

HandJoKe: Joint-Guided Keypoint Denoising Transformer for Depth-Based 3D Hand Pose Estimation

Ji Gan[✉], Lei Chen[✉], Ping Hu, *Member, IEEE*, Jiaxu Leng[✉], *Member, IEEE*, Weisheng Li[✉], *Member, IEEE*, and Xinbo Gao[✉], *Fellow, IEEE*

Abstract—Existing depth-based 3D hand pose estimation methods typically estimate hand joints from either 2D depth images or 3D point clouds, whereas the approaches that fuse multimodal data remain underexplored. Furthermore, previous methods often struggle to learn geometric-facilitated features and precise joint correlations, especially for occluded hands, due to the lack of explicit prior guidance and insufficient cross-dimensional interaction. By taking advantage of multi-modal fusion, cross-dimensional interaction, and prior guidance, we propose a novel joint-guided keypoint denoising Transformer (named HandJoKe) to achieve more precise hand pose estimation, which can iteratively estimate hand poses based on keypoint features from both 2D depth images and 3D point clouds under explicit joint guidance within only several denoising steps. Rather than directly applying existing multi-modal fusion to perform redundant interactions among many background pixels and irrelevant points, HandJoKe focuses on modeling correlations and capturing dependencies among local informative hand regions (i.e., keypoints), thus attaining higher learning capability with lower computation redundancy. Moreover, a novel joint-guided denoising estimation strategy is introduced to adequately fuse cross-modal keypoint features under explicit joint guidance, achieving geometric-facilitated cross-modal keypoint interaction in both 2D and 3D spaces. The effectiveness of joint guidance can be further strengthened through iterative denoising, since it can subsequently update cross-modal keypoint features based on previous denoised hand poses and thus can help better locate confused joints, especially for occluded hands. Extensive experiments show that HandJoKe has achieved state-of-the-art performance on four public challenging benchmarks, including

single-hand datasets NYU and ICVL, and hand-object datasets DexYCB and HO3D.

Index Terms—3D hand pose estimation, denoising transformer, multi-modal fusion, cross-dimensional interaction.

I. INTRODUCTION

AS A fundamental human-computer interaction technique, 3D hand pose estimation aims to locate essential keypoints of human hands from the captured images. Such a technique is applicable to various human-computer interaction applications, including augmented/virtual reality and gaming. With the recent advances in deep learning techniques and the proliferation of low-cost depth cameras, significant progress has been made in depth-based 3D hand pose estimation in recent years [1], [2], [3], [4].

Due to the unique characteristics of depth data, the captured hand data can be regarded as either 2D depth images or 3D point clouds/voxels, and thus, there are two distinct approaches widely adopted for depth-based 3D hand pose estimation: i) 2D image-based methods typically utilize well-explored 2D convolutional neural networks (CNNs) to generate specific spatial structures or representations from depth images, and further decompose the hand into sub-regions or groups of joints to more effectively capture intricate inter-joint relationships; ii) 3D data-based methods usually leverage the advantages of 3D CNNs or point cloud networks (e.g., PointNet and graph neural networks) to directly capture the geometric structures from 3D data and then regress on the obtained 3D observations. However, considering that both 2D depth images and 3D point clouds are beneficial to 3D hand estimation, these two distinct modalities can complement each other. Therefore, it is highly likely to further improve 3D hand pose estimation by capturing the rich 3D geometric features from point clouds and fusing the 2D pixel-wise features from depth images.

Although multi-modal fusion has emerged as a significant trend in visual perception technology, most existing methods for hand pose estimation regress 3D hand poses from a single modality alone, leaving the multi-modal approaches under-explored. Moreover, it suffers from high computation redundancy and low learning capability when directly applying existing multi-modal fusion methods for 3D hand pose estimation. Specifically, multi-modal fusion strategies like pixel-wise

Received 24 April 2025; revised 20 July 2025; accepted 9 August 2025. Date of publication 13 August 2025; date of current version 22 January 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62206035, Grant 62331008, and Grant 62476048; in part by the Science and Technology Innovation Key Research and Development Program of Chongqing under Grant CSTB2024TIAD-STX0023 and Grant CSTB2023TIAD-STX0016; in part by the Natural Science Foundation of Chongqing under Grant CSTB2024NSCQ-QCXM0060, Grant CSTB2023NSCQ-LZX0047, and Grant CSTB2022NSCQ-MSX0547; in part by China Postdoctoral Science Foundation under Grant 2025MD774186; in part by Chongqing Special Postdoctoral Research Funding under Grant 2024CQBSHTB2002; and in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN202200612. This article was recommended by Associate Editor T. Chen. (Corresponding author: Xinbo Gao.)

Ji Gan, Lei Chen, Jiaxu Leng, Weisheng Li, and Xinbo Gao are with the School of Computer Science and Technology and the Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaobx@cqupt.edu.cn).

Ping Hu is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China.

Digital Object Identifier 10.1109/TCSVT.2025.3597981

1051-8215 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19, 2026 at 15:32:37 UTC from IEEE Xplore. Restrictions apply.

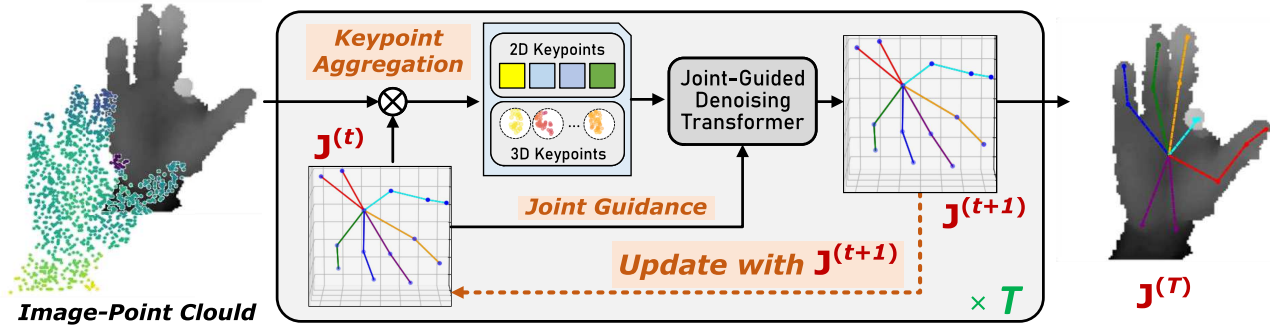


Fig. 1. Concept of HandJoKe. HandJoKe iteratively aggregates cross-modal keypoint features and denoises hand poses under explicit joint guidance.

weighted aggregation or cross-attention mechanism operate fusion over the whole 2D images and 3D data [5], [6], thereby attaining high computation redundancy and memory usage. Moreover, those fusion strategies inevitably perform redundant interactions among many irrelevant pixels and points (such as background pixels and joint-irrelevant points), thus reducing the learning efficiency and capability. Instead, the latent representations of keypoints focus on more informative local regions of hand poses but ignore the invalid pixels or points, thereby reducing the computation redundancy and enhancing the learning efficiency.

Recently, a few methods (such as IPNet [7], HandDAGT [8], and HandDiff [9]) have attempted to integrate the advantages of cross-modal fusion for depth-based hand pose estimation, while they still suffer from learning incorrect geometric features and confused joint correlations, especially for occluded hands due to the lack of explicit prior guidance and insufficient cross-dimensional interaction. For example, IPNet [7] first projected 2D visual features into 3D point features and then iteratively estimated hand poses in the 3D point cloud space; HandDAGT [8] further integrated an adaptive attention module to better enhance geometric features in the 3D space; HandDiff [9] attempted to denoise 3D noisy joints conditioned on normalized point clouds over hundreds of times. However, those methods only fused multi-modal data at the early stage with a 2D-3D projection and mainly refined geometric features in the 3D point space alone, thus making cross-modal fusion and cross-dimensional interaction between the 2D and 3D spaces insufficient. Moreover, previous methods also lack explicit geometric guidance to precisely locate the keypoint features of hands during the intricate cross-modal fusing, thus learning confused geometric structures and joint correlations, especially for occlusions.

To address these challenges, we propose a novel **Joint-guided Keypoint denoising Transformer (HandJoKe)** for depth-based 3D hand pose estimation as shown in Fig. 1, which iteratively estimates and denoises hand poses based on keypoint features extracted from both depth images and point clouds under explicit joint guidance within only several denoising steps. Specifically, 1) we first introduce an **image-point cloud keypoint aggregation** strategy to aggregate multimodal and cross-dimensional keypoint features based on previously estimated hand joints from both 2D depth images and 3D point clouds. This strategy makes the model focus

on modeling correlations and capturing dependencies among local informative regions of hands, rather than processing large numbers of background pixels and irrelevant 3D points, thus attaining higher learning capability with lower computation redundancy. 2) Furthermore, we propose a novel **joint-guided denoising estimation** strategy that effectively fuses cross-modal keypoint features and iteratively denoises the estimated hand pose joints in both 2D and 3D spaces under explicit joint guidance. Particularly, hand joints are explicitly injected into the graph-enhanced attention mechanism of the denoising Transformer as prior geometric guidance, thus achieving geometric-facilitated cross-modal keypoint fusion and interaction. Moreover, the effectiveness of joint guidance can be further strengthened through iterative denoising, since cross-modal keypoint features can be progressively updated based on previously denoised hand poses and thus can better locate confused joints, especially for occluded hands. Finally, HandJoKe follows keypoint-based multi-modal fusion with lower computation redundancy and higher learning efficiency, and it also models hands' topological structures in both 2D and 3D spaces, rather than relying solely on the 3D point cloud space, thereby enabling more efficient cross-dimensional interaction.

