Virtual Reality (2024) 28:143
https://doi.org/10.1007/s10055-024-01039-3

ORIGINAL ARTICLE

Temporally enhanced graph convolutional network for hand tracking from an egocentric camera

Woojin Cho[1] · Taewook Ha[1] · Ikbeom Jeon[1] · Jinwoo Jeon[1] · Tae-Kyun Kim[3,4] · Woontack Woo[1,2]

Received: 17 March 2024 / Accepted: 11 July 2024

Abstract

We propose a robust 3D hand tracking system in various hand action environments, including hand-object interaction, which utilizes a single color image and a previous pose prediction as input. We observe that existing methods deterministically exploit temporal information in motion space, failing to address realistic diverse hand motions. Also, prior methods paid less attention to efficiency as well as robust performance, i.e., the balance issues between time and accuracy. The Temporally Enhanced Graph Convolutional Network (TE-GCN) utilizes a 2-stage framework to encode temporal information adaptively. The system establishes balance by adopting an adaptive GCN, which effectively learns the spatial dependency between hand mesh vertices. Furthermore, the system leverages the previous prediction by estimating the relevance across image features through the attention mechanism. The proposed method achieves state-of-the-art balanced performance on challenging benchmarks and demonstrates robust results on various hand motions in real scenes. Moreover, the hand tracking system is integrated into a recent HMD with an off-loading framework, achieving a real-time framerate while maintaining high performance. Our study improves the usability of a high-performance hand-tracking method, which can be generalized to other algorithms and contributes to the usage of HMD in everyday life. Our code with the HMD project will be available at https://github.com/UVR-WJCHO/TEGCN_on_Hololens2.

Keywords Augmented reality · Computer vision · Deep learning · Tracking · Head mounted displays

1 Introduction

With the rise of Augmented/Virtual Reality(AR/VR) and the commercialization of various Head Mounted Displays(HMD), a volume of research for visually understanding human behaviors, especially hands, is being conducted to provide a better experience to users (Han et al. 2020). It has been reported that a certain level of accuracy and efficiency has been reached for a single hand (Lepetit 2020). However, the problem of understanding human hands in various situations is still an active topic, including hand-object interactions. Recently, studies in the field have been focusing on utilizing RGB input due to the accessibility of RGB cameras and goes beyond simply estimating 3D joint information; they reconstruct a dense hand mesh for its usability in applications (Chen et al. 2022b, 2023; Hasson et al. 2019b, 2020, 2021; Kulon et al. 2020; Lin et al. 2021a, 2023; Ren et al. 2023; Yu et al. 2023; Zuo et al. 2023). However, while the studies focus on accuracy in various situations, they often fail to guarantee real-time

✉ Woontack Woo
wwoo@kaist.ac.kr

Woojin Cho
woojin.cho@kaist.ac.kr

Taewook Ha
hatw95@kaist.ac.kr

Ikbeom Jeon
ikbeomjeon@kaist.ac.kr

Jinwoo Jeon
zkrkwlek@kaist.ac.kr

Tae-Kyun Kim
kimtaekyun@kaist.ac.kr

[1] KAIST UVR Lab, 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[2] KAIST KI-ITC Augmented Reality Research Center, 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[3] KAIST CVL Lab, 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[4] Imperial College London, Exhibition Rd, South Kensington, London SW7 2AZ, UK

Published online: 01 August 2024

performance and temporal coherence, which is crucial for real-world applications.

Few studies (Chen et al. 2022a; Moon and Lee 2020; Park et al. 2020a, b; Tang et al. 2021; Xu et al. 2023; Zheng et al. 2021) have considered limited computational resources and attempted to achieve state-of-the-art performance while ensuring real-time conditions. They proposed a method using a light-weight network structure (Moon and Lee 2020; Xu et al. 2023), adaptive Graph Convolution Network(GCN) (Zheng et al. 2021), additional sensors (Park et al. 2020a, b), or a mobile-friendly pipeline that does not require GPU setup (Chen et al. 2022a). For temporal coherence, recent studies investigate temporal cues from past and future information based on sequential models (Cai et al. 2019; Chen et al. 2021a; Fu et al. 2023; Han et al. 2020; Kocabas et al. 2020; Ye et al. 2023) or adopt global features in a single view as a non-sequential model (Chen et al. 2022a). Overall, the existing methods utilize temporal information mainly as a constraint to penalize the current prediction or estimate the pose candidate by extracting motion information such as optical flow. However, we have observed that these methods utilize temporal information rather deterministically, relying on a constant motion model. Thus, we intend to explore a method that adaptively utilizes temporal information to cover realistic hand motion aspects (Fig. 1).

Fig. 1 Visualization of reconstructed hand mesh with our method from Hololens2 RGB input. The hand mesh is rendered in the off-loaded server

Moreover, several remaining issues must be addressed to apply the findings of hand-tracking research to AR devices. To ensure reliable performance in various situations, it is essential to have access to a significant amount of computational resources, such as a GPU. Also, from an implementation standpoint, there is a potential problem: individual hardware characteristics differ between AR devices, such as the Hologram Processing Unit(HPU) of the Hololens2 and the Qualcomm Snapdragon XR2 processor of the Oculus Quest 2 developed by Meta. Therefore, we determined that a system incorporating an off-loading framework is necessary to leverage standard GPU resources, irrespective of the hardware specifications of different AR devices.

Through these inspirations, we aim to develop a fast hand-tracking system assisted by adaptive temporal cues and integrate the proposed system on AR HMD through the off-loading framework to verify the effectiveness of the system in realistic scenarios where hands interact with objects (Fig. 2). To achieve our goal, we focus on three main points for the hand pose estimation system: First, we leverage GCN which is capable of establishing relationships between the vertices of a hand mesh. Previous studies have demonstrated the significance of GCN-based systems in accuracy and efficiency by adeptly retaining the structure of the hand mesh while capturing attention between vertices. Tang et al. (2021) utilized GCN to refine a rough mesh by incorporating local and global features, while Zheng et al. (2021) proposed a system based on adaptive-GCN that enables spatial-aware regression, which has been adopted for our system. Second, adaptive temporal information utilization is suitable for real-world behavior aspects. As mentioned,

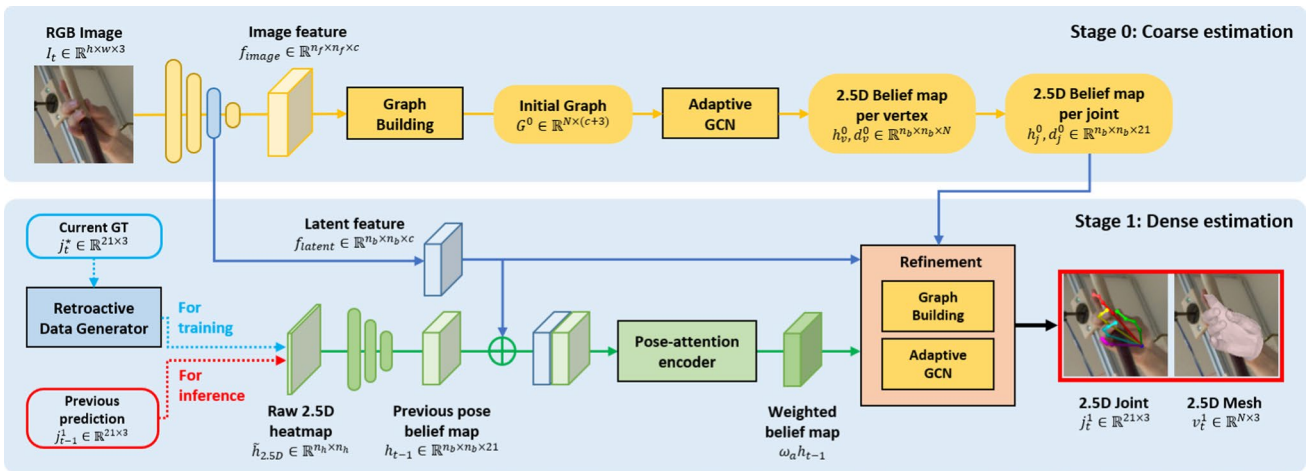

Fig. 2 Schematic of proposed system

most previous works utilizing temporal information assume a constant position or constant velocity model to generate stable predictions. However, there are cases where a specific motion model cannot be applied, such as fast-moving or hand-shifting moments or HMD users' head movement that significantly changes the hand position with respect to the camera view. To address this issue, we propose the pose-attention encoder, which determines the proportion of previous prediction information to be utilized. The module generates a weighted feature by estimating the relevance of hidden states, such as the current image feature and previous pose feature, which is motivated by an attention mechanism. Third, the trainable dataset category is expanded by introducing a retroactive data generator(RDG). The availability of sufficient high-quality datasets is a critical factor affecting the performance of learning-based methods. However, prior approaches incorporating temporal information relied on sequential datasets, typically comprised of real-world video sequences, leading to a more constrained hand pose space than synthetic and non-sequential datasets. Hence, we introduced an RDG that generates a distributed augmented pose, effectively simulating the previous pose by constraining the extent of the common augmentation procedure and categorizing the data type. The approach enables the training of temporal information regardless of the sequentiality of the dataset (Fig. 3).

We train and evaluate our hand tracking system on the FreiHAND, DexYCB dataset and demonstrate our complete system on Hololens2 with the off-loading framework. Our system verifies state-of-the-art performance and reliable results on HMD. The integrated application on HMD will be made publicly available.

Our main contributions are summarized as follows:

1. Propose a novel approach for adaptive temporal information utilization suitable for realistic hand behavior.
2. Real-time 3D hand reconstruction system capable of learning temporal information regardless of the specific characteristics of the dataset.
3. Implement a scalable framework that leverages a distributed offloading system on an HMD.

2 Related work

Our target domain is the reconstruction of 3D hand poses and meshes using monocular RGB images captured from an egocentric viewpoint, including sequences that involve interactions with objects. In this section, we will review recent studies on 3D hand pose and mesh estimation from RGB input. Most of these studies are not limited to hand-only sequences and verified with public hand-object datasets. We will then discuss recent advances in GCN directly relevant to our research and summarize how temporal information is utilized in the pose estimation problem.

