applied sciences

MDPI

Article

A Normalization Strategy for Weakly Supervised 3D Hand Pose Estimation

Zizhao Guo [1], Jinkai Li [2] and Jiyong Tan [3,*]

[1] College of Computer Science, Chengdu University, Chengdu 610106, China; guozzscu25@gmail.com
[2] College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China; smilelijinkai@gmail.com
[3] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 610056, China
* Correspondence: scutjy2015@163.com

Abstract: The effectiveness of deep neural network models is intricately tied to the distribution of training data. However, in pose estimation, potential discrepancies in root joint positions and inherent variability in biomechanical features across datasets are often overlooked in current training strategies. To address these challenges, a novel Hand Pose Biomechanical Model (HPBM) is developed. In contrast to the traditional 3D coordinate-encoded pose, it provides a more intuitive depiction of the anatomical characteristics of the hand. Through this model, a data normalization approach is implemented to align the root joint and unify the biomechanical features of training samples. Furthermore, the HPBM facilitates a weakly supervised strategy for dataset expansion, significantly enhancing the data diversity. The proposed normalized method is evaluated on two widely used 3D hand pose estimation datasets, RHD and STB, demonstrating superior performance compared to the models trained without normalized datasets. Utilizing ground truth 2D keypoints as input, a reduction of 45.1% and 43.4% in error is achieved on the STB and RHD datasets, respectively. When leveraging 2D keypoints from MediaPipe, a reduction in error by 11.3% and 14.3% is observed on the STB and RHD datasets.

Keywords: deep learning; hand pose estimation; biomechanical variability; Hand Pose Biomechanical Model

Citation: Guo, Z.; Li, J.; Tan, J. A Normalization Strategy for Weakly Supervised 3D Hand Pose Estimation. *Appl. Sci.* **2024**, *14*, 3578. https://doi.org/10.3390/app14093578

Academic Editor: Thomas Lindner

Received: 7 March 2024
Revised: 16 April 2024
Accepted: 23 April 2024
Published: 24 April 2024

1. Introduction

Three-dimesional hand pose estimation has become increasingly important in various fields such as virtual reality, augmented reality, human–computer interaction, and sign language recognition. Recent advances [1–3] in deep learning have had a significant impact on this task, resulting in state-of-the-art performance [4–7]. Deep neural networks are optimized adaptively through end-to-end training, which enables them to learn complex representations directly from raw inputs. However, this data-dependent training method also means that the effectiveness and generalization capabilities of the models are closely tied to the distribution of the training data [8]. This presents challenges pertaining to dataset quality.

The inconvenience of 3D pose acquisition presents a significant bottleneck for 3D hand pose estimation [9]. Thus, researchers have proposed a range of techniques to address this issue, including multi-dataset joint training [10–13], weakly supervised learning [12,14–16], and self-supervised learning [9,17]. However, these methods often fail to account for potential disparities in annotating standards across datasets and the variability in biomechanical features (such as bone length or palmar structure) among samples. These factors can directly impact the feature distribution of training data. Given that model performance is heavily influenced by the extent to which the training data capture the underlying data distribution, such challenges can indeed reduce the estimation ability of the model. Therefore, effective preprocessing of the training data is essential for achieving optimal performance.

Most recent approaches for 3D pose estimation rely on the 2D-to-3D lifting pipeline [6], which detects 2D keypoints firstly and then lifts them to 3D. Our work adopts this common practice strategy to detail the proposed normalization strategy. The lifting process is inherently depth ambiguous [6], meaning that one single 2D pose can correspond to multiple potential 3D poses. The mapping rules that determine which 3D pose corresponds to the given 2D pose are learned from the training data. However, current datasets are typically collected by multiple individuals, resulting in potential variability in biomechanical features both within and across datasets. Thus, training models on such inconsistent data can lead to learned mapping rules that generate 3D poses with inconsistent biomechanical features.

Multi-dataset joint training is a widely used method [10–13] for improving model generalization ability. However, disparities in annotating standards across datasets can negatively impact network performance, particularly due to differences in the location of the root joint on the hand. For instance, the Rendered Hand Pose Dataset (RHD) [13] marks the root joint in the wrist area, while the Stereo Hand Pose Tracking Benchmark (STB) [18] dataset marks it in the palmar center. Directly feeding the model with these data, which have significantly different palm structures, can result in the model receiving conflicting knowledge, leading to suboptimal pose estimation performance.

In conclusion, current training methods for hand pose estimation face two main challenges: insufficient accommodation of biomechanical variability among samples and disparities in annotating standards across datasets. To address these issues, this study introduces a data normalization strategy aimed at enhancing model accuracy and generalization. This strategy comprises two main steps: first, aligning the root joint position in training data to a common reference point based on the target annotating standard, and second, normalizing biomechanical features using a Hand Pose Biomechanical Model (HPBM). This model is designed inspired by the biomechanical feature extraction proposed by Spurr et al. [12]. Specifically, the spatial angles of the finger bones are defined as dynamic parameters, while the palmar structure and bone lengths are regarded as static parameters, these designations are contingent upon the spatio-temporal consistency of specific individual parameters. After unifying the static parameters in training dataset with the HPBM, the normalized HPBM-encoded poses are acquired. Subsequently, these poses are mapped to a 3D coordinate-encoded representation through Sequential Least Squares Programming (SLSQP). By employing this approach, the impact of biomechanical variability and annotation standards across different datasets are mitigated.

In addition, the HPBM can facilitate the development of a weakly supervised training methodology that does not require any labeled data but rather relies on prior knowledge, such as skeletal length and articulation range of motion. By fixing the static parameters of the HPBM-encoded pose and generating bone curvatures within the biomechanically feasible range, the 3D normalized hand pose dataset can be expanded without the need for manual labeling. This new dataset offers a greater variety of poses, effectively tackling the challenge of 3D keypoint annotation and thereby enhancing the estimation accuracy of the model. For instance, through weakly supervised training, the error is minimized from 4.85 to 4.57 when the model is fed with ground truth 2D keypoints on the RHD datasets.

The main contribution of this work lies in three folds:

1. In consideration of the anatomical characteristics of the hand, a novel HPBM is introduced, specifically crafted to intuitively depict the biomechanical features of hand poses. This model addresses the limitations associated with traditional 3D coordinate-encoded pose representations.
2. A universal and effective dataset normalization strategy was explored, utilizing the HPBM to unify biomechanical features and standardize root joint positions within and across datasets. In addition, a weakly supervised training methodology was also introduced, utilizing the HPBM to autonomously generate robust 2D–3D pose pairs without the need for manual annotations.
3. Compared to training on non-normalized data, models trained on the proposed normalized dataset demonstrated significant improvements. On the RHD and STB

datasets, notable reductions of 45.1% and 43.4% in error were observed when the model utilized the ground truth 2D pose. Weakly supervised training further achieved an additional 5.7% reduction. The proposed framework proves effective, delivering advanced estimation performance.

2. Related works

2.1. The 2D-to-3D Lifting Pipeline

Compared to end-to-end 3D pose estimation, the 2D-to-3D pipeline approach divides the task into two independent parts: 2D pose estimation from the image and lifting the 2D pose to a 3D pose. Recent 2D-to-3D lifting methods [6,19–23] have demonstrated superior performance compared to end-to-end approaches, owing to the reliable and effective 2D keypoint detection methods developed in previous works [24–27]. To advance 2D-to-3D pose estimation, a Graph Stacked Hourglass Network [22] is introduced. This architecture comprises a repeated encoder–decoder and a graph-structured feature processing approach across multiple scales of skeletal representations, which enhances the ability of the model to learn both local and global feature representations. Chen et al. [28] proposed breaking down the 3D pose estimation task into bone direction prediction and bone length prediction, aiding in addressing depth ambiguity. In addition, domain generalization [29] is applied in hand pose estimation to address the issue of images with characteristics different from the training data. These advances in 2D-to-3D lifting have significantly improved the accuracy of 3D pose estimation.