Experiments are conducted on four challenging benchmarks, including single-hand datasets NYU [10] & ICVL [11] and hand-object datasets DexYCB [12] & HO3D [13], demonstrating that HandJoKe achieves mean distance errors of 6.97mm on NYU, 5.53mm on ICVL, 7.65mm on DexYCB, and 1.75mm on HO3D, outperforming previous state-of-the-art (SOTA) approaches for depth-based 3D hand pose estimation.

Our primary contributions are summarized as follows:

- 1) We propose a joint-guided keypoint denoising Transformer HandJoKe for depth-based 3D hand pose estimation, which iteratively estimates and denoises hand poses based on keypoint features from both depth images and point clouds under explicit joint guidance within only several denoising steps.
- 2) We propose a joint-guided denoising estimation strategy that effectively fuses cross-modal keypoint features under explicit joint guidance and iteratively denoises estimated hand poses in both 2D and 3D spaces, achieving geometric-facilitated cross-modal keypoint fusion and cross-dimensional interaction.

- 3) Extensive experiments demonstrate that HandJoKe achieves state-of-the-art performance on public benchmarks, including single-hand datasets NYU & ICVL and hand-object datasets DexYCB & HO-3D.

II. RELATED WORK

A. 2D Image-Based Hand Pose Estimation

The 2D image-based methods treat the depth data as single-channel depth images and utilize 2D CNNs to regress 3D hand poses. According to the prediction strategy, these methods can be mainly divided into regression-based methods and detection-based methods.

Regression-based methods (such as Pose-REN [3] and REN [14]) typically utilize 2D CNNs to extract global or local features from depth images, and then employ fully connected layers to map these features into the 3D space for regressing the 3D coordinates of hand pose joints. For example, REN [14] proposed a region ensembled network that divides the feature map from the final convolutional layer into several spatial regions, which are then integrated by the subsequent fully connected layers to regress the 3D positions of hand joints. Pose-REN [3] further proposed a hierarchical method that fuses features of different joints based on the topological structures of hand poses. By aggregating the features from all fingers, it predicts final hand poses more accurately.

In contrast, detection-based methods (like DenseReg [15] and AWR [4]) estimate 3D hand poses by generating heatmaps or offset vector fields through dense pixel-wise regression. DenseReg [15] regressed 3D heatmaps and unit vector fields, and combined a mean-shift aggregation strategy to accurately estimate 3D hand joint positions from a single depth image. AWR [4] proposed a simple yet effective adaptive weighting regression method that unifies detection and regression approaches into a single step. Guided by adaptive weight maps, it aggregates different regions by discretely integrating all pixels of the dense representation, thus achieving more accurate and robust estimation.

Due to the inability to perceive the three-dimensional structures of hands as well as the neglect of spatial geometric relationships inherent in depth maps, the performance of 2D image-based methods significantly degrades for complex gestures or low-quality depth maps [1], [2], [3], [4], [10], [14], [15], [16], [17], [18].

B. 3D Data-Based Hand Pose Estimation

To overcome the limitations of depth images, some studies directly utilized 3D spatial information from depth data for hand pose estimation. By directly processing 3D point clouds or voxel data, these methods can more intuitively capture the 3D geometric shapes of hands, providing more accurate spatial positioning information and relationships of hand joints. [19] first encoded 3D hand point clouds as 3D volumes and then adopted a 3D CNN to directly regress 3D hand poses in a single forward pass. [20] reformulated the problem of 3D hand pose estimation as a voxel-to-voxel prediction task by utilizing the 3D voxelized grids and estimated per-voxel likelihoods for each hand keypoint.

In recent years, some methods like [21], [22], [23], and [24] have further leveraged the efficiency of PointNet, especially in processing point cloud data. PointNet has unique advantages in handling point cloud data, as it can maintain permutation invariance and viewpoint robustness. This is crucial for 3D hand pose estimation because point clouds of the hands from different angles should maintain a consistent representation, unaffected by the changes in point orders or viewpoints. [23] is a network architecture based on PointNet that can directly estimate hand joint positions from 3D point clouds, which first converts depth images into a set of 3D points and then processes them with a hierarchical PointNet. This PointNet is trained end-to-end to extract hierarchical hand features and regress 3D hand poses.

C. Multi-Modal Fusion-Based Methods

Multi-modal fusion has been widely studied in many computer vision tasks. Typically, cross-modal features can be calculated through element-wise attention [5] and further fused with cross-attention or Transformers [6]. Nevertheless, depth-based 3D hand estimation from multi-modal data remains under-explored, and there are only a few attempts have been made, like IPNet [7] and HandTAGT [8]. Specifically, IPNet [7] first adopted a 2D CNN to extract visual representations from depth images and then projected them into the 3D space for iterative feature refinement, thus fully exploiting the 3D geometric structures of depth data. In addition, a 3D parametric hand model is introduced to improve the robustness of the model against occlusions. HandTAGT [8] proposed a novel 3D hand pose estimation method based on the adaptive graph attention modules, which can adaptively utilize kinematic and local information of hand poses depending on the degrees of hand occlusions.

Although these methods have made attempts in multi-modal fusion, they generally face a common challenge, i.e., how to effectively handle issues like hand occlusions, pose complexity, and cross-dimensional mapping during the multi-modal fusion. When dealing with occluded hands and complex poses, explicit geometric guidance is critical for the model to accurately learn informative keypoint features and correct joint correlations in both 2D and 3D spaces. Nevertheless, previous methods lack explicit guidance to learn precise keypoint features and correct joint correlations, especially for occlusions.

D. Relationship and Distinctions With Diffusion Models

Diffusion models have recently emerged as one of the most powerful generative models in computer vision. Due to their strong modeling and generative capabilities, they have also been successfully extended from image synthesis [25] to a variety of related tasks, including image enhancement [26], [27], super-resolution [28], depth estimation [29], [30], video anomaly detection [31], and pose estimation [9], [32], [33], [34], [35].

Most existing diffusion-based methods for hand pose estimation adopt a 2D-to-3D lifting paradigm, where 2D keypoints are used as conditional inputs to guide the diffusion process for inferring 3D joint positions. In such a framework, 3D joints

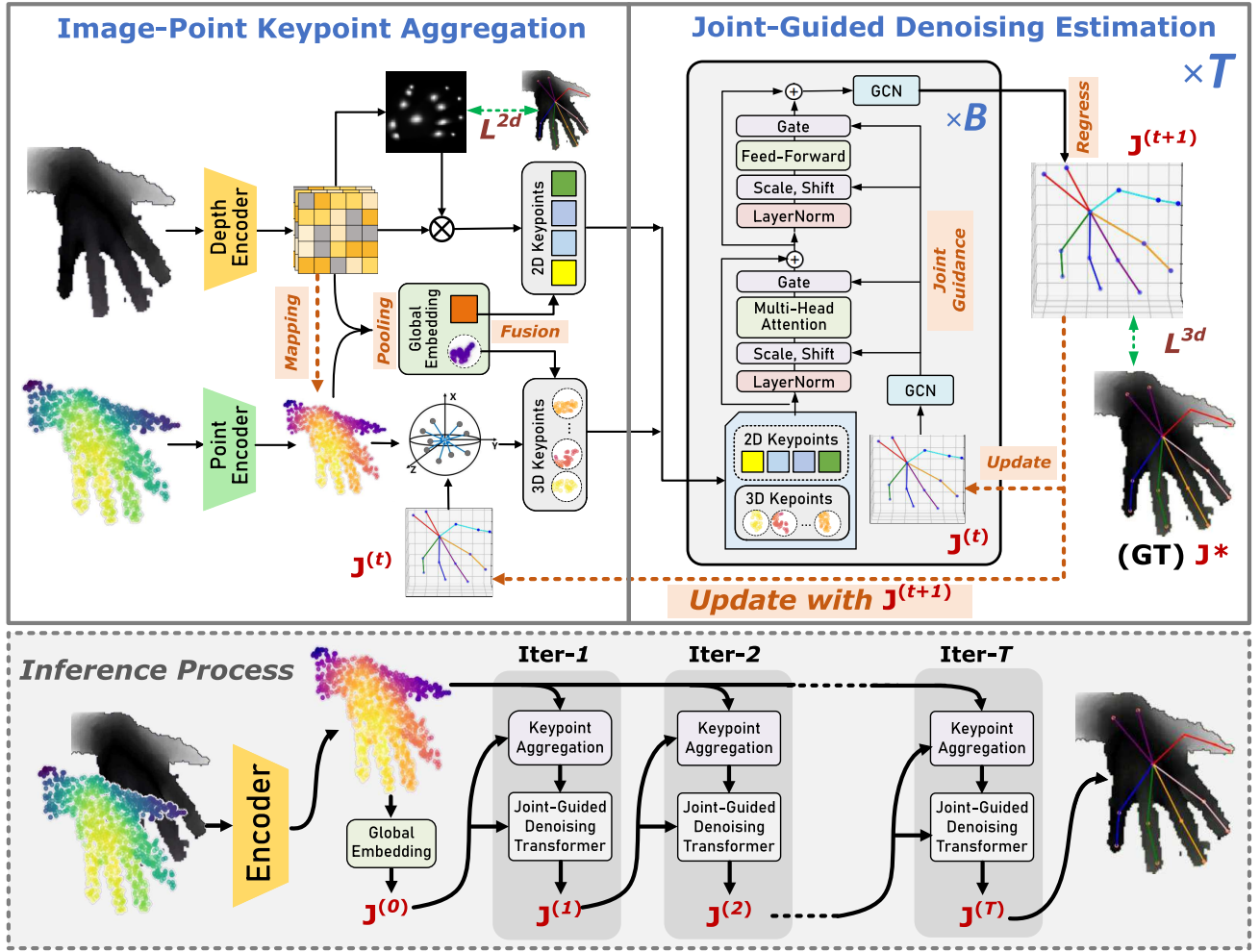


Fig. 2. The overall architecture of HandJoKe. Specifically, the first part learns cross-modal features from both 2D depth images and 3D point clouds and then aggregates local keypoint features around the estimated hand pose joints. The second part introduces a novel joint-guided denoising Transformer to adequately fuse the cross-modal keypoint features and iteratively denoise the estimated hand poses under explicit joint guidance within only several denoising steps.

are treated as noise inputs instead of guidance, and they are iteratively refined over hundreds of denoising steps. Although a few diffusion-based methods have been applied to hand pose estimation and attempted to incorporate multi-modal inputs [9], [35], they typically performed simple feature fusion across modalities, which largely limits their ability to exploit complementary information among different modalities.