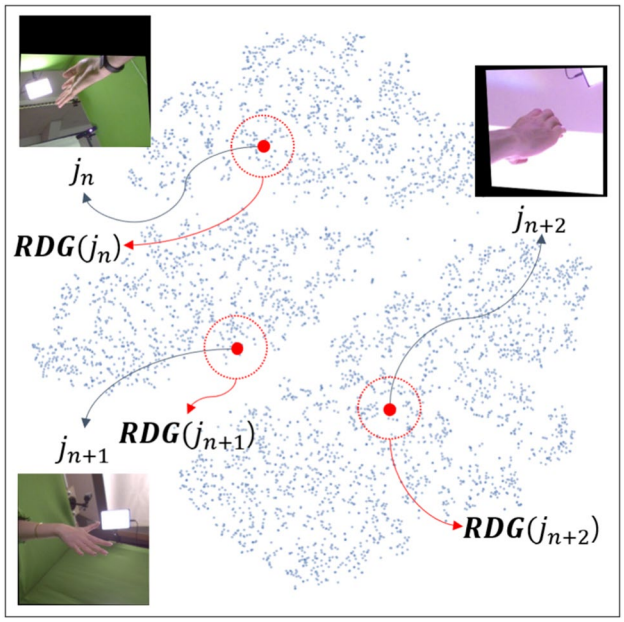

Fig. 3 2D t-SNE visualization of 1000 FreiHAND samples. Each frame has a distinct hand pose, as the dataset is non-sequential. We generate a synthesized previous pose from the ground truth of the current pose within the adjusted distribution of pose space from RDG

2.1 3D hand pose and mesh estimation from RGB

The current state-of-the-art hand pose and mesh estimation is dominated by deep learning-based approaches, which can roughly divide into generative and discriminative approaches.

Generative approaches regress the pose and shape coefficients of the parametric hand model, typically MANO (Romero et al. 2017a), as a differentiable layer in the network. Recent works (Cao et al. 2021; Chen et al. 2022b; Hasson et al. 2019b, 2020; Wang et al. 2020a) propose the work with an autoencoder (Kingma and Welling 2013), which combines an image feature encoder and a model parameter decoder. Additional supervision is often applied using the feature extracted in the intermediate step, such as segmentation map, projected 2D keypoints, etc (Baek et al. 2019; Boukhayma et al. 2019; Chen et al. 2021c; Lin et al. 2023; Zhang et al. 2019b, 2021; Zhou et al. 2020). Among these works, Baek et al. (2020) introduce an end-to-end trainable system utilizing various data sources from hand-only and hand-object domains through a domain adaption using Generative Adversarial Network. In studies that presented various benchmark datasets, regressing the coefficients of this parametric hand model is used as the baseline method (Hampali et al. 2020; Zimmermann et al. 2019). Few

works proposed a semi/self-supervised framework based on the autoencoder (Liu et al. 2021; Tu et al. 2022) and neural rendering-based optimization system (Qu et al. 2023). As the prior information on hand shape is embedded in a parametric model, it is relatively less dependent on training data, and plausible hand pose can be expected in various environments. However, since hand pose and shape are generated with a few parameters, the expressable pose space is limited, and fitting the coefficients is challenging due to the nonlinearity of the model parameters. To overcome the limitation, we adopt a discriminative approach that regresses the coordinates of the mesh or joint without relying on a fixed hand model.

The discriminative approach can be classified based on whether the objective is a joint or a mesh vertex and whether the pose coordinates are directly regressed or regressed as a heatmap. In the case of directly regressing the hand joint, various studies based on autoencoders have been introduced to effectively learn latent space from RGB images (Spurr et al. 2018; Yang and Yao 2019; Yang et al. 2019). Due to the rich information and usability of mesh, regression for each vertex of the hand is also studied, and most of these approaches adopt a GCN or a transformer (Chen et al. 2021b; Ge et al. 2019; Kulon et al. 2020; Lin et al. 2021b). Furthermore, non-direct regression approaches to target coordinates have also been proposed. Zimmermann et al. (Zimmermann and Brox 2017) proposed estimating a 2D heatmap for each joint using a Convolutional Pose Machine (Wei et al. 2016) and lifting it to a 3D pose. Iqbal et al. (2018) and Mueller et al. (2018) have demonstrated that the regression approach of the 2.5D representation, which consists of a 2D heatmap and a relative depth map of the hand, is more effective than direct-coordinate regression. Subsequent studies adopt the framework and indicate promising results (Fan et al. 2021; Spurr et al. 2020). Moon et al. (2020) introduced a method for estimating 1D heatmaps per axis for each human body mesh vertex, including the hand. As such, estimating a 2.5D dense heatmap for each mesh vertex has attracted attention to perform effective mesh reconstruction (Moon et al. 2020; Yang et al. 2021; Zheng et al. 2021). Recent studies have also presented promising results in 3D shape reconstruction using Signed Distance functions(SDFs) (Chen et al. 2022b, 2023; Ye et al. 2023). Since the discriminative method generally adopts a single-frame prediction pipeline, a jittery prediction issue occurs. Also, as the method is highly dependent on the training dataset, there is a risk of overfitting and encountering a generalization problem.

Most of the studies mentioned above do not satisfy the real-time condition as they primarily focus on improving the accuracy of the proposed method. Recently, the balancing issue between accuracy and efficiency has been raised for the practical application of hand pose estimation. To achieve real-time performance, several works have rigorously designed lightweight networks (Lim et al. 2020; Kulon et al. 2020; Moon and Lee 2020; Xu et al. 2023; Zheng et al. 2021; Zhou et al. 2020). MobRecon (Chen et al. 2022a) proposed a system that requires minimal computing resources with mobile-friendly lightweight stacked structures and a novel feature lifting module. Tang et al. (2021) developed an efficient multi-stage framework and mesh refinement using GCN, which satisfies both high accuracy and real-time conditions. Some approaches focus on formulating an efficient loss function for joint optimization or integrating specific constraints to achieve the goal. Zhang et al. (2019a) proposed a unified framework combining LSTM with hand and object joint optimization processes and constructing an efficient system by carefully designing mesh-related losses. Kulon et al. (2020) applied a decoder based on spatial mesh convolutions and utilized a simple loss function for mesh reconstruction. H2ONet (Xu et al. 2023) achieved high performance by decoupling the reconstruction pipeline into lightweight structures but adopted scenario-specific assumptions.

We focus on the 2.5D heatmap regression approach, targeting hand vertices, and construct a pipeline with a module capable of adaptively utilizing temporal information. Furthermore, we achieved high mesh reconstruction accuracy by designing a GCN-based efficient network while maintaining real-time performance. Our method has been successfully utilized in real application scenarios with users wearing an AR HMD.

2.2 Graph-convolution network based pose estimation

Due to the ability to reflect the structural characteristics of the hand or body in the form of a graph, the approach based on the graph convolution network has been steadily gaining attention in the pose estimation problem. GCN can be classified into spectral domain (Bruna et al. 2013; Defferrard et al. 2016; Kipf and Welling 2016) that performs convolutional operation with Fourier transformation, and a spatial domain (Gilmer et al. 2017; Monti et al. 2017; Xu et al. 2018) that outperforms in the field of pose estimation by expanding the spatial definition of a convolution.

In the spectral domain, some works utilize GCN to reconstruct a hand or body mesh in a coarse-to-fine way (Choi et al. 2020; Ge et al. 2019). Further, the coarse-to-fine scheme has extended in a spatial domain based on a simple encoder-decoder architecture (Chen et al. 2021b; Kulon et al. 2020). Lin et al. (2021a) improved the regression accuracy by modeling global vertex-vertex interaction using a transformer (Vaswani et al. 2017). Tse et al. (2022) proposed a hand reconstruction system based on attention-guided graph convolution, which can capture dynamic mesh information. In the two-hand reconstruction domain, Li et al. (2022)

proposed a pyramid image feature attention module to capture local and global patch attention simultaneously. Zheng et al. (2021) pointed out that the initial feature graph-building process used in previous studies caused a loss of spatial information and introduced a framework including a spatial-aware graph-building method. Our study adopted this framework due to the simplified structure and efficiency of the lightweight regression module.

2.3 Temporal coherence on pose estimation

Recent works attempt to satisfy hand pose coherence over time by utilizing temporal information since the prediction of the previous frame includes a dense queue for the current frame's pose. To propagate temporal information across frames, one approach suggested in a study (Hossain and Little 2018) involves employing a sequence-to-sequence model, while another study (Cai et al. 2019) introduces the use of a spatial-temporal graph. Video input is also utilized to exploit temporal features (Cai et al. 2019; Chen et al. 2021a; Fu et al. 2023; Kocabas et al. 2020; Ye et al. 2023; Zhao et al. 2021), but in this case, a frame-to-frame real-time operation is not feasible due to the heavy structural system or the necessity of future information. Han et al. (2020; 2022) demonstrate an integrated hand-tracking system within the Oculus Quest VR headset, utilizing the available inputs that depend on the hardware and proposed a regression network using tracking history. Chen et al. (2022a) design a feature lifting module utilizing a global receptive field for temporal coherence in a single-view method that does not rely on sequential information. Yang et al. (2020) introduce a technique for synthetically generating extensive sequential datasets to utilize the temporal motion information of the hand. In contrast, our approach presents acquiring temporal information from pre-existing non-sequential datasets.

Our critical insight is that adaptive utilization of temporal information enables robust pose prediction possible via imitating realistic hand motion aspects. In contrast, previous studies targeted temporal coherence, an effective indicator when assuming a constant position/velocity model. It provides significant stability of the predictions for datasets in a controlled environment but does not cover unintended hand motions from a real user. Therefore, we propose a pose-attention encoder to address this limitation by estimating the feature-level relevancy of the current image and previous pose.

3 Method

Our goal is to propose a hand tracking system in HMD by estimating a set of hand pose joints $j_t \in \mathbb{R}^{21\times3}$ and mesh vertices $v_t \in \mathbb{R}^{N\times3}$ in 3D space with N vertices for current frame t, given an input RGB image $I_t \in \mathbb{R}^{h\times w\times3}$ and joint pose prediction from the previous frame j_{t-1}. To fully utilize the given information, we designed a two-stage structured approach. In the first stage, we extract a feature map from the input image and estimate coarse pose through adaptive-GCN based module. In the second stage, we refine the prediction by using the regressed coarse pose and the adaptive temporal pose feature induced from the latent feature map in the intermediate stage and the information of the previous frame. In the following, we will describe the details of each stage.