2.2. Techniques for Addressing Limited Annotated 3D Pose Training Data

Self-supervised learning has been used in various methods to address the challenge of annotating 3D pose data, achieving notable results [9,17,30–32]. For instance, a self-supervised module [17] is proposed that utilizes 2D spatial relationships and 3D geometric knowledge to overcome domain gaps in hand pose estimation and reduce the need for abundant 3D pose labeled data. In Ren et al. [33], depth images are utilized for 3D pose estimation, which uses an image-to-image translation technology for pre-training and a dual-branch network for pixel-wise estimation in a decoupled way. Additionally, S2HAND [32] was proposed to jointly estimate pose, shape, texture, and camera viewpoint. It utilizes 2D detected keypoints to obtain geometric cues from the input image and learns an accurate hand reconstruction model through the consistency between 2D and 3D representations.

Weakly supervised learning is another widely used approach to enhance the generalization capability and reduce the reliance on expensive annotations in 3D pose estimation tasks [12,14–16,34–37]. For example, depth images obtained from commodity RGB-D cameras are utilized during training [35], thus reducing the necessity for 3D annotations. A method [15] is proposed to extract weak 3D information directly from 2D images without 3D pose supervision, which uses 2D pose annotations and perspective prior knowledge to generate relative depth labels, and a weakly-supervised pre-training strategy based on a 2D pose dataset to distinguish the depth relationship between two points in an image. In Spurr et al. [12], a set of losses is developed to constrain the prediction to lie within the range of biomechanically feasible 3D hand configurations and improve the accuracy in such a weakly supervised manner. Our work draws inspiration from their approach to extract biomechanical features. In addition to those weakly supervised techniques, a Multi-View Video-Based 3D Hand dataset, MuViHand [38], has been proposed. It uses synthetic images and provides a valuable resource for improving the generalization ability of pose estimation models.

In addition to weakly supervised and self-learning techniques, many studies [10–13,39] have utilized multi-datasets to train models jointly. These studies focus on improving performance by exploring unconstrained monocular 3D hand pose estimation [10], developing novel compressed latent distribution representations [11,39], or leveraging prior hand biomechanical knowledge [12]. Meanwhile, they utilize multi-dataset joint training to

enhance the performance of the network. Although it is a straightforward approach, it has demonstrated efficacy in improving pose estimation performance.

3. Method

In this section, we provide a detailed description of the proposed framework, which consists of four components: (1) an overview, (2)the Hand Pose Biomechanical Model (HPBM), (3) the normalization method, and (4) the weakly supervised training strategy.

3.1. Overview

The hand structure, as depicted in Figure 1a, is composed of a set of 21 joints denoted as $[\mathbf{j}_0, \ldots, \mathbf{j}_{20}] \in \mathbb{R}^{21\times 3}$. The datasets, such as the STB dataset [18], provide annotations for the root joint $\mathbf{j}_0^{\mathrm{P}}$ at the center of the palm, while other datasets, such as the HO-3D [40] and RHD datasets [13], annotate the root joint, $\mathbf{j}_0^{\mathrm{W}}$, at the wrist location.

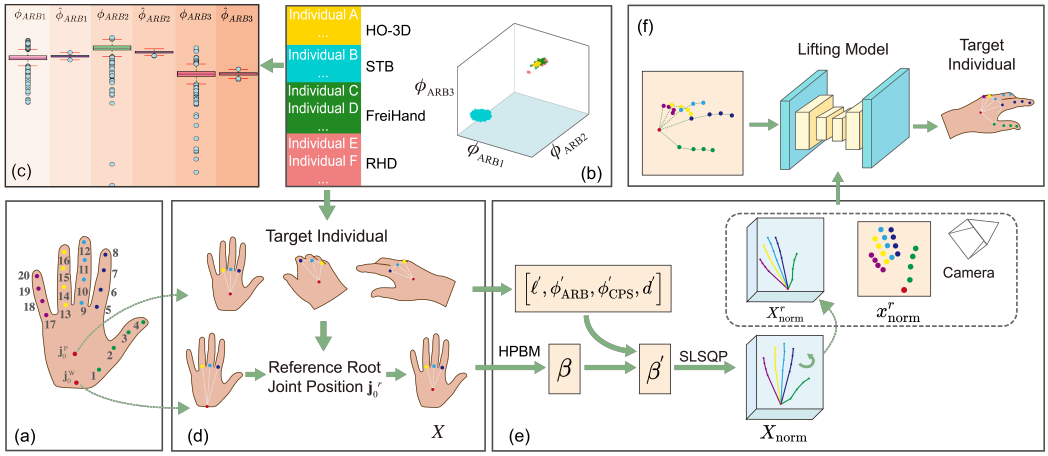

Figure 1. Overview of the normalization strategy. (**a**) Hand structure and joint index. $\mathbf{j}_0^{\mathrm{P}}$ and $\mathbf{j}_0^{\mathrm{W}}$ denote the root joint located in the palmar center and the wrist, respectively. (**b**) The distributions of biomechanical features ϕ_{ARB1}, ϕ_{ARB2}, and ϕ_{ARB3} in the FreiHand, HO-3D, RHD, and STB datasets; their specific meaning can be found in Section 3.2. (**c**) Box plots illustrating the distribution of estimated pose features (ϕ_{ARB1}, ϕ_{ARB2}, ϕ_{ARB3}) and their corresponding ground truth values ($\hat{\phi}_{\mathrm{ARB1}}$, $\hat{\phi}_{\mathrm{ARB2}}$, $\hat{\phi}_{\mathrm{ARB3}}$) derived from the STB dataset. (**d**) Relocating $\mathbf{j}_0$ position. (**e**) Computation of normalized 2D–3D pose pairs. (**f**) Illustrating the mapping rules for lifting model.

Figure 1b depicts the dispersion of distinct individuals from diverse datasets within a tri-dimensional feature space characterized by variables denoted as ϕ_{ARB1}, ϕ_{ARB2}, and ϕ_{ARB3}. Notably, this illustration highlights a clustering tendency among individuals of the same group, whereas considerable variations in features are evident among individuals of different groups. The box plot in Figure 1c illustrates the discrepancies in features between the predicted hand pose generated by a model trained on non-normalized datasets and the ground truth pose. This highlights the inconsistency in biomechanical features resulting from training the model on non-normalized datasets.

To mitigate disparities in $\mathbf{j}_0$ positions and variations in biomechanical features, a normalization strategy comprising two steps was introduced. First, for a specific detected individual pose, adjustment of the $\mathbf{j}_0$ position of the training data was made based on the reference $\mathbf{j}_0$ location, as illustrated in Figure 1d. Second, normalization of the biomechanical features of the training data was performed by replacing the static biomechanical parameters with those of the target individual, as demonstrated in Figure 1e. Additionally, data augmentation was introduced by randomly rotating the normalized 3D pose, resulting in increased data diversity. By projecting these poses onto a 2D plane, 2D–3D pose pairs were obtained. These pairs represented the authentic mapping rules for the given individual and maintained consistent biomechanical features. Training a lifting model using such data enabled the generation of poses aligned with the features of the target individual.