In contrast, this paper proposes a joint-guided denoising framework that directly estimates hand joints from both 2D depth images and 3D point clouds within only several denoising steps. Rather than treating 3D hand joints as noisy inputs, HandJoKe estimates from coarse hand joints, and the joint-guided strategy explicitly leverages hand joints as guidance to help the model focus on extracting informative cross-modal features from joint-relevant regions, effectively enhancing the joint-specific feature fusion and improving both structural awareness and estimation accuracy.

Nevertheless, the proposed framework indeed has relationships with conditional diffusion models, since they both generate the target contents by iteratively removing noises conditioned on specific inputs and progressively reconstructing the original data. Moreover, our framework has also achieved

satisfactory results with fewer denoising steps than conventional diffusion models. This may bring some insights to existing conditional generation paradigms for analyzing and understanding images and videos, such as the scene-dependent diffusion model [31] for video anomaly detection.

III. METHODOLOGY

Fig. 2 illustrates the overall architecture of HandJoKe for depth-based 3D hand pose estimation, which consists of two key parts: image-point cloud keypoint aggregation and joint-guided denoising estimation. The details are given as follows.

A. Image-Point Cloud Keypoint Aggregation

a) *Cross-modality feature extraction*: It is essential to first map the input cross-modal data into the shared latent spaces. Specifically, we adopt a ConvNeXT-based auto-encoder [36] to transform the given 2D depth image $\mathbf{D} \in \mathbb{R}^{H \times W}$ into a depth feature map $\mathbf{F}^{2d} \in \mathbb{R}^{C \times H' \times W'}$, where C denotes the channel of the depth feature map. Simultaneously, we also utilize PointNet++ [37] to process the input 3D point cloud $\mathbf{P} \in \mathbb{R}^{N \times 3}$ for obtaining the 3D point features $\mathbf{F}^{3d} \in \mathbb{R}^{N \times D}$,

where N and D denote the number of 3D points and the dimension of point features. It is worth noting that 3D point clouds can feasibly be obtained from depth images, since each pixel in the depth image corresponds to a point in 3D space, under which we can back-project each pixel to a 3D point using the pinhole camera model and the intrinsic matrix.

b) Super point mapping on image-point cloud: Inspired by prior works [7], [8], we further perform an early cross-modal fusion to project 2D depth features $\mathbf{F}^{2d}$ into 3D point features $\mathbf{F}^{2d \rightarrow 3d} \in \mathbb{R}^{N \times D}$, thus obtaining a super point feature map $\tilde{\mathbf{F}}^{3d}$ that integrates the cross-modal 2D-3D features. Specifically, for each 3D point $p_n \in \mathbf{P}$, we select its K nearest neighbours $\mathcal{N}^K(p_n)$ projected from the depth feature map $\mathbf{F}^{2d}$ and aggregate the 2D depth features as

$$\mathbf{F}_n^{2d \rightarrow 3d} = \frac{\sum_{(u,v) \in \mathcal{N}^K(p_n)} \mathbf{F}^{2d}(u,v) / (\|p_{(u,v)}^{3d} - p_n\|_2 + \epsilon)}{\sum_{(u,v) \in \mathcal{N}^K(p_n)} 1 / (\|p_{(u,v)}^{3d} - p_n\|_2 + \epsilon)}, \quad (1)$$

where (u,v) is the 2D coordinate of each pixel in the depth feature map, $\mathbf{F}^{2d}(u,v)$ is the corresponding pixel feature, $p_{(u,v)}^{3d}$ is the corresponding transformed 3D coordinate through the camera intrinsics, $\mathcal{N}^K(p_n)$ contains all the 2D coordinates of K nearest neighbours of p_n , and ϵ is a decimal close to 0. Finally, the super point features $\tilde{\mathbf{F}}^{3d} \in \mathbb{R}^{N \times D}$ by concatenating the 3D point features $\mathbf{F}_n^{2d \rightarrow 3d}$ and $\mathbf{F}^{3d}$ as

$$\tilde{\mathbf{F}}^{3d} = [\mathbf{F}^{3d} \parallel \mathbf{F}^{2d \rightarrow 3d}], \quad (2)$$

where $[\cdot \parallel \cdot]$ denotes the concatenation operation.

c) Dual-modal keypoint selection: We further perform a multi-modal keypoint selection conditioned on the estimated 3D hand joints $\mathbf{J} \in \mathbb{R}^{J \times 3}$ in the 3D space, which focuses on the most informative regions of hands but ignores the irrelevant pixels or points. To extract 2D keypoints from the depth images, we explicitly encourage the model to generate 2D heatmaps $\mathcal{H}_j^{2d}$ that focus on local regions near the j -th hand joint $\mathbf{J}_j$ under the weighted average regression [4] as

$$\mathbf{C}_j = \frac{\sum_{(u,v) \in \Omega} (u,v) \cdot \exp(\mathcal{H}_j^{2d}(u,v))}{\sum_{(u,v) \in \Omega} \exp(\mathcal{H}_j^{2d}(u,v))}, \quad (3)$$

$$\mathcal{L}_{uv}^{2d} = \frac{1}{J} \sum_j \|\mathbf{C}_j - \mathbf{C}_j^*\|_1, \quad (4)$$

where $\mathbf{C}^* \in \mathbb{R}^{J \times 2}$ denotes the ground-truth 2D coordinates of 3D hand joints projected on the feature map, $\mathbf{C}_j$ is the estimated 2D coordinate of the j -th joint, $\Omega \in \mathbb{R}^{H' \times W' \times 2}$ is the 2D coordinate grid of the feature map, and $\mathcal{L}_{uv}^{2d}$ is the 2D ground-truth supervision. It is worth noting that our primary task is to estimate 3D hand pose joints $\mathbf{J}$ instead of their 2D coordinates $\mathbf{C}$, and the motivation of 2D ground-truth supervision $\mathcal{L}_{uv}^{2d}$ is to help the model extract accurate 2D keypoint features. Then, the 2D keypoint feature of the j -th pose joint is aggregated from the depth feature map as

$$\mathbf{K}_j^{2d} = \sum_{(u,v) \in \Omega} \mathcal{H}_j^{2d}(u,v) \cdot \mathbf{F}^{2d}(u,v). \quad (5)$$

Simultaneously, we further obtain the 3D keypoint feature $\mathbf{K}_j^{3d}$ of the j -th joint $\mathbf{J}_j$ by aggregating the features of local

neighbours $\mathcal{N}^{\leq r}(\mathbf{J}_j)$ within the perceived Euclidean distance (i.e. the radius r) from the super point features $\tilde{\mathbf{F}}^{3d}$ as

$$\mathbf{K}_j^{3d} = \frac{\sum_{p_n \in \mathcal{N}^{\leq r}(\mathbf{J}_j)} \tilde{\mathbf{F}}^{3d} / (\|p_n - \mathbf{J}_j\|_2 + \epsilon)}{\sum_{p_n \in \mathcal{N}^{\leq r}(\mathbf{J}_j)} 1 / (\|p_n - \mathbf{J}_j\|_2 + \epsilon)}, \quad (6)$$

where p_n is a 3D neighbour point near the joint $\mathbf{J}_j$. Eventually, cross-modal keypoint features $\mathbf{K}^{2d} \in \mathbb{R}^{J \times D}$ and $\mathbf{K}^{3d} \in \mathbb{R}^{J \times D}$ are calculated conditioned on the 3D pose joints $\mathbf{J}$.

d) Global-local interaction of keypoints: Although the obtained cross-modal keypoint features attain fine-grained local details, they overlook the importance of the global geometric information. Therefore, we further abstract the global embeddings of each input modality and enhance the local keypoint features through the global-local interaction. Specifically, the global embeddings of 2D depth images and 3D point clouds are respectively calculated through the global pooling (probably followed by a multi-layered perception for the feature projection) as

$$\mathbf{k}_g^{2d} = \mathcal{P}ool(\mathbf{F}^{2d}), \quad \mathbf{k}_g^{3d} = \mathcal{P}ool(\tilde{\mathbf{F}}^{3d}), \quad (7)$$

where $\mathcal{P}ool$ denotes the global pooling. Then, global features are modulated with bias-induced layers [38] and fused with cross-modal keypoint features for global-local interaction as

$$\tilde{\mathbf{K}}^{2d} = \mathcal{B}(\mathbf{k}_g^{2d}) + \mathbf{K}^{2d}, \quad \tilde{\mathbf{K}}^{3d} = \mathcal{B}(\mathbf{k}_g^{3d}) + \mathbf{K}^{3d}, \quad (8)$$

where $\mathcal{B}$ denotes the bias-induced layer. Eventually, we obtain the cross-modal keypoint features $\tilde{\mathbf{K}}$ as

$$\tilde{\mathbf{K}} = [\tilde{\mathbf{K}}^{2d} \parallel \tilde{\mathbf{K}}^{3d}] \quad (9)$$

which attains the topology structures with the global-local information in both 2D and 3D spaces.