3.1 Coarse pose estimation

3.1.1 Initial graph building

The first coarse estimation step is based on the one proposed by Zheng et al. (2021). The input image I passes convolution blocks to extract the image feature map $f_{image} \in \mathbb{R}^{n_f\times n_f\times c}$ with a feature size n_f and dimension of c. During the initial feature extraction process, the latent feature map $f_{latent} \in \mathbb{R}^{n_b\times n_b\times21}$ with a feature size of n_b is generated and fed into the pose-attention encoder for the dense estimation step. The initial feature graph $G^0 \in \mathbb{R}^{N\times(c+3)}$ is constructed by initial graph building module SAIGB from (Zheng et al. 2021), while 0 indicates the stage 0. It uniformly distributed each portion of the feature map to every vertex of the feature graph and concatenated template coordinates of each vertex from a parametric hand model MANO (Romero et al. 2017b). It has been shown that the approach significantly improves network performance by effectively transferring the extracted spatial information.

3.1.2 Adaptive GCN-based coarse regression

In the coarse pose estimation step, temporal information is not utilized to leverage the discriminative method's strength, which is highly effective in single-shot detection. Therefore, the first pose regression is based solely on the image features extracted from the initial graph building. The graph convolution operation to regress the vertex interactions can be represented as below refer to (Doosti et al. 2020):

$$\hat{G}^i = \sigma(\tilde{A}G^iW) \tag{1}$$

where σ is the activation function, $\tilde{A} \in \mathbb{R}^{n\times n}$ is the row-normalized adjacency matrix per graph with n nodes, and W is the trainable weights matrix, i indicates the stage of the system. As shown in (Zheng et al. 2021), a trainable adjacency matrix $\tilde{A}$ that is initialized with the identity matrix effectively constructs the vertex interactions and allows the capture of flexible range dependencies. We stack two layers of adaptive-GCN with a LeakyReLu and Dropout considering the balance of performance and computation speed. The

ablation study in Sect. 4.6 proves that the designed module shows the best performance improvement without compromising real-time operating conditions.

3.1.3 2.5D pose representation

Previous studies (Iqbal et al. 2018; Moon and Lee 2020; Zheng et al. 2021) on various pose regression problems indicate that predicting the target in the form of a belief map, rather than directly estimating the target coordinate, improves the accuracy of the estimation. Following the (Iqbal et al. 2018), we apply spatial softmax normalization and Hadamard product to the latent feature map from the last layer of GCN to generate a 2.5D belief map $h_v^0, d_v^0 \in \mathbb{R}^{n_b \times n_b \times N}$, where h_v indicates a 2D belief map and d_v is a relative depth map, each with the belief map size n_b. For dense estimation, a 2.5D belief map $h_j^0, d_j^0 \in \mathbb{R}^{n_b \times n_b \times 21}$ is extracted through fully connected layers, which includes compressed information on hand pose.

3.2 Dense pose estimation

In contrast to Zheng et al. (2021), the baseline for the coarse estimation step, our network is engineered to adaptively leverage temporal information to handle various hand movement scenarios. Specifically, in cases where the current hand pose is similar to the previous pose and the hand moves gradually, the network intensively utilizes information from the previous prediction. On the other hand, in cases where there is a significant difference in poses between frames, such as when the hand is moving quickly or undergoing rapid changes, the network prioritizes information from the current image instead of the previous prediction.

For this purpose, we introduced the pose-attention encoder(PAE), which takes the previous frame's pose heatmap $\tilde{h}_{2.5D}$ and the input image's latent feature map f_{latent} to estimate the pose-attention weight ω_a. This weight is designed to indicate how much the network should focus on prior pose information. The PAE allows the network to selectively utilize the previous pose information while learning discriminative features between the image and pose in the embedded space. The weighted previous pose heatmap and 2.5D belief maps of joints extracted from coarse prediction are then fed into the distinct GCN module to regress the final dense pose/mesh j^1, v^1, where 1 indicates stage 1. To fully utilize given resources, we develop a Retroactive Data Generator(RDG) applied only for the training process, which produces data that can be interpreted as a ground-truth hand pose of the previous frame in any dataset. The module allows us to generate valuable temporal data that resembles a real-world environment. Thus, the RDG directly contributes to improving the generalization performance of the network. The following sections will provide a detailed description of each module in the order of their operation within the system.

3.2.1 Retroactive data generator

Based on the insight that the previous pose does not necessarily have to match the true previous pose precisely, we developed a limited range of augmented poses, termed *retroactive data*, treated as the ground-truth for the previous pose. To fulfill the intended purpose of generated data, we carefully design the existing augmentation process; instead of a random pose generator role, we propose a module capable of representing the distribution of previous poses observed in real-world scenarios. However, directly feeding only the retroactive data as the previous pose to the model was found to lead to learning a dependency on the previous pose. To mitigate this, we set RDG to results poses with a value of 0, thus enhancing the realism of the previous pose dataset. By including a zero-value pose that significantly deviates from the actual previous pose, the subsequent PAE phase avoids simply depending on the previous pose. Instead, it recognizes the possibility of incorrect values and selectively utilizes the provided RDG data based on the assessed importance of the previous pose data. Thus, RDG can induce a robust prediction result even in the absence of a previous pose, which contributes to re-initialization due to intermittent tracking failure. Furthermore, a notable advantage of synthesizing the previous pose is the ability to assess the quantitive similarity between the previous and current poses. Accordingly, we derived the attention weight from the RDG and incorporated it into the subsequent stages as a metric indicating the degree of attention that should be allocated to the previous pose.

We categorized the generated pose data into three types: static, perturbed, and zero-value. The static type refers to generating previous pose data identical to the ground-truth current pose, while the perturbed type involves adding noise to the current pose. We conducted an ablation study to evaluate the impact of each data type and the performance based on the distribution ratio of each type. Regarding the perturbed type, we apply normal-distributed noise independently to the entire hand and each joint. This combination of global and local translation leads to realistic data augmentation. We generate a raw 2.5D pose heatmap $\tilde{h}_{2.5D} \in \mathbb{R}^{n_h \times n_h}$ from the augmented 2.5D joint pose by assigning a relative depth value to the 2D pixel location of each joint with heatmap size n_h. Also, a 2D Gaussian kernel is applied for a plausible heatmap representation. With the sampled type and generated noise, we compute the ground-truth attention weight $\omega^\star$ as follows:

$$\omega^{\star} = \cos\left(\frac{\pi}{2} \cdot \frac{\omega}{c}\right) \tag{2}$$

ω denotes latent weight in range $\omega \in (0, c)$ due to the pose type, c is a constant value manually set to 10. The equation yields $\omega^{\star}$, which equals 1 when the previous pose is identical to the ground-truth of the current frame, approaches 0 as the magnitude of the applied noise grows, and becomes 0 in the case of a zero-value pose. In the inference, a dense joint pose from the previous frame j^1_{t-1} has transferred instead of generated pose from RDG, and only the raw 2.5D heatmap generation process is performed. The organization and distribution of RDG data will be extensively discussed in Sect. 4.6. Notably, we do not distinguish between sequential and non-sequential datasets provided in the training process. In all cases, the RDG receives the ground-truth pose of the current frame $j^{\star}_t \in \mathbb{R}^{21\times 3}$ as input and generates a raw 2.5D heatmap $\tilde{h}_{2.5D}$ and ground-truth attention weight $\omega^{\star}$ as output.

3.2.2 Pose-attention encoder

The main purpose of the PAE is to estimate the relevance of the previous frame's information compared to the current frame's image by extracting the attention weight ω_a. First, in order to generate a more realistic pose heatmap from the raw 2.5D heatmap $\tilde{h}_{2.5D}$, a previous pose belief map $h_{t-1} \in \mathbb{R}^{n_b \times n_b \times 21}$ is generated using a small convolutional block that has the same resolution as the latent image feature f_{latent}. Then, the attention weight ω_a is computed and multiplied to the previous belief map h_{t-1}:

$$\omega_a h_{t-1} = \mathrm{PAE}(Concat(h_{t-1}, f_{latent})) \tag{3}$$

PAE performs a comparable function to previously suggested attention-based modules as it updates the corresponding hand pose feature with an attention-based input image feature. However, unlike existing systems that had to derive the correct hand pose directly from the image input, we require a pose feature only for refining the coarse pose, which plays a more indirect role. In general, existing attention-based modules such as a transformer require a substantial number of parameters and expensive computations, so we designed PAE rather explicitly. Using a configuration that includes a depthwise separable convolution layer, a LeakyReLU activation layer, and a max pooling layer, we extract a single attention weight through a lightweight network. This approach takes into account both spatial and channel-wise features. We also tested with a network architecture based on transformers, drawing inspiration from IntagHand (Li et al. 2022) and FastMETRO (Cho et al. 2022), which introduced GCN-based attention modules. Nevertheless, even though it yielded slightly lower performance, we observed a substantially reduced inference speed(84.3 FPS for same setup in Table 1). Consequently, we opted for the explicit encoder structure that we currently employ.

3.2.3 Refinement module

As indicated in Zheng et al. (2021), utilizing all the mesh vertex information estimated in the coarse pose estimation step is inefficient. Thus we use only per joint 2.5D belief maps h^0_j, d^0_j for dense estimation. In the same way as coarse estimation, the SAIGB module is used, but in this case, the initial graph G^1 is constructed by incorporating not only image features but also the output of coarse pose estimation and weighted belief map extracted from the previous prediction.

$$G^1 = \mathrm{SAIGB}(Concat(h^0_j, d^0_j, f_{latent}, \omega_a h_{t-1})) \tag{4}$$

The final graph $\hat{G}^1$ estimated with Eq. 1, and 2.5D belief maps per vertex h^1_v, d^1_v are extracted in the same way. Finally, the 2D location of the kth vertex of mesh v is computed with a weighted average of the 2D belief map h^1_v, and the corresponding depth is calculated as a summation of the relative depth map d^1_v. We denote Ω as the set of all pixel locations in the input image and each pixel location as p, then the 2D location of kth vertex u_k, v_k and relative depth d_k is computed as below:

$$(u_k, v_k) = \sum_{p\in\Omega} h_v(p) \cdot p \tag{5}$$

$$d_k = \sum_{p\in\Omega} d_v(p) \tag{6}$$

The training process of our system is conducted end-to-end. The coarse hand pose is estimated using only the image features of the current frame, and the previous frame's prediction is adaptively utilized to refine the pose densely. This structure enables the production of temporally consistent poses as well as robust single-shot estimation. This system improves the user experience by providing low-jitter 3D hand poses and operates effectively in a real environment with complex hand motion aspects.