3.2. Hand Pose Biomechanical Model

To provide a clearer explanation, the following predefined notation and rules will be set:

1. The kinematic chain starting from the root joint $\mathbf{j}_0$ and extending to the fingertips.
2. The parent joint of a given joint $\mathbf{j}_i$ is denoted as $\mathbf{j}_{p(i)}$.
3. A bone $\mathbf{b}_i$ is defined as the vector pointing from the parent joint to its child joint, computed by $\mathbf{b}_i = \mathbf{j}_{i+1} - \mathbf{j}_{p(i+1)}$.
4. The root bones vectors are represented as $\mathbf{B_{root}} = \{\mathbf{b}_i\}_{i\in\{0,4,8,12,16\}}$, as indicated by the solid line in Figure 2a.
5. The angular distance between two vectors, v_1 and v_2, is denoted by $\alpha(v_1, v_2)$ and is calculated using the arccosine function as $\alpha(v_1, v_2) = \arccos\left(\frac{v_1^T v_2}{\|v_1\|_2\|v_2\|_2}\right)$.
6. The normalized vector is defined as $\text{norm}(x) = \frac{x}{\|x\|_2}$.
7. The operator $\mathbf{P_{xy}}(\mathbf{v})$ is used to project a vector $\mathbf{v}$ orthogonally onto the $\mathbf{x} - \mathbf{y}$ plane where $\mathbf{x}$, $\mathbf{y}$ are vectors.

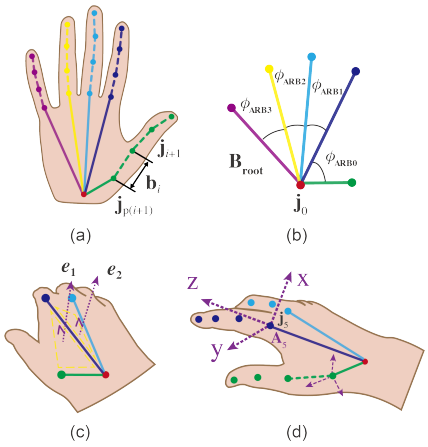

Figure 2. Visualization of the biomechanical architecture of the hand. (**a**) Hand structure. (**b**) The angular distances between root bones. (**c**) The curvatures of the palmar surface. (**d**) The local coordinate system.

The construction of the HPBM was divided into two distinct components: palm modeling and finger modeling. The palm section encompasses joints $\{\mathbf{j}_0, \mathbf{j}_1, \mathbf{j}_5, \mathbf{j}_9, \mathbf{j}_{13}, \mathbf{j}_{17}\}$, while the finger section comprises the remaining part of the hand. The approach defines the palm structure using biomechanical features like root bone length and angles, with $\mathbf{j}_0$ as the reference position. Then the palm is utilized as the foundation to construct coordinate systems to determine finger flexion.

The palmar structure of a particular individual was observed to remain relatively static in the spatio-temporal domain, representing a consistent biomechanical feature. To define the palmar structure, three parameters were used: the length of the root bones ℓ_{root}, the Angular Distances between the Root Bones (ARB) ϕ_{ARB}, and the Curvatures of the Palmar Surface (CPS) ϕ_{CPS}.

ℓ_{root} can be determined by calculating the Euclidean distance between the palmar joints, specifically, $\ell_{root} = [\|\mathbf{b}_0\|_2, \|\mathbf{b}_4\|_2, \|\mathbf{b}_8\|_2, \|\mathbf{b}_{12}\|_2, \|\mathbf{b}_{16}\|_2]$. The ARB features are illustrated in Figure 2b, and are computed in the following equation:

$$\phi_{\text{ARB}i} = \alpha\left(\mathbf{b}_{i\cdot 4}, \mathbf{b}_{(i+1)\cdot 4}\right), \quad \text{for } i \in \{0, 1, 2, 3\}. \tag{1}$$

The CPS features can represent the curvature characteristics of neighboring bone surfaces, as illustrated by the two yellow triangles highlighted by dashed lines in Figure 2c. Based on the physical structure of the hand, ϕ_{CPS} can be obtained:

$$\phi_{\mathrm{CPS}i} = \arccos(e_i \cdot e_{i+1}), \quad \text{for } i \in \{0,1,2\}. \tag{2}$$

where e_i is normal vector of the surface:

$$e_i = \mathrm{norm}\left(\mathbf{b}_{i\cdot 4}, \mathbf{b}_{(i+1)\cdot 4}\right), \quad \text{for } i \in \{0,1,2,3\}. \tag{3}$$

During the transformation of the HPBM-encoded pose to a 3D-coordinate representation, the arccos function generates two-directional surfaces for each $\phi_{\mathrm{CPS}i}$, potentially resulting in multiple structural solutions for the palm. To address this, the relative surface direction d_i is introduced as a constraint to ensure that the unique pose is defined:

$$d_i = \begin{cases} 1, & \text{if } e_i \cdot \left(\mathbf{b}_{(i+2)\cdot 4} - \mathbf{b}_{(i+1)\cdot 4}\right) > 0 \\ -1, & \text{if } e_i \cdot \left(\mathbf{b}_{(i+2)\cdot 4} - \mathbf{b}_{(i+1)\cdot 4}\right) < 0 \end{cases}, \tag{4}$$

where $i \in \{0,1,2\}$.

The morphological arrangement of the finger, delineated by the stippled delineation, is shown in Figure 2a. The incipient stages of the modeling necessitate the instantiation of the orientation ascribed to each bone element. To expedite this process, a foundational requirement entails the instantiation of distinct local reference frames A_i, for each finger, as shown in Figure 2d.

The establishment of the coordinate system utilizes the palm plane as a reference framework. Specifically, the root bones $\mathbf{b}_4, \mathbf{b}_8, \mathbf{b}_{12}, \mathbf{b}_{16}$ are deemed to be situated within approximate proximity to a common plane. The hyperplane P_h that ostensibly maintains a predominantly orthogonal orientation concerning the basis vectors e_1, e_2, e_3 is derived by minimizing the following function:

$$\mathbf{J} = \sum \|v_h e_i - 1\|_2, \tag{5}$$

the v_h is the normal vector of P_h. Describing the spatial orientation of $\mathbf{b}_i$ involves creating a local frame called A_i. Within this frame the x, y, and z-axes of A_i can be defined as:

$$\begin{cases} A_i^x = \mathrm{norm}(v_h) \\ A_i^z = \mathrm{norm}(b_{p(i)}) \\ A_i^y = \mathrm{norm}\left(A_i^z \times A_i^x\right) \end{cases}. \tag{6}$$

Based on the established coordinate, the spatial orientation of $\mathbf{b}_i$ can be defined by:

$$\begin{matrix} \theta_i^f = \alpha\left(\mathbf{P}_{xz}(\mathbf{b}_i), \mathbf{A}_i^z\right) \\ \theta_i^a = \alpha(\mathbf{P}_{xz}(\mathbf{b}_i), \mathbf{b}_i) \end{matrix}, \tag{7}$$

where θ_i^f and θ_i^a represent the flexion and abduction angles, respectively. In combination with the length of each finger, these parameters can define the structure of the hand–finger part.

In conclusion, the proposed HPBM-encoded pose can be represented as

$$\beta = \left[\ell, \phi_{\mathrm{ARB}}, \phi_{\mathrm{CPS}}, d, \theta^f, \theta^a\right], \tag{8}$$

where $\ell = \{\|\mathbf{b}_i\|_2\}_{i=0}^{20}$ represents the length of the bones. The parameters $[\ell, \phi_{\mathrm{ARB}}, \phi_{\mathrm{CPS}}, d]$ remain constant across the training data for the same individual, while $\left[\theta^f, \theta^a\right]$ vary with

pose changes. The former is categorized as static biomechanical parameters and the latter as dynamic biomechanical parameters.

3.3. Normalization Strategy

The normalization strategy comprises two steps, namely relocating $\mathbf{j}_0$ to the reference position and normalization of biomechanical features, as depicted in Figure 3. In a subsequent paper, each step will be discussed in detail.