B. Joint-Guided Denoising Estimation

a) Iterative denoising estimation: HandJoKe can iteratively estimate and denoise hand poses based on keypoint features from both the 2D depth image $\mathbf{D} \in \mathbb{R}^{H \times W}$ and 3D point cloud $\mathbf{P} \in \mathbb{R}^{N \times 3}$ under explicit joint guidance. Specifically, HandJoKe first aggregates the cross-modal keypoint features from the cross-modal inputs $\mathbf{D}$ and $\mathbf{P}$ conditioned on the estimated pose joint $\mathbf{J}^{(t)}$ of the t -th iteration,

$$\tilde{\mathbf{K}}^{(t)} = \mathcal{S}(\mathbf{J}^{(t)}, \mathbf{D}, \mathbf{P}), \quad (10)$$

where $\mathcal{S}$ denotes the cross-modal keypoint aggregation sampler as described in the section III-A, and $\tilde{\mathbf{K}}^{(t)}$ denotes the aggregated cross-modal keypoint features of the t -th iteration. After that, a joint-guided denoising Transformer $\mathcal{T}$ is introduced to denoise $\mathbf{J}^{(t)}$ under explicit joint guidance as

$$\mathbf{J}^{(t+1)} = \mathcal{T}(\tilde{\mathbf{K}}^{(t)}, \mathbf{J}^{(t)}), \quad (11)$$

where $\mathbf{J}^{(t+1)}$ denotes the denoised joints. Particularly, the initial 3D pose joints $\mathbf{J}^{(0)}$ can be directly estimated from the global embedding $\mathbf{k}_g^{3d}$ of 3D super points $\tilde{\mathbf{F}}^{3d}$ through the linear projection as $\mathbf{J}^{(0)} = \delta(\mathbf{k}_g^{3d})$ (where δ denotes a multi-layered perception), and we can denoise the initial hand joints $\mathbf{J}^{(0)}$ for T iterations to obtain the final hand pose joints $\mathbf{J}^{(T)}$.

b) *Joint-guided cross-modal interaction*: The denoising Transformer $\mathcal{T}$ is stacked by B blocks, where each Transformer block adopts the explicit joint guidance in Eq. (11). Specifically, given the estimated hand pose joints $\mathbf{J}$ and the corresponding cross-modal keypoint features $\tilde{\mathbf{K}}$, we first introduce graph convolutional layers to model the kinematic topology information of hand keypoints and then calculate the joint guidance through the multi-layered perception, i.e.,

$$\mathbf{E}^{gn} = \mathcal{G}(\mathbf{J}), \quad (12)$$

$$\mathbf{G}^{scale}, \mathbf{G}^{shift}, \mathbf{G}^{gate} = \delta(\mathbf{E}^{gn}), \quad (13)$$

where $\mathcal{G}$ is a graph convolutional layer and δ is a multi-layered perception. Then, we adaptively modulate the cross-modal keypoint features $\tilde{\mathbf{K}}$ through the multi-head self-attention mechanism under the specific joint guidance as

$$\mathbf{K}^{mod} = (1 + \mathbf{G}^{scale})\tilde{\mathbf{K}} + \mathbf{G}^{shift}, \quad (14)$$

$$\mathbf{K}^{attn} = \mathcal{Attn}(\mathbf{K}^{mod}) \quad (15)$$

$$\mathbf{K}^{guide} = (1 + \mathbf{G}^{gate})\mathbf{K}^{attn}, \quad (16)$$

where $\mathcal{Attn}$ denotes the multi-head attention module. Moreover, we further introduce the adaptive graph convolution layer in each Transformer block to fully model the kinematic topology information of hand keypoints as

$$\hat{\mathbf{K}} = \sigma(\mathbf{A}\mathbf{K}^{guide}\mathbf{\Theta}_K) + \mathbf{K}^{guide}, \quad (17)$$

where $\mathbf{A} \in \mathbb{R}^{J \times J \times D}$ is the learnable graph adjacent matrix that reflects the connection strengths between hand keypoints of different channels, $\mathbf{\Theta}_K \in \mathbb{R}^{D \times D}$ is the transformation weights, and σ is the non-linear activation function. Eventually, the cross-modal keypoint features $\tilde{\mathbf{K}}$ are adequately fused and modulated into $\hat{\mathbf{K}}$ by injecting the explicit joint guidance $\mathbf{J}$ in each Transformer block, which is more effective for denoising the previously estimated hand pose joints.

C. Overall Training Process and Objectives

For each training step, the joint-guided denoising Transformer $\mathcal{T}$ iteratively denoises the initial hand pose joints $\mathbf{J}^{(0)}$ for T iterations to obtain the final prediction $\mathbf{J}^{(T)}$, and the 3D ground-truth joints $\mathbf{J}^*$ are adopted to supervise the estimated joints of each denoising iteration $\mathbf{J}^{(t)}$. The whole network is trained end-to-end with the overall training objective as

$$\mathcal{L}_{all} = \lambda_{2d}\mathcal{L}_{uv}^{2d} + \lambda_{3d}\|\mathbf{J}^{(0)} - \hat{\mathbf{J}}\|_2 + \sum_{t=0}^{T-1} \|\mathcal{T}(\hat{\mathbf{K}}^{(t)}, \mathbf{J}^{(t)}) - \mathbf{J}^*\|_2, \quad (18)$$

where $\mathbf{J}^*$ denotes the annotated 3D joints, and $\mathcal{L}_{uv}^{2d}$ is the 2D supervision in Eq. (4). Empirically, λ_{3d} and λ_{2d} are set to 1.

IV. EXPERIMENTS

A. Experiential Settings

1) *Datasets*: All evaluations are conducted on both **Single-Hand** and **Hand-Object** datasets. Specifically, single-hand datasets provide depth images of single hands captured by 3D sensors, and hand-object datasets are more challenging benchmarks that contain depth images of precise hand-object poses during the interaction. Evaluated datasets are as follows:

Single-Hand Dataset:

- **NYU [10]** consists of 73K training frames and 8.3K test frames of hand poses. The dataset contains various views and annotations, making it suitable for evaluating hand pose estimation models. Additionally, the dataset annotations cover 36 different joints, and we only select 14 joints for evaluation, as in previous works [7], [8].
- **ICVL [11]** consists of 22K training frames and 1.6K testing depth frames. Each frame is annotated with 16 joints, including one palm joint and three joints for each finger. The dataset comprises 10 subjects, each recorded from different participants. For each subject, 26 different hand poses are included.

Hand-Object Dataset:

- **DexYCB [12]** is a large-scale dataset designed for capturing hand-object grasping, which consists of 582K frames from more than 1000 sequences captured from 8 different viewpoints with 20 distinct objects for grasping tasks. The dataset is officially divided into four configurations: **S0** (default) includes all viewpoints and subjects, **S1** (unseen subjects) splits by subjects, **S2** (unseen views) splits by different camera viewpoints, and **S3** (unseen grasping) splits by the grasped objects. We conducted experiments on all four splits to comprehensively evaluate the estimation performance.
- **HO3D [13]** is another widely used hand-object dataset, which is composed of sequences in which different subjects manipulate various objects. Similar to previous works [8], we only use the HO3D_v2 version that annotates more than 65 sequences captured with 10 different subjects and 10 objects, resulting in approximately 80K images. Moreover, the dataset is divided into approximately 70K training images and 10K testing images. Evaluation of different methods on HO3D was conducted on the official online submission platform.

2) *Evaluation Metrics*: The Mean Per Joint Position Error (MPJPE) is a widely used metric in hand pose estimation, and therefore, we employ it as the main metric.

3) *Implementation Details*: All experiments were conducted on a workstation with an NVIDIA RTX 4090 GPU, and the whole architecture was implemented with PyTorch. The default configuration of HandJoKe on NYU, DexYCB, and HO3D is with the denoising iteration of 3 and block depth of 3, i.e., ($T=3, B=3$), and the configuration on ICVL is with ($T=1, B=2$). During training, the model was optimized with AdamW [39] with a learning rate of 0.0005 and $(\beta_1, \beta_2) = (0.5, 0.999)$. The input depth images were resized to a resolution of 128×128 , and the point clouds were randomly sampled to 1024 points. Moreover, online data augmentation techniques were employed to address overfitting, including random rotations ($[-180, 180]$), 3D scaling ($[0.8, 1.2]$), and 3D translations ($[-10, 10]$). We set the batch size to 32 and optimized the model for 60 epochs on NYU & ICVL and 30 epochs on DexYCB & HO-3D to reach convergence.