3.3 Loss

We define the loss function in two configurations. *Hand loss* to minimize the discrepancy between the predicted hand mesh and ground-truth data, and *Attention loss* to induce the network to reflect the properties of real sequential data as we intended.

Hand Loss We apply the L1 norm between prediction and ground-truth of hand vertex and joints.

$$\mathscr{L}_{vert} = \sum_{n=1}^{N} \left\| v_n - v_n^{\star} \right\|_1 \tag{7}$$

$$\mathscr{L}_{joint} = \sum_{n=1}^{N} \left\| j_n - j_n^{\star} \right\|_1 \tag{8}$$

where v_n and j_n indicates a n th vertex and joint of the mesh, respectively; $\star$ denotes the ground-truth.

We also adopt the mesh smoothness term following (Chen et al. 2022a; Wang et al. 2018; Zheng et al. 2021), as it demonstrated promising results for mesh quality. The loss function for normal vector and edge length is defined as follows:

$$\mathscr{L}_{normal} = \sum_{c \in C} \sum_{(i,j) \in c} \left\| \frac{v_i - v_j}{\left\| v_i - v_j \right\|_2} \cdot \mathbf{n}_c^{\star} \right\| \tag{9}$$

$$\mathscr{L}_{edge} = \sum_{c \in C} \sum_{(i,j) \in c} \left\| \left\| v_i - v_j \right\|_2 - \left\| v_i^{\star} - v_j^{\star} \right\|_2 \right\| \tag{10}$$

where C is face sets of mesh and $\mathbf{n}_c$ indicates a unit normal vector of face c.

We iteratively apply the hand loss for each stage, so the total loss for hand is defined as:

$$\mathscr{L}_{hand} = \sum_{t=1}^{T} \mathscr{L}_{vert} + \mathscr{L}_{joint} + \lambda_1 * \mathscr{L}_{normal} + \mathscr{L}_{edge} \tag{11}$$

where $\mathscr{L}$ represents the loss in stage t and λ_1 is set to 0.1 for scaling the normal loss term.

Attention Loss To transfer the previous pose information to the refinement module, we use a weighted feature map $\omega_a h_{t-1}$, which has the same resolution as the 2D belief map per joint h_j^0, the output of coarse pose estimation. However, since we generate $\tilde{h}_{2.5D}$ from raw 2.5D joint j_{t-1}, it may consist of a different distribution pattern than the pose distribution of the belief map generated through GCN. Therefore, we proposed a heatmap loss to induce the distribution of latent heatmap h_{t-1} to be similar to h_j^0, only when the ground-truth for the result of coarse pose estimation becomes identical to the previous pose data. In other words, we applied the L2 loss between heatmaps only when the attention weight $w^{\star}$ is 1.

$$\mathscr{L}_{hm} = \left\| h_j^0 - h_{t-1} \right\|, \text{ only if } \quad w^{\star} = 1.0 \tag{12}$$

This caused the refinement module to incorporate the previous frame's pose information more effectively, and we verified the effect in the ablation study.

We use L2 norm for the attention weight w_a from Pose-attention encoder and ground-truth similarity $w^{\star}$.

$$\mathscr{L}_{weight} = \left\| w_a - w^{\star} \right\| \tag{13}$$

Thus the total loss for the Pose-attention encoder is defined as:

$$\mathscr{L}_{attention} = \mathscr{L}_{weight} + \lambda_2 * \mathscr{L}_{hm} \tag{14}$$

λ_2 is set to 10 according to preliminary experiments.

We train our network for each dataset and report cross-trained models utilized for real HMD applications. In all cases, the model trained based on the total loss $\mathscr{L}$:

$$\mathscr{L} = \mathscr{L}_{hand} + \mathscr{L}_{attention} \tag{15}$$

3.4 Offloading framework for HMD

To overcome the challenge of deploying a deep learning network onto the HPU of the Hololens 2, we devised the offloading framework shown in Fig. 4. The image data captured from the Hololens 2 is transmitted to a server and processed by our system to predict the hand pose, and only the predicted hand pose data is sent back to the Hololens 2. Our system utilizes the WebSocket communication protocol. Unlike the request-response model of HTTP, Websocket maintains a persistent connection after being established, resulting in lower latency for data transmission by eliminating unnecessary header data. Also, to account for network bandwidth and latency, the Hololens2 transmits the compressed image to the server over the network. We adopt the lossy compression algorithm based on its compression performance and encoding time. Utilizing lossy compression reduces network traffic drastically with negligible latency increases.

After the transmission, the server decodes the data to restore the image, which loses slight information compared to the original image. For pre-processing, we adopt a detection-by-tracking approach described in (Han et al. 2020) to crop the hand region. The input is cropped using a predefined bounding box initially, and the next center of the hand is extrapolated from the previous two tracked poses. Subsequently, the proposed system is executed, yielding 2.5D hand poses in image coordinates, which are then transmitted back to the Hololens2. On the device side, the root depth is extracted from the depth image in the HMD through 2D pose

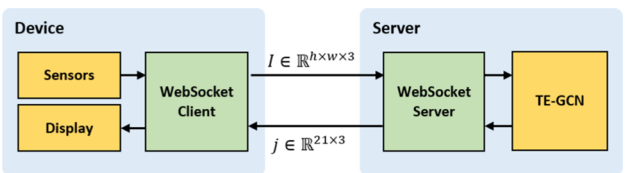

Fig. 4 Diagram for HMD application

prediction on the wrist. Finally, the 2.5D hand pose is lifted to a 3D hand pose in world coordinates using the camera parameters of the HMD and the root depth.

4 Experiment

We evaluate our hand tracking system on recent hand-object datasets: FreiHAND and DexYCB. Also, the performance of our full framework on HMD is verified with qualitative results in the real world. Since we intend the system to be trainable with all features of various datasets, we evaluate both the non-sequential dataset(FreiHAND) and the sequential dataset(DexYCB). Please note that both datasets were *treated as non-sequential* during the training phase. This indicates that the continuous frame information from the DexYCB was not utilized. Instead, we aimed to replicate the learning of temporal information solely using data generated by RDG.

4.1 Datasets

FreiHAND (Zimmermann et al. 2019) is a large-scale hand dataset that consists of non-sequential hand poses with real-world objects and various synthesized backgrounds. It comprises monocular color images with pose/mesh annotations for 130,240 training samples, while the evaluation set has 3960 samples with real-world backgrounds and supports an online-evaluation system.

DexYCB (Chao et al. 2021) is currently the most recent benchmark for real hand-object interaction sequences. This dataset comprises 582K RGB-Depth frames of a hand interacting with 20 YCB objects, including 1000 grasping sequences with a black background. We mainly used the official S0 split dataset and considered both right and left-hand samples while filtering out samples whose hand center is outside the image. Evaluation of the remaining official split S1 (unseen subjects), S2 (unseen views), and S3(unseen grasping) sets are also conducted.

4.2 Evaluation metrics

As our output of the model is the 3D hand joint/vertex, we evaluate both joints and vertices in FreiHAND and only joints in DexYCB according to the ground-truth provided by each dataset.

Mean per Joint/Vertex Position Error (MPJPE/MPVPE) is computed by Euclidean distances(mm) between the predicted 2.5D joint/vertices and ground-truth data after performing a 3D alignment with Procrustes analysis.

F-Score states the harmonic mean between recall and precision between two data w.r.t a threshold. We report F@5 mm and F@15 mm.

Area Under Curve (AUC) is the area under the percentage of correct keypoints(PCK) curve with error thresholds.

Mean keypoint acceleration (MKA) is computed following Han et al. (Han et al. 2020) and reported only on the DexYCB, as the metric requires sequential ground-truth information. The metric indicates the temporal smoothness of the predicted pose set, which is computed as below:

$$\mathrm{MKA}_t = \mathrm{mean}(j_{t-1} + j_{t+1} - 2j_t) \tag{16}$$

where MKA_t and j_t are the mean acceleration and joint position of the hand in the frame t. Despite the acceleration showing erroneous results when there is sudden hand movement, it has been employed as a valuable metric in previous studies (Han et al. 2020; Kanazawa et al. 2019; Tu et al. 2023) as it is a metric that approximates temporal smoothness in general AR/VR usage scenarios.

4.3 Implementation details

4.3.1 Preprocessing

The input images are cropped with a hand-bounding box and resized to 256×256. Then the augmentation process consists with scaling(±25%), rotation(±60°), random horizontal flip, color jittering(±20 of RGB in 8-bit) is performed. To acquire the bounding box, we crop the image with a fixed box size from the center of the image on the FreiHAND and adopt refined root depth provided by I2L-MeshNet (Moon and Lee 2020). In the case of DexYCB, since the recorded hand position is not centered, we crop an appropriate bounding box based on the provided 2D hand pose ground-truth. Additionally, we obtain the ground-truth hand vertices pose from MANO (Romero et al. 2017a) pose coefficients of DexYCB through the MANO layer facilitated by Manopth (Hasson et al. 2019a).

4.3.2 Training setup

We utilize the weight of ResNet-34 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) as our backbone. We trained 50 epochs in an end-to-end manner for the FreiHAND. For the DexYCB, we used the pre-trained weights with 10 epochs on FreiHAND for regularization. It demonstrated better performance than training on DexYCB alone at preliminary experiments. For generalization, we also verified cross-train setup of two datasets. In this setup, the model has trained 5 epochs with only FreiHAND and cross-trained for 45 epochs using a batch configuration distributed based on the dataset sizes' ratio. We use AdamW solver (Loshchilov and Hutter 2017) with a batch size of 64, initial learning rate of $3e{-}4$, and the rest are set to default value(β_1 of 0.9, β_2 of 0.999, and weight decay of $1e{-}2$). For hyperparameters,

we set the image input resolution to 256×256, the channel dimension of feature map c to 512, and the size of the belief map and heatmap n_b, n_h to 32 and 64, respectively.