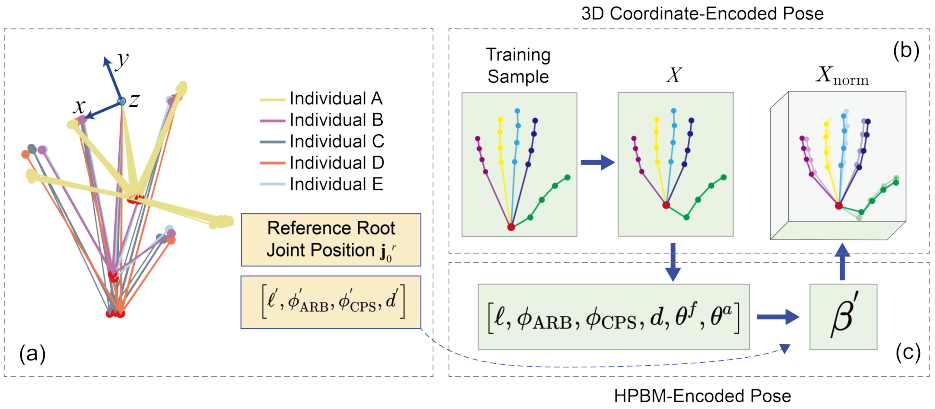

Figure 3. Flowchart of the normalization process. (**a**) Aligning the joint $\mathbf{j}_9$ across various individual palm and superimposing the planes containing joints $\mathbf{j}_9$, $\mathbf{j}_{13}$, and $\mathbf{j}_0$, the distribution of the $\mathbf{j}_0$ is depicted by the red dots. (**b**) 3D Coordinate-Encoded Pose. (**c**) HPBM-Encoded Pose. Figures (**b**,**c**) depict the normalized process.

The current state of 3D pose annotation standardization is inadequate, as demonstrated in Figure 1c; the root joint $\mathbf{j}_0$ can be marked either at the palmar center or the wrist, necessitating the initial normalization of $\mathbf{j}_0$ position across all training sets. To accomplish this, a reference coordinate system, denoted as $\mathbf{A}_{\text{root}}$, is established with $\mathbf{j}_9$ as the origin. Specifically,

$$\begin{aligned} \mathbf{A}_{\text{root}}^{x} &= \text{norm}(\mathbf{j}_{13} - \mathbf{j}_9) \\ \mathbf{A}_{\text{root}}^{z} &= \text{norm}((\mathbf{j}_5 - \mathbf{j}_9) \times \mathbf{A}_{\text{root}}^{x}). \\ \mathbf{A}_{\text{root}}^{y} &= \mathbf{A}_{\text{root}}^{z} \times \mathbf{A}_{\text{root}}^{x} \end{aligned} \tag{9}$$

In the first step, depicted in Figure 3, the reference position of the root joint, denoted as $\mathbf{j}_0^r$, was derived by averaging all target individual samples within the reference coordinate system $\mathbf{A}_{\text{root}}$. Subsequently, all $\mathbf{j}_0$ within the training samples were adjusted to align with the location of $\mathbf{j}_0^r$. This procedure ensured that the training data adhered to a uniform annotation standard.

Considering the challenge of neural networks acquiring accurate 2D–3D mapping knowledge from diverse training sets. An effort is made to preserve the dynamic biomechanical parameters while normalizing the static biomechanical features of training samples to the specific target individual. This process, outlined in Figure 3, involves mapping the 3D coordinate-encoded pose X to the HPBM-encoded pose β. The parameters $\left[\ell', \phi'_{\text{ARB}}, \phi'_{\text{CPS}}, d'\right]$ are used to characterize the static parameters of the target individual. In β, the dynamic parameters θ^f and θ^a remain unchanged, while the static parameters are replaced with the corresponding features of the target individual. The resulting normalized pose is expressed as:

$$\begin{aligned} \beta' &= \text{norm}(\beta) \\ &= \left[\ell', \phi'_{\text{ARB}}, \phi'_{\text{CPS}}, d', \theta^f, \theta^a\right]. \end{aligned} \tag{10}$$

Next, the normalized pose β' is converted into a 3D coordinate-encoded pose X_{norm}, ensuring consistency in biomechanical features across the training set. This procedure

can be regarded as the inverse operation of mapping X_{norm} to β', essentially solving the solution of the nonlinear function $\boldsymbol{f} : X_{\text{norm}} \mapsto \beta'$. In this study, the SLSQP algorithm is employed to solve it. It should be noted that BHPM only represents the biomechanical features and spatial structure of the hand but does not preserve the absolute orientation information. This is because 2D–3D pose pairs can be obtained by randomly rotating the 3D pose in the spatial domain and projecting it onto the 2D plane. Therefore, incorporation of absolute pose orientation information into the HPBM-encoded pose is deemed unnecessary for training.

It is possible to randomly rotate X_{norm} along the x-axis, y-axis, and z-axis to obtain X^r_{norm}, from which a 2D projection can be acquired, resulting in pairs denoted as $(X^r_{\text{norm}}, x^r_{\text{norm}})$, where x^r_{norm} represents the 2D pose projection of X^r_{norm}. This process is illustrated in Figure 1e.

3.4. Weakly Supervised Training

As previously discussed, the HPBM can represent biomechanical features and finger flexion states in an intuitive manner. By fixing static parameters $[\ell, \phi_{\text{ARB}}, \phi_{\text{CPS}}, d]$ and randomly generating dynamic parameters $[\theta^f, \theta^a]$ within a reasonable range, it is possible to create an infinite number of 2D–3D pose pairs for training. The generated training dataset is denoted as $\mathcal{D}$.

We summarize the proposed weakly supervised strategy in the following steps:

Step 1: Determine potential threshold ranges, denoted as $[\Theta_{\text{min}}, \Theta_{\text{max}}]$, for dynamic parameters θ^f and θ^a. These thresholds, established as prior knowledge, can be set based on experience or calculated from training datasets, specifically $\Theta_{\text{min}} = \min(\Theta^{ta})$ and $\Theta_{\text{max}} = \max(\Theta^{ta})$, where Θ^{ta} are sets of $[\theta^f, \theta^a]$ derived from the target individual.

Step 2: Randomly generate HPBM-encoded poses within the established threshold ranges.

Step 3: Transform the randomly generated sample into a 3D coordinate-encoded pose X_{norm}.

By utilizing this method, a 2D–3D lifting dataset can be created without the need for any manually labeled data. Subsequently, weakly supervised training can be performed based on this dataset.

4. Experiments

4.1. Experimental Setup

Experiments are conducted on several commonly used benchmarks as listed below.

FreiHAND [41] was created to facilitate research on 3D hand pose estimation. The dataset includes 130,240 training images, each with corresponding 3D annotations for joint locations. And an additional 3960 evaluation samples are provided for performance evaluation.

HO-3D [40] comprises color images of human hands interacting with objects. The dataset consists of 68 sequences, encompassing a total of 77,558 frames. The training set includes 66,034 images, while the test set comprises 11,524 images. This large-scale dataset provides a valuable resource for research in hand-object interaction and related computer vision tasks.

RHD [13] contains 41,258 training samples and 2728 testing samples, each of which is a synthetic image of a hand in a specific pose. It provides information on keypoint visibility and occlusion/cropping, as well as camera parameters, segmentation maps, and depth maps for algorithm development.

STB [18] provides both 2D and 3D annotations for 18,000 stereo pairs. For the evaluation process, 12,000 images captured using the Point Grey Bumblebee2 stereo camera from the "B1Counting", "B1Random", "B2Counting", and "B2Random" categories are utilized. This set of stereo image pairs includes both left and right images.