B. Comparison With State-of-the-Arts

1) *Results on Single-Hand Datasets*: As listed in Table I, we compared the proposed HandJoKe with previous SOTA

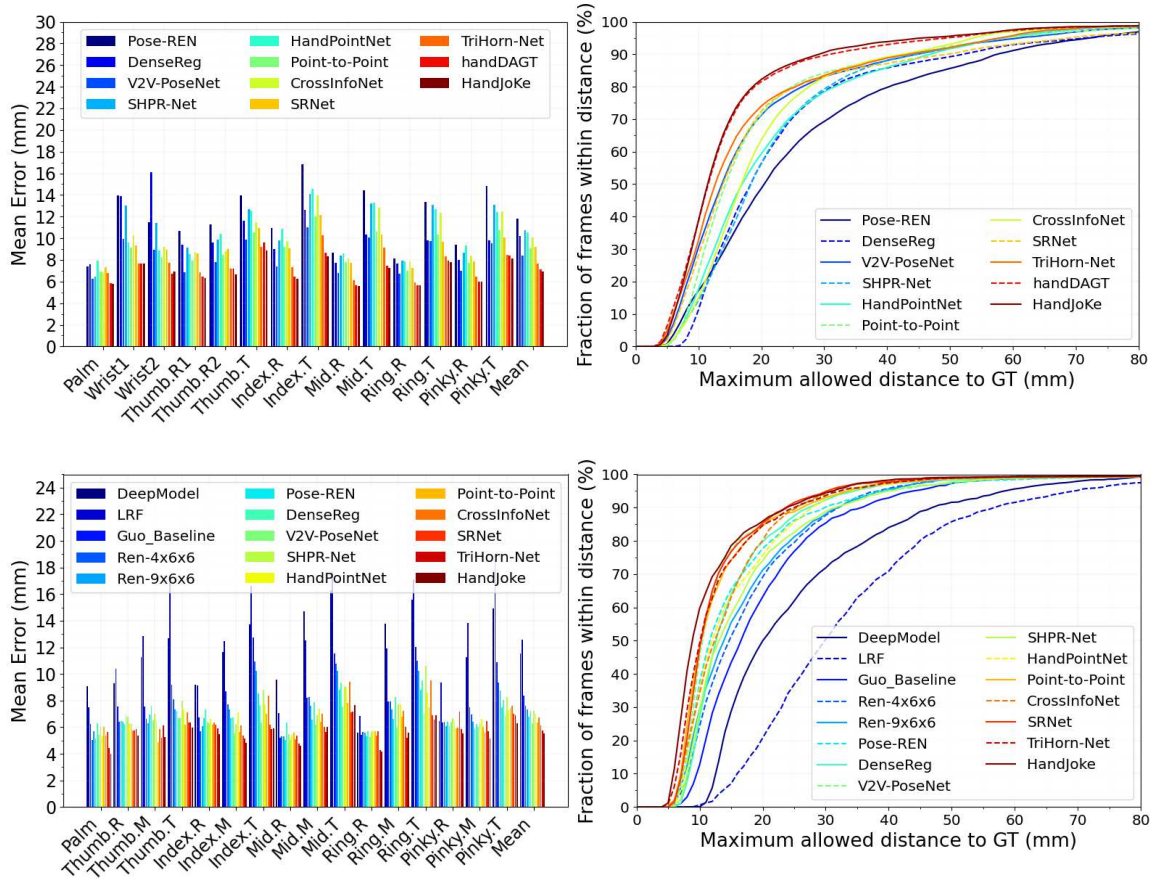


Fig. 3. Comparison with SOTA methods of the per-keypoint error and success rate under different thresholds on NYU (top) and ICVL (bottom).

TABLE I
COMPARISON WITH SOTA METHODS ON THE SINGLE-HAND DATASETS

Method	Input	MPJPE (mm) ↓	
		ICVL	NYU
DepthPrior++ [18]	Depth	8.10	12.24
Pose-REN [3]	Depth	6.79	11.81
DenseReg [15]	Depth	7.30	10.21
SRN [17]	Depth	6.26	7.79
A2J [40]	Depth	6.46	8.61
A2J-Transformer [41]	Depth	-	8.43
JGR-P2O [2]	Depth	6.02	8.29
TriHorn-Net [42]	Depth	5.73	7.68
AWR [4]	Depth	5.98	7.48
HandVoxNet [43]	Voxel	-	8.72
V2V-PoseNet [20]	Voxel	6.28	8.41
HandPointNet [23]	Point	6.94	10.54
HandTransformer [44]	Point	6.47	9.80
Point-to-Point [24]	Point	6.30	9.10
HandFolding [21]	Point	5.95	8.58
HandR2N2 [45]	Point	5.70	7.27
HandDiff [9]	Depth&Point	5.72	7.38
IPNET [7]	Depth&Point	5.76	7.17
HandDAGT [8]	Depth&Point	5.66	7.12
HandJoKe (Ours)	Depth&Point	5.53	6.97

methods on single-hand datasets NYU and ICVL. Moreover, we also drew the per-keypoint error and success rate under

different thresholds of different methods on NYU & ICVL as shown in Fig. 3. Additionally, we omitted some methods in Fig. 3 due to the absence of result files of those methods. It can be observed that the cross-modal methods significantly outperform the traditional unimodal methods (that adopted the 2D depth images or 3D data alone), demonstrating the effectiveness of multi-modal fusion for the 3D hand pose estimation. Moreover, the proposed HandJoKe also outperforms previous cross-modal methods and achieves the best performance with the lowest MPJPE, demonstrating the effectiveness of our joint-guided denoising estimation.

2) *Results on Hand-Object Datasets:* To further demonstrate the effectiveness of our method, we evaluated HandJoKe on more challenging hand object datasets. Specifically,

a) *DexYCB:* As listed in Table II, we compared HandJoKe with previous SOTA methods on the hand-object dataset DexYCB. For a fair comparison, we only compared the methods that were experimented with under the official dataset split protocols. It can be observed that HandJoKe largely outperforms previous SOTA methods under all protocols, demonstrating that HandJoKe can accurately estimate 3D hand poses for more challenging hand-object interaction scenarios with various occlusions and different objects.

b) *HO3D:* As listed in Table III, we compared HandJoKe with previous SOTA methods that have reported MPJPE on the HO3D dataset. Additionally, we also listed the results of

TABLE II
COMPARISON WITH SOTA METHODS OF THE MPJPE (MM) ON THE HAND-OBJECT DATASET DEXYCB [12]

Method	Input	MPJPE (mm) ↓				
		S0	S1	S2	S3	Average
A2J [40]	Depth	23.93	25.57	27.65	24.92	25.52
BC [46]	RGB	17.34	22.26	25.49	18.44	18.44
METRO [47]	RGB	15.24	-	-	-	-
AttnGconv [48]	RGB	16.05	21.22	27.01	17.93	20.55
HandOccNet [49]	RGB	14.04	-	-	-	-
IPNET [7]	Depth&Point	8.03	9.01	8.60	7.80	8.36
HandDAGT [8]	Depth&Point	7.72	8.68	8.22	7.52	8.03
HandDiff [9]	Depth&Point	7.66	8.73	8.40	7.53	8.07
HandJoKe (Ours)	Depth&Point	7.28	8.42	7.62	7.27	7.65

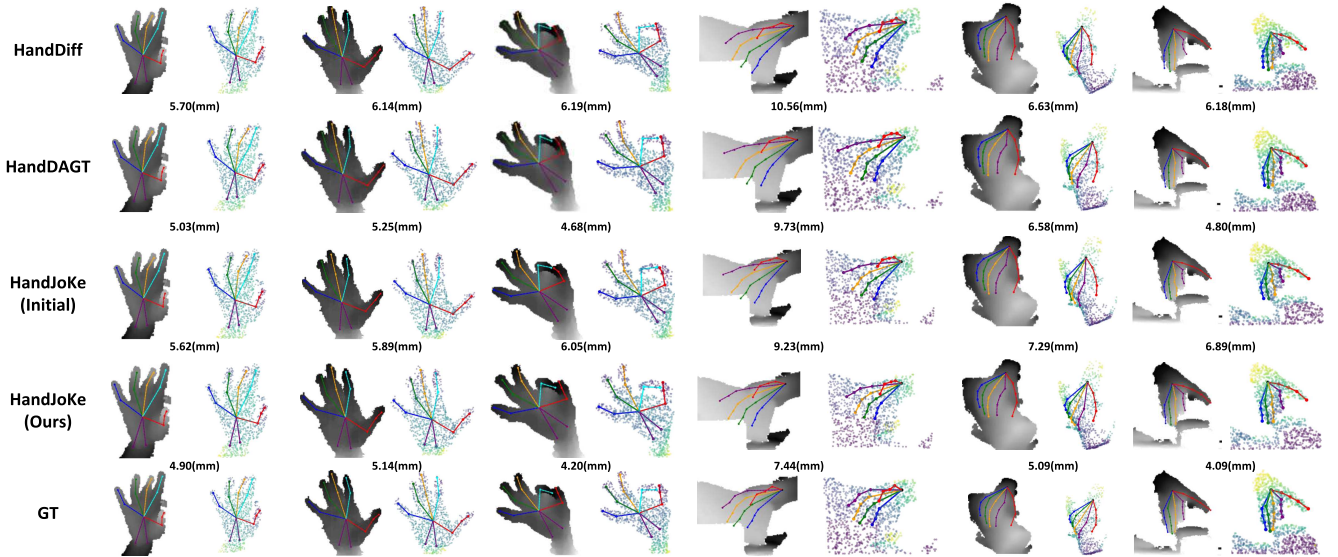


Fig. 4. Qualitative comparison of different methods on NYU and DexYCB.