Since the proposed HMD framework includes an image compression process, we apply an additional process during the training for the model integrated into the HMD framework. Details of image compression will be described in the section below. Dodge and Karam (2016) report that jpeg image compression does not significantly affect the inference performance of various networks. Nevertheless, to remove minor adverse effects, we forced image quality degradation by compressing the portion of image sets through random sampling. The sampling and compression ratios are set to 0.2 and 50%, respectively.

4.3.3 HMD system configuration

The proposed HMD application with Hololens2 is implemented in a server with an AMD Ryzen 9 7950X 16-Core CPU@4.50GHz, 64GB of RAM, and an Nvidia GeForce RTX 4090 GPU. The server is also utilized for training and evaluation of the network. The Wi-Fi specification is 5 G wireless, theoretically communicating data at up to 867 Mbps. As a result of measuring network quality, the upload speed of Wi-Fi is 122.21 Mbps, and the download speed is 50.23 Mbps.

4.4 Quantitative results

FreiHAND For the FreiHAND, we follow existing works (Boukhayma et al. 2019; Chen et al. 2021b, 2022a; Choi et al. 2020; Hasson et al. 2019b; Kulon et al. 2020; Lin et al. 2021a; Moon and Lee 2020; Zheng et al. 2021; Zimmermann et al. 2019) to conduct a comprehensive comparison with our method. Quantitative results with the official test set of FreiHAND are summarized in Table 1. Since FreiHAND is non-sequential, we assumed all previous poses to be zero-valued in our method. Despite this, our results are comparable to the state-of-the-art performance shown by the baseline method (Zheng et al. 2021). In other words, our method demonstrated notable accuracy even in the absence of temporal information, indicating its robustness in scenarios such as re-initialization due to tracking failure. While our findings demonstrate superior performance in terms of framerate compared to existing methods, it would be unfair to directly compare them to approaches that utilize less powerful GPUs. Therefore, we additionally report framerates on our hardware setup(†) if available.

In the case of MobRecon (Chen et al. 2022a), we referred to the results of the DenseStack backbone, which simultaneously reported performance and speed among various network settings. Since the result is tested on an Apple A14 CPU, we report the model's performance speed within our hardware setup(†). Lin et al. (2021a), a transformer-based method, showed performance close to state-of-the-art, but inference speed of slow due to the heavy HRNet (Wang et al. 2020b) and transfomer (Vaswani et al. 2017) adopted as encoder and decoder, respectively. Referring to Fig. 5, our method achieves nearly state-of-the-art performance in 3D PCK and performs better from an error threshold of 10 mm or higher. We also report the performance of our cross-trained model for real-time demo in Table 1, which shows relatively low performance, but still comparable. Regarding the generalization performance, which is its primary purpose of cross-training, the model exhibits better performance in our environment. A related result is shown in Fig. 9.

DexYCB For the DexYCB with official split S0 test set, we compare our method with existing works (Chao et al. 2021;

Table 1 Comparisons with state-of-the-art on FreiHAND

Method	MPJPE(↓)	AUC_J (↑)	MPVPE (↓)	AUC_V (↑)	F@5 mm (↑)	F@15 mm (↑)	FPS (↑)
ICCV19-MANO Fit (Zimmermann et al. 2019)	13.7	0.730	13.7	0.729	0.439	0.892	–
ICCV19-Hasson et al. (2019b)	13.3	0.737	13.3	0.736	0.429	0.907	–
ICCV19-MANO CNN (Zimmermann et al. 2019)	10.9	0.783	11.0	0.783	0.516	0.934	–
CVPR20-YoutubeHand (Kulon et al. 2020)	8.4	0.834	8.6	0.830	0.614	0.966	–
ECCV20-I2L-Mesh (Moon and Lee 2020)	7.4	–	7.6	–	0.681	0.973	53
ECCV20-Pose2Mesh (Choi et al. 2020)	7.7	–	7.8	–	0.674	0.969	8
CVPR23-Yu et al. (2023)	7.3	–	7.3	–	–	–	–
CVPR21-Chen et al. (2021b)	6.9	0.863	7.0	0.861	0.715	0.977	64
CVPR22-MobRecon (Chen et al. 2022a)	6.9	–	7.2	0.856	0.694	0.979	67 / 93†
CVPR21-Lin et al. (2021a)	6.8	–	6.7	–	0.717	0.981	18
ISMAR21-Zheng et al. (2021)	6.5	0.871	6.7	0.867	0.722	0.982	109†
Ours	6.5	0.870	6.7	0.866	0.723	0.981	140
Ours*	8.6	0.829	8.5	0.830	0.611	0.967	140

*Denotes cross-trained model for HMD. The previous pose for our method is all set to zero-value, as the FreiHAND is non-sequential

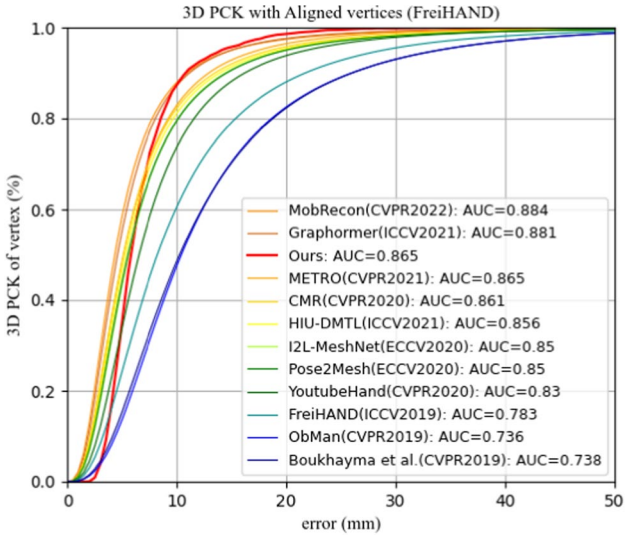

Fig. 5 Comparison of 3D PCK on FreiHAND

Chen et al. 2022b; Li et al. 2021; Lin et al. 2023; Spurr et al. 2020; Tse et al. 2022; Xu et al. 2023) that utilize monocular input as ours. Since DexYCB is a sequential dataset and does not provide ground-truth for vertices, we compute only hand joint accuracy and MKA. As shown in Table 2, the proposed method demonstrates superior computational efficiency compared to other methods. Recent studies evaluated with DexYCB (Chen et al. 2022b; Li et al. 2021; Lin et al. 2023; Tse et al. 2022; Yu et al. 2023) all aimed at the simultaneous reconstruction of hands and objects, so real-time performance is not guaranteed. Among the existing studies, the most recent work, H2ONet (Xu et al. 2023) shows a significant improvement in accuracy compared to previous works. However, H2ONet leverages the substantial assumption that hand pose transformation remains rigid across nearby frames. Since DexYCB consists of a sequence of fixed hand poses after grasping an object, the assumption is suitable for the dataset and leads to significant performance improvement. However, it is not applicable in real-world scenarios. Moreover, this method only achieved marginal real-time performance, making it unsuitable for applications involving HMDs. Lin et al. (2023) also showed high performance, but the computation speed is not reported as a method consisting of a heavy network backbone with a self-attention layer. The proposed method prioritizes notably superior speed over optimal accuracy, which offers significant advantages for HMD-based frameworks. These frameworks must account for not only the inference speed of the model but also other factors, such as pre/post-processing and communication speed. A comprehensive analysis of the MKA value will be provided in Sect. 4.6, as limited comparable research findings available.

We perform experiments for four distinct split sets to validate the model's generalization performance across various criteria. However, there is a limited number of cases where results for all DexYCB split sets have been reported, resulting in a relatively small comparison group. Referring to Table 3, all split sets, except for S1, achieved state-of-the-art performance in quantitative metrics. A notable observation is that when evaluating unseen viewpoints using the S2 split, the model's performance exceeded the default split's, and it performed less when testing the S1 split with unseen subjects. This difference may be attributed to the initial training with FreiHAND, which facilitated learning across all camera perspectives but had less effect on generalizing subject-related aspects. Consequently, the proposed system's performance is not solely dependent on dataset size and demonstrates superior generalization capabilities across different criteria such as subjects, viewpoints, and grasping scenarios.

Performance on Hololens2 In the proposed offloading framework, we evaluated the average latency for each step, as shown in Fig. 6. The process involved 10.23 ms for transmitting images from Hololens 2 to the server, 3.56 ms for preprocessing, 7.21 ms for TE-GCN inference, 0.13 ms for post-processing, and finally, 8.04 ms for returning hand

Table 2 Comparisons with state-of-the-arts on DexYCB

Method	MPJPE (↓)	AUC_J (↑)	MKA (↓)	FPS (↑)
ECCV22-Chen et al. (2022b)	19.00	–	–	–
ECCV20-Spurr et al. (2020)	17.34	0.698	–	–
CVPR22-Tse et al. (2022)	16.05	0.722	–	–
CVPR22-Li et al. (2021)	12.80	–	–	–
CVPR23-Yu et al. (2023)	8.92	–	–	–
CVPR21-Chao et al. (2021)	6.83	0.864	–	–
CVPR23-Lin et al. (2023)	5.47	–	–	–
CVPR23-H2ONet (Xu et al. 2023)	5.30	0.894	–	35
Ours	5.81	0.884	5.197	140
Ours*	5.94	0.881	5.482	140

*Denotes cross-trained model for HMD

Table 3 Comparison of official split types of DexYCB based on MPJPE(mm)

Method	S0 (sequences)	S1 (subjects)	S2 (views)	S3 (grasping)
ECCV20-Anil et al. (Armagan et al. 2020)	17.34	22.26	25.49	18.44
CVPR22-Tse et al. (2022)	16.05	21.22	27.01	17.93
Ours	**5.81**	7.08	**5.51**	**6.50**
Ours*	5.94	**6.94**	5.54	6.60

* denotes cross-trained model for HMD

Bold values denote the best performance for each aspect

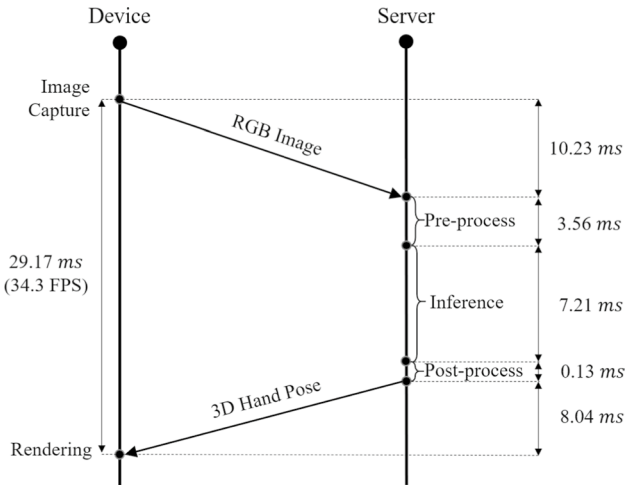

Fig. 6 Diagram of the latency for each step within our HMD-Server offloading framework

pose to Hololens 2. Thus, a total of 29.17 ms is required to process a single image. Since we processed the latest image received on the server, the framerate perceivable by users remains consistent with the reported latency at 34.3 FPS. Furthermore, additional experiments showed that WiFi specifications and connection environment directly influence data transmission and reception latency. Hence, it is expected that the overall framerate of the proposed module will increase in the future with improved communication environments.