To evaluate the effectiveness, we adopt a cross-dataset training and evaluation methodology. Models are trained on the HO-3D and FreiHAND official training datasets and

assessed on the RHD and STB datasets. This approach allows us to assess generalization of the models across new datasets, considering the diverse sources of training and testing data.

For evaluation, the STB dataset consists of samples from a single individual, while the RHD evaluation dataset includes samples from four distinct individuals, designated as RHD-A, RHD-B, RHD-C, and RHD-D, corresponding to individuals A, B, C, and D, and specifically covering hand poses numbered 923, 898, 979, and 981.

Quantitative evaluations employ the following metrics:

MPJPE: Mean Per Joint Position Error (MPJPE) is a metric that quantifies the error, measured in Euclidean distance (mm), between the predicted joint positions and the corresponding ground truth joint positions, with respect to the root-relative coordinates. The computational methodology aligns with Zimmermann et al. [13].

PA-MPJPE: Procrustes Analysis Mean Per Joint Position Error (PA-MPJPE) is a variant of MPJPE that employs Procrustes analysis [42] to align the predicted and ground truth poses while ignoring global variation.

PCK: Percentage of Correct Keypoints (PCK) quantifies the percentage of correctly predicted keypoints that fall within a specified threshold from the ground truth keypoints.

AUC: Area under the curve (AUC) quantifies the overall accuracy by calculating the area under the curve of the Percentage of Correct Keypoints (PCK) versus error thresholds.

4.2. Implementation Details

At evaluation, the 2D keypoints were obtained from both ground truth values and the MediaPipe framework for comparative analysis. MediaPipe is an open-source platform developed by Google that offers advanced and robust hand pose estimation capabilities. The proposed normalization and weakly supervised method are generic and adaptable to different 3D pose estimators. To demonstrate this, two representative 3D pose estimators were utilized as backbones: (1) ST-GCN [3] (1-frame), a pioneering network that leverages GCN-based architecture to encode global and local joint relations, and (2) VPose3D [23] (1-frame), a fully-convolutional advanced network. The specific model settings were referenced from PoseAug [43].

The models were trained by the PyTorch deep learning framework. During the training process, the AdamW optimizer with a weight decay of 0.05 was used to optimize the network. The input image size was 224 × 224, and the initial learning rate was set to 0.001. An RTX3090 GPU was employed for training the model, utilizing a batch size of 512 across a total of 500 epochs. Mean Squared Error (MSE) was employed as the loss function for model training. These settings remain consistent for both the weakly supervised training and the normalization strategy. Regarding the weakly supervised training, 200,000 samples were generated for each individual.

4.3. Evaluation of Normalization Strategy

Table 1 provides a comprehensive comparison of the effects of the normalization strategy on the VPose3D and ST-GCN models across different 2D pose sources and datasets. In terms of **PJ** errors, both VPose3D and ST-GCN consistently exhibit lower errors on the STB and RHD datasets when trained with normalization (w/norm). Specifically, by employing the normalization strategy and utilizing GT 2D pose inputs, a substantial reduction in joint position errors of 45.1% and 43.4% is observed in the STB and RHD datasets, respectively. When MediaPipe 2D poses are used as input, the estimation errors of VPose3D decrease by 11.3% and 14.3% in the STB and RHD datasets, respectively. Higher AUC values are also observed across all scenarios, further emphasizing the effectiveness of the normalization approach. This consistent enhancement in performance metrics underscores the importance of the normalization strategy in improving the accuracy of lifting models.

Table 1. The effect of normalization strategy. At the inference process, both ground truth (GT) 2D keypoints and those generated by the MediaPipe framework were fed to the lifting model. "w/ norm" denotes the training dataset with normalization, while "w/o norm" signifies the dataset without normalization. "PJ" denotes PA-MPJPE, the "↑"/"↓" denote the lower/higher, the better. The AUC has a threshold ranging from 20 to 50 mm.

2D Pose	Model	Datasets	w/o norm		w/ norm	
			PJ ↓	AUC ↑	PJ ↓	AUC ↑
GT	VPose3D	STB	12.37	0.960	6.78	0.995
		RHD-A	9.91	0.984	5.21	0.996
		RHD-B	9.03	0.988	5.24	0.996
		RHD-C	7.87	0.992	4.70	0.997
		RHD-D	7.62	0.993	4.33	0.998
	ST-GCN	STB	13.52	0.958	6.77	0.995
		RHD-A	11.18	0.970	4.48	0.995
		RHD-B	10.35	0.977	4.76	0.994
		RHD-C	7.89	0.988	4.07	0.996
		RHD-D	7.47	0.990	3.71	0.998
MediaPipe	VPose3D	STB	11.39	0.964	10.10	0.989
		RHD-A	13.53	0.945	11.35	0.949
		RHD-B	13.07	0.947	11.43	0.946
		RHD-C	10.83	0.971	9.18	0.971
		RHD-D	10.92	0.968	9.45	0.966
	ST-GCN	STB	12.42	0.963	12.33	0.951
		RHD-A	14.70	0.929	12.23	0.931
		RHD-B	14.29	0.930	12.11	0.930
		RHD-C	11.58	0.958	9.50	0.962
		RHD-D	11.54	0.957	9.66	0.959

The lifting models serve as projection functions that capture mapping rules between 2D and 3D joints. Therefore, evaluating model accuracy based on GT 2D keypoints is more reliable, as the 2D–3D pose pairs $\{X_{GT}, x_{GT}\}$ adhere to true mapping rules, while $\{X_{GT}, x_{mp}\}$ pairs with 2D keypoints estimated by MediaPipe may not reflect accurate 2D–3D mappings. This signifies that there are inherent 2D errors from the keypoints detector, and experiments using 2D pose x_{GT} as input can more effectively assess the model performance. In Table 1, when the model is provided with GT 2D keypoints, the estimation accuracy significantly decreases, demonstrating that our model closely adheres to the actual mapping rules. When the model fed with MediaPipe 2D keypoints, the accuracy also improves, albeit to a lesser extent compared to when fed with GT data. It can be inferred that the errors primarily arise from inaccuracies in the 2D keypoints, given the strong performance of the model when using GT 2D keypoints.

To comprehensively assess the efficacy of the proposed normalization method, the distribution of static biomechanical features in both the normalized and non-normalized datasets is visually depicted. By utilizing PCA to project features onto a 2D plane, as illustrated in Figure 4, a clear alignment emerges between the distribution in the normalized training dataset and the testing datasets, while the distribution of training data without normalization exhibits significant differences from that of the testing datasets. This illustrates the successful normalization of static biomechanical features to match the target individual.

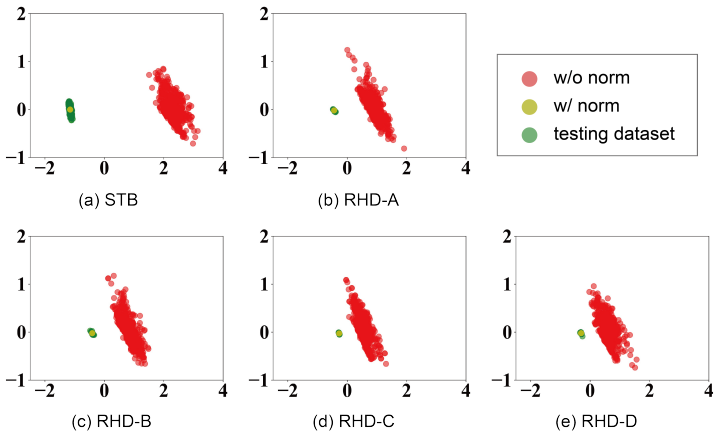

Figure 4. Illustration of the distribution of static biomechanical features in non-normalized dataset, normalized dataset, and the testing dataset. The visualization is achieved using principal component analysis (PCA). A random selection of 500 samples from each dataset was conducted, where each sample encompasses static features $\{\ell, \phi_{\text{ARB}}, \phi_{\text{CPS}}\} \in \mathbb{R}^{27}$.