TABLE III
COMPARISON WITH SOTA METHODS ON HO3D [13]

Method	Input	MPJPE (mm) ↓
Hybrik [50]	RGB	2.89
ArtiBoost [51]	RGB	2.53
HandOccNet [49]	RGB	2.49
HandVoxNet++ [52]	Voxel	2.46
IPNET [7]	Depth&Point	1.81
HandDAGT [8]	Depth&Point	1.81
HandJoKe (Ours)	Depth&point	1.75
HandJoKe* (Ours)	Depth&point	1.33

* denotes the results with annotated cropped regions.

HandJoKe that were predicted from the inputs using both estimated and ground-truth cropped hand regions in Table III. It can be observed that HandJoKe outperforms the existing methods and achieves the new SOTA performance, further demonstrating its effectiveness for depth-based 3D hand estimation.

3) *Qualitative Comparison of Different Methods:* As shown in Fig. 4, we presented the qualitative comparison of different

methods. We can observe that: i) for fine-grained poses (e.g., the 1st and 3rd columns) and poses with finger occlusions (e.g., the 5th and 9th columns), HandJoKe can accurately estimate joint positions with a stronger perception of 3D geometric structures; ii) for complex poses with severe hand-object occlusions (e.g., the 7th and 9th columns), previous SOTA methods are inferior to our HandJoKe, since our method achieves better geometric-facilitated cross-modal key-point fusion and interaction under explicit joint guidance; iii) furthermore, the 3rd and 4th rows show that the iterative denoising significantly improves the estimation accuracy, since the initial hand joints can be subsequently updated under explicit joint guidance and thus can better locate confused joints, especially for occluded hands. Overall, experimental results demonstrate the effectiveness of HandJoKe.

C. Ablation Study

1) *Effectiveness and Time Cost of Cross-Modal Fusion:* Both 2D depth images and 3D point clouds can benefit 3D hand estimation, and these two distinct modalities can complement each other. To demonstrate the effectiveness of multi-modal fusion, we compared the hand estimation results

TABLE IV

EFFECTIVENESS AND TIME COST OF CROSS-MODAL FUSION ON NYU

ID	Input Modality		Network [†] Architecture	Speed* (FPS) $\uparrow$		MPJPE (mm) $\downarrow$
	Depth	Point		Train	Test	
1	✓	×	Enc.	183.8	277.4	8.71
2	×	✓	Enc.	990.5	1459.1	9.98
3	✓	✓	Enc.	151.6	147.0	7.49
4	✓	×	Enc. + Dec.	137.3	64.5	8.44
5	×	✓	Enc. + Dec.	326.8	76.6	7.69
6	✓	✓	Enc. + Dec.	112.7	52.7	6.97

* During comparison, the training batch size is 32 and testing is 1.

[†] “Enc.” corresponds to estimating with the feature extractor only, and “Dec.” to the joint-guided denoising decoder with ($T=3$, $B=3$).

TABLE V

ABLATION RESULTS OF DIFFERENT NETWORK ARCHITECTURES ON NYU

ID	Global-Local Interaction	Multi-Head Attention	Joint Guidance	Iterative Denoising	MPJPE (mm) $\downarrow$
1	×	×	×	×	10.68
2	✓	×	×	×	8.18
3	✓	✓	×	×	7.49
4	✓	✓	×	✓	7.37
5	✓	✓	✓	×	7.27
6	×	✓	✓	✓	7.14
7	✓	✓	✓	✓	6.97

of using different input modality data in Table IV. It can be observed that multi-modal approaches largely outperform uni-modal approaches (that adopt either 2D depth images or 3D point clouds alone). Multi-modal approaches can achieve better performance by further integrating the proposed joint-guided denoising strategy, which can more adequately fuse the cross-modal features. Moreover, comparison with previous SOTAs in Table I~III also demonstrates the superiority of the cross-modal fusion for depth-based 3D hand pose estimation.

Although multi-modal fusion improves the estimation precision, it inevitably increases computational time cost. As shown in Table IV, we compared the training and inference FPS (Frames Per Second) regarding whether leveraging the multi-modal data. The evaluation was conducted on an NVIDIA RTX 4090 GPU with a training batch size of 32 and a test batch size of 1. We can observe that i) it is more computationally expensive to process depth images than point clouds, since the feature extractor of images is much larger than that of 3D points; ii) it achieves much higher precision but introduces more time cost when leveraging the multi-modal data; and iii) the difference between unimodal and multi-modal scenarios becomes less significant after introducing the joint-guided denoising Transformer with iterative estimation. Nevertheless, our cross-modal approaches can meet the requirement for processing real-time videos, achieving 52.7 FPS and a much higher precision of 6.97 mm.

2) *Ablation Study on Network Architecture*: To verify the effectiveness of each key component in HandJoKe, we have conducted an ablation study on different network architectures in Table V. We can observe that the performance gradually improves by integrating different key components.

TABLE VI

ABLATION RESULTS OF MPJPE (MM) WITH DIFFERENT INITIALIZED $\mathbf{J}^{(0)}$

Initial 3D joints $\mathbf{J}^{(0)}$	NYU	ICVL	DexYCB				
			S0	S1	S2	S3	Avg
Randomly Estimated	7.32	5.85	7.62	8.91	7.98	7.58	8.02
	6.97	5.53	7.28	8.42	7.62	7.27	7.65

Specifically, i) ID 1~2 & 6~7 reveal that the global-local interaction improves the estimation performance by adequately fusing the input data with richer informative information; ii) ID 4~5 & 7 indicates that the joint guidance is extremely effective for extracting precise geometric features and learning correct joint correlations of hands by explicitly injecting the geometric information; iii) ID 5 & 7 show that HandJoKe is compatible with iterative denoising strategy and can achieve better performance by iteratively refining the estimated hands.

Furthermore, we presented the qualitative results of iterative denoising in Fig. 5. It can be observed that: i) it consistently refines previously estimated hand poses by gradually increasing the denoising iterations, and the improvement is more apparent for complex poses with severe self-occlusions (e.g., the 3rd column) or hand-object interaction (e.g., the 7th and 9th columns), thus demonstrating the effectiveness of our joint guidance strategy; ii) however, the improvement becomes marginal for the last denoising iteration, since the estimated hand poses have already been accurate enough, leaving limited room for improvement.

As shown in Fig. 6, we also provided the qualitative results of using attention and joint guidance. We can observe that: i) the attention mechanism consistently improves the estimation precision (e.g., the 1st and 2nd columns), since the attention mechanism is largely effective in fusing and enhancing cross-modal features; ii) the joint guidance can further largely refine incorrect hand poses, especially for the complex scenarios with self-occlusions (e.g., the 2nd and 3rd rows). This is because hand joints are explicitly injected as prior geometric guidance, thus achieving geometric-facilitated fusion and interaction.

3) *Importance of Initial 3D Joint Estimation*: As shown in Table VI, we further demonstrate that a precise initial 3D joint estimation $\mathbf{J}^{(0)}$ is crucial for HandJoKe to achieve satisfactory performance. This is because HandJoKe not only aggregates keypoint features conditioned on the previous estimated hand joints $\mathbf{J}^{(i)}$ but also iteratively denoises the initial hand joints $\mathbf{J}^{(0)}$ for the final pose estimation as shown in Fig. 2. Therefore, a randomly initialized $\mathbf{J}^{(0)}$ will lead HandJoKe to aggregate incorrect keypoint features and denoise from an inferior starting point, which eventually causes inefficient feature learning and falls into worse local optima. Experimental results in Table VI demonstrate the importance of initial 3D joint estimation.

4) *Efficiency Comparison and Analysis*: We further evaluate and analyze the efficiency of our proposed framework to provide a comprehensive assessment.

a) *Efficiency of cross-modal fusion over different scales*: To demonstrate the efficiency of keypoint-based cross-modal

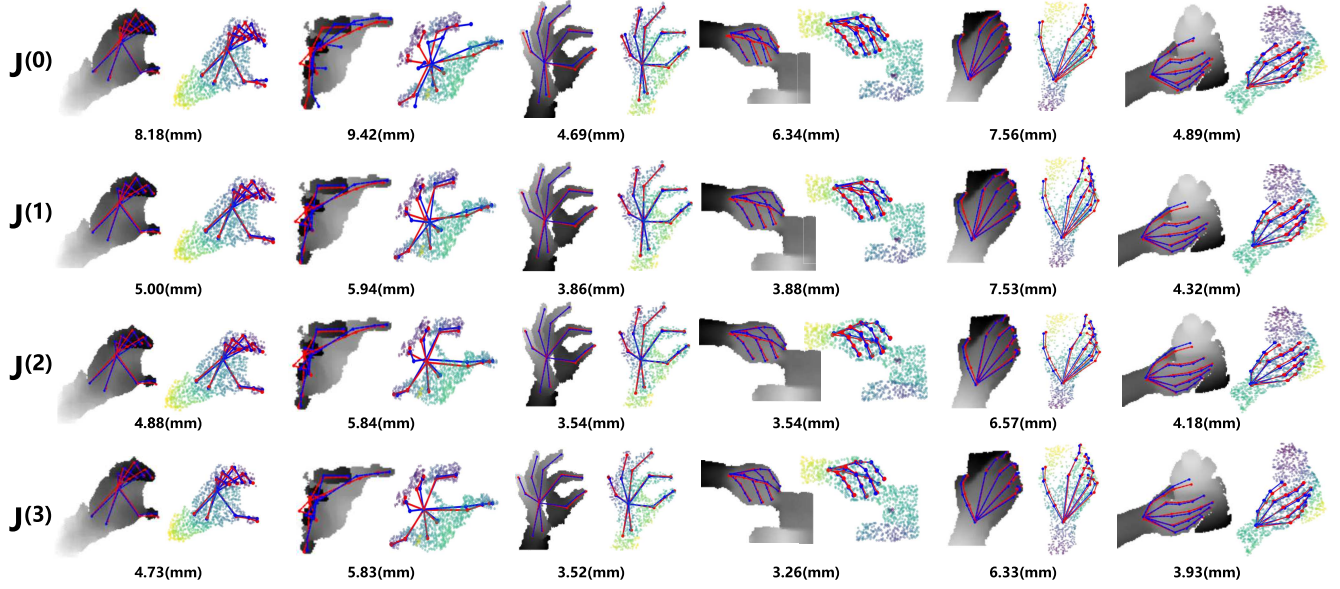


Fig. 5. Qualitative results of iterative denoising with different iterations on NYU and DexYCB.