4.5 Qualitative results

We collect challenging cases in Fig. 7, which include cases of significant self-occlusion or object occlusion. The results for both datasets exhibit robust joint and mesh predictions, and some sequences show results that appear to fit better visually than the ground-truth mesh. In Fig. 8, we present some failure cases and notable successes to demonstrate the detailed performance of our model. Our system struggles with unnatural and artificial hand poses, excessive self-occlusion, and ambiguous RGB input on depth, leading to incorrect predictions. However, our model performs well in cases where the hand is partially out of view, or the RGB input is blurred due to fast motion. These successful cases are challenging to achieve using deterministic temporal information utilization, indicating that our system performs temporally adaptive pose regression without relying too heavily on current or previous frames. Figure 9 shows the sample results of the proposed system on HMD. We qualitatively evaluated the system by streaming RGB images from Hololens2, estimating the hand information on the server, and transferring the data to the HMD. We utilized a cross-trained version of our model, which exhibited much better generalization performance. Due to latency issues, mesh rendering is performed only on the server for visualization purposes. Since we used the default MANO right-hand model, there are discrepancies in shape between the user's hand and the prediction model, resulting in an incorrect pose at the root position or the edge of the hand. To address this problem, we plan to implement online personalization of the hand model in future work.

4.6 Ablation study

We report the ablation studies on DexYCB with major loss components configuration in Table 4. We apply the same training setup for the DexYCB described in Sect. 4.3. Although the individual loss configurations did not have a notable impact on MKA, the results with the full loss configuration showed significantly lower MKA values. This indicates that the proposed loss set contributes to temporal coherence in a complementary manner, even when trained without ground-truth previous pose information of DexYCB.

To support our system design, we conduct a ablation study with various factors on network design. Table 5 shows performance according to different backbones and the number of stacked layers of adaptive-GCN. When using the ResNet-18 backbone, the inference speed increased due to the relatively light model, but the accuracy decreased due to the simplified model complexity. We found that using two stacked GCN layers resulted in better accuracy and speed, and supposed that adding more layers increased the optimization difficulty.

Next, we evaluate the impact of each data type defined in the Retroactive Data Generator on performance. Since