The comparison in Figure 4 also reveals that static biomechanical features within the same individual exhibit minimal variation, with fluctuations largely attributed to the annotation process. Notably, the disparities between non-normalized training data and STB datasets exhibit a greater margin compared to those of RHD datasets. This observation substantiates why models trained on non-normalized datasets exhibit better performance on the RHD dataset than on the STB dataset.

Table 2 compares the errors in static biomechanical features between models trained with and without normalization on different datasets. Across all datasets (STB, RHD-A, RHD-B, RHD-C, RHD-D), the model trained with normalization (w/ norm) consistently demonstrates lower errors in biomechanical features compared to the model trained without normalization (w/o norm). For instance, in the STB dataset, the error in ℓ is reduced from 71.05 to 8.43, showcasing a substantial improvement. Similar trends are observed in other biomechanical features, such as ϕ_{ARB} and ϕ_{CPS}. These results emphasize the effectiveness of the normalization method in enhancing the accuracy.

Table 2. Comparison of errors in static biomechanical features. The model VPose3D employed for this analysis was trained on both non-normalized and normalized datasets. The comparison encompasses biomechanical features $\{\ell, \phi_{\text{ARB}}, \phi_{\text{CPS}}\}$, with the errors $\{\mathbf{e}_\ell, \mathbf{e}_{\mathbf{CPS}}, \mathbf{e}_{\mathbf{ARB}}\}$ computed by evaluating the averaged Euclidean distance between the predicted biomechanical features and their corresponding ground truth values.

Datasets	w/ norm			w/o norm		
	$\mathbf{e}_\ell$	$\mathbf{e}_{\mathbf{CPS}}$	$\mathbf{e}_{\mathbf{ARB}}$	$\mathbf{e}_\ell$	$\mathbf{e}_{\mathbf{CPS}}$	$\mathbf{e}_{\mathbf{ARB}}$
STB	8.43	0.0001	0.036	71.05	0.245	0.818
RHD-A	9.89	0.0169	0.006	33.75	0.079	0.062
RHD-B	10.10	0.0162	0.007	31.51	0.067	0.053
RHD-C	8.69	0.0083	0.007	22.73	0.122	0.044
RHD-D	8.08	0.0057	0.007	21.03	0.118	0.035

The consistency in the palmar structure of the training data, constrained by $[\ell, \phi_{\text{ARB}}, \phi_{\text{CPS}}, d]$, leads to significantly improved accuracy in estimating the palmar joints $[\mathbf{j}_0, \mathbf{j}_1, \mathbf{j}_5, \mathbf{j}_9, \mathbf{j}_{13}, \mathbf{j}_{17}]$, as evident in the heatmap in Figure 5. This improvement is particularly pronounced for $\mathbf{j}_0$. Similarly, when trained on a dataset constrained by $\left[\ell, \theta^f, \theta^a\right]$, the prediction accuracy of finger joints is also improved. For instance, the position of $\mathbf{j}_5$ is constrained by the predefined palmar structure, and the position distribution of $[\mathbf{j}_6, \mathbf{j}_7, \mathbf{j}_8]$ is closer to that of the target individual due to the constraints $|b_5|, |b_6|, |b_7|$.

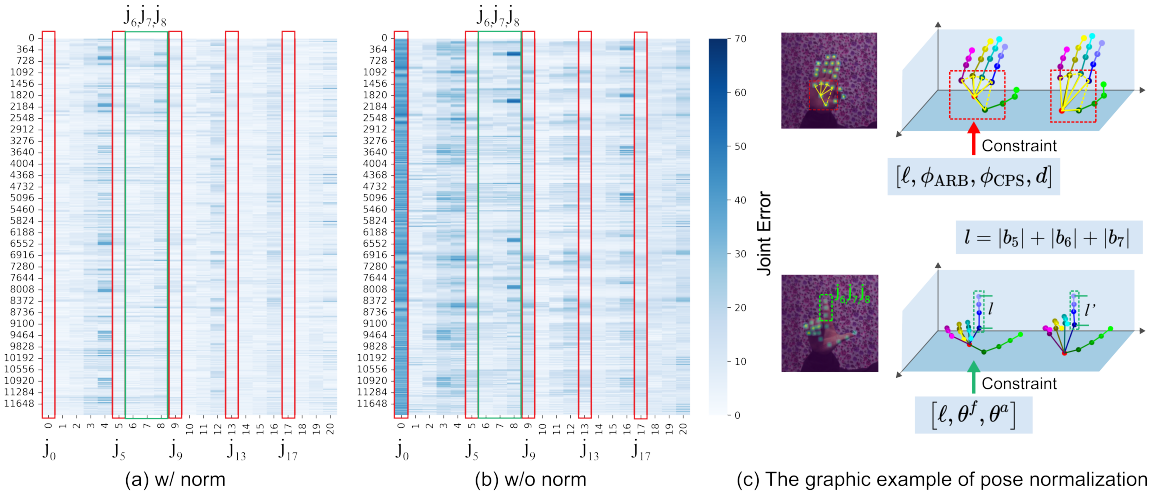

Figure 5. Joint error heatmaps in the STB dataset and the graphic example of pose normalization. (**a**) Joint error diagram with normalization. (**b**) Joint error diagram without normalization. (**c**) Normalization diagram illustrating key points localization of the palm and finger joints.

Figure 6 presents qualitative results illustrating the impact of the normalization strategy. The estimated poses (green line) on both the STB and RHD datasets show a closer alignment with the ground truth (GT) poses (red line) compared to the model trained without the normalized dataset (blue line).

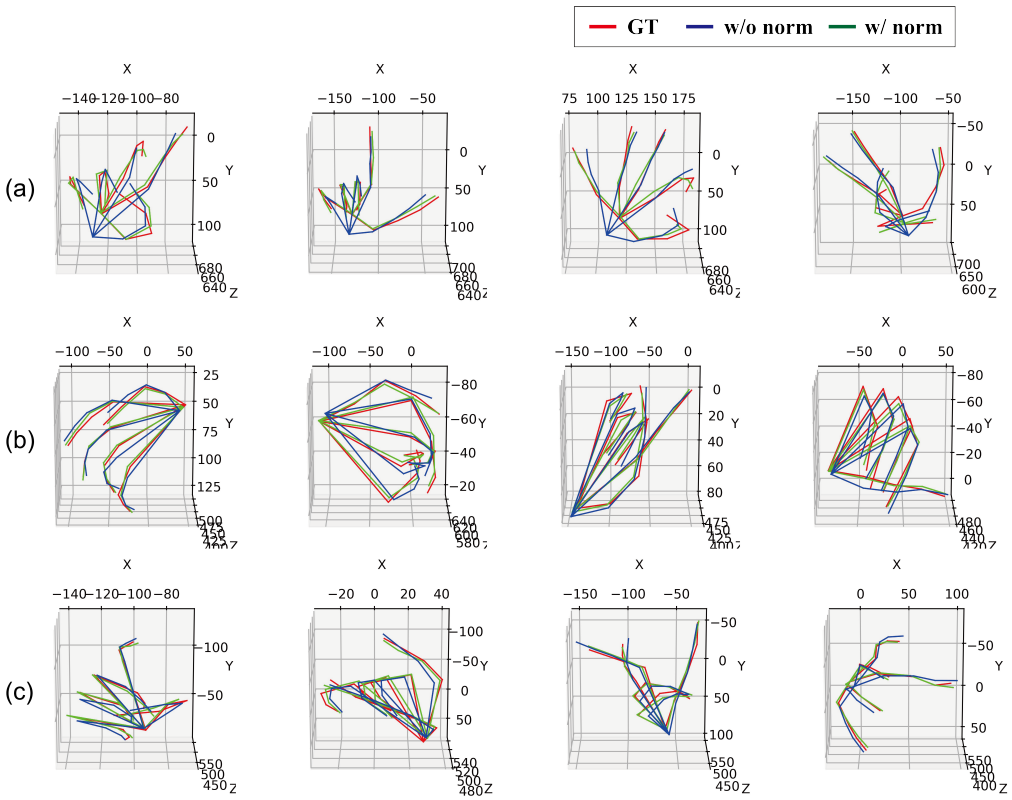

Figure 6. Qualitative results of poses generated by the models trained on normalized dataset, the models trained without normalized dataset, and GT poses. (**a**) Results from the STB dataset. (**b**) Results from the RHD-A dataset. (**c**) Results from the RHD-B dataset.