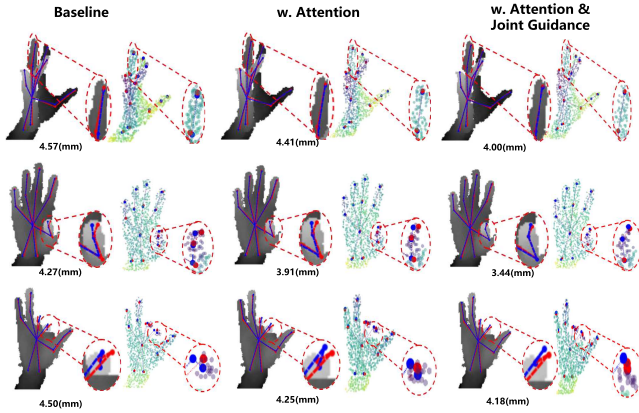


Fig. 6. Qualitative results of using attention and joint guidance on NYU.

TABLE VII
EFFICIENCY OF CROSS-MODAL FUSION OVER DIFFERENT SCALES

ID	Fusion Scale	Denoising Steps	GPU Mem. (GB)		Speed (FPS) ↑	MPJPE (mm) ↓
			Train [†]	Test		
1	Whole	1	OOM*	2.6	75.9	×
2	Whole	3	OOM*	3.9	45.8	×
3	Keypoints	1	3.7	1.1	93.5	7.27
4	Keypoints	3	5.1	1.3	52.7	6.97

[†] During comparison, the training batch size is 4 and testing is 1.
* “OOM” denotes “out of memory”, which exceeds the physical memory of RTX 4090 GPU (i.e., > 24.0 GB).

fusion, we compared the efficiency between fusion over the whole data and keypoints as shown in Table VII. We observed that cross-modal fusion over keypoints attains faster inference speed and lower training and testing GPU memory usage than fusion over the whole data. More critically, fusion over the whole data suffers from Out-of-Memory (OOM) issues, even with a training batch size of 4, and its training GPU memory

TABLE VIII
EFFICIENCY COMPARISON WITH PREVIOUS MODELS ON NYU

Method	Input Modality	Denoising Iterations	Speed (FPS) ↑	MPJPE (mm) ↓
A2J-Transformer [40]	Depth	×	48.2	8.43
V2V-PoseNet [20]	Voxel	×	21.3	8.41
Point-to-Point [24]	Point	×	41.8	9.10
HandFolding [21]	Point	3	84.0	8.58
HandDiff [9]	Depth&Point	$10 \times 2^*$	19.8	7.42
HandDiff [9]	Depth&Point	$10 \times 10^*$	4.8	7.38
HandDAGT ($T=1$) [8]	Depth&Point	1	90.1	8.02
HandDAGT ($T=3$) [8]	Depth&Point	3	79.0	7.12
HandJoKe ($T=1$)	Depth&Point	1	93.5	7.27
HandJoKe ($T=3$)	Depth&Point	3	52.7	6.97

* $m \times n$ denotes n hypotheses with m iterations for each in HandDiff.

is at least 4 times greater than that of keypoint-based fusion. This is because it inevitably introduces redundant interactions among many irrelevant pixels and points when fusing over the whole data; in contrast, keypoint-based fusion focuses on more informative regions of hands and ignores invalid pixels and points, thereby significantly reducing the computational redundancy.

b) *Efficiency Comparison with previous models:* As shown in Table VIII, we also compared HandJoKe with previous models regarding the computation efficiency and accuracy. Additionally, we only compared with several typical and SOTA methods that provided official implementations or reported computation costs. Specifically, we can observe that: i) compared with unimodal approaches (like A2J-Transformer, V2V-PoseNet, etc.), HandJoKe achieves much higher precision with slightly faster inference speed, where cross-modal methods fuse more data with higher accuracy but also introduce additional computation cost. ii) Compared with diffusion models like HandDiff, our method attains much faster inference

TABLE IX
RESULTS OF MPJPE (MM) WITH DIFFERENT DENOISING ITERATIONS

Different Iterations	Large		Small
	NYU	YCB-S0	ICVL
$T=1$	7.27	7.81	5.53
$T=2$	7.12	7.41	5.79
$T=3$	6.97	7.28	6.10
$T=4$	7.11	7.32	6.22

TABLE X
RESULTS OF MPJPE (MM) WITH DIFFERENT BLOCK DEPTHS

Different Block Depths	Large		Small
	NYU	YCB-S0	ICVL
$B=1$	7.30	7.45	5.74
$B=2$	7.15	7.39	5.53
$B=3$	6.97	7.28	5.91
$B=4$	7.01	7.44	6.02

speed and higher precision. This is because HandDiff needs to generate multiple hypotheses with dozens of denoising steps for each, ultimately resulting in hundreds of denoising iterations. iii) Compared with previous SOTA methods like HandDAGT, HandJoKe achieves considerably higher precision but with comparable inference speed, and we notice that there is a trade-off between precision and efficiency. Nevertheless, our proposed HandJoKe can meet the requirements for processing real-time videos.

5) *Ablation Study on Network Configurations*: We further evaluated the ablation results of HandJoKe with different configurations on the datasets with different scales, e.g., the large datasets NYU and DexYCB and the small one ICVL.

a) *Different iterations of joint-guided denoising*: As listed in Table IX, we evaluated the results of HandJoKe with different denoising iterations. Specifically, i) for the large datasets, it can be observed that the performance can be gradually improved by increasing the denoising iterations (under which the estimated hands can be better refined), while further increasing the denoising iterations over a threshold may hurt the overall performance; ii) For the small dataset ICVL, we can observe that increasing the denoising iterations is more likely to decrease the estimation performance. Nevertheless, larger denoising iterations also lead to higher computation costs and slower inference speeds.

b) *Different transformer block depths*: As listed in Table X, we also evaluated the results with different numbers of Transformer blocks. We observed that i) for large datasets, deeper Transformers will lead to better performance, since deeper networks can learn more intricate features. Nevertheless, the improvement becomes marginal when stacking over three blocks. ii) For the small dataset ICVL, it is difficult to further improve estimation performance by stacking deeper networks, probably due to overfitting issues.

c) *Result Analysis regarding datasets scales*: It is worth noting that ICVL is significantly smaller than NYU and DexYCB (as described in section IV-A.1). Particularly, we can observe that increasing the denoising iterations and block depths is more likely to i) gradually improve the performance

on large datasets NYU and DexYCB ii) but decrease the precision on the small dataset ICVL, as shown in Tables IX&X. The major reason is that small datasets are easy to overfit, thus making the improvement and difference of various models on ICVL marginal. Instead, large datasets like NYU and DexYCB can better evaluate the effectiveness of different models.

V. CONCLUSION

This paper proposes a novel joint-guided keypoint denoising Transformer HandJoKe for depth-based 3D hand pose estimation. The proposed HandJoKe can iteratively estimate and denoise hand poses based on keypoint features from both 2D depth images and 3D point clouds under explicit joint guidance within only several denoising steps. Specifically, we first introduce an image-point cloud keypoint aggregation strategy to locate keypoint features from the cross-modal inputs based on the estimated hand pose joints. Secondly, we further propose a novel joint-guided denoising estimation strategy to adequately fuse cross-modal keypoint features through explicit joint guidance and iteratively denoise estimated hands with cross-dimensional interaction in both 2D and 3D spaces. Extensive experiments on public benchmarks demonstrate that HandJoKe outperforms previous SOTA methods for depth-based 3D hand pose estimation. Finally, we hope the proposed HandJoKe can bring some valuable insights into depth-based 3D hand pose estimation, including the explicit joint guidance, iterative denoising estimation, and cross-modal fusion. In our future work, we plan to further extend our method to more complex scenarios, such as multi-hand interactions in dynamic environments.