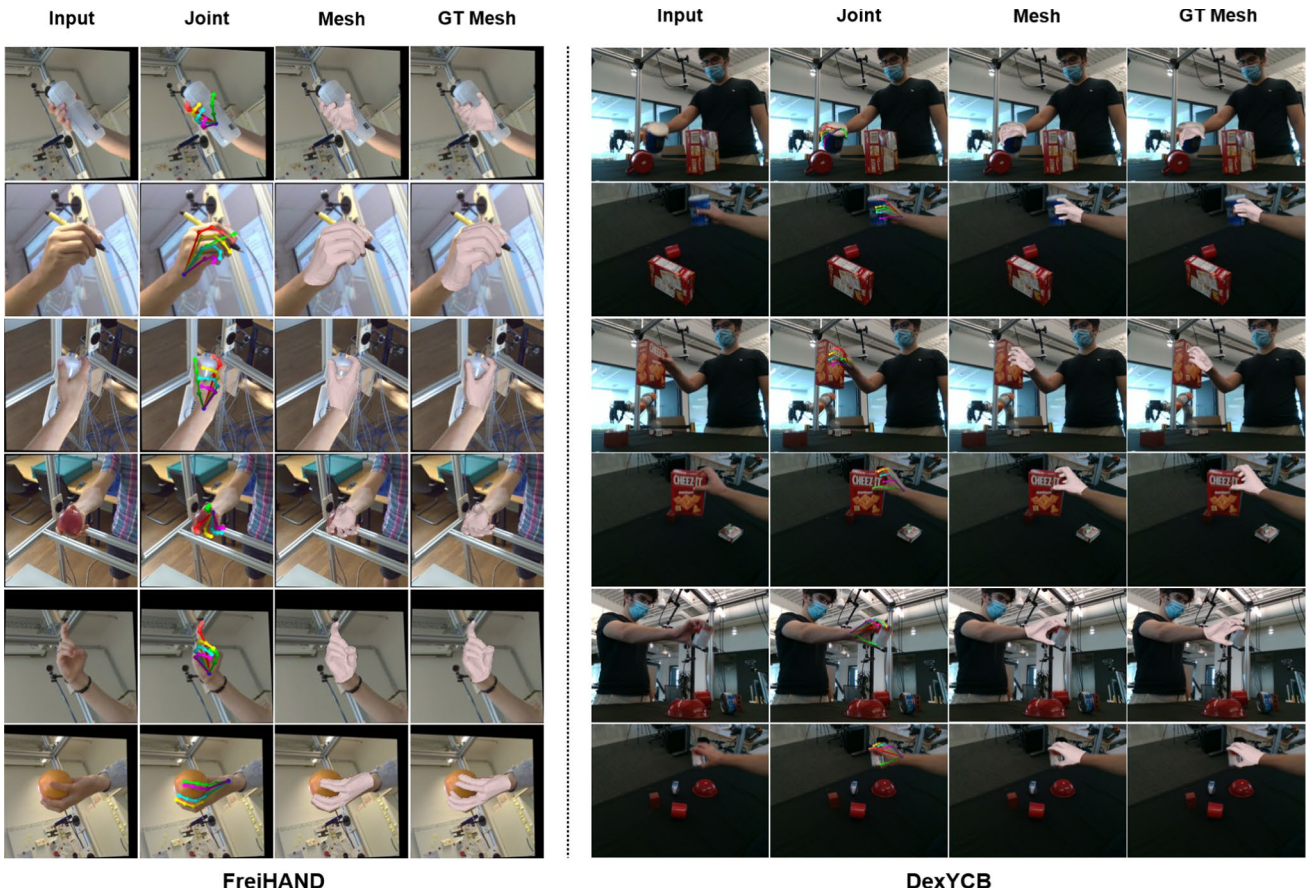

Fig. 7 Qualitative results on FreiHAND and DexYCB

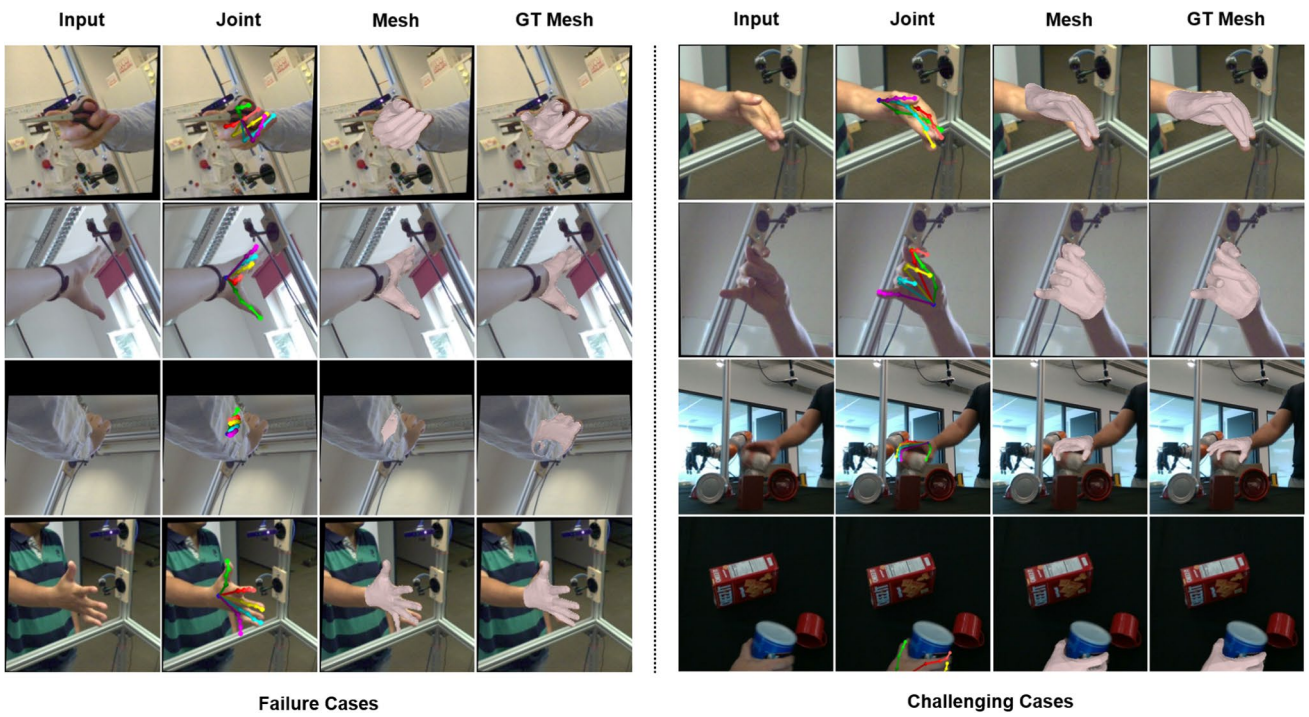

Fig. 8 Failure and challenging cases result in FreiHAND and DexYCB

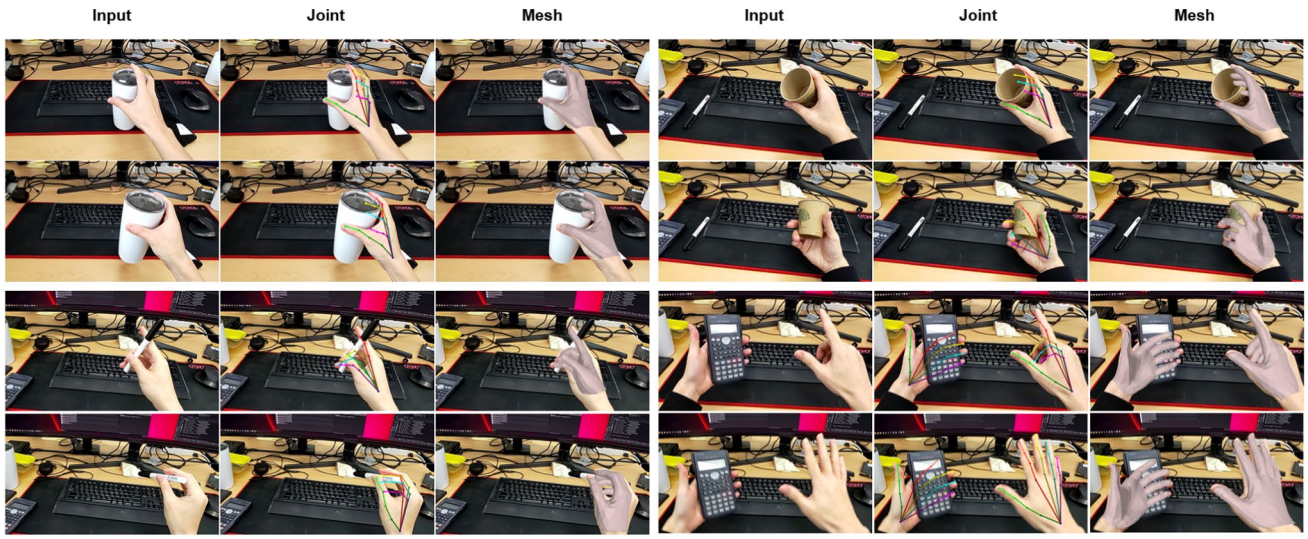

Fig. 9 Visual results on Hololens2 viewpoint. The scene is captured on Hololens2, and hand mesh is rendered in the off-loaded server

Table 4 Ablation study on Loss components

Method	MPJPE (↓)	AUC_J (↑)	MKA (↓)
Ground-truth	–	–	4.094
$\mathcal{L}_{hand}$	6.17	0.877	5.643
$\mathcal{L}_{hand} + \mathcal{L}_{weight}$	5.89	0.882	5.666
$\mathcal{L}_{hand} + \mathcal{L}_{hm}$	5.83	0.883	5.667
$\mathcal{L}_{hand} + \mathcal{L}_{weight} + \mathcal{L}_{hm}$	**5.81**	**0.884**	**5.197**

Ground-truth MKA is computed on DexYCB testset

Bold values denote the best performance for each aspect

Table 5 Ablation study of different backbones and GCN layer configuration

Backbone	GCN layer	MPJPE (↓)	AUC_J (↑)	MKA (↓)	FPS (↑)
ResNet-18	2	5.94	0.881	5.406	**187**
ResNet-18	3	6.01	0.880	5.291	176
ResNet-34	2	**5.81**	**0.884**	**5.197**	140
ResNet-34	3	5.92	0.882	5.201	133

Bold values denote the best performance for each aspect

ground-truth data of the previous pose is required for the non-RDG case, we perform training and testing only on the DexYCB. The results are presented in Table 6; due to the limitations of experimenting with all possible distributions while adjusting control variables, the reported results are limited to a few selected ratio combinations. As the influence of several variables acted simultaneously, making it challenging to draw intuitive conclusions for each variable individually. The MKA value shows the lowest value in the case of r1, where only the Static type from RDG is available. The result occurs as the network uniformly considers the current frame to be identical to the pose of the previous frame, leading to r1 displaying the most temporally smooth outcomes, irrespective of pose accuracy. The result from r2, comprising only perturbed data, emphasizes the disparity between conventional augmentation methods and our proposed RDG module. It indicates that data augmented only with random noise failed to effectively serve as the previous pose, resulting in inadequate temporal coherence and accuracy.

Despite the complexity of various influencing factors, through multiple experiments, the proposed system was able to improve temporal smoothness by achieving the highest pose estimation performance and the second-best MKA value at the r0 ratio. It should be noted that temporal smoothness is one aspect of the representable motion by the proposed system. As shown in Table 1, it demonstrates state-of-the-art performance even in non-sequential dataset situations, i.e., re-initialization scenarios with no previous pose.

We also examine the effect of image quality degradation on the performance of our proposed system. As shown in Table 7, the performance of the model significantly decreases when the image compression is not considered during training. On the other hand, when the model is trained with quality-degraded images, it achieves a performance close to the optimal case. This implies that the model is trained to operate regardless of image quality by degrading the partial training image set. As a result, we can maintain the state-of-the-art performance of the trained model while taking advantage of image compression in the HMD framework.

Table 6 Ablation study of Retroactive Data Generator

	Static	Perturbed	Zero-value	MPJPE	AUC	MKA
wo/RDG	–	–	–	10.374	0.79	9.692
r1	1.0	0.0	0.0	18.354	0.639	**4.412**
r2	0.0	1.0	0.0	7.776	0.845	5.914
r3	0.0	0.0	1.0	6.022	0.880	5.409
r4	0.5	0.5	0.0	8.127	0.838	5.616
r5	0.5	0.0	0.5	5.918	0.882	5.475
r6	0.0	0.5	0.5	5.909	0.882	5.398
r7	0.65	0.25	0.1	6.008	0.880	5.431
r8	0.45	0.35	0.2	5.926	0.882	5.353
r9	0.2	0.45	0.35	5.939	0.881	5.317
Ours (r0)	0.2	0.65	0.15	**5.879**	**0.882**	5.271

Each model is trained on the DexYCB only

Bold values denote the best performance for each aspect

Table 7 Ablation study of image quality degradation process for HMD framework

Compress. in train	Compress. in test	MPJPE (↓)	AUC_J (↑)	MPVPE (↓)	AUC_V (↑)	MKA (↓)
FreiHAND						
No	No	**6.5**	**0.870**	**6.7**	**0.866**	–
No	Yes	8.1	0.840	8.2	0.836	–
Yes	No	6.7	0.568	6.9	0.864	
Yes	Yes	6.7	0.866	6.9	0.862	–
DexYCB						
No	No	**5.9**	**0.882**	–	–	5.271
No	Yes	6.0	0.880	–	–	6.654
Yes	No	5.9	0.882	–	–	**5.108**
Yes	Yes	5.9	0.882	–	–	5.395

Each model is trained on the target dataset only

Bold values denote the best performance for each aspect

5 Conclusion

We proposed a real-time novel hand-tracking system that enables robust hand pose estimation under various real environmental conditions and utilizes it in HMD as an offloading framework. Our system consists of a 2-stage GCN-based lightweight network that balances accuracy and real-time performance. Furthermore, we have developed a novel approach to incorporating temporal information via the attention mechanism, which we have validated on recent benchmark datasets. The proposed system achieves state-of-the-art balanced performance in FreiHAND and DexYCB and has demonstrated the importance of adaptive utilization of temporal information for real-world scenarios. Moreover, integrating our system into an HMD through the offloading framework has expanded its potential for practical applications for any other mobile devices, and we have demonstrated the model's generalization performance.

Limitation One notable limitation is that the integrated HMD system did not achieve a desirable level of runtime speed. The primary factor contributing to increased latency in the proposed HMD framework was preprocessing and communication speed, which can be enhanced through system optimization. Another limitation is the hand tracker's vulnerability when dealing with unseen subjects. We plan to address this limitation by leveraging the system's capability to learn from various public datasets, irrespective of the sequential nature of the data.

Future work Since we solely assessed the real-time performance of the hand tracker on an HMD, our next step involves implementing practical applications on HMDs and conducting user studies to demonstrate the actual usability and effectiveness for subjects. From a system perspective, our plan is to delve into advanced methods that leverage temporal information and employ self-supervised learning approaches, aiming to enhance the overall generalizability of the system.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10055-024-01039-3.

Author contributions Cho was involved in the overall research structure and all processes, and Ha, IK.Jeon, and JW.Jeon performed HMD-based implementation and experiments. Kim and Woo provided advice on study composition and reviewed the manuscript.

Funding This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01270, and RS-2024-00397663).

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Armagan A, Garcia-Hernando G, Baek S, Hampali S, Rad M, Zhang Z, Xie S, Chen M, Zhang B, Xiong F et al. (2020) Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3D hand pose estimation under hand-object interaction. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, Springer, pp 85–101

Baek S, Kim KI, Kim T-K (2019) Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1067–1076

Baek S, Kim KI, Kim T-K (2020) Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6121–6131

Boukhayma A, Bem Rd, Torr PH (2019) 3D hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10843–10852

Bruna J, Zaremba W, Szlam A, LeCun Y (2013) Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203

Cai Y, Ge L, Liu J, Cai J, Cham T-J, Yuan J, Thalmann NM (2019) Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2272–2281

Cao Z, Radosavovic I, Kanazawa A, Malik J (2021) Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 12417–12426

Chao Y-W, Yang W, Xiang Y, Molchanov P, Handa A, Tremblay J, Narang YS, Van Wyk K, Iqbal U, Birchfield S et al. (2021) Dexycb: a benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9044–9053

Chen L, Lin S-Y, Xie Y, Lin Y-Y, Xie X (2021a) Temporal-aware self-supervised learning for 3D hand pose and mesh estimation in videos. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1050–1059

Chen X, Liu Y, Ma C, Chang J, Wang H, Chen T, Guo X, Wan P, Zheng W (2021b) Camera-space hand mesh recovery via semantic aggregation and adaptive 2D-1D registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13274–13283

Chen X, Liu Y, Dong Y, Zhang X, Ma C, Xiong Y, Zhang Y, Guo X (2022a) Mobrecon: mobile-friendly hand mesh reconstruction from monocular image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20544–20554

Chen Y, Tu Z, Kang D, Bao L, Zhang Y, Zhe X, Chen R, Yuan J (2021c) Model-based 3D hand reconstruction via self-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10451–10460

Chen Z, Hasson Y, Schmid C, Laptev I (2022b) Alignsdf: pose-aligned signed distance fields for hand-object reconstruction. In: European conference on computer vision, Springer, pp 231–248