4.4. Evaluation of Weakly Supervised Approach

The evaluation results of the model trained on $\mathcal{D}$ are presented in Table 3. The dataset $\mathcal{D}$ is constructed solely based on prior knowledge, without any manual labeling. Notably, the model attained satisfactory performance despite the absence of manually annotated data. When the model was fed with GT and MediaPipe 2D keypoints, it both achieved higher accuracy compared to the model trained on non-normalized datasets. Here, the $\mathcal{D}_n$ is used to denote the normalized HO-3D and FreiHAND datasets. Furthermore, weakly supervised training was conducted on the joint dataset $\mathcal{D}$+$\mathcal{D}_n$, resulting in even more accurate estimation results, as shown in Table 3; the errors reduced from 7.19 to 4.57 and from 11.29 to 10.03 for GT and MediaPipe 2D keypoints, respectively.

Table 3. Comparison of the weakly supervised method with advanced techniques using the PA-MPJPE metric. $\mathcal{D}_n$ denotes the normalized training dataset. "GT" and "MediaPipe" represent 2D keypoint sources, and VPose3D is employed as lifting model in this evaluation.

Methods		Datasets				
		RHD-A	**RHD-B**	**RHD-C**	**RHD-D**	**RHD**
	LDR [11]	-	-	-	-	11.63
	SS-GCN [17]	-	-	-	-	13.29
	RC2CHP [44]	-	-	-	-	13.14
GT	$\mathcal{D}$	5.97	6.52	7.67	8.51	7.19
	$\mathcal{D} + \mathcal{D}_n$	4.73	4.93	4.42	4.25	**4.57**
MediaPipe	$\mathcal{D}$	11.19	11.70	10.69	11.63	11.29
	$\mathcal{D} + \mathcal{D}_n$	10.92	11.14	9.01	9.22	**10.03**

It should be noted that the generated weakly supervised data may not strictly adhere to physical plausibility. Despite the dynamic parameters $\{\theta^f, \theta^a\}$ being confined within the range $[\Theta_{\min}, \Theta_{\max}]$, there exists interdependence among these parameters. For instance, when the index finger is angled to a certain degree, it consequently affects the permissible range of angles for the other fingers. Restricting $\{\theta^f, \theta^a\}$ to $[\Theta_{\min}, \Theta_{\max}]$ does not necessarily guarantee physical plausibility. Currently, the method for quantitatively assessing the plausibility of poses has not yet been developed, and it will be addressed in future work.

From Tables 1 and 3, it can be observed that despite the weakly supervised data not being strictly physically plausible, their incorporation with $\mathcal{D}_n$ leads to further improvements in model performance compared to a model solely trained on $\mathcal{D}_n$. For instance, with GT 2D keypoints as input, the **PJ** errors for individual RHD-A, RHD-B, RHD-C, and RHD-D exhibit reductions of 0.48, 0.31, 0.28, and 0.08, respectively. This improvement can be attributed to the broader range and diversity of poses included in the training data.

To provide a more comprehensive understanding of the efficacy of the proposed method, a thorough comparison is conducted with numerous other state-of-the-art techniques. This comparison is presented in Table 3 and Figure 7. Compared to advanced methods [13,31,35,44–47], in terms of **PJ** and AUC metrics, the proposed method showed fewer errors. This achievement is accomplished without the need to incorporate training data from the RHD dataset. Solely leveraging the static biomechanical features of individuals enables the approach to achieve higher estimation accuracy.

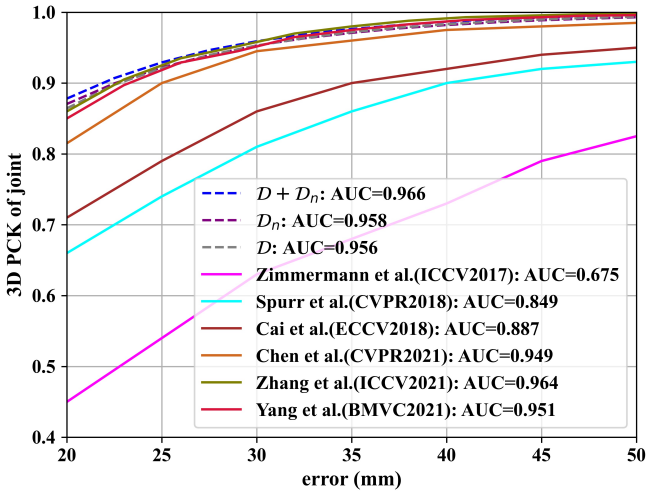

Figure 7. Comparative analysis with state-of-the-art methods [13,31,35,45–47] on the RHD dataset. The x-axis represents the threshold, while the y-axis denotes the PCK at the corresponding threshold.

5. Conclusions

This study addresses critical challenges in the domain of 3D hand pose estimation, specifically focusing on discrepancies in root joint positions and variability in biomechanical features of training samples. The proposed concept of HPBM proves to be an effective method in overcoming these challenges.

Through the normalization strategy, significant improvements in accuracy have been achieved, evidenced by evaluations on the RHD and STB datasets. Specifically, two benchmark models, VPose3D and ST-GCN, were assessed across both normalized and non-normalized datasets. The findings reveal a noteworthy decrease in PJ errors. When using GT 2D keypoints, substantial reductions of 45.1% and 43.4% in errors are evident for the STB and RHD datasets, respectively.

The static biomechanical features of the estimated pose predicted by the model trained on normalized datasets also show more accurate results. On the STB dataset, the $\{\mathbf{e}_{\ell}, \mathbf{e}_{\mathbf{CPS}}, \mathbf{e}_{\mathbf{ARB}}\}$ are reduced from 71.05, 0.245, 0.818 to 8.43, 0.0001, 0.036, respectively, indicating a significant improvement in hand structure accuracy for the target individual.

Moreover, the deployment of a weakly supervised methodology is facilitated by the HPBM, allowing for dataset expansion and improved estimation capabilities. Evaluation results of the model trained on $\mathcal{D}$ show that even without manual labeling, using MediaPipe 2D keypoints as input, the model obtains a PJ of 11.29, which is competitive with other advanced methods. Joint training with dataset $\mathcal{D}_n$ yields even more accurate estimation results, further reducing from 11.29 to 10.03.

The rigorous experiments reveal a substantial enhancement in model performance compared to baseline methods. It demonstrates that the proposed normalization strategy is effective in addressing existing non-normalized problems and enhancing the performance of the model. While this study encodes hand poses from a biomechanical perspective, it does not assess their validity within human biomechanical constraints. This aspect warrants further investigation in future research.