REFERENCES

- [1] P. Ren, H. Sun, J. Hao, Q. Qi, J. Wang, and J. Liao, "Pose-guided hierarchical graph reasoning for 3-D hand pose estimation from a single depth image," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 315–328, Jan. 2023.
- [2] L. Fang, X. Liu, L. Liu, H. Xu, and W. Kang, "JGR-P2O: Joint graph reasoning based pixel-to-offset prediction network for 3D hand pose estimation from a single depth image," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 120–137.
- [3] X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neuro-computing*, vol. 395, pp. 138–149, Jun. 2020.
- [4] W. Huang, P. Ren, J. Wang, Q. Qi, and H. Sun, "AWR: Adaptive weighting regression for 3D hand pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, pp. 11061–11068.
- [5] Z. Chen et al., "AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection," 2022, *arXiv:2201.06493*.
- [6] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," 2022, *arXiv:2203.04838*.
- [7] P. Ren et al., "Two heads are better than one: Image-point cloud network for depth-based 3D hand pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, pp. 2163–2171.
- [8] W. Cheng, E. Kim, and J. H. Ko, "HandDAGT: A denoising adaptive graph transformer for 3D hand pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 35–52.
- [9] W. Cheng, H. Tang, L. V. Gool, and J. H. Ko, "HandDiff: 3D hand pose estimation with diffusion on image-point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 2274–2284.
- [10] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Trans. Graph.*, vol. 33, no. 5, pp. 1–10, Sep. 2014.
- [11] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3D articulated hand posture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3786–3793.

- [12] Y.-W. Chao et al., "DexYCB: A benchmark for capturing hand grasping of objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 9044–9053.
- [13] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "HONotate: A method for 3D annotation of hand and object poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3193–3203.
- [14] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4512–4516.
- [15] C. Wan, T. Probst, L. V. Gool, and A. Yao, "Dense 3D regression for hand pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2018, pp. 5147–5156.
- [16] K. Du, X. Lin, Y. Sun, and X. Ma, "CrossInfoNet: Multi-task information sharing based hand pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9888–9897.
- [17] P. Ren et al., "Spatial-aware stacked regression network for real-time 3D hand pose estimation," *Neurocomputing*, vol. 437, pp. 42–57, May 2021.
- [18] M. Oberweger and V. Lepetit, "DeepPrior++: Improving fast and accurate 3D hand pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 585–594.
- [19] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3D convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5679–5688.
- [20] J. Y. Chang, G. Moon, and K. M. Lee, "V2V-PoseNet: Voxel-to-Voxel prediction network for accurate 3D hand and human pose estimation from a single depth map," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5079–5088.
- [21] W. Cheng, J. H. Park, and J. H. Ko, "HandFoldingNet: A 3D hand pose estimation network using multiscale-feature guided folding of a 2D hand skeleton," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2021, pp. 11240–11249.
- [22] X. Chen, G. Wang, C. Zhang, T.-K. Kim, and X. Ji, "SHPR-net: Deep semantic hand pose regression from point clouds," *IEEE Access*, vol. 6, pp. 43425–43439, 2018.
- [23] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand PointNet: 3D hand pose estimation using point sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8417–8426.
- [24] L. Ge, Z. Ren, and J. Yuan, "Point-to-point regression PointNet for 3D hand pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 489–505.
- [25] Y. Que, L. Xiong, W. Wan, X. Xia, and Z. Liu, "Denoising diffusion probabilistic model for face sketch-to-photo synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 10424–10436, Oct. 2024.
- [26] S. Kang, S. Gao, W. Wu, X. Wang, S. Wang, and G. Qiu, "Image intrinsic components guided conditional diffusion model for low-light image enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 12, pp. 13244–13256, Dec. 2024.
- [27] Q. Yan et al., "Efficient image enhancement with a diffusion-based frequency prior," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Mar. 7, 2025, doi: [10.1109/TCSVT.2025.3549351](https://doi.org/10.1109/TCSVT.2025.3549351).
- [28] M. Zhang et al., "Spatial-spectral aggregation transformer with diffusion prior for hyperspectral image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3557–3572, Apr. 2025.
- [29] L. Papa, P. Russo, and I. Amerini, "D4D: An RGBD diffusion model to boost monocular depth estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9852–9865, Oct. 2024.
- [30] S. Shao, Z. Pei, W. Chen, D. Sun, P. C. Y. Chen, and Z. Li, "MonoDiffusion: Self-supervised monocular depth estimation using diffusion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 4, pp. 3664–3678, Apr. 2025.
- [31] C. Cao, H. Zhang, Y. Lu, P. Wang, and Y. Zhang, "Scene-dependent prediction in latent space for video anomaly detection and anticipation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 224–239, Jan. 2025.
- [32] W. Shan et al., "Diffusion-based 3D human pose estimation with multi-hypothesis aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 14715–14725.
- [33] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "DiffPose: Toward more reliable 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 13041–13051.
- [34] J. Choi, D. Shim, and H. J. Kim, "DiffuPose: Monocular 3D human pose estimation via denoising diffusion probabilistic model," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Aug. 2023, pp. 3773–3780.
- [35] L. Li, L. Zhuo, B. Zhang, L. Bo, and C. Chen, "DiffHand: End-to-end hand mesh reconstruction via diffusion models," 2023, *arXiv:2305.13705*.
- [36] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.
- [37] C. R. Qi, Y. Li, H. Su, and L. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–12.
- [38] W. Cheng and S. Lee, "Point auto-encoder and its application to 2D-3D transformation," in *Proc. Adv. Vis. Computing: Int. Symp. Vis. Comput.*, 2019, pp. 66–78.
- [39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
- [40] F. Xiong et al., "A2J: Anchor-to-Joint regression network for 3D articulated pose estimation from a single depth image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jul. 2019, pp. 793–802.
- [41] C. Jiang et al., "A2J-transformer: Anchor-to-Joint transformer network for 3D interacting hand pose estimation from a single RGB image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 8846–8855.
- [42] M. Rezaei, R. Rastgoo, and V. Athitsos, "TriHorn-net: A model for accurate depth-based 3D hand pose estimation," *Expert Syst. Appl.*, vol. 223, Aug. 2023, Art. no. 119922.
- [43] J. Malik et al., "HandVoxNet: Deep voxel-based network for 3D hand shape and pose estimation from a single depth map," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7113–7122.
- [44] L. Huang, J. Tan, J. Liu, and J. Yuan, "Hand-transformer: Non-autoregressive structured modeling for 3D hand pose estimation," in *Proc. 16th ECCV*. Cham, Switzerland: Springer, 2020, pp. 17–33.
- [45] W. Cheng and J. H. Ko, "HandR2N2: Iterative 3D hand pose estimation using a residual recurrent neural network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jul. 2023, pp. 20904–20913.
- [46] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz, "Weakly supervised 3D hand pose estimation via biomechanical constraints," in *Proc. Eur. Conf. Comput. Vis.*, Jun. 2020, pp. 211–228.
- [47] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1954–1963.
- [48] T. H. E. Tse, K. I. Kim, A. Leonardis, and H. J. Chang, "Collaborative learning for hand and object reconstruction with attention-guided graph convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1654–1664.
- [49] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "HandOccNet: Occlusion-robust 3D hand mesh estimation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1496–1505.
- [50] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybriK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 3383–3393.
- [51] L. Yang et al., "ArtiBoost: Boosting articulated 3D hand-object pose estimation via online exploration and synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2750–2760.
- [52] J. Malik et al., "HandVoxNet++: 3D hand shape and pose estimation using voxel-based neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8962–8974, Dec. 2022.



Ji Gan received the B.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 2015, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include pattern recognition, human-computer interaction, handwriting recognition, and generation.



Lei Chen received the B.E. degree. He is currently pursuing the master's degree from Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include computer vision and 3D hand pose estimation.



Weisheng Li (Member, IEEE) received the B.E. and M.S. degrees from the School of Electronics and Mechanical Engineering, Xidian University, Xi'an, China, in July 1997 and July 2000, respectively, and the Ph.D. degree from the School of Computer Science and Technology, Xidian University, in July 2004. He is currently a Professor with Chongqing University of Posts and Telecommunications. His research interests include intelligent information processing and pattern recognition.



area chair for CVPR 2024.

Ping Hu (Member, IEEE) received the B.E. degree from Sichuan University, China, the M.Eng. degree from the University of Chinese Academy of Sciences, China, and the Ph.D. degree from Boston University, USA. He is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, and a Post-Doctoral researcher with Boston University. His research interests include machine learning and computer vision. He serves as a senior program committee member for IJCAI 2023 and an



Xinbo Gao (Fellow, IEEE) received the B.E., M.S., and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Elec-

tronic Engineering with Xidian University. He is currently a Cheung Kong Professor of the Ministry of Education of China and a Professor of pattern recognition and intelligent systems with Xidian University. Since 2020, he has also worked at Chongqing University of Posts and Telecommunications. He has published seven books and around 300 technical articles in refereed journals and proceedings. His research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition. He is a fellow of the Institute of Engineering and Technology, Chinese Institute of Electronics, China Computer Federation, and Chinese Association for Artificial Intelligence. He also served as the general chair/co-chair, the Program committee chair/co-chair, or a PC member for around 30 major international conferences. He is also on the editorial boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He also served as the general chair/co-chair, the Program committee chair/co-chair, or a PC member for around 30 major international conferences.



Jiaxu Leng (Member, IEEE) received the B.E. degree in electronic information science and technology from Tianshui Normal University, Tianshui, China, in 2012, the M.S. degree in electrical engineering from the University of Electronic Science and Technology, Sichuan, China, in 2015, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2020. After receiving his master's degree, he was with Hisense, Qingdao, China, for two years as an algorithm engineer. He is currently with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include computer vision, deep learning, object detection, and stereo vision.