

for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 1

[77] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*, pages 250–264. Springer, 2021. 2, 5, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 14:43:09 UTC from IEEE Xplore. Restrictions apply.