Author Contributions: Conceptualization, Z.G.; methodology, Z.G.; software, J.L.; validation, Z.G. and J.L.; investigation, J.T.; writing—original draft preparation, Z.G.; writing—review and editing, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC2001601 and in part by the National Natural Science Foundation of China under Grant U1913205 and Grant 62103180.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 10012–10022.
2. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
3. Cai, Y.; Ge, L.; Liu, J.; Cai, J.; Cham, T.J.; Yuan, J.; Thalmann, N.M. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2272–2281.
4. Jiang, C.; Xiao, Y.; Wu, C.; Zhang, M.; Zheng, J.; Cao, Z.; Zhou, J.T. A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation from a Single RGB Image. *arXiv* **2023**, arXiv:2304.03635.
5. Karunratanakul, K.; Prokudin, S.; Hilliges, O.; Tang, S. HARP: Personalized Hand Reconstruction from a Monocular RGB Video. *arXiv* **2022**, arXiv:2212.09530.
6. Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; Yuan, J. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13232–13242.
7. Deng, X.; Zuo, D.; Zhang, Y.; Cui, Z.; Cheng, J.; Tan, P.; Chang, L.; Pollefeys, M.; Fanello, S.; Wang, H. Recurrent 3D Hand Pose Estimation Using Cascaded Pose-guided 3D Alignments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 932–945. [CrossRef] [PubMed]
8. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **2021**, *64*, 107–115. [CrossRef]
9. Kundu, J.N.; Seth, S.; YM, P.; Jampani, V.; Chakraborty, A.; Babu, R.V. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20448–20459.
10. Sharma, S.; Huang, S. An end-to-end framework for unconstrained monocular 3D hand pose estimation. *Pattern Recognit.* **2021**, *115*, 107892. [CrossRef]
11. Li, M.; Wang, J.; Sang, N. Latent distribution-based 3D hand pose estimation from monocular RGB images. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4883–4894. [CrossRef]
12. Spurr, A.; Iqbal, U.; Molchanov, P.; Hilliges, O.; Kautz, J. Weakly supervised 3d hand pose estimation via biomechanical constraints. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 211–228.
13. Zimmermann, C.; Brox, T. Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4903–4911.
14. Kulon, D.; Guler, R.A.; Kokkinos, I.; Bronstein, M.M.; Zafeiriou, S. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4990–5000.
15. Qiu, Z.; Qiu, K.; Fu, J.; Fu, D. Weakly-supervised pre-training for 3D human pose estimation via perspective knowledge. *Pattern Recognit.* **2023**, *139*, 109497. [CrossRef]
16. Yang, C.Y.; Luo, J.; Xia, L.; Sun, Y.; Qiao, N.; Zhang, K.; Jiang, Z.; Hwang, J.N.; Kuo, C.H. CameraPose: Weakly-Supervised Monocular 3D Human Pose Estimation by Leveraging In-the-wild 2D Annotations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 2924–2933.
17. Guo, S.; Rigall, E.; Qi, L.; Dong, X.; Li, H.; Dong, J. Graph-based CNNs with self-supervised module for 3D hand pose estimation from monocular RGB. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 1514–1525. [CrossRef]
18. Zhang, J.; Jiao, J.; Chen, M.; Qu, L.; Xu, X.; Yang, Q. A hand pose tracking benchmark from stereo matching. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 982–986.
19. Ci, H.; Wang, C.; Ma, X.; Wang, Y. Optimizing network structure for 3d human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2262–2271.
20. Liu, K.; Ding, R.; Zou, Z.; Wang, L.; Tang, W. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 318–334.

21. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
22. Xu, T.; Takano, W. Graph stacked hourglass networks for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16105–16114.
23. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7753–7762.
24. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
25. Ma, X.; Rahmani, H.; Fan, Z.; Yang, B.; Chen, J.; Liu, J. Remote: Reinforced motion transformation network for semi-supervised 2d pose estimation in videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 1944–1952.
26. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
27. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.L.; Grundmann, M. Mediapipe hands: On-device real-time hand tracking. *arXiv* **2020**, arXiv:2006.10214.
28. Chen, T.; Fang, C.; Shen, X.; Zhu, Y.; Chen, Z.; Luo, J. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 198–209. [CrossRef]
29. Lee, S.; Park, H.; Kim, D.U.; Kim, J.; Boboev, M.; Baek, S. Image-free Domain Generalization via CLIP for 3D Hand Pose Estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 2934–2944.
30. Wan, C.; Probst, T.; Gool, L.V.; Yao, A. Self-supervised 3d hand pose estimation through training by fitting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10853–10862.
31. Spurr, A.; Dahiya, A.; Wang, X.; Zhang, X.; Hilliges, O. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 11230–11239.
32. Chen, Y.; Tu, Z.; Kang, D.; Bao, L.; Zhang, Y.; Zhe, X.; Chen, R.; Yuan, J. Model-based 3d hand reconstruction via self-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 10451–10460.
33. Ren, P.; Sun, H.; Hao, J.; Qi, Q.; Wang, J.; Liao, J. A Dual-Branch Self-Boosting Framework for Self-Supervised 3D Hand Pose Estimation. *IEEE Trans. Image Process.* **2022**, *31*, 5052–5066. [CrossRef] [PubMed]
34. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 398–407.
35. Cai, Y.; Ge, L.; Cai, J.; Yuan, J. Weakly-supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 666–682.
36. Hua, G.; Liu, H.; Li, W.; Zhang, Q.; Ding, R.; Xu, X. Weakly-supervised 3D human pose estimation with cross-view U-shaped graph convolutional network. *IEEE Trans. Multimed.* **2022**, *25*, 1832–1843. [CrossRef]
37. Cai, Y.; Ge, L.; Cai, J.; Thalmann, N.M.; Yuan, J. 3D hand pose estimation using synthetic data and weakly labeled RGB images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3739–3753. [CrossRef] [PubMed]
38. Khaleghi, L.; Sepas-Moghaddam, A.; Marshall, J.; Etemad, A. Multi-view video-based 3D hand pose estimation. *IEEE Trans. Artif. Intell.* **2022**, *4*, 896–909. [CrossRef]
39. Iqbal, U.; Molchanov, P.; Gall, T.B.J.; Kautz, J. Hand pose estimation via latent 2.5 d heatmap regression. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 118–134.
40. Hampali, S.; Rad, M.; Oberweger, M.; Lepetit, V. Honnotate: A method for 3d annotation of hand and object poses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3196–3206.
41. Zimmermann, C.; Ceylan, D.; Yang, J.; Russell, B.; Argus, M.; Brox, T. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 813–822.
42. Gower, J.C. Generalized procrustes analysis. *Psychometrika* **1975**, *40*, 33–51. [CrossRef]
43. Gong, K.; Zhang, J.; Feng, J. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 8575–8584.
44. Yang, L.; Li, S.; Lee, D.; Yao, A. Aligning latent spaces for 3d hand pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2335–2343.
45. Chen, X.; Liu, Y.; Ma, C.; Chang, J.; Wang, H.; Chen, T.; Guo, X.; Wan, P.; Zheng, W. 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, Nashville, TN, USA, 19–25 June 2021; pp. 13274–13283.

46. Zhang, B.; Wang, Y.; Deng, X.; Zhang, Y.; Tan, P.; Ma, C.; Wang, H. Interacting two-hand 3d pose and shape reconstruction from single color image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 11354–11363.
47. Yang, L.; Li, J.; Xu, W.; Diao, Y.; Lu, C. BiHand: Recovering Hand Mesh with Multi-stage Bisected Hourglass Networks. In Proceedings of the British Machine Vision Conference, Manchester, UK, 7–10 September 2020.