Chen Z, Chen S, Schmid C, Laptev I (2023) gsdf: geometry-driven signed distance functions for 3D hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12890–12900

Cho J, Youwang K, Oh T-H (2022) Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In: European conference on computer vision, Springer, pp 342–359

Choi H, Moon G, Lee KM (2020) Pose2mesh: graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In: Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, Springer, pp 769–787

Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. Adv Neural Inf Process Syst 29

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255

Dodge S, Karam L (2016) Understanding how image quality affects deep neural networks. In: 2016 eighth international conference on quality of multimedia experience (QoMEX), IEEE, pp 1–6

Doosti B, Naha S, Mirbagheri M, Crandall DJ (2020) Hope-net: a graph-based model for hand-object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6608–6617

Fan Z, Spurr A, Kocabas M, Tang S, Black MJ, Hilliges O (2021) Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In: 2021 International Conference on 3D Vision (3DV), IEEE, pp 1–10

Fu Q, Liu X, Xu R, Niebles JC, Kitani KM (2023) Deformer: dynamic fusion transformer for robust hand pose estimation. arXiv preprint arXiv:2303.04991

Ge L, Ren Z, Li Y, Xue Z, Wang Y, Cai J, Yuan J (2019) 3D hand shape and pose estimation from a single RGB image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10833–10842

Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: International conference on machine learning, PMLR, pp 1263–1272

Hampali S, Rad M, Oberweger M, Lepetit V (2020) Honnotate: a method for 3D annotation of hand and object poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3196–3206

Han S, Liu B, Cabezas R, Twigg CD, Zhang P, Petkau J, Yu T-H, Tai C-J, Akbay M, Wang Z et al (2020) Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Trans Graph (ToG) 39(4):87–1

Han S, Wu P-c, Zhang Y, Liu B, Zhang L, Wang Z, Si W, Zhang P, Cai Y, Hodan T, et al. (2022) Umetrack: unified multi-view end-to-end hand tracking for vr. In: SIGGRAPH Asia 2022 conference papers, pp 1–9

Hasson Y, Varol G, Tzionas D, Kalevatykh I, Black MJ, Laptev I, Schmid C (2019a) Learning joint reconstruction of hands and manipulated objects. In: CVPR

Hasson Y, Varol G, Tzionas D, Kalevatykh I, Black MJ, Laptev I, Schmid C (2019b) Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11807–11816

Hasson Y, Tekin B, Bogo F, Laptev I, Pollefeys M, Schmid C (2020) Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 571–580

Hasson Y, Varol G, Schmid C, Laptev I (2021) Towards unconstrained joint hand-object reconstruction from RGB videos. In: 2021 International conference on 3D vision (3DV), IEEE, pp 659–668

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hossain MRI, Little JJ (2018) Exploiting temporal information for 3D human pose estimation. In: Proceedings of the European conference on computer vision (ECCV), pp 68–84

Iqbal U, Molchanov P, Gall TBJ, Kautz J (2018) Hand pose estimation via latent 2.5 d heatmap regression. In: Proceedings of the European conference on computer vision (ECCV), pp 118–134

Kanazawa A, Zhang JY, Felsen P, Malik J (2019) Learning 3D human dynamics from video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5614–5623

Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907

Kocabas M, Athanasiou N, Black MJ (2020) Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5253–5263

Kulon D, Guler RA, Kokkinos I, Bronstein MM, Zafeiriou S (2020) Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4990–5000

Lepetit V (2020) Recent advances in 3d object and hand pose estimation. arXiv preprint arXiv:2006.05927

Li K, Yang L, Zhan X, Lv J, Xu W, Li J, Lu C (2021) Artiboost: boosting articulated 3D hand-object pose estimation via online exploration and synthesis. arXiv preprint arXiv:2109.05488

Li M, An L, Zhang H, Wu L, Chen F, Yu T, Liu Y (2022) Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2761–2770

Lim GM, Jatesiktat P, Ang WT (2020) Mobilehand: Real-time 3d hand shape and pose estimation from color image. In: Neural information processing: 27th international conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV, Springer, pp 450–459

Lin K, Wang L, Liu Z (2021a) End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1954–1963

Lin K, Wang L, Liu Z (2021b) Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 12939–12948

Lin Z, Ding C, Yao H, Kuang Z, Huang S (2023) Harmonious feature learning for interactive hand-object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12989–12998

Liu S, Jiang H, Xu J, Liu S, Wang X (2021) Semi-supervised 3d hand-object poses estimation with interactions in time. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14687–14697

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101

Monti F, Boscaini D, Masci J, Rodola E, Svoboda J, Bronstein MM (2017) Geometric deep learning on graphs and manifolds using mixture model CNNs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5115–5124

Moon G, Lee KM (2020) I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single RGB image. In: European conference on computer vision, Springer, pp 752–768

Moon G, Yu S-I, Wen H, Shiratori T, Lee KM (2020) Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single TGB image. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, Springer, pp 548–564

Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C (2018) Ganerated hands for real-time 3D hand tracking from monocular RGB. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 49–59

Park G, Argyros A, Lee J, Woo W (2020a) 3d hand tracking in the presence of excessive motion blur. IEEE Trans Vis Comput Graph 26(5):1891–1901

Park G, Kim T-K, Woo W (2020b) 3d hand pose estimation with a single infrared camera via domain transfer learning. In: 2020 IEEE International symposium on mixed and augmented reality (ISMAR), IEEE, pp 588–599

Qu W, Cui Z, Zhang Y, Meng C, Ma C, Deng X, Wang H (2023) Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15100–15111

Ren P, Wen C, Zheng X, Xue Z, Sun H, Qi Q, Wang J, Liao J (2023) Decoupled iterative refinement framework for interacting hands reconstruction from a single RGB image. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 8014–8025

Romero J, Tzionas D, Black MJ (Nov. 2017a) Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia). URL http://doi.acm.org/10.1145/3130800.3130883

Romero J, Tzionas D, Black MJ (2017b) Embodied hands: modeling and capturing hands and bodies together. ACM Trans Graph (TOG) 36(6):1–17

Spurr A, Song J, Park S, Hilliges O (2018) Cross-modal deep variational hand pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition pp 89–98

Spurr A, Iqbal U, Molchanov P, Hilliges O, Kautz J (2020) Weakly supervised 3d hand pose estimation via biomechanical constraints. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16, Springer, pp 211–228

Tang X, Wang T, Fu C-W (2021) Towards accurate alignment in real-time 3D hand-mesh reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11698–11707

Tse THE, Kim KI, Leonardis A, Chang HJ (2022) Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1664–1674

Tu Z, Huang Z, Chen Y, Kang D, Bao L, Yang B, Yuan J (2022) Consistent 3d hand reconstruction in video via self-supervised learning. arXiv preprint arXiv:2201.09548

Tu Z, Huang Z, Chen Y, Kang D, Bao L, Yang B, Yuan J (2023) Consistent 3D hand reconstruction in video via self-supervised learning. IEEE Tran Patt Anal Mach Intell 45(8):9469–9485

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30:261–272

Wang J, Mueller F, Bernard F, Sorli S, Sotnychenko O, Qian N, Otaduy MA, Casas D, Theobalt C (2020a) Rgb2hands: real-time tracking of 3d hand interactions from monocular RGB video. ACM Trans Graph (ToG) 39(6):1–16

Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X et al (2020b) Deep high-resolution representation learning for visual recognition. IEEE Trans Patt Anal Mach Intell 43(10):3349–3364

Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang Y-G (2018) Pixel2mesh: generating 3d mesh models from single RGB images. In: Proceedings of the European conference on computer vision (ECCV), pp 52–67

Wei S-E, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4724–4732

Xu H, Wang T, Tang X, Fu C-W (2023) H2onet: Hand-occlusion-and-orientation-aware network for real-time 3D hand mesh reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 17048–17058

Xu K, Hu W, Leskovec J, Jegelka S (2018) How powerful are graph neural networks? arXiv preprint arXiv:1810.00826

Yang J, Chang HJ, Lee S, Kwak N (2020) Seqhand: RGB-sequence-based 3d hand pose and shape estimation. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, Springer, pp 122–139

Yang L, Yao A (2019) Disentangling latent hands for image synthesis and pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9877–9886

Yang L, Li S, Lee D, Yao A (2019) Aligning latent spaces for 3d hand pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2335–2343

Yang L, Chen S, Yao A (2021) Semihand: Semi-supervised hand pose estimation with consistency. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11364–11373

Ye Y, Hebbar P, Gupta A, Tulsiani S (2023) Diffusion-guided reconstruction of everyday hand-object interaction clips. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 19717–19728

Yu Z, Li C, Yang L, Zheng X, Mi MB, Lee GH, Yao A (2023) Overcoming the trade-off between accuracy and plausibility in 3D hand shape reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 544–553

Zhang H, Bo Z-H, Yong J-H, Xu F (2019a) Interactionfusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. ACM Trans Graph (TOG) 38(4):1–11

Zhang X, Li Q, Mo H, Zhang W, Zheng W (2019b) End-to-end hand mesh recovery from a monocular RGB image. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2354–2364

Zhang X, Huang H, Tan J, Xu H, Yang C, Peng G, Wang L, Liu J (2021) Hand image understanding via deep multi-task learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11281–11292

Zhao Z, Zhao X, Wang Y (2021) Travelnet: self-supervised physically plausible hand motion learning from monocular color images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11666–11676

Zheng X, Ren P, Sun H, Wang J, Qi Q, Liao J (2021) Sar: spatial-aware regression for 3D hand pose and mesh reconstruction from a monocular RGB image. In: 2021 IEEE international symposium on mixed and augmented reality (ISMAR), IEEE, pp 99–108

Zhou Y, Habermann M, Xu W, Habibie I, Theobalt C, Xu F (2020) Monocular real-time hand shape and motion capture using multi-modal data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5346–5355

Zimmermann C, Brox T (2017) Learning to estimate 3d hand pose from single RGB images. In: Proceedings of the IEEE international conference on computer vision, pp 4903–4911

Zimmermann C, Ceylan D, Yang J, Russell B, Argus M, Brox T (2019) Freihand: a dataset for markerless capture of hand pose and shape from single RGB images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 813–822

Zuo B, Zhao Z, Sun W, Xie W, Xue Z, Wang Y (2023) Reconstructing interacting hands with interaction prior from monocular images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9054–9064